5-2018

Evidentiary Reasoning: An Examination of Elementary and Middle School Students' Knowledge of Scientific Evidence in Biology

Jamison Wills
Purdue University

EVIDENTIARY REASONING: AN EXAMINATION OF ELEMENTARY AND MIDDLE SCHOOL STUDENTS' KNOWLEDGE OF SCIENTIFIC EVIDENCE IN BIOLOGY

by

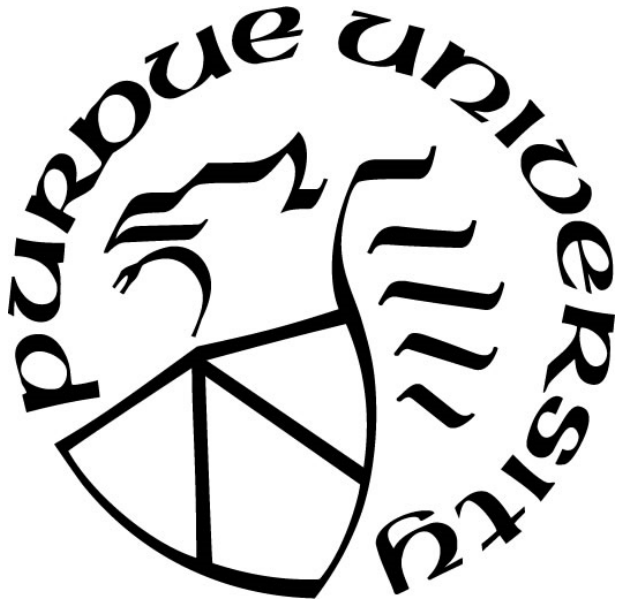
Jamison Wills

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy



Department of Educational Studies

West Lafayette, Indiana

May 2018

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Ala S. Samarapungavan
Department of Educational Studies
Dr. Lynn A. Bryan
Department of Curriculum and Instruction
Dr. Toni K. Rogat
Department of Educational Studies
Dr. David A. Sears
Department of Educational Studies

Approved by:

Dr. F. Richard Olenchak
Head of the Graduate Program

This dissertation is dedicated to my family. To my wife, Sloan, without your unending support and encouragement, this would not have been possible. To my kids, Jaden and Maya, you will always be my greatest accomplishment.

ACKNOWLEDGMENTS

I would like to thank my advisor, Ala Samarapungavan, for her invaluable contributions to my thinking and for her support and encouragement. To my other committee members, Drs. Lynn Bryan, Toni Rogat, and David Sears, I would like to thank you for helping to shape and refine my ideas about learning and teaching.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Author: Wills, Jamison. PhD
Institution: Purdue University
Degree Received: May 2018
Title: Evidentiary Reasoning: An Examination of Elementary and Middle School Students' Knowledge of Scientific Evidence in Biology.
Major Professor: Ala S. Samarapungavan

This project examines upper elementary and middle school students' knowledge of scientific evidence. Informed by literature in cognitive development, learning sciences, and science education, this proposal argues that science educators have typically treated evidence as a simple and unitary construct that is disconnected from other phases of scientific activity. Evidence in the philosophy and history of science, on the other hand, is multifaceted, sophisticated, and involves the coordination of disciplinary knowledge and methodological practices. Based on a conceptual analysis of evidence in this literature, I developed a framework of evidential dimensions that important to scientific reasoning. Two fifth and two seventh-grade classrooms in suburban Midwestern public schools completed one of two science narratives containing a subset of dimensions from the framework. High and low performing students on the narratives were interviewed. Semi-structured interviews were conducted with participating teachers as well as classroom observations. Teachers also provided descriptions of inquiry tasks used in the classroom. Results indicate students struggled reasoning with and about aspects of evidence from the framework. Further, teacher interviews, classroom observations and tasks reveal oversimplified notions of evidence at play in the classroom, and I suggest these instructional facets are associated with student performance.

CHAPTER 1. INTRODUCTION

Epistemology, the branch of philosophy concerned with the origins, nature, and validation of knowledge, has occupied a central place in the history of western philosophy. Plato wrestled with various definitions of knowledge in the *Theaetetus* (Chappell, 2013). Following Plato, Aristotle articulated an early form of empiricism, proposing that universal truths about the natural world could be obtained by way of observation and induction (Baofu, 2012). In the time since, the field of epistemology has continued to contend with formulating the characteristics and the nature of scientific knowledge, especially with respect to how evidence relates its construction and verification. General epistemology has focused broadly on the relation between evidence and knowledge across disciplines (Bod, 2014; Goldman, 1986; T. Kelly, 2014). However, the focus of this dissertation is on the nature of knowledge-evidence relationships in science.

The epistemology of science focuses on the nature and role of empirical evidence in relation to scientific theories, principles, and models (Chalmers, 1999; Franklin, 1986). Scientific evidence is generated from a complex and interconnected web of practices which involves the coordination of conceptual and methodological knowledge (Chalmers, 1999; Franklin, 1986). Research in education acknowledges the centrality of evidence in science and the importance of creating instructional spaces where students can reason with scientific evidence. For instance, recent reform documents (National Research Council, 2012; NGSS Lead States, 2013) highlight the importance of crafting educational spaces where disciplinary knowledge and scientific practice are interwoven in order for students to develop a robust knowledge base of the sciences.

Decades of research in science education and learning sciences demonstrate that students are able to evaluate evidence as well as incorporate relevant epistemological considerations such as how experimental error can influence evidentiary results in some contexts (Lubben & Millar, 1996). Additional studies have established that students are able to construct accurate interpretations of phenomena based on evidence (Schauble, 1996) and they can distinguish between their own theoretical commitments and the set of evidence in favor of them (Koslowski, 1996). However, for every example of student ability, there seem to be two others highlighting the obstacles many students continue to face when reasoning with evidence in science. Examples

include the lingering effects of pre-existing beliefs or students' reliance on superficial or inconsequential information when evaluating evidence (Chinn & Malhotra, 2002; Zimmerman & Glaser, 2001; Zohar, 1995).

It is my view that these persistent issues with students' evidentiary reasoning stem, in part, from the notions of scientific evidence currently at play in science education. Specifically, scholarship in science education has traditionally utilized straightforward and under-analyzed notions of evidence in their investigations. For example, a number of studies examine students' evidentiary reasoning with simplistic and knowledge lean covariation tasks in non-scientific contexts where participants are asked to evaluate instances of perfect, imperfect, and mixed covariation examples and form the correct causal attributions. While students can successfully perform in these contexts, the evidence they are asked to evaluate is simple and largely disconnected from other phases of scientific activity.

Studies of science, however, conceptualize the relationship between evidence, data, and theory in science as more intricate and sophisticated. Additionally, the complexities of this relationship are organically constructed according to disciplinary standards and norms (Weber, 2012). Thus, evidentiary knowledge is discipline specific and involves the acquisition and coordination of both content knowledge and contextualized sets of inquiry practices (e.g., methodological procedures, tools, etc.). Consider the role of mechanism in science as an example. As a primary goal of a number of scientific disciplines (Woodward, 2011), explanations of the mechanisms involved in natural phenomena can represent features such as parts, causes, and organization (Craver & Tabery, 2016). However, mechanisms of evolution, plate tectonics, and the stability of planetary orbits are qualitatively different, and their discovery and articulation require both sufficient amounts of disciplinary content knowledge and methodological knowledge. Research in the philosophy and history of science supports the view that methodological norms and standards are born from disciplinary contexts (Brandon, 1996; Franklin & Perovic, 2015; Mayr, 2004). For example, research in the complexity of biological systems incorporates a methodological approach that takes a different form than the traditional theory-experiment methodologies of other disciplines such as physics (Rheinberger, 1997).

Currently, there are few studies in psychology, the learning sciences, or science education that systematically examines how science learners construe the many facets of scientific evidence in a domain either developmentally or in the context of instruction. As a result, the field

does not have a full picture of 1) students' evidentiary reasoning abilities in a particular discipline or 2) how their understanding of evidence develops over time and with instruction. For these reasons, it is crucial to obtain a more complete understanding of student's evidentiary abilities.

Based on a conceptual analysis of evidence in the history and philosophy of science (Wills & Samarapungavan, 2017), the theoretical framework for this study posits that scientific evidence is complex, multifaceted, and intimately connected to other aspects of scientific activity such as the quality of the experimental design and data collection procedures (Heilbron, 2003; Staley, 2004). Further, the conceptual analysis was instrumental in the development of a framework of evidential dimensions relevant to scientific inquiry. Although the framework is organized around distinct phases of scientific inquiry, it does not draw sharp boundaries between them and recognizes their interconnected nature. The conceptual analysis identifies the following aspects of evidentiary knowledge as critical to scientific inquiry: 1) knowledge of variables (e.g., identifying, providing a rationale, and operationalizing relevant variables), 2) an awareness of sound procedures for collecting data (e.g., evaluating the accuracy of tools employed, sample considerations, etc.), 3) ability to make informed judgments regarding the quality of evidence (e.g., are the interpretations objective and thorough), and 4) sufficient grasp of the social features of representing data in communicable forms and developing evidence-based explanations, models, and arguments.

The disciplinary context of this study is biology. Biology, as an important part of the life sciences, is a particularly fertile ground for exploration. It represents a diverse spectrum of foci ranging from broad areas of interest (e.g., the origins of life) to more nuanced considerations (e.g., cellular processes and functions). Additionally, many areas of biology are comprised of complex systems such as replication, growth, and hierarchical organizations that operate across numerous planes of an organism (Mayr, 2004). According to the Indiana Science Standards, both fifth and seventh-grade students are expected to have knowledge of content in biology as well as knowledge about the nature of science and design processes (Education, 2010). This includes an understanding about how to formulate testable questions, design a test, plan and carry out an investigation, and identify patterns, examine causes, and propose explanations. The purpose of the current work is to explore and describe students' understanding of scientific evidence in

biology. Specifically, this project proposes to discover the evidentiary knowledge of upper elementary and middle school students.

Research Questions

1. What evidentiary knowledge do fifth and seventh-grade students possess about dimensions of evidence contained in the conceptual framework?
2. How do fifth and seventh-grade students differ in their performance across the dimensions of evidence?

CHAPTER 2. REVIEW OF RELEVANT LITERATURE

The following sections are devoted to exploring the relevant literatures on scientific evidence. Beginning with scholarship in the history and philosophy of science, this project examines both historical and contemporary scholarship to conduct a conceptual analysis of evidence in science. The section culminates with a brief discussion of the findings and presents a table of evidential dimensions derived from the analysis.

The remaining sections are comprised of research germane to science education. The first contains important developmental scholarship that highlights key information about what students can do and when. The final section is organized around the theoretical framework and encompasses science education studies examining each evidential theme. The results demonstrate the nature of evidence in science education is in need of revision if students are to gain the evidentiary underpinnings central to knowledge acquisition in the sciences.

For example, prior to engaging in any experimental tests of a phenomenon, it is essential for each domain to determine what sorts of things will count as evidence (T. Kelly, 2008). Among others, this includes concerns such as the reliability of human observation to the use and precision of experimental tools. Attempts to detail how scientific evidence relates to scientific principles, theories, or models have traditionally been the territory of philosophy; specifically, the epistemology of science. The discourse generated from the varying schools of thought is part of rich past in the philosophy of science and represent a spectrum of ideas from a focus on particular processes thought to capture the essence of scientific activity to more descriptive accounts constructed from historical examples. The following section presents a brief overview of the major perspectives on the nature of scientific evidence that have emerged from the literature.

Confirmation and Falsification

A significant theme in the philosophy of science literature is whether the objective of science is to confirm theories or to falsify them. Confirmationists claimed that the evidentiary chain began with the senses and continued through a meticulous use of logic and experimental methodologies to achieve confirmation of the theory. This characterization of the epistemology

of science can be seen across a number of models from early forms of inductivism (e.g., Bacon's *Novum Organum*) to models focused purely on the formal relations of hypothesis and evidence statements (Hempel, 1965). The inductivist approach has been subject to several critiques including Descartes's questioning of the reliability of perceptual data (Descartes, 1989) and the Humean critique, which in short form, states no amount of sensory/perceptual evidence can fully insulate a universal statement from rejection because the relationship between the data and the statement is dependent on the assumption that there will continue to be future regularity in nature; an assumption which cannot be construed in a non-circular way (Russell, 1912).

Falsificationists, on the other hand, argued the epistemology of science was grounded in disconfirming theories rather than the opposite. For these theorists, the process of scientific activity is best described as one where 1) a potentially falsifiable theory is proposed, 2) the theory is then subjected to severe experimental tests and 3) the accumulation of falsifying evidence serves to discredit the proposed theory. In the instance a theory survives the tests, the falsificationist submits they represent an approximation of reality rather than a demonstration of truth and thus are in no way immune to future falsification (Popper, 1992).

Falsificationism construed knowledge validation (Popper distinguished processes of discovery from those of validation – falsification was a theory about the validation of knowledge) in science as a process of establishing the falsity of laws and theories by way of deduction, thereby avoiding the Humean criticism. Nevertheless, with the advent of naturalistic approaches falsificationists would eventually run into trouble on historical grounds. According to Chalmers (1999), examinations of Newton's gravitational theory, Bohr's theory of the atom, and Maxwell's kinetic theory of gases all show instances of falsification in their experimental record yet the theories were not discarded. Moreover, the theoretical entities falsificationists aimed to discredit are more complex than simple hypothetical statements so if a hypothesis turns out to be false, the falsificationist cannot pinpoint exactly what the evidence has falsified.

Another epistemic challenge to knowledge in science is the underdetermination of scientific theories by evidence (Stanford, 2013). The origins of underdetermination (e.g., the holist thesis) are attributed to the work of Duhem (1954) and Quine (1951). Duhem, who proposed the modern scientific version of the underdetermination argument, was himself a physicist. He suggested that scientists resolve problems of underdetermination by relying on their "good sense" or their disciplinary knowledge of the likelihood and plausibility of various

sources of experimental error as well as of mechanisms by which observed effects might have been produced (Duhem, 1954). At the heart of this problem is the idea that the evidence available at any given point may be insufficient to establish what sets of beliefs we should form as a result. Consider a test of a hypothesis concerning the effectiveness of a medical treatment such as a vaccine to protect/confer immunity against an infectious agent such as the human immunodeficiency virus (HIV). In order to carry out this test, we must first presuppose a number of additional beliefs about what an immune system is, how the immune system and vaccines interact generally, as well as how other factors may or may not impact the results, and so on. If we conduct an experiment and obtain results indicating the vaccine did not protect against the infectious agent, how sure can we be the evidence demonstrates the impotence of the vaccine itself instead of some other equally reasonable explanation (e.g., dosage, patient compliance, variations in patients' prior health, errors in measuring effects of vaccine etc.)? Since no empirical evidence is generated in isolation from theoretical constructs or the complex network of supplementary assumptions associated, the experimental results (regardless of outcome) cannot be definitive.

The articulation and development of these challenges have led to the advancement of methods to illustrate how evidence in support of a particular hypothesis or theory can be considered confirmatory. For example, Bayesian versions focus on the ways in which the accumulation of confirmatory evidence probabilistically authenticates or justifies theoretical knowledge in science. Other scholars have articulated mathematical models for estimating how such factors as the weight, specificity, and relevance of evidence confer support for scientific knowledge (Crupi, Tentori, & Gonzalez, 2007; Glymour, 1980a, 1980b; Joyce, 2005). While acknowledging the potential underdetermination of theories by evidence, many scholars have continued to hold that scientific evidence can confer differential support for competing scientific models or explanations. For example, Laudan (1990) acknowledges the issue of underdetermination but suggests that it comes in degrees and is situationally dependent. Using the historical debate between the Cartesians and the Newtonians concerning the shape of the earth as a representative case, Laudan demonstrates that by carefully evaluating the specifics of the competing theories and their respective evidence scientists can successfully use evidence to discriminate between competing theoretical claims.

Developing an epistemology of science grounded in analyses of historical examples of radical theory change, a new crop of scholars (e.g., Kuhn, 1962 & Lakotos, 1970) broke ties with the confirmation/falsification dichotomy. Their investigations focused specifically on points in time where seismic shifts in thought were taking place (e.g., the Copernican Revolution), and developed descriptive accounts which placed the theoretical structures of science at the forefront. The advances of naturalistic views on the epistemology of science generated a host of additional insights about how it operates. The epistemology of science was seen to be much more fluid and dynamic than either the confirmationists or falsificationists had previously acknowledged. Rather than a singular commitment to processes of confirmation or its opposite, detailed episodes of scientific transformation such as the transition from Ptolemaic to Copernican astronomy encompassed a variety of practices which served to both confirm and falsify competing theories and hypotheses. Additionally, naturalistic epistemologies of science characterized science as an inherently social enterprise where vital aspects of a domain (e.g., content & methodology) are continuously constructed, debated, and revised over time according to the community of practitioners.

While these naturalistic approaches have been influential, the theoretical particulars of their epistemology have been subject to scrutiny. Two of the more significant to materialize are the problem of theory-ladenness and the process by which theoretical structures or paradigms (to use a Kuhnian term) are revised or replaced. The former charges that scientists operating within a particular paradigm are unable to divorce their theoretical commitments from the experimental apparatus calling into question the objectivity of the results. Many have challenged the claims advanced by the theory-ladenness position (e.g., Fodor, 1984), yet others have documented cases of the way theoretical commitments or theory color both perception and methodology. For example, Brewer and Lambert (2001) present the case of N-rays from the history of science. After the discovery of X-rays, the French physicist, Blondlot, reported the discovery of a new form of radiation, N-rays. However, a visiting physicist uncovered that Blondlot and his co-workers were able to detect the radiation even though the apparatus used to discover them was altered such that no N-rays should have been detectable. From a methodological perspective, Schindler (2011) argues that theory is directly related to experimental results. Using the scientific case where zebra patterned magnetic anomalies were found on the ocean floor, Schindler

illustrates how the scientists would not be able to interpret the signals they received as meaningful without relying on some theoretical account.

The culmination of the discourse thus far creates an image of science that is made up of important theoretical structures, methodological standards for obtaining, analyzing, evaluating, and presenting data, and frames the social practices of the scientific community as paramount to evaluating bodies of evidence to generate knowledge claims. Despite providing methodology a spot on the platform of science, early accounts of science rarely gave it anything more than a cursory mention. In fact, its perfunctory treatment eventually led Hacking (1982, 1983) to proclaim that no other field in philosophy had been neglected the way experiment had. The following section reviews important work about the role of experimental methodologies in science.

Contemporary Methodology in the Epistemology of Science

Prior to the latter part of the twentieth century, experimental methodology in science was subsumed under the representing or theoretical umbrella and was thus accorded little attention. However, post-Kuhnian critiques of radical relativism have directed focus on the role of shared methodological norms and standards in developing scientific consensus. As scholars turned their powers of analysis to methodological considerations, it become clear the nature of experimentation was a complex concept comprised of a set of analytical tactics and procedural methods through which the empirical sciences actively intervenes with the material world to create new processes, objects, and substances (Hacking, 1983; Radder, 2009b). Moreover, the epistemic activity of experimentation was determined to be discipline specific in that each respective field commissioned their own ensembles of practices and technologies (Galison, 1987). Traditional discussions in and about how science intervenes with the natural world have been framed around variations of the following questions: 1) what is the role of experiment in deciding between rival theories or hypothesis, 2) what function does experiment perform in the confirmation or support of theories or hypotheses, and 3) how can we rationally believe in the results of experiments.

Answers to these questions have come in a variety of stripes focusing on a number of unique aspects of experimentation. Two works in physics that capture the influential nature of experimental methodologies in particular, however, stand out as principally influential towards

the development of current conceptions of experimental methodology. The core of the first is an in-depth analysis of experiments conducted in particle physics during the 1950s and 1960s. In each presented example, Franklin (1986) traces the way a series of experiments were utilized to demolish long standing beliefs about physical laws. In doing so, Franklin captures the various positions and methodological strategies scientists took to reason through the dilemma of how two particles (the theta and the tau) could have the same charge, mass, and lifespan yet exhibit a varied pattern of decay. Likewise, Galison (1987), also using a case study approach, provides a wealth of historical specifics relating to experimental episodes in electromagnetism, detection of the muon, and the discovery of weak neutral currents. Galison demonstrates through sedulous attention to detail the way these various experimental episodes progressed to become more exact and by extension delivered results with greater consistency and reliability.

An additional theme that emerges is the tools and instruments scientist use to generate the data brought to bear on theories and hypotheses. For example, Mayo (1996, 2005) demonstrates how statistical models designed to detect error can be utilized to determine the "truth" or "falsity" of theoretical entities. The general idea is theories are subjected to rigorous statistical tests that are aimed at uncovering error and through this process favorable evidence towards one of the theories will be generated. Scientists can, then, use the outcome to rationally select between competing theories. With respect to technology, Radder (2009a) argues contemporary scientific experiments utilize technology extensively, and the reciprocal nature of their relationship can lead to technological innovations as well as novel experimental techniques. The role of technology in science can range from simple applications such as measuring the mass of an object to the generation of complex, three-dimensional models of the universe. For example, Craig Venter and his colleagues extensively used technology and developed a way to sequence DNA molecules in a more organized configuration that allowed them and eventually other groups of scientists to "see the genetic world" in ways previously unavailable (Anton, 2000). Similarly, Hoffert et al. (2002) provide lucid discussions about how technology can aid in developing future solution options for macroclimate stabilization and Corot, Robert, Idée, and Port (2006) highlights the way advances in medical imaging technology utilizing iron oxide nanocrystals aids medical professionals and researchers alike.

What these important works and many others (e.g., Hacking, 1983; Staley, 2004) suggest is that not only do experiments and their tools and instrument have a life of their own outside of

theory but they are also intimately related to important judgments about the evidence generated, such as its accuracy, precision, and quality. Post-positivist scholarship in the area of science studies (Giere, 1984; T. S. Kuhn, 1962; Lakatos, 1970; Laudan, 1996) has discussed the reciprocal influence or co-evolution of theoretical knowledge and experimental procedures and methods to generate evidence for that knowledge in communities of scientific practice. Contemporary scholarship on experimentation acknowledges the discrete but interdependent relationship between theory, data, and evidence in science. Moreover, despite the discipline-specific nature of experiment, each branch of science is concerned with issues such as causal inference and data reliability (Weber, 2012) as well as the utility of statistical arguments to further validate experimental findings (Franklin & Perovic, 2015). In sum, experiment in science involves the coordination of disciplinary knowledge and methodological standards and norms. The experimental component includes: 1) knowledge of variables (e.g., identifying, providing a rationale, and operationalizing relevant variables), 2) an awareness of sound procedures for collecting data such as evaluating the accuracy of tools employed, sample considerations, etc., 3) accounting for potential sources of error, and 4) the collection of diverse sources of relevant data.

The Social Nature of Science

In each of the previous sections, the social aspects of science have been integral to the development of research fields and the overall advance of knowledge acquisition. From the brief discussion of naturalistic views to way the renewed focus on experimental methodologies highlighted their discipline-specific nature, the cultural and social dimensions of science have been front and center. The incorporation of the social has extinguished long-held visions of the lone scientist toiling away in a lab insulated from the world. The generation of theories, designing and conducting of experiments, and the ensuing evaluation of evidence do not occur in a vacuum nor do they stand only on the shoulders of a few. Instead, contemporary scholarship recognizes science as comprised of a host of social practices such as collaboration and the development of cultural standards and norms (Cetina, 1999; Latour & Woolgar, 1986; National Research Council, 2015). Consider the recent advances in particle physics concerning the existence of the Higgs boson in which CERN orchestrated some of the largest collaborative partnerships in the history of science. The ATLAS and CMS collaborations were each comprised of more than 3000 scientists representing a diverse spectrum of disciplines and nationalities

(CERN, 2015). Without agreed upon disciplinary content, methodologies, technological integration, and standards for communication these discoveries would not be possible.

Additional examples include the creation of scientific concepts and models. For example, the process of labeling parts of the natural world or explaining phenomena (e.g., neutrinos, electrons, photosynthesis, gravitational force, etc.) is an inherently social practice. According to Holger (2013), theoretical terms such as the ones cited above are born and derive their meaning from the scientific community. Likewise, the proliferation of models in science is also imbued with social dynamics. From a gas, to the solar system, to the atom, to the double helix of DNA, models are a form of distributed cognition created from the mental workings of particular groups in specific settings and then shared with the community (Nersessian, 2006, 2008). The culmination of these views has helped to shape revisions in thinking about the structure of scientific knowledge to account for how a plurality of inputs can successfully lead to knowledge. H. Longino (2015), for instance, makes use of a map metaphor to elucidate how scientific knowledge focused on solving specific "puzzles" can be incomplete on grand scale yet still yield accurate knowledge about natural processes.

The integration of the social into the scientific account has raised concerns about long-held characteristics attributed to science such as the degree of truth contained in its knowledge as well as the rationality of its methodology. One issue in particular concerns the ability of science to remain objective, which can be stated in the following way: if scientific knowledge is the result of collaboration and cooperation among differing groups of scientists, how can its knowledge be objective in any traditional sense? Many theorists have responded to this criticism on the grounds that science, as a social enterprise, maintains its objective and rational character through applied mechanisms such as the critical evaluation of research or the specific methodological standards and codes of conduct adopted by scientific communities (Latour & Woolgar, 1986; H. E. Longino, 1990).

Based on the above, the social nature of science is conceptualized as integral towards each phase of scientific inquiry where the knowledge generated, regardless of discipline, are articulated, and legitimated by the scientific community and can be seen to represent what Roth (2005) refers to as the socially-negotiated products of science. Thus, it is just as important to recognize the social dynamics of science as it is to acknowledge the theoretical components and experimental methodologies discussed previously.

What is Scientific Evidence?

The scholarship in science studies discussed above demonstrates the complexities of examining scientific evidence as a single construct. For example, no single uniform account of the scientific enterprise exists. Rather each discipline organically develops its own set of practices which manifest in areas such as methods, standards of evidence, and norms for communication and interaction. These disciplinary-centric aspects lead to diverse sets of commitments with respect to the overarching aims of each field. According to Hoffman (2007), for instance, the field of Chemistry is not focused on the testing of theories or examinations of alternative hypothesis. Instead chemists are working more on making things (e.g., sulfuric acid) placing them closer to engineers than the traditional view of the scientist. Climate scientists, on the other hand, are more representative of traditional views associated with scientific practice in that they generate and test theories and make extensive use of models (both theoretical and applied) for explaining natural processes such as temperature fluctuations occurring in the upper atmosphere of the earth or predicting future climate situations (Lloyd, 2010). Despite the discipline-specific nature of scientific theories, experimental methodologies, and conceptualizations of evidence, there are multiple points of similarity.

Evidentiary reasoning in science (the generation, evaluation, and use of evidence in relation to knowledge claims) is complex, multifaceted, and contextualized to other aspects of scientific practice. It involves the simultaneous coordination of disciplinary knowledge, models (e.g., of phenomena, data, etc.), methodological considerations, data, and procedures for analysis. Consider a simple experiment designed to determine if weight is a causal factor in the time it takes an object to fall to the ground. Antecedent to the experiment is the formation of sufficient background knowledge to form the theoretical underpinnings from which the hypotheses are developed. Simple forms of background knowledge would consist of the fact that things appear to fall at the same rate of acceleration when dropped regardless of their size or mass. A more nuanced understanding would include the inclusion of concepts such as free fall (a special type of motion where gravity is the only force operating on an object) and the acceleration of gravity. To properly test whether weight is a causal factor, the data collection procedures will need to include appropriate controls such as dropping items from the same height and ensuring the time component is used appropriately to ensure consistent levels of accuracy. It will also be vital to incorporate a representative sample of objects to generate a firm basis for

conclusions. Without these important considerations, obtained results could lead to the erroneous conclusion that weight is a causal factor (i.e., the greater the weight = the faster it drops) in the time it takes an object to fall to the ground. This example, while rudimentary, depicts how the quality and accuracy of evidence is directly tied to other phases of scientific practice.

Conceptual Framework for Thinking About Scientific Evidence

Table 1 summarizes the key elements of a framework that contains components or aspects of evidentiary knowledge. The framework is comprised of three primary categories in which each encompasses a subset of topics that correspond to the phases of scientific inquiry.

Planning, Design, and Collection. This category is marked by the evaluation of interrelated processes surrounding the initial formulations of an empirical study and extending through to its completion. Examples include assessing the connection between the research question(s) and the stated conclusion(s) as well as examining the studied variables. If a particular study includes explanations about phenomena outside of the research questions or focuses on the wrong variables, the accuracy of the results automatically become dubious. Additional considerations include assessing the justification for "why" the variables of interest are targeted and reviewing specific variable information such as definitions, sampling intervals (e.g., how often), range (e.g., how long), and scale (e.g., nominal or ordinal). Without an appropriate sampling interval or range, for instance, the conclusions reached are without the necessary ingredients to be credible. Similarly, the particular procedures selected for collecting data can have a significant impact on evidence as a finished product.

Evaluating the Quality of Evidence. Similar to issues related to validity and reliability, this category contains considerations that are not insulated from each other or other thematic categories. For example, determining the relative credibility of the source and the objectivity of the analyses are intimately related. If either is found to be lacking, it immediately calls into question the merit or worth of the other. Comparably, the analysis of the collected data and the procedures employed to collect it are also related. If the procedures are conducted in a haphazard way or are missing important pieces to the sample, the validity of the conclusions are brought into question. Interpretations of evidence generated without attention to alternative explanations, data (e.g., its representational form, transformations, etc.) are going to lead to reductions in the force of the other dimensions even if they have been afforded adequate attention in the analysis.

Table 1

Conceptual Framework for Thinking about Evidence

	Description
Planning, Design, and Collection	
Question Generation	• Based on what is known and are shaped by potential/anticipated evidence and in turn delineate what will count as evidence
Variable Selection and Operationalization	• Relevant variables are identified/selected and justified • Are variables: Continuous/categorical • What is the sampling interval /range/ frequency
Quality of design & data collection procedures	Is the design appropriate for the purposes of the study? Does it target the variables in an unconfounded way? Are the methods of data collection appropriate and trusted? • Technical precision and sensitivity of measurement tools/devices: Do they have acceptable accuracy and sensitivity for measuring the variables of interest and are they used properly • Sampling: Are the data collected in an unbiased way, representative of the population, and of sufficient range • Are there diverse kinds/sources of relevant data collected? • Are there appropriate models for aggregating and analyzing primary data that guide collection? • Accounting for potential sources of error in data collection
Analysis, Interpretation, & Explanation	
Analyses of Data	Do examinations of data meet accepted standards • Descriptive statistics vs more complex analyses • Examinations of error • How are anomalies (e.g., outliers) resolved • Graphical representations to organize data/illuminate patterns
Interpretations / Conclusions	• Are claims supported by evidence? • Are the results consistent with past research? • Alternative explanations explicitly addressed? • Free from bias/conflicts of interest? • Were limits discussed?
Social Factors	
	Scientific evidence and its communication relies on: • Expertise/training (researcher) • Reporting of results to community • Peer-review of work (proposal, publication) o Expert feedback and evaluation o Journal quality

The Social Dimensions of Evidentiary Knowledge. The social make-up of science is found across all levels of scientific activity. Communities of scientific practice not only work within the boundaries of their own discipline, but they also frequently rely on and collaborate with other disciplines to generate pivotal contributions. For example, developmental scientists have incorporated recent ideas from biology and physics (Greenberg, 2014) and discoveries in

ecology have been generated from a host of interdisciplinary collaborations (Anton, 2000). Moreover, disciplinary practitioners engage in the continual development of ideas and research techniques that generate multifaceted sets of evidence about natural phenomena (Franklin & Perovic, 2015). The social aspects of science are fundamental towards developing disciplinary content, methodologies, technological integration, and standards for communication.

This theoretical framework captures the multifaceted nature of scientific evidence. It also demonstrates that each phase of scientific activity is tightly connected to other evidential dimensions. The following section reviews literature highlighting the knowledge students have to reason with evidence.

Cognitive/Developmental Research

The following section contains important developmental research about students' ability to think scientifically. This body of scholarship is comprised of both theoretical and practically oriented research and is comprised of key information about 1) present-day interpretations about the nature of cognition, 2) the types of reasoning abilities students have that are directly applicable to science, and 3) a brief examination of how similar students' thinking is to the thinking of practicing scientists.

Cognition. Contemporary understanding in cognition posits that human knowledge is organized in domain-specific structures such as objects, language, and number (Carey, 2009; Spelke & Kinzler, 2007). Developmental researchers characterize domain-specific structures as being comprised of functions independent and distinct from one another. Thus, compared to the global nature of knowledge in domain-general approaches, domain-specific theorists suggest that knowledge is continual, gradual, and dependent upon context (Carey & Spelke, 1994; Fischer, 1980; R. Gelman, 1996). This body of research supports the stance that the nature of each domain may reflect and require different paths of development, thus advances in evidence evaluation, and scientific thinking more generally, may be quite different from other domains such as language or math.

Theory Construction. Research in development demonstrates that by the time a child is ready to begin their K-12 education, they already have a number of mental abilities considered to be prerequisites for scientific thinking and, by extension, evidence evaluation. Young students possess an abundance of knowledge about causal relations (Carey, 2009; Gopnik & Schulz,

2004; Schulz & Gopnik, 2004), have formed a number of theories about features of the natural world (Stavy, 1991; Vosniadou & Brewer, 1992), and are able to revise their theories when exposed to opposing evidence and can select theories that are more consistent with the available evidence (Samarapungavan, 1992). For example, Bonawitz, van Schijndel, Friel, and Schulz (2012) investigated the relationship between existing beliefs and the discovery of novel evidence opposing those beliefs. These researchers obtained pre-school and early elementary children's theories about object balance and exposed students to toys that both confirmed and contradicted their stated theory choice. When given a choice between the two toys results showed children were more likely to explore the belief-violating toys and were able to revise their theories accordingly. Similarly, Legare (2012) demonstrates when children are *encouraged* to generate an account for unexpected results, they lean towards exploratory, hypothesis-testing behavior in an effort to discover the discrepancy between their initial ideas and the anomalous outcome. These studies suggest that even very young children will engage in discovery-based behavior when confronted with anomalous outcomes and revise their theories accordingly when the context is designed to promote analysis and reflection.

Theory of Mind. Scholarship in young children's theory of mind has demonstrated preschool children have an understanding of their own mental contents, the mental contents of others, and the difference between content in the mind and reality (Corriveau, Pasquini, & Harris, 2005; Flavell, 2000; Flavell, Flavell, Green, & Moses, 1990; Lane, Wellman, & Evans, 2010; Wellman & Lagattuta, 2004; Ziv & Frye, 2004). They are able to mentally generate alternative accounts of situations that have already occurred (Guajardo & Turley-Ames, 2004), and they can accurately categorize external objects as natural or artificial as well as identify and discuss their properties (S. A. Gelman, 2004; S. A. Gelman & Kremer, 1991).

Research has also revealed preschool-aged children's ability to attend to and analyze the sources of their beliefs (Bright-Paul, Jarrold, & Wright, 2008; O'Neill & Gopnik, 1991), and that they recognize access to information plays a pivotal role in the generation of knowledge (O'Neil, Astington, & Flavell, 1992). The rise of these skills are fundamental towards generating an understanding of the special status afforded to evidence in science as well as developing the reflective mechanisms so central to the evaluation and justification of beliefs (D. Kuhn & Pearsall, 2000; O'Neill & Gopnik, 1991).

Children as Scientists. Another key set of studies have sought to determine the extent to which children's thinking mirrors that of scientists. This scholarship has compared mature scientific thinking with young students' scientific thinking and results show children share a number of similarities with "real" scientists in the way they approach thinking about and understanding the natural world. For example, young children combine general knowledge of the world with contextually relevant knowledge to construct coherent and consistent explanatory frameworks (Blown & Bryce, 2010; Samarapungavan & Wiers, 1997), their explanations about the natural world are comprised of the same general form as those employed by scientists (Brewer, Chinn, & Samarapungavan, 1998; Gopnik, 2012), and they not only exhibit a preference for empirical evidence when making judgments but are also sensitive to whether alternative possibilities exist (Sandoval & Cam, 2010).

Conclusions. The aggregate of these developmental findings establish that by the time young students are of age to enter compulsory schooling, they have formed a number of theories about how portions of their world works, recognize important epistemological distinctions between the mind and reality, recognize the way informational accuracy is related to access, ascribe a greater value to empirical evidence than other forms of evidence, and are amenable to revising their theories when confronted with contradictory information. These characteristics mirror accounts detailing the way scientists approach problems and suggest that young children approach making sense of their world in many of the same ways as mature scientists (Council, 2007). While this is not to say that young children are capable of generating complex theories with high degrees of predictive accuracy, it does contradict the long held view that young children, especially those at the beginning of their formal education, have an impoverished cognitive skill set and are not ready to engage in scientific content and practice (Metz, 2008; Sandoval, Sodian, Koerber, & Wong, 2014).

From an applied perspective, developmentally oriented research and scholarship in science education has revealed the powerful influence opportunities to engage with science as a body of content and science as a set of practices has on the development of scientific thinking. That is, chances to participate in the knowledge building practices of science have been shown to positively impact not only knowledge acquisition in science but also the various strategies students employ to solve scientific problems (Schauble, 1996). Thus in order for students to successfully learn to think scientifically and reason with evidence, they will need to be immersed

in learning environments incorporating scientific content and practice (Lehrer, Schauble, & Lucas, 2008). Another key observation of this research is that it suggests the observed differences between children and scientists with respect to their applicable scientific knowledge may be due to differences in acquired knowledge (e.g., conceptual & inquiry practices) instead of differences in core cognitive equipment.

Science Education and Learning Sciences Research

Research in the evaluation of evidence has been conducted on important areas such as the development of scientific reasoning and evidence evaluation, the ability to generate inferences from data, and how students' analyses are influenced by evidential characteristics such as whether empirical data is present. Additional research has investigated students' understanding of methodological issues of measurement and error as well as their ability to participate in social practices such as using evidence in argumentation to evaluate or justify explanation. The following sections discuss studies in science education that target the evidential dimensions listed in Table 1. Additionally, the organizational structure of the review coincides with the framework. The first section, then, will correspond to examinations of reliability and validity centered dimensions.

Theory, Existing Beliefs, & Cause. Due to the centrality of theoretical structures in science, Kuhn, Amsel, and O'Loughlin (1988) examined whether subjects could coordinate or differentiate between theory and evidence. Specifically, these researchers investigated subjects' ability to reconcile existing beliefs (theories) about causal variables in the face of covariation evidence to the contrary. Using situations constructed outside scientific disciplines (e.g., variables that contributed to a person catching a cold), a range of participants (grade 6, 9, and adults) were initially presented with questions about their causal beliefs regarding catching colds. From these preexisting beliefs, the researchers identified a subset of variables participants' believed to be causal in whether a person caught a cold and constructed participant-specific manipulations of covariation data. If a participant pointed to the patterns of data as the justification for their response to a question, they were coded as evidence-based. If, on the other hand, they referenced their beliefs (i.e., the theory) regarding an outcome, they were coded as theory-based.

Analyses demonstrated several emerging strategies. First, evidence which violated expectations was either dismissed or was accepted only in part. Second, participants unknowingly modified a theory such that the evidence would be in support of it. Finally, participants exhibited difficulty identifying the correct relationship between both covariation and noncovariation events with respect to causality. This surfaced when participants were asked to create a pattern of evidence illustrating the influence of a factor. Taken together, results showed that children and adults had difficulty differentiating between their theories and the evidence, especially when the evidence violated pre-existing beliefs. Ultimately, this led Kuhn et al. to suggest that children were developmentally deficient in their ability to reason scientifically.

Many researchers have questioned the conclusions of Kuhn et al. (1988) on both methodological and conceptual grounds. Samarapungavan (1992), for instance, demonstrated elementary-aged school children are able to use similar considerations as scientists when asked to choose between competing explanations of natural phenomena. Using theory choice criteria found in the philosophy of science literature, Samarapungavan examined student ability to select among alternative accounts based on four criteria: range of explanation, non-ad hocness, empirical consistency, and logical consistency. Once students were categorized as holding a geo or heliocentric framework, they were exposed to observations designed to be neutral towards their existing beliefs and were then provided two opposing explanations focused on one of the four metaconceptual criteria to choose from. While the results revealed an age x performance interaction, all students were able to utilize the same sorts of criteria as practicing scientists to evaluate rival theories when domain-knowledge is taken into account.

Amsel and Brock (1996) investigated both students and adults in their ability to evaluate covariation data independently of beliefs. Using tasks designed to be less complex than Kuhn et al., these researchers presented subjects with data sets about plant health containing either the presence of a variable participants strongly believed to have a causal influence on healthy plants or a variable strongly believed to have no causal influence. Results showed children, just like adults, were able to accurately judge variables as causal when they covaried with plant health and non-causal when covariation was absent. The performance differences that did emerge, however, occurred when children were asked to make the correct causal judgments in belief-violating scenarios. Leach (1999) objected to the domain general nature many of these studies utilized and sought to examine how students of different ages coordinate theory and evidence in

scientific contexts. Groups of students in elementary, middle, and high school completed an instrument comprised of scenarios in electrical circuits and floating and sinking. Scenarios were accompanied by a set of explanations and students, working in collaborative pairs, were asked to choose one to predict future behavior. Overall, participants were able to hold theory and evidence in separate epistemic categories but, similar to Amsel and Brock, results showed a number of instances across grade levels where students contradicted their previous statements and generated ad hoc modifications to their explanations when observations were unexpected.

The difficulty students' exhibit overcoming their pre-existing beliefs when faced with contradictory evidence has been shown to coincide with decisions made throughout history by practicing scientists. For instance, when confirming evidence began to emerge for the Copernican Model, many scientists rejected the findings and continued to adhere to the Ptolemaic Model (T. S. Kuhn, 2003). Further, as Koslowski (1996) argues there are a number of instances in the history of science where a theory, especially in its early form, is treated more like a working hypothesis that can easily be modified or revised to account for the encountered evidence. Therefore, the deficiencies attributed to students, when viewed through the lens of historical science, reflect similar patterns of decision making as the practicing scientist.

Methodological critiques have addressed issues such as task complexity. In two separate studies using less complicated tasks, Sodian, Zaitchik, and Carey (1991) and Koerber, Sodian, Thoermer, and Nett (2005) found that first and second grade students were remarkably competent (55% & 86% respectively) in choosing the correct empirical test to conclusively show which hypotheses was correct. Even when asked to generate a test of hypotheses rather than select one, students were able to distinguish between simple conclusive and inconclusive experimental tests. Moreover, results from the 2005 study established children as young as four are capable of holding beliefs and evidence in separate mental categories and understood the role evidence can play in belief revision. Likewise, Piekny, Grube, and Maehler (2014) found a similar interaction between age and performance on covariation tasks and concluded the ability to evaluate perfect and imperfect covariation develops during the latter preschool and early primary school years but proficiency in evaluating imperfect covariation requires more time to develop due to the inherent ambiguity of the task.

The culmination of this research suggests that while students can differentiate between theory and evidence, the separation, especially with younger populations, is fragile and its

development is not something that occurs as a single transformational act. Rather its trajectory is best represented as a dynamic series of transformations over time where less effective strategies are supplanted by more effective ones (D. Kuhn, 2000; Siegler, 2000). Additionally, these studies frame the struggles exhibited by students as knowledge-based deficiencies rather than an inability to think or reason scientifically.

Quality of Design & Data Collection Procedures. The ability to design experiments and then collect and analyze data is a constituent practice of science. These practices are comprised of methodological knowledge as well as judgments concerning which procedures to adopt. Data collection procedures relate to a range of topics such as appraising the quality of the tools used for taking measurements and seeking out and evaluating possible sources of error. Research into the ideas students hold about scientific experimentation demonstrates a delicate understanding. For example, both Carey, Evans, Honda, Jay, and Unger (1989) and Schauble, Glaser, Duschl, Schulze, and John (1995) discovered many students believe the purpose of experimentation is to generate favorable conclusions and failed to view them as a vehicle for understanding the relations that exist between variables. Students' ability to design and carry out experiments have also been shown to be influenced by situational factors of the task such as whether the experimental activity is perceived to be positive or negative (Zimmerman & Glaser, 2001).

Variables. A principal feature of scientific practice is identifying and understanding variables relevant to the purposes of a study. Knowledge of and about variables is particularly central to experimentation in science. Some variable centered studies have focused on students' ability to correctly label and/or design unconfounded experiments. These studies exposed students to instruction centered on controlling variables. The Control of Variables Strategy (CVS) is grounded in the logic of experimentation. It instructs students to differentiate between confounded and unconfounded experiments and underscores how the accuracy of conclusions derived from unconfounded experiments is qualitatively different than those developed from confounded experiments (Strand-Cary & Klahr, 2008). Students trained in CVS have been shown to significantly outperform control groups when no differences in skills were evident in pre-instruction testing, and CVS students have demonstrated higher achievement on measures of transfer (both near and far) and have been shown to retain their ability over time(Klahr & Li, 2005).

According to Zohar (1995), studies like the one above focus on tasks which overemphasize simple variable control at the expense of more complex understanding. For example, knowledge about the variables that contribute causally to a car achieving good gas mileage is merely a portion of the required understanding. Equally important is possessing an understanding of how variables such as tires, engine size, weight, and individual driving habits combine to directly affect the number of miles the car will travel on a gallon of gasoline. Results showed that although the undergraduate students were able to successfully make causal attributions, they encountered difficulties reasoning about interactions between variables. Similarly, D. Kuhn, Iordanou, Pease, and Wirkala (2008) constructed a multivariable prediction task (MVP) and hypothesized that student mastery of a control of variables strategy (COV) should transfer to more complex multi-variable situations. The study presented sixth grade students who had mastered COV with an avalanche task containing five dichotomous variables (slope angle, soil type, cloud cover, snow pollution, & wind speed) and asked them to predict avalanche risk based on the variables they felt most likely to cause an avalanche.

A potential point of contention with the task concerns the ability of the students to cognitively deal with the multivariable nature of the exercise. That is, there may be developmental constraints of cognitive load. In anticipation of this, the researchers targeted middle school students (an older population) and the task incorporated a chart that identified both visually and textually the causal and non-causal effects in the problem. Results demonstrated that contrary to the original hypothesis, the students struggled to incorporate multiple variables in constructing their predictions instead preferring to focus on one explanatory variable at a time leading these researchers to suggest that skill development and transfer is complex and does not progress linearly.

Working with Data. Recognizing the tendency for science education research to ask students to reason from designed outcomes, Kanari and Millar (2004) exposed students to two separate investigations where they reasoned from data. The tasks were comprised of one where an independent variable covaried with the dependent variable and one where an independent variable did not covary. The objective was to identify commonalities in the applied strategies employed by 10-14-year old students in a pendulum and box task as well as to assess performance variations as a function of age, education, and the type of task (i.e., those where covariation was present versus those where it was not). These researchers also examined the

hypotheses students generated about the task as well as the way they approached the relationship between the results and their initial hypotheses. The outcomes from the fifth, seventh, and eighth grade students showed significant differences between tasks where the IV covaried when compared to tasks where the IV did not covary. For example, all students generated accurate conclusions when covariation was present but only half were able to perform at the same level in the absence of covariation. Moreover, while students were more likely to repeat measurements in the absence of covariation to try and sort out puzzling results, they selectively recorded data, lacked an awareness of measurement error, and exhibited a tendency to hold on to their original hypotheses when facing disconfirming evidence.

Other studies sharing a focus on the quantitative aspects of data have investigated how features of data such as sample size and variability influenced student evaluations and their confidence in generating conclusions about the data. In an examination of third, sixth grade students and adults, Masnick and Morris (2008) presented subjects with one of two constructed stories. The first cover story contained information about a group engineers testing the quality of sports balls by using a robotic launcher (quality in these examples referred to the length the ball would travel when hit). The second story was structured around two athletes who were asked to participate in tasks (e.g., hitting golf ball) to determine their respective fit for team. Participants were asked to assess each scenario and specify what conclusions could be generated from the data and to justify their decisions. Analyses show all age groups exhibited sensitivity to the way larger samples impact confidence about conclusions, and even the youngest population displayed an emergent ability to attend to between group variability when presented with data sets containing enough numerical variance.

Data Collection, Measurement, & Error. Executing reliable measurement procedures bears a central relationship to evidential accuracy. A particularly important component of this relationship concerns an understanding about the uncertainty (i.e., error) inherent to all measurements in experimental designs (D. E. Penner & Klahr, 1996). Investigations in elementary students' procedural and conceptual knowledge in science reveal a range of understanding about empirical data and its collection and evaluation. Lubben and Millar (1996) revealed a general developmental progression in students ages seven, nine, and eleven regarding knowledge about empirical data (e.g., its compilation, functions, and analysis). For example, in questions targeting student knowledge regarding the relationship between the spread of values in

a data set and the reliability of an average value, only 15% of seven year olds made use of this information compared to almost 40% of the nine and eleven year olds. Metz (2004) surveyed second, fourth and fifth-grade elementary students understanding of uncertainty in their own designed studies, and results showed similar age-related differences relative to performance. Older students, for instance, were able to trace experimental uncertainty to issues such as insufficient data or design errors unlike their younger counterparts. Metz also found that more than 50% of each grade group could identify multiple sources of uncertainty from their designs.

Masnick and Klahr (2003) engaged second and fourth grade students in experiments with ramps in order to assess their ability to design an unconfounded experiment, identify potential sources of error, understand the role of error in measurement outcomes, and recognize alternative explanations for variation in repeated measurements. Experiments were staged on two ramps where students could vary the incline, surface, and the length of the run. Students designed four experiments in all and were asked to make predictions prior to each test. Sources of error were provided to students and their ability to reason about their influence was evaluated. Performance measures revealed a general progression towards older students, but both grade levels could identify potential sources of error prior to experimentation despite the fact they did not receive regular science instruction. Moreover, Masnick and Klahr's discovery that second grade children could discuss various ways in which experimental outcomes can be influenced suggests they have some understanding of aspects that can contribute to experimental error.

A common theme through this research is the surprising ability young students' exhibit about abstract concepts such as experimental error. However, it is important to note that the tasks used were tightly bounded and students were provided with sources of error in the Masnick and Klahr study. The use of restricted investigative contexts is a departure from the types of environments practicing scientists navigate. Acknowledging this limitation, Schauble (1996) sought to investigate middle school students performance in contexts designed to be more representative of science, which contain numerous variables and mechanisms which may possess causal force but are not easily observed. Students participated in a water canal task and were asked to examine observable variables (e.g., size, shape, weight, etc.) and to attempt to understand causal mechanisms not readily observable (e.g., turbulence & buoyancy).

The design required individuals to approach their investigations systematically and underscored the importance of evidence-based observation in generating explanations about

experimental results. Students were evaluated on their ability to conduct an experiment, the causal beliefs they held about the mechanisms in the task, and the relationship between how their theories influenced their experimentation as well as how their experiments influenced their theories. Similar to the results obtained in previous examinations, the group of non-college educated adults outperformed the fifth and sixth grade students both from a process perspective (e.g., general approaches and the applications of task-specific strategies) and their beliefs about the causal structure of the tasks. The adults were more systematic and comprehensive in their strategies and thus better equipped to generate valid inferences. Younger students, however, did improve at approaching the experimental context systematically and constructing explanations from evidence, thus suggesting that student improvement and understanding can be obtained through opportunities to practice.

Quality of Evidence. Scientists regularly form judgments about the quality of their own evidence and explanations as well as that of others in the field. To do this, they evaluate specific features of evidence such as whether it was produced from a single study or replicated many times over. They consider the source(s) and inspect for objectivity – a guiding principle in which both the scientist and the study are expected to be free from bias or conflicts of interest. For example, scientists evaluate the affiliations a respective scientist may have and whether those previously established relationships could color their work. Scientists also place considerable value on the analysis of data and perform focused investigations into aspects of their own experimental data such as statistical tests designed to enumerate relationships between variables in the study or to establish evidence in favor of a particular hypothesis. Further, practicing scientists assess the chains of reasoning used to establish the connection between evidence and conclusions.

Students' knowledge of scientific evidence has been examined across the elementary, middle, and post-secondary levels in areas such as constructing explanations and arguments (Driver, Newton, & Osborne, 2000; Jiménez-Aleixandre, Rodríguez, & Duschl, 2000; McNeill, 2011; Osborne, Erduran, & Simon, 2004), and applying evidence to generate models of various phenomena (David E. Penner, Giles, Lehrer, & Schauble, 1997; Stratford, Krajcik, & Soloway, 1998; Windschitl, Thompson, & Braaten, 2008). Similar to the results reported in other sections of this review, students' evidentiary knowledge represents a mixture of success and struggle. For example, Chinn and Malhotra (2002) found that fourth grade students can successfully reason

with evidence from experimental situations even when it required them to revise their pre-existing beliefs. Tullos and Woolley (2009) demonstrated that five and six-year old children can correctly decide between different types of evidence (e.g., supporting, irrelevant, or no evidence) to make inferences about the reality status of a novel being. Results from a study focused on modeling revealed first and second grade students can integrate instruction to create evidence-based models about a human elbow that exhibit a functional understanding that incorporates features of motion as well as constraints (David E. Penner et al., 1997).

In an effort to identify which reason (authority, plausible causal mechanism, or data) students find salient in justifying causal claims, Sandoval and Cam (2010) asked third and fourth grade elementary students to evaluate opposing claims made by two different characters, and participants were asked to identify which character they viewed as providing the better reason for deciding a claim. Results demonstrated that the children did not accept claims simply on the basis of authority and most children were found to weakly order the status of justification from data (i.e., empirical evidence) to plausible mechanisms being the most preferable and ambiguous data and appeals to authority being the least preferable. Participants were also sensitive to the strength of evidence as well as the existence of alternative explanations. Overall, these students exhibited similar preferences towards as scientists towards opposing claims. They evaluated the nature of the supporting evidence (e.g., empirical, authority, etc.), and then sought to discover more granular features of the evidence.

Conversely, McNeill and Krajcik (2007) report on a curriculum that engages middle school students in the study of substances and properties of "real-world" items such as soap and found students struggle to provide evidence for their claims and will often generate them without any justification. In the discipline of biology, Duncan and Reiser (2007) found high school students had difficulty reasoning about the interactions between genes on one organizational level and the proteins, cells, and tissues that take place on another organizational level. Jeong, Songer, and Lee (2007) found that middle school students struggled with tasks designed to assess their evidentiary knowledge across six distinct concepts of evidence (priority, relevance, objectivity, replicability, and example and table interpretations). The questions were grounded in everyday experiences with the weather (e.g., individual experience with a tornado) rather than an intervention about the concepts and processes of the atmosphere. Each of the twelve questions (two questions for each concept) presented students with a problem or situation proposed by a

peer student, and results showed that students' knowledge of scientific evidence was tenuous. For example, students had difficulty discriminating between relevant and irrelevant evidence and failed to recognize the importance of reliable and objective observations.

Across these studies, many of the tasks presented to students contain examples of evidence that are problematic. For example, although Sandoval and Cam (2010) determined that third and fourth grade students placed an emphasis on empirical evidence when judging between competing claims, the tasks presented evidence in the form of simple covariation. Further, question eight (Jeong et al., 2007, p. 95) asks students to reason about the connection between precipitation and humidity based on a small dataset. Not only does this question ignore other important factors related to precipitation and humidity but it asks participants to evaluate a set of evidence generated from a week of observations thereby disregarding the time needed to develop evidence of sufficient quality. Moreover, the evidence students were expected to evaluate in both of these examples was disconnected from important methodological standards and norms related to a discipline. Acquiring disciplinary knowledge of important aspects such as content and methods are vital towards developing the evidentiary underpinnings of a domain. For example, Aikenhead (2005) exhibits the way these factors are interrelated in his study on critical care nurses. Before information was transformed into evidence, the nurses looked for multiple sources of evidence (e.g., blood pressure, temperature, etc.) to corroborate a conclusion, analyzed data to an effort to identify trends that converged on a conclusion, and assessed the context (i.e., medical history, current condition, etc.) surrounding their patients. Without the nurses receiving adequate training in the content and practices of the discipline, the patients care would likely be compromised. In order to acquire the knowledge to reason with evidence, students require the same exposure to the content and practices of a discipline.

Science and the Social. Practicing scientists engage in the construction and revision of scientific knowledge through a host of socio-cultural practices (Cetina, 1999; Latour & Woolgar, 1986) such as collaboration, argumentation and debate, and by providing substantive critiques of other finished work according to disciplinary standards and norms. Researchers have examined students' evidential reasoning by incorporating the social practices of science in areas such as the nature of science (Norman G. Lederman, 1992; N. G. Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002), collaboration (Chinn, O'Donnell, & Jinks, 2000; R. Gelman & Brenneman, 2004; Tao & Gunstone, 1999), and model generation and argumentation (Driver et al., 2000;

Jiménez-Aleixandre et al., 2000; David E. Penner et al., 1997; Raghavan & Glaser, 1995; Stratford et al., 1998; White, 1993).

Both Samarapungavan, Mantzicopoulos, and Patrick (2008) and Bouillion and Gomez (2001), demonstrate the ways the social dimensions of scientific inquiry practices can be integrated into science instruction to support student learning. The former structured kindergartners' model construction and refinement around practices of group collaboration. In a curricular unit based in the life sciences, students created models of the life cycle of a monarch butterfly and engaged with members of their group to present and justify their respective models and debate the strengths and weaknesses of members' constructions. This structured participation emphasized important facets of scientific knowledge building and helped to facilitate kindergartners' ability to generate and refine questions and predictions regarding the structure and traits of living things.

The latter study engaged groups of fifth-grade students in a curriculum which underscored similar social practices of collaboration and group discussions and debate. This study added an additional social component and incorporated a team-based approach towards solving problems. The students worked together as a class to identify a local problem in need of a solution. The class decided on the issue of river pollution in their immediate neighborhood, and worked with other project partners (e.g., Chicago Academy of Sciences, parents, Forest Preserve, and community organizations) who were interested in solving the pollution problem to form one large collaborative group. Through problem-based discussions, students engaged in an exchange of ideas with their classmates and the other partners. Results showed that in addition to science content learning, students expanded their ability to consider other perspectives, form questions, and analyze and compare various solution proposals.

As with each of the previous sections in this review, students have been found to exhibit difficulty when engaging in the above practices. For example, students have been shown to rarely identify weaknesses in their opponents' positions when engaging in collaborative argumentation and tend to concentrate solely on support of their own position (D. Kuhn & Udell, 2007) and will rely on and articulate unaccepted forms of evidence in group discussions such as anecdotal experiences or personal opinions (G. J. Kelly & Chen, 1999). The combination of these difficulties has led science education researchers to generate instructional strategies and supports to assist students' knowledge acquisition. These can take the form of technological

integrations (Jackson, Krajcik, & Soloway, 2000; Varma & Linn, 2012) as well as other scaffolds designed to provide students with a combination of metacognitive, discipline-specific, and cognitive supports.

Metacognitive supports can be integrated in the form of prompts where students are asked to articulate and then assess their own particular strategies of knowledge acquisition. These supports also encourage students to actively monitor the formation of their ideas and to compare and contrast them with scientifically accepted versions (Quintana et al., 2004). For example, ThinkerTools (White, 1993) exposed students to increasingly complex models of how forces influence the motion of various objects. The activities embedded in the software provided consistent opportunities for students to view the construction of their own knowledge by comparing their ideas with those of their classmates as well as accepted scientific understanding at strategic intervals. Discipline specific supports provide students with opportunities to participate in the practices and norms of a domain (e.g., the language, tools and methods) and to generate an overarching understanding of the way the social activities such as collaboration contribute to the construction and revision of scientific knowledge (Linn, Clark, & Slotta, 2003).

Students' knowledge acquisition can also be supported cognitively. These supports are tailored to provide structure to problems in the form constraining the scope of content and organizing information in functional ways, thereby masking unimportant features of a problem space while simultaneously highlighting its relevant features (Quintana et al., 2004). For example, the BioKIDS (Songer, 2006) software exposed students to important scientific practices such as generating hypotheses, analyzing data, and creating evidence based explanations. A focus of the technology was to provide students with simple icons and content hints to focus student attention on the salient information. Similarly, Wu, Krajcik, and Soloway (2002) designed eChem to support student learning by restricting the scope of content thereby lowering the cognitive burden placed on students.

Learning environments such as BGuILE (Reiser, Tabak, & Sandoval, 2001; Sandoval & Reiser, 2004) structure curricular content from the perspective of the discipline by making domain-specific strategies (e.g., argumentation standards, theories and investigative approaches) explicit for students and incorporate a number of cognitive and metacognitive supports to scaffold students' knowledge acquisition in biology. Targeting middle and high school aged grade levels these researchers provided students opportunities to participate in discipline specific

methodological and evidentiary practices. For example, in the unit on ecosystems and natural selection, students are exposed to a crisis in the Galapagos Islands where they complete investigations that incorporate interconnected aspects of a complex ecosystem such as the relationship between climate and plants and animals. The curriculum focuses on a dataset containing both physical and behavioral features of a finch population that inhabits the island. This information is paired with a crisis threatening their survival, and students examine data about the finches in order to develop evidence-based explanations for why some finches are able to survive while others die.

Across these studies, most groups were able to generate sound explanations and could provide descriptions and examples of the evidence used in their construction. From a grade level perspective, middle school students were able to advance explanations of the finches' survival or death using the theory of natural selection. For example, students were able to correctly identify characteristic features of the surviving finches (e.g., longer beaks) that gave them a competitive advantage over other members of the species. Likewise, high school students exhibited greater proficiency at writing evolutionary explanations and increased their performance on a transfer task where they are asked to explain a natural selection result (Sandoval, 1998).

Despite incorporating a rich set of scaffolds to aid students in acquiring evidentiary knowledge in biology, evidence in this study is similar to the evidence examined previously. For example, a computer-based image displaying the final journal of a group of high school students' explanations about natural selection (Sandoval & Reiser, 2004, p. 350) correctly notes that surviving finches have longer beaks which allow them to consume harder seeds thereby increasing their fit as a result of a selection pressure introduced to the environment. However, their scaffolded explanation does not take into consideration the time required to produce such changes in the finch population. Time is a pivotal factor in evolutionary processes, and any set of evidence that overlooks its role will be incomplete.

Conclusion. While this review provides rich data on students' science learning and the acquisition of evidentiary knowledge, the notions of evidence students are presented with are simplistic and knowledge lean. Moreover, many of the investigative spaces students navigate are isolated and divorced from their characteristically interrelated nature. That is to say, research on students' understanding of variables or the differences between theory and evidence are captured in compartmentalized ways. Consequently, there exists little research that directly targets

students' understanding of the complexities of evidence or the way evidence is intimately connected to other phases of scientific activity. What is more, many curricular interventions in science research still leave important aspects of scientific epistemology implicit. For example, G. J. Kelly and Chen (1999) examined the extent to which the discourse practices in a high school physics class mirrored those found in scientific communities. During the analysis, the researchers discovered there was no lesson detailing the scientific norms of communication (e.g., the centrality of empirical evidence, etc.). Thus, despite the substantial scientific knowledge of the teachers, students were left to determine what counts as evidence when forming explanations.

Due to these issues, the field does not have a full picture of 1) students' evidentiary knowledge in a particular domain or 2) how their understanding develops with instruction over time. Following current thinking in the philosophy and history of science, this project views the relationship between evidence, data, and theory as multifaceted and interconnected. Evaluations of evidence in science require a combination of discipline-specific content knowledge, an understanding of experimental methodology, and a grasp of accepted procedures for data analysis. Thus evidence in science is more complex than covariation and understanding how social practices of science contribute to knowledge generation are just as important to students' developing mature notions of evidence as acquiring an understanding of other evidential dimensions.

CHAPTER 3. METHODOLOGY

As noted previously, science education research has traditionally relied on simple notions of evidence, which are not representative of the multifaceted, complex, and interrelated nature of evidence operating across the empirical sciences. This has led to a paucity of research across the elementary and middle school grade bands on students' evidentiary reasoning and how it develops with instruction. Drawing upon cognitive science (Chi, 1997; Vosniadou & Brewer, 1992) as well as interpretive techniques (Boland, 1985) for gathering and analyzing data, this research integrates both quantitative and qualitative methods in a cross-sectional design to generate multiple sources of data about students' evidentiary knowledge and its development. The research questions for this research were:

1. What evidentiary knowledge do fifth and seventh-grade students possess about aspects of evidence contained in the conceptual framework?
2. How do fifth and seventh-grade students differ in their performance across varied dimensions of evidence?

Participants and Selection Rationale

A combination of convenience and maximum variation sampling procedures were used (Johnson & Christensen, 2014). Convenience sampling procedures were applied to recruit the samples of fifth and seventh-grade classrooms from the three suburban public schools in the Midwest. The school corporations were selected due to previously established relationships with teachers in these areas. Demographic and ISTEP data for participating schools is provided below in Table 2.

The rationale for the chosen grade bands is grounded in the developmental literature, which establishes strong experiential trends. This is evident in studies investigating students' understanding of the nature of science (Khishfe & Abd-El-Khalick, 2002) and research on the relationship between instruction and conceptual change (Raghavan & Glaser, 1995; Stratford et al., 1998). Further, there is research detailing a number of competencies students have to think scientifically at ages where regular science instruction is absent. For example, young students

Table 2

Demographic Data

		Elementary School 1	Elementary School 2	Middle School
Demographics	% Asian	1.6	6.5	1.6
	% Black	2.7	9.3	2.0
	% Hispanic	4.8	7.2	6.1
	% Multiracial	2.5	6.1	3.2
	% White	88.3	70.2	87.1
	% Free or reduced lunch	25.9	20.2	19.8
ISTEP Data	ISTEP LA Passing Rate	77.5	73.60	78.9
	ISTEP Math Passing Rate	80.0	80.20	73.6
	ISTEP Science Passing Rate	82.5	89.2	84.5
Total Student Enrollment by Grade		(5^{th}) N=103	(5th) N=467	(7^{th}) N=449

enter school with relatively complex theories of the natural world (Brewer & Samarapungavan, 1991; Gopnik, 2012). The combination of these results underscores the value of examining populations across upper elementary and secondary grade bands and this project will add key data regarding students' evidentiary reasoning.

Participants were regular education students (e.g., no focus or special education) from two fifth and two seventh-grade classrooms in suburban mid-western schools. To preserve confidentiality, all students and teachers have been given pseudonyms. All participating teachers were white females. The seventh-grade teachers had taught for an average of 11.5 years (Mrs. Murray = 18 years, Mrs. Carter = 5 years). The fifth-grade teachers had been teaching for an average of 27 years (Mrs. Keck = 14 years, Mrs. Samuels = 40 years). There was a total of 67 students in the study. Thirty-five seventh-grade (Mrs. Murray = 19, Mrs. Carter = 16) and thirty-two fifth-grade students (Mrs. Keck = 16, Mrs. Samuels = 16). The seventh-grade participants were 100% white and three percent were on free or reduced lunch. The fifth-grade classes were 88% white, 9% Black and 3% Asian. Twelve percent of the fifth-grade students were on free or reduced lunch.

Data Sources and Coding

Data sources for this project included: 1) classroom-based assessment, 2) the reading level of participating students, 4) assessment of science interest, 4) audio tape transcriptions of semi-structured interviews with high and low performing students, 5) audio tape transcriptions of semi-structured interviews with teachers, 6) classroom observations of science instruction, and 7) copies of science activities provided by the teachers.

Evidentiary Reasoning Assessment (ERA)

ERA Procedure. All participating students (N=67) were assessed using a researcher developed assessment with eight constructed response questions. Students were randomly assigned one of two science stories, which they completed in their normal classroom setting. Once the ERA was handed out to participating students, procedural instructions were provided by the researcher and their classroom teacher. Completion time ranged from 30-40 minutes. All ERA assessments were digitized and stored in an electronic data base for analysis.

ERA Structure and Content. The assessment was comprised of eight questions targeting aspects of a science story designed to elicit students' evidentiary knowledge. ERA items were developed based on the conceptual framework for thinking about evidence and content validity was established by expert review.

Science Stories. As mentioned previously, the ERA consisted of two science stories. The first task was based on research about how mosquitos find food. In this study, van Bruegel, Riffell, Fairhall, and Dickinson (2015) detail how mosquitos integrate the sensory cues of smell, visual features, and heat signatures to locate potential hosts. The article was published in the journal Current Biology. The second task was based on research about plant defenses. Ford et al. (2014) examined how predation risk and plant defenses combine to influence distributions of Acacia trees that were well-defended (trees with long thorns) and poorly defended (trees with short thorns). This article was published in Science magazine. Both stories utilize experimental designs and performed multiple tests.

The published research examples were modified to reduce the overall complexity. Grade level appropriateness was established by two licensed teachers in the state of Indiana. Additionally, both tasks were reviewed by participating teachers prior to student completion. Final versions of the science stories were isomorphic and consisted of the following structure:

1) brief introduction to the problem, 2) overview of past research, 3) purpose of the current research, 4) outline of experimental design, 5) test descriptions followed by their corresponding results, and 6) summary and conclusions. A descriptive breakdown of the task structure follows. One paragraph comprised of about eight sentences was devoted to the introduction and overview of past research. The research purposes and experimental design were detailed in two paragraphs that consisted of around fifteen sentences. Both tasks consisted of four experimental tests listed separately by number with the title of each in bold and underlined. These were followed by brief descriptions of each test. Test results were listed directly underneath their corresponding test descriptions. The science stories concluded with a summary and conclusions totaling about five sentences each.

The ERA questions immediately followed their respective science story and were based on the conceptual framework for thinking about scientific evidence. The items were structured as problem scenarios where two similarly aged students from another class were debating aspects of the science story. Questions of this form have been shown to be an effective way to elicit beliefs, perceptions, and understanding across a wide range of disciplines (Brown, 2000; Peabody, Luck, Glassman, & et al., 2004; Stecher et al., 2006; Veal, 2002). Students in the debates were presented as holding opposing positions. Students in these problem scenarios represented both genders, and the order of debate was alternated across questions. For example, if a male student's position was listed first on a given question, a female's position would be listed first on the very next question. Participants were then asked to construct a response where they identified which of the students they agreed with and to provide an explanation detailing why they agreed with them. This particular answer format was selected due to its ability to elicit complex reasoning processes and evidence-based explanations (Hee-Sun, Liu, & Linn, 2011; Rodriguez, 2002).

While ERA items were based on aspects contained in the conceptual framework for thinking about scientific evidence, the complexity of scientific evidence prohibited designing an assessment that targeted its features comprehensively. Due to this constraint and other limitations such as the amount of time needed to read and complete the tasks, the ERA was constrained to eight questions that targeted components of evidence within each science narrative. This included questions about variables, the experimental design, the conclusions and interpretations that were derived from the test outcomes and the connections between these distinct phases of

scientific activity (see Table 3). The items were divided into pairs so that each of the included evidential aspect included two questions.

Table 3

ERA Question Distribution

Question	Item Descriptions	Connection to Conceptual Framework
1	Question is designed to elicit students' thinking about the experimental design and its connection to the evidence.	*Quality of design & data collection procedures*
2	Asks students to evaluate the design to determine if an additional test would benefit the study.	*Quality of design & data collection procedures*
3	Examines students' thinking about the selected variables of the study.	*Variable Selection and Operationalization*
4	Explores students' thinking about the benefit of supplementary examinations of variables (Mosquito task) or the inclusion of other environmental variables (Acacia task).	*Variable Selection and Operationalization*
5	Asks students to evaluate the overall quality of the studies' conclusions with respect to sample characteristics (Mt) or the plausibility of other explanations (At).	*Interpretations / Conclusions*
6	Question is designed to elicit students' thinking about the conclusions based on the accuracy of tools or sample characteristics.	*Interpretations / Conclusions*
7	Explores students' thinking about the relationship between test design and the evidence produced to form conclusions.	*Quality of design & data collection procedures and Interpretations / Conclusions*
8	Asks students to consider the relationship between the addition of new variables and the experimental design.	*Variable Selection and Operationalization and Quality of design & data collection procedures*

Experimental Design. This pair of questions asked students to reason about the design of their assigned narrative. The first experimental design question targeted students' understanding about how experimental tests in the science narrative connect and build on each other to create an explanation about how mosquitos find food or how the Acacia trees adapted to predation threats in their environment. Both narratives presented students in a debate where one advocated for the entire test set while the other suggested the scientists could have obtained the same results from the last test. In each case, students had to consider the role each test played in contributing evidence towards the narrative's conclusions. For example, in the mosquito narrative, the data from the test set illustrated that mosquitos become active upon smelling CO_2, search for visual

targets to approach, and if heat is detected once close enough, mosquitos land in an attempt to feed. In the final test, all three sense variables were present simultaneously, so eliminating earlier tests would prevent the creation of this model. Moreover, the narrative states that past research demonstrated mosquitos depend on their sense of smell and the presence of heat to locate food and that the scientists hypothesized they also used visual information. Without the entire test set, there would be no evidence showing that mosquitos also rely on visual data to locate food. High quality responses will highlight the way these aspects contribute developing the explanation of how mosquitos find food.

The second question, in both tasks, presented students with a debate about the benefit of including an additional test that targeted existing aspects of the respective studies. One student in the debate suggesting the scientists should have added an additional test to the set. In each case, the added test was irrelevant. Students, then, had to consider the tests along with the evidence and decide whether the additional test would provide beneficial information. For example, in the Acacia narrative, data from the tests showed that the distribution of longer thorns was related to the presence of the impalas (predator) and the impala grazed in open areas where trees with the longer thorns were found because there were fewer places for predators of the impala to hide. The student in the mock debate suggested adding a test where poorly defended trees (short thorns) were moved from their wooded surroundings and placed in open areas where well-defended trees were found. This suggestion is irrelevant because the narrative made clear that the only difference between the trees with long and short thorns was the particular environment they were in. High quality answers will incorporate this information as justification for why the additional tests is unnecessary.

Variables. This pair of questions asked students to reason about the chosen variables in the context of the science story. The first item presented students, in both narratives, on opposing sides about whether additional focus variables should be added to the study. One of the students argued in favor of adding the variable, while the other argued the scientists were justified to exclude it. In both cases, the suggested addition was irrelevant. Students had to consider whether adding the variable would be justified and represent a contribution to the collection of existing evidence. For example, in the mosquito narrative, data from the tests revealed that mosquitos begin their search for food upon detecting CO_2. Without its presence, the mosquitos were relatively inactive. Further, the decision to select CO_2 as a focus variable was articulated in the

section detailing past research. The student in the debate suggested the scientist focus on one gas was limiting and suggested adding oxygen as focus variable. Given the information on previous research and the evidence from the narrative, the design of tests to examine the influence of oxygen is unnecessary. High quality responses will point to this information in their justifications for rejecting oxygen as a focus variable.

The second question in this series also suggested the addition of focus variables. However, this time the suggestions were relevant in both narratives. One of the students in the mock debates argued in favor of their addition, while the other flagged them as irrelevant. Similar to the previous question, students had to consider whether adding the variables would represent an improvement to the evidence and the narratives' conclusions. For example, in the Acacia narrative, the suggested variables were the soil the trees with different thorns were in and the amount of sunlight they received. Both variables represent influential factors to plant growth patterns. Responses that acknowledge the value these suggested variables and their explanatory potential for why only some of the trees grew longer thorns would be considered high quality.

Interpretations/Conclusions. This question set required students to reason about what was claimed in their respective narratives. The first question targeted students' understanding of how the characteristics of the sample are related to the evidential quality and scope contained in claims about how mosquitos utilize sense data to find food or how predation risk influence the distribution of well-protected Acacia trees. Both narratives presented students debating about the limits of the evidence and conclusions based on the representativeness of the sample. One of the students presented the sample from the narrative as a problem and the other student claimed it was not important. Students had to consider whether these sample limitations were a legitimate concern and if they were, did they carry over to the evidence and the conclusions that could be drawn. For example, in the mosquito narrative, one of the students calls the narrative's conclusions into question on the basis that the scientists only examined one type of mosquito. The implication being that different mosquitos may respond differently to sense data or that some types may search for food in an entirely different way. High quality answers will incorporate sample considerations such as these into their evidential evaluations.

The second question, in both tasks, portrayed a student debate about the quality of the evidence based on limitations. One student in the debate claims the limitations are sufficient enough to call the conclusions into questions, while the other minimizes their importance.

Similar to the preceding question, students had to consider whether the proposed limitation was indeed sufficient enough to impact the evidence in a way that impacted the narrative's conclusions. For example, in the Acacia tree narrative, the data showed the distribution of well-protected trees was related to the presence of the impala, however, the student in the mock debate questions suggests that because the scientists did not examine whether birds and insects fed on the leaves, the conclusions are questionable. While there was no mention of these considerations in the narrative, the text did outline the scientists spent an abundance of time studying the trees and their environments and their findings pointed to the impala's feeding on the Acacia leaves as the reason for longer thorns on some trees. High quality answers will note this and will question how longer thorns would deter something as small as a bird or insect from continuing feed on the Acacia leaves.

Relationships. The final pair of questions were aimed at students' understanding of the interrelatedness between the phases scientific inquiry. The first question in this set explicitly targeted the relationship between contextual features of the experimental design and the evidence itself. Both narratives presented a slightly different question format. Rather than two students debating, these questions asked students to think about how they would respond to a teacher asking the class to consider whether changes in the location of the experiment would result in changes to test outcomes. Students had to consider if the change would impact test outcomes and if so, to reflect on how they would be impacted. For example, in the mosquito task, the suggested change was to conduct the experiment in the mosquitos' natural environment instead of a lab. Students were then asked to reason about whether this would result in changes to the evidence. High quality answers will capture numerous issues that would arise such as how the focus variables could be introduced and controlled, or how the mosquitos could be tracked with accuracy.

The final question asked students to think about connections between planning aspects of their narrative and its experimental design. Just like the previous question, this item was framed from the perspective of a teacher asking a class to think about if the design of the tests would change if the scientists thought other factors were contributing to how mosquitos locate food or the distribution of Acacia trees. Students had to consider if the change would impact the experimental design and if so, how would the design be influenced. For example, in the Acacia narrative, the additional factor was thought to also play a role in the Acacia trees growing longer

thorns. Students were then asked to reason about whether this would necessitate a change to experimental tests. Answers that acknowledge changes will occur and are also able to relate those changes to the experimental design will be considered high quality.

Coding

ERA Item Scoring. ERA responses were coded based on cognitive science techniques for the analysis of verbal protocol data (Boland, 1985; Chi, 1997). The initial scoring rubric was developed from the conceptual framework and refined and revised inductively as needed based on the set of responses obtained. The original coding scheme was developed based on recommendations made by the Indiana Science Standards with respect to fifth and seventh-grade students' knowledge of the nature of science and the practices of science. Participating teachers evaluated the ERA to determine whether their students would be able to reason with the dimensions of evidence contained in the eight questions, and the coding scheme was then revised to incorporate teachers' suggestions. This version of the coding scheme was used to code a subset of fifth and seventh grade participants (N=16, N=17) and revised based on responses that emerged from the data. Item scores ranged from 0 to a value of 4 (see Appendix C). Across all items, students that provided no response or recorded that they did not know were given a zero. There were four item level codes developed for scoring:

1) No understanding: Responses assigned a score of one either restated information provided in the text or provided an answer that does not demonstrate an understanding of the evidential aspect addressed in the question.
2) Beginning understanding: Scores of a two were assigned to answers that focused on the aspect of evidence but addressed it in the form of simple rules (e.g., more (tests, research, etc.) = more information = better) or low-level justifications.
3) Intermediate understanding: A score of three was assigned to answers that engaged with the evidential aspect in question and provided one piece of relevant support from their science narrative.
4) Advancing understanding: Responses assigned a score of four presented a greater number of relevant pieces of support and furnished a greater level of detail about the aspect of evidence addressed.

Detailed descriptions of the codes were tailored to fit both their respective science narratives and the dimension of evidence probed in each question. Additionally, the final coding scheme also evaluated the quality of students' evidentiary reasoning. For example, a student could overlook contextual features of their respective narrative and still have their answer coded at a higher level. Chris, a seventh-grader, wrote the following for his ERA answer for question two on the mosquito narrative: "Howard, because this test would explain if mosquitos have a size preference and or they can see big animals easier than smaller or they can see small animals easier than big." When compared to the details of the narrative, Chris' answer does not take into account that the host-seeking behavior of mosquitos is activated by CO_2, and a test of how mosquitos respond to visual data that did not also include CO_2 would be unproductive. However, Chris does illustrate the added benefit of knowing the extent to which mosquitos rely on visual data to locate potential food sources and this would represent an improvement to the knowledge acquired in the study. Due to this, Chris' ERA answer was coded at a higher level.

The reliability of the final coding scheme was established by an independent rater who coded 25% of randomly selected ERA responses (r = .93). All disagreements were resolved through discussions. The following section presents examples of the coding scheme by question. For context, a brief description of the targeted aspect of evidence is included along with a table containing the question as it appears on both tasks and exemplars of responses for each of the item codes.

Experimental Design: Coordinating Evidence for Alternative Models Across a Test Set. Question one on both tasks asks students to evaluate whether the entire test set was needed as opposed to a single test that one student portrayed as containing all the information needed. The second question presented students with a debate about the benefit of including an additional test that targeted an existing aspect of the study. In both cases, the suggested test additions were irrelevant. Table 4 contains the question text for both tasks and examples of student responses and the assigned code for the first and second item.

Variable Selection: Differentiating between Plausible and Causally Implausible Variables in Setting Up Experimental Designs to Collect Evidence. Question three examined students' judgments about including an additional variable to the study. In both cases the additional variable was irrelevant. The second item in this set suggested including two additional variables relevant to their respective science narratives. Due to structural differences between the

Table 4

Experimental Design Items and Samples of Coded Responses

Code	Acacia Narrative	Mosquito Narrative
	Q1: Serena says that test four was the only experiment needed to show that the longer thorns were a survival response of the Acacia trees. Jaden thinks that all of the tests are important because they each provide unique information about the Acacia trees environment.	Q1: Michele questions the number of tests the scientists did. She says that test four was the only experiment needed to show that mosquitos use a combination of senses to locate food and bite. Howard thinks that all the tests are important because they each provide unique information about how mosquitos find food.
1	*Jaden, because the environment it where it can grow and defend itself but if it was in a different one it might be really different.*	*Howard, because I think that it does provide unique information on how mosquitos find food.*
2	*Jaden, because you need to take many tests to see all of the information and to see if you were right or wrong.*	*Howard, because the more tests & data they collect the more accurate the experiment will be.*
3	*Jaden, because to be able to get to test 4, you have to know the prior knowledge gained from the previous test*	*Howard. I agree that all the test were necessary because with each test you can see how the mosquitos react to the different components. If you only performed the last test, you wouldn't know what really affected them.*
4	*Jaden, all of the tests are important because they each provide unique information about the Acacia trees environment. I know this because each of the test had different & new information. Such as test 3 it showed that impalas had a preference for leaves on the branches with the short thorns which led to test 4 showing that the trees only had long thorns for protection when the impalas were there.*	*Howard, because to be able to understand the reaction of mosquitos depending on what surrounds them. In test 1 we saw CO2 with the mosquitos & active movement occurred so that helps back up the results of test 4.*
Code	Acacia Narrative	Mosquito Narrative
	Q2: Jaden thinks the scientists should have done a test where the Acacia trees with short thorns were placed in open areas with the impalas to see if thorn length would change. Serena said that doing this test was not necessary because tests 3 & 4 show that thorn length was a response to environmental threats.	Q2: Howard thinks the scientists should have done a test where the mosquitos were given only visual information to see if they use it to find food. Michele said that doing a test with only visual information was not necessary because tests 3 & 4 show that mosquitos use visual information to find food?
1	*Jaden, because maybe if the tree was in an open placed area it might grow.*	*Howard, because she has a good point.*
2	*Jaden, because you can never have too much data so why not do the test as it looks to me they don't have anything stopping them.*	*Howard, because 1 more test would have made the test more accurate*
3	*Serena, because tests 3 & 4 showed that the environment changed the length of the thorns. So you don't have to run the complete opposite it will show the same results.*	*Michele, because test 3 showed they saw the cows and flew to them.*
4	*No entry*	*Howard, because in tests 3 & 4 other variables caused the mosquitos to move. Having just a visual test with no other variables, would help determine how mosquitos find food.*

studies the narratives were based on, the specific questions varied by task. Table 5 provides the question text for each task and examples of student responses and the assigned code.

Interpretations and Conclusions: Generalizability of Conclusions from Samples, Sufficiency of Evidence and Plausible Causal Explanations, and Sufficiency of Evidence and Instrumentation Error. The first of these items exhibited students debating the merits of the final conclusions of the science narratives. In the both tasks, students were asked to consider the generalizability of the narrative based on sample characteristics. For the second item, the questions diverge due to methodological differences in the respective narratives. In the Acacia narrative, students considered the sufficiency of evidence compared to an alternative causal explanation. In the mosquito narrative, students were presented with a debate about the sufficiency of evidence instrumentation error (see Table 6).

Replication: Ecological Validity and Replication from a Constrained to Rich Environment. The first item in the set asked students to consider the impact of altering aspects of the experimental design would have on the outcomes. The mosquito narrative presented students with a proposed change to the location of the experiment (lab vs nature), while the Acacia narrative asked students to consider a change from the trees natural environment to a recreated one (see Table 7).

Discovery: Additional Causal Variables and the Design of Experimental Tests. The final item targeted student understanding about the connections between planning aspects of a study (e.g., variable identification) and the experimental design. The mosquito narrative asked students to reason about whether the discovery of another sense factor mosquitos utilized to find food would influence test design. The Acacia narrative asked students to consider whether an additional factor thought to contribute to thorn length would impact the test design (see Table 7).

Student Interviews

Interview Procedure. Maximum variation sampling procedures (Johnson & Christensen, 2014) was used to collect qualitative interview data on high and low performing students. Two students per classroom (one high, one low) were selected for semi-structured interviews about their answers on the ERA within two weeks of completing the assessment. Each student was questioned individually using an interview protocol developed by the researcher. Each of the interviews was scheduled at the discretion of the classroom teacher and did not interfere with

Table 5

Variable Items and Samples of Coded Responses

Code	Acacia Narrative	Mosquito Narrative
	Q3: Kevin thinks the scientists should also look at how the leopards and wild dogs influences the types of thorns the Acacia trees grow. Rachel thinks the scientists had good reasons to only focus on the impalas.	Q3: Brian thinks the scientists should also look at how oxygen, the gas people and animals breathe in, influences mosquitos' search for food. Jordan thinks the scientists had good reasons to only focus on carbon dioxide.
1	*Rachel, because they did have good reasons.*	*Jordan, because he thinks the scientists had good reasons to only focus on carbon dioxide.*
2	*Kevin, it could be good to study more to see different effects.*	*Brian, because they should do tests on oxygen because if they never try they will never know.*
3	*Rachel, because the impalas were the only animals that ate the leaves.*	*Jordan, because before they put carbon dioxide in the room there had to be oxygen and they barely moved.*
4	*Rachel, the scientists had good reasons to only focus on the impalas because impalas spent most of their time in open areas where they would feed on acacia trees & other animals like leopards and wild dogs spent their time in the wooded areas not near the acacia trees.*	*Jordan, because inhaling doesn't affect how mosquitos find their meal because it doesn't produce a source of location. When mosquitos find CO2 & heat they know it is their prey. The oxy doesn't effect that as shown in the flight test.*
Code	Acacia Narrative	Mosquito Narrative
	Q4: Rachel thinks the scientists should have examined the soil and the amount of sunlight received for trees with both types of thorns. Kevin asked Rachel how investigating the soil and the amount of sunlight the trees receive helps to answer the question of whether the Acacia trees grew longer thorns as a way to defend themselves.	Q4: Jordan thinks the scientists should have varied the size of fake animals and the amount of heat they gave off. Brian asked Jordan how changing the size of the fake animals or the amount of heat they gave off helps to answer the question of how mosquitos use sense information to find food.
1	*because it has the most reasonable answer that I think is in the answer.*	*because mosquitos use heat and carbon dioxide to find food.*
2	*because the more components they focus on the more facts they would figure out to help their experiment.*	*because it supports Jordan's claim & it would give the scientists more knowledge on what mosquitos prefer & what they would go for in the wild.*
3	*testing the soil and sunlight they received would help because if they weren't getting the right nutrient they might not be growing right.*	*because changing the size & amount of heat would and could reflect of how different kinds of mosquitos react to the differences in size & heat amount. For example, one kind might dive right in while the other goes in slowly or not at all*
4	*this would show if the thorns were different sizes to defend themselves or because they had different growing habits.*	*in response to how the mosquitos reacted to the change in heat, it would in fact answer the main question. The mosquitos approached the fake cow because of the co2 & came closer to the prop. If the heat is changed, they will either come closer or move away.*

students' instructional time. Each interview was conducted on school grounds during normal hours in an area provided by the student's classroom teacher. Before each interview began,

Table 6

Interpretation and Conclusion Items and Samples of Coded Responses

Code	Acacia Narrative	Mosquito Narrative
	Q5: Michael thinks the scientists should have reported how many Acacia trees of each thorn size were in the study. Without this information, Michael has doubts about the quality of the evidence. Alicia thinks the number of trees with long and short thorns have nothing to do with the quality of the evidence.	Q5: Since the scientists didn't experiment with different types of mosquitos, Olivia thinks their evidence is limited to the mosquitos used in the study. Jackson thinks the evidence from the study is NOT limited.
1	*Alicia, because that fact has nothing to do with thorn growth.*	*Olivia, because we don't know what size the cage was or how much money to use in the experiment was.*
2	*Michael, because if you fined more trees then you can do more research and can have more for more resorses for the scientists.*	*Jackson, because they can always find out more about the miscetos and how the find food. They could do so many different tests and or studys on how miscetos find food.*
3	*Michael, because without providing #s how are people supposed to believe that the scientists didn't just do this test on a few trees - instead of multiple trees.*	*Olivia, because there are different kinds of mosquitos that could be attracted to different things.*
4	*No entry*	*Olivia, testing different types of mosquitos would help. If scientists tested different types of mosquitos, they would see if heat, smell, & visual information affects all types.*
Code	Q6: Since the scientists didn't consider whether other plant-eating organisms like insects or birds also fed on the Acacia trees leaves, Alicia thinks the scientists' evaluation of the evidence is incomplete. Michael thinks the evidence from the study is NOT incomplete.	Q6: Jackson thinks the scientists should have reported how accurate the computer was at recording the mosquitos. Without this information, Jackson has doubts about the quality of the evidence. Olivia thinks the accuracy of the computer doesn't have anything to do with the quality of the evidence.
1	*Michael, because the study isn't complete.*	*Olivia, because she thinks the accuracy of the computer doesn't have anything to do with the quality of the evidence.*
2	*Alicia, because there are more animals and they need more info*	*Jackson, because not all computers are 100% accurate.*
3	*Michael, because I think the evidence from the study is not incomplete because the scientists' conclusion gave a clear reasoning on why the trees thorns grew longer as a response to the plant-eating impalas. The trees grew longer thorns as a protection against the impalas.*	*Jackson, I agree with him because if the computer is messed up or a sensor wasn't working it could mess up the entire experiment.*
4	*No entry*	*No entry*

students were afforded time to review their ERA assessment and answers. The interviews took between 15 and 20 minutes to complete. The researcher read each individual ERA question aloud to students prior to interview prompts. The interviews were audiotaped and transcribed for analysis.

Table 7

Relationship Items and Samples of Coded Responses

Code	Acacia Narrative	Mosquito Narrative
	Q7: The teacher asked the class to imagine that the scientists decided to plant some Acacia trees at a local zoo that had some impalas, leopards, and wild dogs instead of observing the trees in their natural environment. The teacher then asked the class to think carefully about whether changing the experimental tests in this way would influence the results of the study.	Q7: The teacher asked the class to imagine that the scientists decided to watch the mosquitos in nature instead of using a lab with a computer to record them. The teacher then asked the class to think carefully about whether changing the experimental tests in this way would influence the results of the study.
1	*if you change your test I don't think it would help.*	*They would not influence the results because when you find out a part of information that leads to more info.*
2	*it would influence the results because it would be in a different environment with lots of people & they could see if the trees reacted differently w/ people as a factor.*	*They would because in nature there are more animals outside so they could've gotten more answers.*
3	*I think it would because the location of the trees and their surroundings are part of the reason they grow a certain way.*	*Yes, it would change the results because in a different habitat, the insects would react differently. In the story, it said they flew back up to the walls & ceiling and in a natural environment they would probably retreat & go try to feed off something else.*
4	*No entry*	*Yes, because the amount of heat & co2 levels would change tremendously, the visual sightings would vary greatly, and they wouldn't be able to control the senses they wanted mosquitos to use.*
Code	Acacia Narrative	Mosquito Narrative
	Q8: The teacher asked the class to imagine that the scientists thought there were other factors in addition to the impalas that contributed to the Acacia trees growing longer thorns. The teacher then asked the class to think carefully about whether this information would change the tests the scientists decided to do.	Q8: The teacher asked the class to imagine that the scientists thought there were other sense cues in addition to smell, heat, and visual information that mosquitos relied on to find food. The teacher then asked the class to think carefully about whether this information would change the tests the scientists decided to do.
1	*It wouldn't change the answers I don't think because those people should be very smart.*	*Yes, because the teacher asked the class to think carefully about whether this information would change the tests the scientists decide to do.*
2	*I do not because it would be more research and more evidence in what they are looking for and that it would be better to not restart form all their hard work*	*It would influence them to find more the more the merrier right*
3	*I think it would because testing new and more factors gives you different information on the trees.*	*I think these additions would alter the tests the scientists decided to conduct because they would have to test the added abilities importance in finding food as well*
4	*No entry*	*I think it would influence the tests because if there were more senses the mosquitos had the more tests they would have to do. And if they had to do more tests there would be more outcomes from the tests that they would have to find.*

Interview Structure and Content. The protocol (see Appendix D) consisted of eight follow up questions that asked student to elaborate on their written response. The prompts were isomorphic in that they were structured the following way: on question ___, you noted that you agreed with _____, can you tell me more about why you agreed with _____ (see sample question).

Example of Interview Question

Question 1: Michele questions the number of tests the scientists did. She says that test four was the only experiment needed to show that mosquitos use a combination of senses to locate food and bite. Howard thinks that all of the tests are important because they each provide unique information about how mosquitos find food.

You answered (**show student their response**)

Can you tell me more about why you agreed with [Michele or Howard]?

Is there anything else?

Interview Goals. The interview prompts were designed to allow students to further articulate or elaborate upon their thinking on their ERA answer and gain insight into how they thought about the dimension of evidence in the respective questions. As noted, students had the benefit of their reviewing their responses before the interview began and during each of the prompts. Depending on students' responses to the initial prompt, a follow-up question would or would not be presented. In cases where the student provided an interview response that aligned with their ERA answer and indicated they had nothing else to add, no other prompts were presented. On the other hand, in cases where a student provided an interview response that diverged from their ERA answer without student acknowledgement, their response required clarification, or if they appeared unsure (e.g., I think, maybe, etc.), additional question prompts were presented. For example, during his interview response about coordinating evidence for alternative models across a test set (question 1), Michael notes that the alternative model of one test is insufficient but uses vague language to indicate about what would represent a reasonable number. He was then prompted about how the acceptable number of tests is determined. See Table 8 for the exchange.

Table 8

Example of Additional Interview Prompt

Student	ERA	Interview Response
Michael	Jaden, because you won't get enough information back without doing the tests	uhh I agreed with Jaden because like you know how like if you only have 1 test done like you need to keep doing more to find out different answers and like how they vary back and forth between each other and like you can't just do 1 test you need to like keep doing more but not too many like you just gotta do like a good amount [seems uncertain] so you get your-the right information *Researcher: how do you figure out what a good amount is?* I don't know.

In his interview response, Michael becomes unsure when he begins to talk about how the appropriate number of experimental tests are determined. Noticing this, Michaels is specifically asked how this part of the design process is established to which he replied he did not know. As noted, only in the special cases described above did the researcher engage further with students during the interviews.

Coding

Student Interviews. Student interviews were organized and coded using qualitative content analysis techniques (Chi, 1997; Vosniadou & Brewer, 1992) to highlight key differences between high and low performing students. Initially, the ERA and interview responses of the eight students' answers were joined to create a complete response set. These sets were then arranged by ERA score to form high (score of 3 or 4) and low (score of 1 or 2) performing groups. Three coding categories were developed directly from these sets: 1) mirrored, 2) elaborated, and 3) changed. Students that repeated their ERA answer in the interviews, even if using different words, were coded as mirrored. When reasoning through question one about whether a reduction to the experimental test set is warranted, the first row in Table 8 illustrates the similarity between the ERA answer and the students interview response. Responding to the same question, the second row displays an elaborated response where the student goes into more detail about why the test set is necessary. The final example illustrates a student changing their ERA answer during the interview discussion process (see Table 9).

Table 9

Examples of Response Sets and Codes

ERA Answer	Interview Response	Code
Because all the tests gave very important & unique information. I think they were all important because all four test have different actions from the mosquitos. Like how ex 3 & 4 were not very different from each other but the actions from the mosquitos were very different.	Because if you only ran a few tests you wouldn't get all the details on how mosquitos react to all different kinds of reasons…all different kinds of like elements of life.	M
Because you won't get enough information back without doing the tests.	Because like you know how like if you only have 1 test done like you need to keep doing more to find out different answers and like how they vary back and forth between each other and like you can't just do 1 test you need to like keep doing more but not too many like you just gotta do like a good amount so you get your-the right information.	E
Scientists are focusing on what instinct they use to find food. Not to see what size animal they prefer.	Well because if they only change the size the size doesn't really matter cause but the body heat it-it says in the message that body heat was like one of the main things that mosquitos-cause if they can smell that and the-they can see the body heat they would want a lot of body heat *Researcher: oh so that would be different from what you circled here, right?* I: yeah *Researcher: so, are you changing your mind so that now you think possibly the amount of heat an animal puts off would be good information to know* I: yeah, yeah. *Researcher: can you give me an example of how you think* I: yeah, if you put in a small cow...you use the 2 fake cows 1 of them is really and the other is really big and the small one has very little body heat and the big one has a lot of body heat. *Researcher: Can I stop your just a second. Did you notice that in your example you're adding size and heat?* I: yeah...mmm-hmm *Researcher: So now size and heat?* I: yeah, a little bit...yeah... yeah, that is also (important?) *Researcher: Can you tell me what happened that made you change your mind?* I: Yeah... well if you look at it from different angles like you just read all the (couldn't make out) but see if changing the fake animal and the amount of heat it puts off would help to answer the question of how mosquitos find food, which I thought yeah that is true.	C

Contextual Variables

Students.

Assessment of Science Interest. Student interest has been shown to have a significant impact on outcome performance across multiple domains (Bransford, Brown, & Cocking, 1999). Due to this, there were three questions placed at the end of the ERA aimed at assessing students' interest (see Appendix B). The questions probe students' views of science, an evaluation of their own science ability, and their interest in the specific ERA task assigned to them. The response categories ranged from one to five in an interval, Likert-based scale.

Reading Ability. Due to the specialized nature of scientific language and the potential difficulty this presents to the design of scientific assessments (National Research Council, 2014), data on the reading level of the participants was obtained from the classroom teacher. I was not, however, able to access standardized test scores of reading achievement. Teachers classified students as below, at, or above grade level reading. Of the thirty-five seventh-grade students, all were rated as at grade level reading or above. In the fifth-grade population, 66% were rated at grade level reading or above and 34% were rated as below.

Instructional.

Teacher Interview and Procedure. All teachers participated in semi-structured interviews using an interview protocol developed by the researcher. Interviews were scheduled at the teacher's discretion, and they took place on school grounds either after school or during the teacher's prep period. The interviews took between 20 and 30 minutes to complete. The interviews were audiotaped and transcribed for analysis.

Interview Structure and Content. The interview protocol (see Appendix E) was comprised of a set of questions that targeted information such as: a) how much time teachers spend each week on science instruction, b) the various science topics presented to the class, c) the teachers' views of science and scientific evidence, d) the nature of investigative activities (e.g., are students exposed to experimental design considerations, etc.) and f) the extent to which instruction is aimed at engaging students to think about and evaluate evidence.

Coding

Teacher Interviews. The analytic approach taken towards the teacher interviews was more descriptive than the ERA or student interviews. This was due, in part, to the structure of the questions, which highlighted key components of instruction such as how often students engaged

in activities that exposed them to the practices of science as well as how often they worked with scientific evidence. Additionally, the interviews probed teachers' science background and training along with their methods of instruction. Several interview questions, however, were analyzed using similar means as the other sections. These questions addressed aspects of instruction such as what teachers wanted students to know about science, and what they wanted students to know about science evidence. Across these questions, teachers' responses were categorized as general, functional, or complex. For example, in answering the first question, a response highlighting basic ideas about science (e.g., fun, active, etc.) was coded as general. Responses that emphasized career pathways were coded as functional, and those that drew attention to scientific processes and scientific thinking were coded as complex (see Table 10).

Table 10

Teacher Interview Sample of Coded Responses

Question: What do you want students to know about science?		
Teacher	**Response**	**Code**
Keck	I want them to know that they're all scientists umm and then I want them just to be interested in science and to try to discover things on their own…and discover things and I try to tell them there's probably many things tha-that are not discovered out there it could be you you could be the one who finds some things.	General
Samuels	I hope they get an interest in it and pursue a career because the future's going to be technology and uh there's um there's data out there I don't-I can't give the exact percentage it changes all the time but the jobs that these kids are going to have when they're older will first of all be many jobs and secondly may not have been invented yet. So, they have to be able to-to grasp those concepts whatever they need for their job learn those things.	Functional
Murray	Science processes and skills. And I can show you this (directs me to mini posters in room that list what appear to be practices from k-12 framework). Mostly the science processes and skills. If they learn the processes and skills, how to think like a scientist, how to ask questions, how to make observations, how to collect data, differentiate that data between qualitative and quantitative. Then you just take the content and apply all those skills with the content. And then I probably say making it applicable or integrating it with the other subjects.	Complex

Classroom Observations and Procedure. The observations were conducted on teacher-identified days where substantive instruction about scientific evidence was to be taking place. The observations took place on school grounds in the teacher's normal classroom. Prior to beginning the lesson, the researcher was introduced to the class as an observer and was then provided a place to sit where classroom instruction and activities could easily be observed

without disturbing student learning. The observations ranged from thirty to fifty minutes in length.

Observation Structure and Content. The observation document is based on the conceptual framework for thinking about scientific evidence. Along with transcripts of which dimensions of scientific evidence were part of class instruction and how they were constructed, field notes were taken about task details such as the overarching purpose, content area, structure (e.g., whole-class, small groups, individual), and time on task. This includes key descriptions of how the lesson was delivered (e.g., lecture or interactive and student-centered), the types of examples used, and the extent to which students were afforded opportunities to engage with scientific evidence. While not formally scored and analyzed, this component of the study provides descriptive data about the topics of scientific evidence addressed during these instructional sessions and provides insight into the ways students are thinking and working with evidence in classroom settings.

Coding

Class Observations. This portion of the project is purely at the descriptive level. Notes from observations were transformed into a transcript, annotated, and then analyzed to determine which aspects of the conceptual framework for thinking about scientific evidence were enacted in the classroom.

CHAPTER 4. RESULTS

This section presents and interprets the results with respect to the research questions:

1. What evidentiary knowledge do fifth and seventh-grade students possess about dimensions of evidence contained in the conceptual framework?
2. How do fifth and seventh-grade students differ in their performance across the dimensions of evidence.

ERA Items

Key to answering the research questions above was the development of an assessment targeting specific features of evidence from the framework. Based on expert evaluation and teacher review, the adapted tasks were appropriate for fifth and seventh-grade students. Item analysis tests (Raykov & Marcoulides, 2011) were conducted on a pilot sample to further examine the quality of ERA items. The initial sample was comprised of 16 fifth and 17 seventh-grade students and difficulty and discrimination indices were created (see Table 11).

Table 11

Pilot Difficulty and Discrimination Indices Across ERA Items

Grade	Indices	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
5	Df	.48	.47	.44	.50	.42	.41	.47	.50
	D	.25	.20	.30	.40	.30	.30	.45	.55
7	Df	.72	.72	.71	.63	.74	.60	.76	.71
	D	.40	.50	.50	.40	.35	.30	.35	.40

With respect to difficulty (Df), these data illustrate ERA questions are near the optimal value of .50 for constructed-response items (Lord, 1952). As expected, items were less difficult for seventh-grade students, in part, due to their increased reading ability. Based on guidelines offered by Ebel (1965), Discrimination was considered adequate ($.30 < D < .40$) or good ($D > .40$) across all items except for the first two and this applied only for the fifth-grade sample. Closer examination revealed these students consistently provided answers on items one and two

that demonstrated limited success reasoning about evidential aspects related to the experimental design. The quality of answers on these items suggests a lack of knowledge as instrumental to the lower discrimination indices rather than an issue with these specific items. Additionally, discrimination indices were within acceptable parameters for the seventh-grade students.

Evidentiary Knowledge and Patterns of Reasoning

The research questions centered on the evidentiary knowledge students possessed and how it varied between grade levels. Analyses of ERA and student interview data illustrate differences across the four pairs of questions addressing evidentiary aspects in the following categories: 1) quality of design and data collection procedures, 2) variable selection and operationalization, 3) analysis, interpretation, and explanation, and 4) the relationship between these varied evidential categories. The maximum score possible was 32 points. Descriptive statistics for fifth and seventh-grade students are presented in Table 12. The mean score for seventh-grade students was 20.6857 and 15.1563 for fifth-grade. With respect to the item-level coding scheme, these averages equate to slightly better than the midway point between beginning and intermediate understanding for seventh-graders (2.59) and right below a beginning understanding for fifth-graders (1.89). Mean scores of ERA items by grade are displayed in Figure 1.

To assess whether these mean differences were significant, a statistical analysis of variance (ANOVA) was conducted using reading as a covariate with the total score as the dependent variable. Results indicated there was a statistically significant difference between seventh and fifth-grade students' mean scores $F(1,62) = 67.060$, $p < .01$ at the $\alpha = .05$ level. To further explore the knowledge fifth and seventh-grade students exhibited across ERA items and their performance differences, item level investigations were conducted. The following section presents and discusses the results by question pair.

Experimental Design and Evidence. The first item pair assessed students' knowledge of how an experimental test set is connected and builds evidence. Question one targeted students' understanding about how experimental tests in the science narrative relate to and build on each other to create evidence related to the purposes of the study. In the mosquito narrative, this corresponded to how mosquitos utilize sense data to find food. For the Acacia narrative,

Table 12

Descriptive Statistics

Task	5th or 7th	Mean	Std. Deviation	N
Mosquitos	Fifth Grade	15.5625	2.96578	16
	Seventh Grade	22.4118	2.06334	17
	Total	19.0909	4.28197	33
Acacia trees	Fifth Grade	14.7500	2.01660	16
	Seventh Grade	19.0556	2.01384	18
	Total	17.0294	2.94891	34
Total	Fifth Grade	15.1563	2.52867	32
	Seventh Grade	20.6857	2.63206	35
	Total	18.0448	3.78367	67

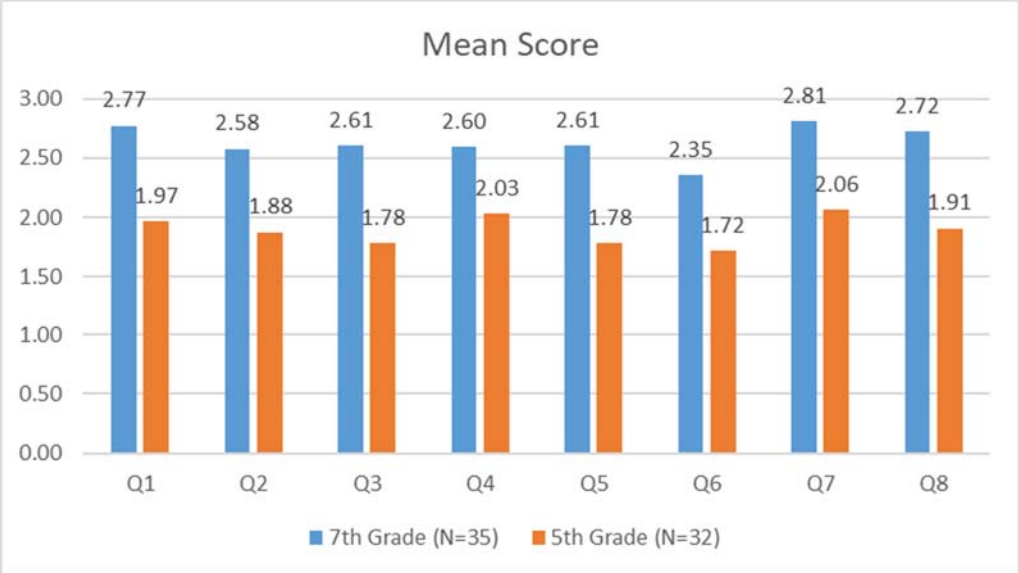


Figure 1

Mean Score by Question and Grade

the issue at hand was whether longer thorns around the leaves of some of the trees were an example of plant defenses. In the question text for both narratives, one student advocated that the

last test in the set was the only one needed, whereas position two supported the whole test set by referencing the unique information each test provides. The second item was based on a debate about whether an additional test should have been undertaken. The idea here being that the suggested test would represent an improvement to the study. In both cases, students had to consider the role of the tests or the addition of a test played in contributing evidence towards the narrative's conclusions. Figure 2 contains the average scores on this item pair by grade.

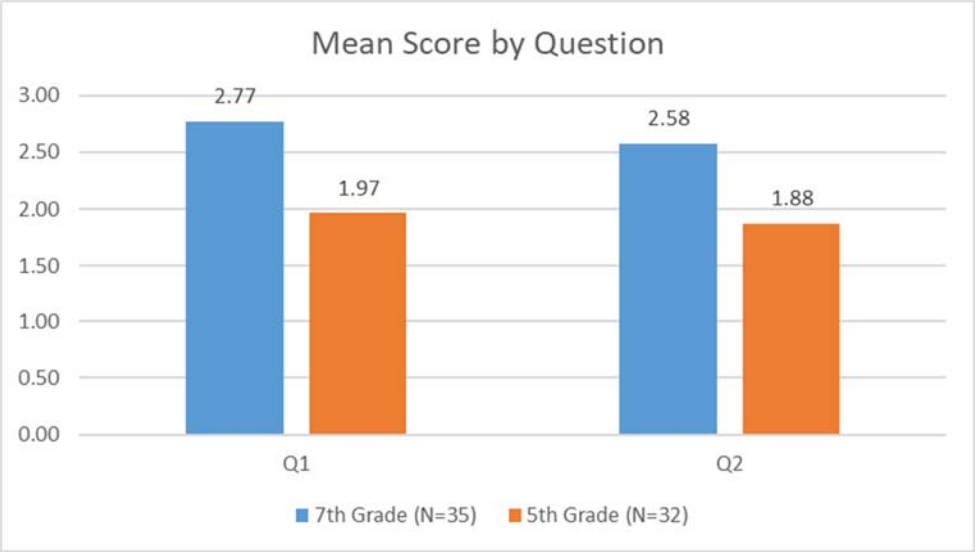


Figure 2

Mean Score by Question and Grade

On questions one and two, seventh-grade students averaged close to an intermediate understanding on the first item and between a beginning and intermediate understanding on the second. Fifth-graders averaged right below a beginning level understanding on both items. The distribution of item level codes reveals seventh-grade students demonstrated greater overall understanding about how each of the tests contributed to the evidence on question one. Fifth-graders tended to view the question as a simple numerical comparison where the larger number is preferable with no articulation about the coordination of evidence for alternative models across a test set (see Table 13). These differences can be seen in the ERA answers provided by Blake and Farah. Blake suggests, "…because you need to test all situations & make sure that 1 variable does not affect another variable or if a variable is important at all." On the other hand, Farah writes, "…because all the testing will lead up to a good answer." These answers highlight the

importance of experimental tests, but Blake's emphasis on the importance of testing variables to detect relationships or to determine significance reveals a more developed understanding about the way evidence is coordinated across the test set.

Table 13

Coding Distribution for Question 1

Code	Description	Fifth	%	Seventh	%
1	No Understanding	7	22%	0	0%
2	Beginning	19	59%	14	40%
3	Intermediate	6	19%	14	40%
4	Advancing	0	0%	7	20%
	Total	32	100%	35	100%

For question two, the differences between the grade levels was less stark with over half of the seventh-grade students coded as no or beginning understanding (see Table 14). Similar to fifth-graders on question one, seventh-graders supported the addition of the irrelevant test with statements based in simple judgment of more equals better. This can be seen in Tatum's ERA response, "…because you can never have too much data so why not do the test as it looks to me they don't have anything stopping them." This response overlooks details about how the proposed test will add value to the evidence and advances a form of the simple judgment more equals better. Students that provided higher level answers were able to coordinate the evidence to reject the alternative model. These narrative based justifications incorporated specific evidence from aspects of the science narrative. For example, Kenley (5th grade) focused her ERA answer on the details of a test where mosquitos did use vision to respond to environmental stimuli: "…because the mosquitos used their eyes when finding the cows & then flying away because there was no heat. In her answer, Kenley rejects the alternative model and supports her position by referring to a specific test and its corresponding evidence.

Variable Selection and Evidence. The next two questions focused on differentiating between plausible and causally implausible variables in setting up experimental designs to collect evidence. Question three presented students on opposing sides about whether an additional focus variable should be included. The implication being the evidence in the science

Table 14

Coding Distribution for Question 2

Code	Description	Fifth	%	Seventh	%
1	No Understanding	8	25%	1	3%
2	Beginning	20	63%	19	54%
3	Intermediate	4	13%	9	26%
4	Advancing	0	0%	6	17%
	Total	32	100%	35	100%

narrative would be strengthened as a result. However, the focus variables suggested in these question scenarios were causally implausible. In the mosquito narrative, the proposed variable was oxygen. For the Acacia narrative, the variable was carnivorous leopards or wild dogs that also lived in the environment. Question four was also based on a debate about the addition of variables. This time, though, the suggested variables were causally plausible, and their inclusion would represent an increase in the quality of the evidence obtained from the tests. In both narratives, students had to consider whether the suggested variables and the tests required to examine them would contribute to the evidence in the study. Average scores on these items by grade can be seen in Figure 3.

Fifth-graders provided answers on question three and four that demonstrated between no understanding and beginning understanding when reasoning about variables. On question three, fifth-graders scored below a beginning understanding and demonstrated a beginning understanding on question four. Seventh-graders scored between a beginning and intermediate understanding on both questions. The distribution of item level codes (see Table 15) reveals seventh-grade students were able to recognize the implausible causal nature of the proposed focus variable and used the experimental context of their respective narrative to justify their position. For example, Mandy's ERA answer contrasts the two gases by focusing on the way each communicates the location of potential food sources: "…because oxygen is all around us and it could lead the mosquitos in many places when carbon dioxide almost pinpoints the source of food. Here, Mandy identifies pinpoints that oxygen's ubiquitous nature is what renders it causally implausible. The large majority of fifth-graders, on the other hand, viewed the correct position to be one where more information is always good regardless of whether adding a focus

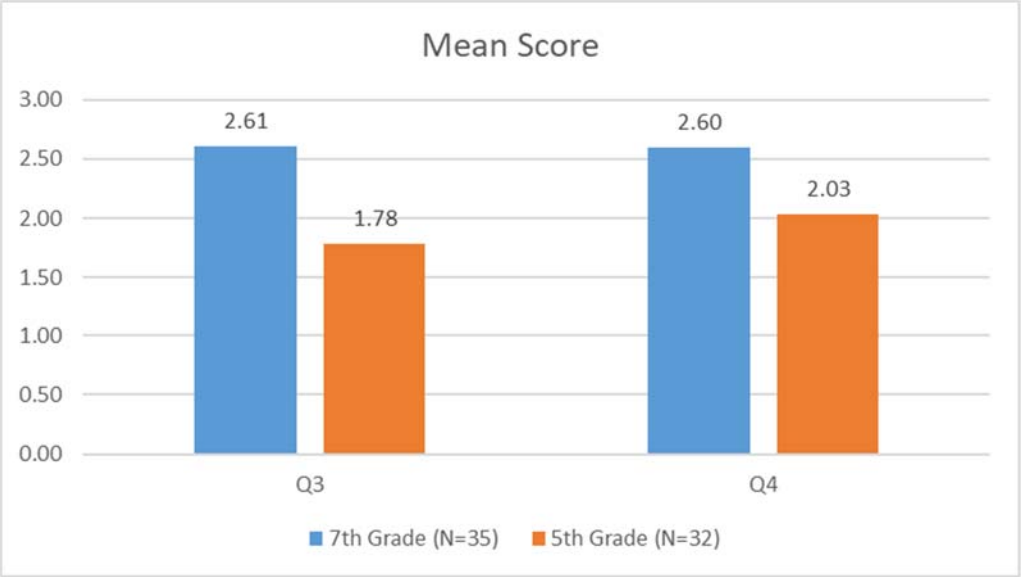


Figure 3

Mean Score by Question and Grade

Table 15

Coding Distribution for Question 3

Code	Description	Fifth	%	Seventh	%
1	No Understanding	8	25%	1	3%
2	Beginning	23	72%	18	51%
3	Intermediate	1	3%	10	29%
4	Advancing	0	0%	6	17%
	Total	32	100%	35	100%

variable actually constitutes evidential improvement. Kris exemplifies this when he states, "…you can never have too much data. Also, there might be a change to either short or long thorns. Kris's ERA answer appears to frame the debate as a simple comparison of quantity and overlooks the whether the proposed variable is casually plausible or implausible.

For the second item in this pair, fifth-graders were able to improve, while seventh-graders remained relatively stable. The key development for the differences in younger students score distribution was the increase in the number of answers coded as intermediate (see Table 16). This

can be explained, at least in part, by the structure of this question. As mentioned previously, question four was the only item comprised of multiple choice and constructed response. The decision to structure the question this way was due to the number of possible choices created from the suggested variables (four on each question). Fifth graders appeared to benefit from the cognitive load reduction as evidenced by their overall performance increase. These students were able to recognize and provide examples of the causally plausible nature of the variables. For example, Camden chooses the multiple-choice selection in his ERA answer that agrees the variables size and the amount of heat are causally plausible and would help to understand how mosquitos utilize information to locate food: "…I think that because if they did the test it would see if they would like a large or small animal. It would also see how much heat they like. In his answer, Camden illustrates how knowledge about the way size and heat impact mosquitos' search for potential food sources would be beneficial. Other quality answers include Lexi's (7^{th}) ERA answer from Acacia narrative where she supports the causally plausible focus variables of soil and sunlight. In her explanation, this Lexi capitalizes on how including the variables increases the evidential quality used to generate conclusions regardless of the outcome: "…this would show if the thorns were different sizes to defend themselves or because they had different growing habits." As with previous questions, there was a high number of students that relied on simple justifications of more equals better and did not engage any further.

Table 8

Coding Distribution for Question 4

Code	Description	Fifth	%	Seventh	%
1	No Understanding	8	25%	0	0%
2	Beginning	16	50%	20	57%
3	Intermediate	7	22%	9	26%
4	Advancing	1	3%	6	17%
	Total	32	100%	35	100%

Interpretations and Conclusions. This item pair asked students to reason about the scientists' final judgments in their respective narrative. The first question presented students either with a debate about alternative explanations that were left unexamined or the evidential

limits based on sample representativeness. The Acacia narrative explored students' reasoning about the sufficiency of evidence and plausible causal explanations. The mosquito narrative presented students with evidential issues related to generalizability of conclusions from samples. In both cases, students had to consider whether these topics were a legitimate concern and if they were, did they carry over to the evidence and the conclusions that were drawn. Question six also targeted distinct aspects of evidence depending on narrative. The mosquito narrative centered on the sufficiency of evidence and instrumentation error. For the Acacia narrative, students confronted a debate about the generalizability of conclusions from samples. Figure 4 contains the average scores on this item pair by grade.

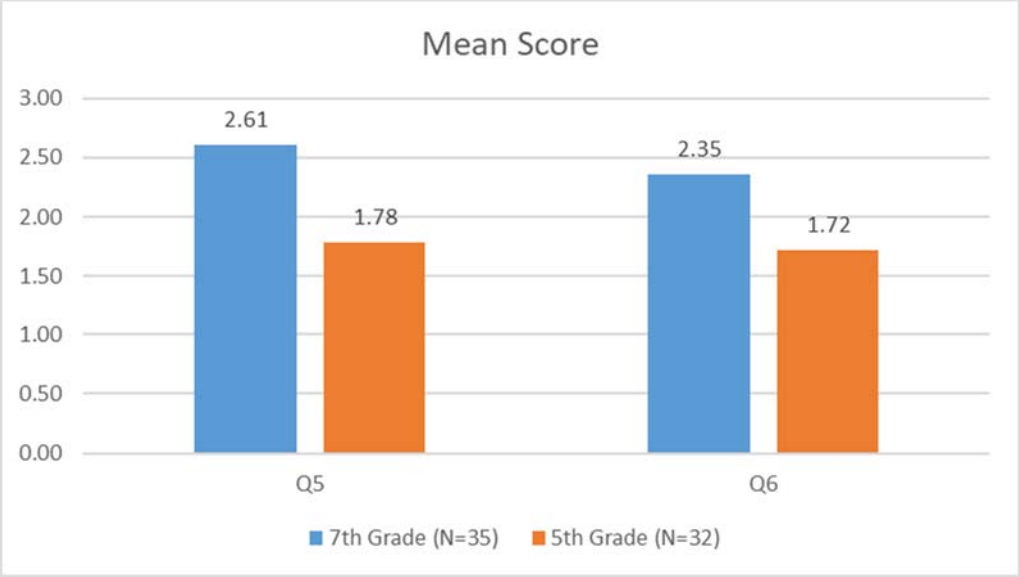


Figure 4

Mean Score by Question and Grade

Seventh-graders scored slightly lower on question six than on question five. Taken together, this group performed between a beginning and intermediate understanding. Fifth-graders remained about a quarter of a point away from averaging a beginning understanding. The distribution of item level codes illustrates that nearly half of the seventh-graders provided answers coded at the intermediate or advancing level, while approximately 80% of fifth graders were coded as beginning or no understanding (see Table 17). This group along with the half of seventh-graders coded at the beginning level focused again on the judgment that more equals

better. For the mosquito narrative, students were persuaded that the sample was problematic. However, their justifications suggested only that scientists could have done more or ran more tests. Carla's ERA answer captures this response type fully: "…because they can always find out more about the miscetos and how the find food. They could do so many different tests and or studys on how miscetos find food." These types of answer overlook the issue of generalizability based on the sample and apply a simple judgment instead. Similar justifications were provided on the Acacia narrative where students advocated that scientists could have conducted more experiments without considering how running more tests would solve the issue raised in the mock debate.

Table 17

Coding Distribution for Question 5

Code	Description	Fifth	%	Seventh	%
0	Don't know/No answer	1	3%	1	3%
1	No Understanding	11	34%	2	6%
2	Beginning	15	47%	12	34%
3	Intermediate	5	16%	17	49%
4	Advancing	0	0%	3	9%
	Total	32		35	

The 16% of fifth-grade and 58% of seventh-grade responses coded at the intermediate level and above were able to illustrate an understanding of how a representative sample is related to evidence. For example, Ray (5th) observes in his ERA answer the restriction on generalizability that results from a limited sample: "…because if you only test with 1 type of mosquito you will only know information of 1 type of mosquito." This is a point Ginny (5th) recognizes in her ERA answer, but she also includes the connection to potential to obtain different results: "…because that is only one kind of mosquito and it could be different test result from a different mosquito. The test result could be totally different." Here, Ginny connects the way a sample influences results. This group was also able to reason about the sufficiency of evidence and plausible causal explanations in the Acacia narrative. Students referred to the duration of the study as evidence calling into question the legitimacy of the proposed alternatives or they engaged directly with the suggested alternatives in the question. For example, Nance (7th)

refutes the suggestion in his ERA answer that the longer thorns were a response to insects eating the leaves by pointing out their small size would be undeterred by the thorns: "…because insects are too small to be effected by the thorns." Here, Nance proposes a counter to the notion that insects or birds were causally plausible alternatives.

Overall results were similar for question six. Almost 60% of seventh-graders and 88% of fifth-graders scored at the beginning level of understanding or lower (see Table 18). These students exhibited difficulty reasoning about the concept of error portrayed in the mosquito narrative as well as the sample issues presented in the Acacia narrative. For the most part, their answers disregarded how the important of the proposed issue or how it was related to evidence. For example, Chris (7th) is unmoved in his ERA answer that technologically-based error is worthy of consideration: "…because if the computer was recording that that means it caught the evidence that obviously shows how the mosquitos reacted." Chris ignores the possibility that the accuracy of the computer's recording could be called into question based on error. In the Acacia narrative, Liz (5th) writes in her ERA answer, "…because the number of trees with the long and short thorns have nothing to do with the quality of the evidence. I honestly didn't think it matters how many. Why? Well, because it wouldn't matter how many trees had short or long thorns, it just matters why some have longer or shorter lengths of thorns. Here, Liz overlooks how generalizability and the sample are related. Contrast this with Lori's (7th) ERA answer that

Table 9

Coding Distribution for Question 6

Code	Description	Fifth	%	Seventh	%
0	No Answer/Don't know	0	0%	1	3%
1	No Understanding	13	41%	3	9%
2	Beginning	15	47%	16	46%
3	Intermediate	4	13%	15	43%
4	Advancing	0	0%	0	0%
	Total	32	100%	35	100%

references the limits the sample can place generalizability: "…because if you do the test with two trees it will prevent your answers from being correct. While she does not articulate the

relationship in detail, there is clear recognition that the experimental sample and the quality and generalizability of the evidence are connected.

Relationships. The final pair of items were designed to probe students understanding about connections between the evidential dimensions of the conceptual framework. In both narratives for question seven, students had to consider if changes in the design would impact the test outcomes and if so, to reflect on how the evidence would be influenced. For the mosquito narrative, the suggested change was conducting the experiment in a natural environment instead of a lab. The Acacia narrative proposed planting a small sample of trees at a local zoo that contained the same main animals found in their natural environment. In the last item, students reasoned about the connection between the identification and selection of causal variables and the design of experiments. This was expressed in the mosquito narrative as scientists had discovered an additional sense cue thought to influence how mosquitos located food and for students to consider if this would impact the design. The Acacia narrative was structured the same way. Students were asked to consider how the experimental design would be impacted if scientists thought there were other factors besides the impalas that the trees were defending themselves against. In both cases, students were asked to think about the relationship between identifying and selecting variables and aspects of the experimental design. Figure 5 displays the mean scores by grade on this item pair.

Seventh-graders averaged close to an intermediate understanding on the final items, and fifth-graders held a beginning understanding on question seven and slightly under that on question eight. The distribution of item level codes demonstrates 68% of seventh-graders were coded as intermediate or above compared to 31% of fifth-graders (see Table 19). Additionally, 69% of fifth-graders scored at the beginning level or below compared to 26% of seventh-graders. The two (6%) remaining seventh-graders were unable to finish the question in the time allotted. Students coded at the higher levels not only acknowledged changes in the design would result in changes to outcomes or evidence, they often identified challenges that ranged in specificity. For example, Sam notes in her ERA answer the difficulty that would result with the variables in the study: “It would influence the results because they wouldn't be able to control the variables.” Building on these ideas, Charlotte also agrees changes would take place and writes, “…the amount of heat and CO2 levels would change tremendously, the visual sightings would vary greatly, and they wouldn't be able to control the senses they wanted mosquitos to use. Both

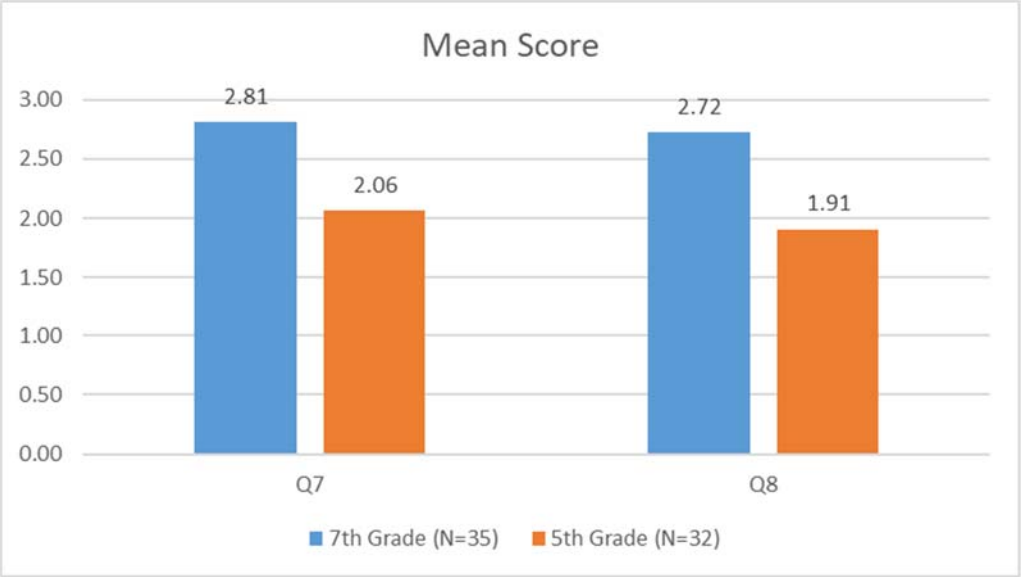


Figure 5

Mean Score by Question and Grade

Table 19

Coding Distribution for Question 7

Code	Description	Fifth	%	Seventh	%
0	No Answer/Don't know	0	0%	2	6%
1	No Understanding	8	25%	1	3%
2	Beginning	14	44%	8	23%
3	Intermediate	10	31%	20	57%
4	Advancing	0	0%	4	11%
	Total	32	100%	35	100%

answers reject the similarity between the rich environment of nature and a constrained one like a lab and highlight the complexities of such a change. Like many of the previous questions, answers coded at a beginning level agreed changes to the design would result but provided a simple justification such as changes here changes there.

Fifth-graders were coded at similar percentages on question eight for beginning understanding and below (68%) as they were on question seven. Seventh-graders ticked up slightly in the number of answers coded at beginning understanding on this item (see Table 20). When reasoning about whether the addition of focus variables would necessitate experimental tests to further investigate, beginning level responses acknowledged changes would take place

Table 20

Coding Distribution for Question 8

Code	Description	Fifth	%	Seventh	%
0	No Answer/Don't know	1	3%	2	6%
1	No Understanding	10	31%	0	0%
2	Beginning	11	34%	12	34%
3	Intermediate	10	31%	18	51%
4	Advancing	0	0%	3	9%
	Total	32	100%	35	100%

but the justifications focused on the benefits of the additional information without demonstrating an understanding of how the extra factors would impact the study. For example, Kaden (5^{th}) agrees in his ERA answer that changes to the tests would take place, but only emphasizes an increase to the overall body of information: “They would because then you know more about what you’ve learned, and you get even more info.” This answer is representative of the more equals better justification. Contrast this with Alexa (7^{th}) who also notes how the presence of additional factors would require the generation of new questions as well as tests to determine influence: “I think it would [change] because there would be different senses so that means different and new questions and tests.” By connecting the development of questions to the design of experimental tests, Alexa demonstrates the interrelated nature of evidential dimension. Answers such as Alexa’s characterized the 60% of seventh-grade responses as well as the 30% of fifth-graders coded as intermediate or above.

The culmination of these analyses suggests students have limited understanding about the aspects of scientific evidence contained in the framework. To further explore differences in knowledge and patterns of reasoning, interviews were conducted with high and low performing students.

Student Interviews

Maximum variation sampling procedures (Johnson & Christensen, 2014) were used to identify and select high and low performing students on the evidentiary reasoning assessment (ERA) for semi-structured interviews to gain additional understanding about the aspects of evidence these students apprehended. Two students from each classroom were selected based on their total ERA scores. The selection of high and low performing students was made from groups comprised of the top and bottom 25% of scorers. The scores of interview students ranged from 28 to 13 (see Table 21). None of the high scoring interview volunteers completed task two, thus the following analysis focuses on students completing task one. All seventh-grade students were reported by their teacher as reading at grade level or above. Three of the four fifth-grade students were reported by their teacher as reading at grade level or above.

Table 21

Interview participants information

Name	Gender	Grade	Grade Level Reading	Task	Teacher	ERA Score
Elijah	M	7	At	Mosquito	Murray	27
Charlotte	F	7	Above	Mosquito	Murray	22
Harper	F	7	Above	Mosquito	Carter	22
Michael	M	7	At	Acacia	Carter	18
Sophia	F	5	Above	Mosquito	Samuels	19
Ethan	M	5	At	Acacia	Samuels	13
Isabella	F	5	At	Mosquito	Keck	17
William	M	5	Below	Acacia	Keck	13

As noted in the preceding section, the ERA was comprised of eight questions targeting specific features of evidence contained in the theoretical framework. Across the eight questions, high scoring seventh-graders were relatively even in the number of times they were coded as beginning (4), intermediate (6), or advancing (6). The low scoring seventh and high scoring fifth-graders largely provided answers coded as beginning (10 of 16). The bulk of low scoring fifth-graders provided answers coded as no understanding (5) or a beginning understanding (9). Figure 6 contains the respective frequencies of the four item codes for interview participants by grade.

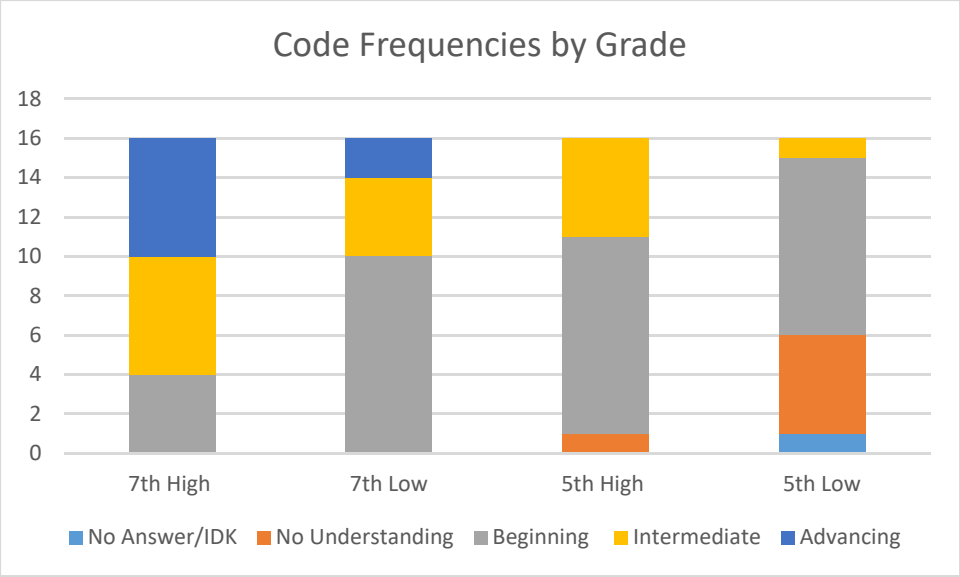


Figure 6

Coding Distribution for Interview Participants by Grade

The following section presents the results of the student interviews by question. To illustrate key differences between high (score of 3 or 4) and low (score of 1 or 2) performing students across ERA questions, examples of high and low scoring students are provided. In cases where no student was coded as high performing, only data from the low performing student is presented. Data from the interviews was initially combined with the corresponding ERA answer to create a response combination for each interview participant. Interview responses were then compared to ERA answers. Analyses revealed students either mirrored their ERA answer, elaborated, or changed their answer altogether and these variations are presented and discussed below.

Experimental Design and Evidence. In this question, students had to consider the role each test played in contributing evidence towards the narrative's conclusions. Students coded as intermediate or advancing cited the value of the test set in their answer and acknowledged the way the test outcomes combined to contribute essential information in a way the single suggested test could not. Students coded as no understanding or beginning understanding provided answers that either restated information found in the text or supported their answer by referencing surface level judgments such as the group of tests is better than a single test.

High scoring students on this item demonstrated an understanding of the way the individual tests converged to supply evidence about how mosquitos find food. The seventh-grade student notes the varied reactions from the mosquitos the tests produced. The ERA response of the fifth-grader also references the test combination and notes two of the studies variables (heat & CO_2) in their justification (see Table 22 and 23). In the interview portion, these students largely mirrored their ERA answers. The fifth-grader's response was coded as an elaboration even though it included considerations not immediately relevant to the study. The low scoring

Table 10

Examples of 7th Grade Responses on ERA and Interview for High and Low Scorers

Student	ERA Answer	Interview
Elijah (High ERA)	Because all the tests gave very important & unique information. I think they were all important because all four test have different actions from the mosquitos. Like how ex 3 & 4 were not very different from each other but the actions from the mosquitos were very different. (3)	Because if you only ran a few tests you wouldn't get all the details on how mosquitos react to all different kinds of reasons…all different kinds of like elements of life. (M)
Michael (Low ERA)	Because you won't get enough information back without doing the tests. (2)	Because like you know how like if you only have 1 test done like you need to keep doing more to find out different answers and like how they vary back and forth between each other and like you can't just do 1 test you need to like keep doing more but not too many like you just gotta do like a good amount so you get your-the right information. (E)

Note. Codes assigned to ERA and Interview responses respectively are provided in parenthesis at the end of each response.

students also referenced a relationship between experimental tests and information or evidence. However, their answers about this relationship remained at a surface level and were not connected to the study. For example, the low scoring seventh-graders' ERA and interview elaboration highlights the relationship between the number of tests and the amount of information collected, but responses like this do not demonstrate an understanding of the cumulative or converging effect tests contribute to evidence in the study. Rather they express this relationship by stressing the importance of obtaining enough data or answers, but the articulation is in the form of a rule-of-thumb rather than bound to specifics of the narrative.

Table 23

Examples of 5th Grade Responses on ERA and Interview for High and Low Scorers

Student	ERA Answer	Interview
Sophia (High ERA)	Because I also believe it is important to know how the mosquitos find food. I also believe all the tests were important because they all gave scientists a clue to how a mosquito eat & react to heat & carbon dioxide etc. (3)	Because like you need to know how like they react to all the things like how like they would react to like carbon dioxide and all like the other things and like umm if they would like change in like a different environment like a rainy or something environment. (E)
Ethan (Low ERA)	Because they needed to understand the other tests to move on to the next ones. (2)	Umm you need to do more tests instead of just one...so umm because you might not have all the data and stuff to find out. (E)

The second item was also aimed at the relationship between the way the test set is connected and builds evidence, and students had to consider whether the additional test would provide beneficial information in light of the existing tests and evidence. Scores of a three or four were assigned to answers that referenced aspects of a specific test that illustrated mosquitos use visual information already. However, students that argued for the value of the added test based on the knowledge gained about the extent mosquitos rely on visual information were also scored in this range. We felt the focus on the depth or degree mosquitos used vision represented a unique perspective on the question that warranted a higher score. Answers coded as no understanding or beginning understanding either restated information found in the text or supported their answer by referencing surface level judgments such as the additional test is better because the test set would be comprised of a greater number.

The lone high-scoring student specifically referenced the importance of understanding how much mosquitos rely on visual information as the basis for their support of the additional test. As can be seen from Table 20, the ERA answer and the interview response both target the informational value of the additional test. The elaborations in the interview extend to also incorporate the benefit of understanding the degree to which mosquitos depend on each of the targeted senses. However, this point resulted directly from a prompt by the researcher, so it is unclear if this information would have still surfaced without it. Low scoring students provided answers that remained at a surface level. The seventh-grade student rejected the benefit of a vision test but provides a level of information that does not appear to extend beyond the question text. Their interview elaborations support this, but the process of discussion also shows some positional vulnerability when the student briefly struggles between supporting and rejecting the

additional test. The fifth-grade student also makes elaborations in their interview response (see Table 24) that support their ERA answer, and while they correctly note variations to the test set would result in changes to outcomes, their response is not connected to the study. Rather these students appear to be advancing a general rule (more tests equals more results equals better information) without any connection to the context of the science narrative.

Table 24

Examples of 7th Grade Responses on ERA and Interview for High and Low Scorers

Student	ERA Answer	Interview
Charlotte (High ERA)	Because the scientists didn't give the mosquitos a chance to just use sight to see their reaction. The mosquitos might not use eyesight very much or they might but they won't know because there wasn't an individual test. (3)	Yes, umm…the scientists didn't give the mosquitos a chance to just use the sight to see their reaction. The mosquitos might not use eyesight very much or they might use it much but they won't know because there wasn't an individual test for just eyesight. *Researcher*: how would a test of vision contribute? umm I think it could contribute cause then they would be able to know if umm the mosquitos based their senses primarily off just smell and heat or if they had to use eyesight to be able to I guess target their uhh prey. (E)
Michael (Low ERA)	Because you wouldn't need to do another test if has already been proven. (2)	It would be unnecessary because you already have all those facts proven you just wouldn't-well it's kind of safe to do another test but like if you already have the facts proven you really wouldn't want to do another test. (E)

Variables Selection and Evidence. This item specifically targets the evidential aspect of variable selection and justification, and students had to consider whether adding the variable would be justified and represent a contribution to the collection of existing evidence. Scores of a three or four were assigned to answers that referenced a specific test outcome as a defense for the focus on CO_2 or provided an answer incorporating other aspects of the narrative such as the outline of past research that established CO_2 as a relevant focus variable. Scores of one or two were given to answers that restated textual information or provided answers advocating for other gases, which indicated a lack of understanding for why the focus variables were chosen and their connection to the evidence.

Only one interview student met the conditions to be considered high scoring. This seventh-grader dismissed the addition of oxygen as a focus variable based on mosquitos'

biologically-based sensory abilities (see Table 25). Their interview response elaborated on this by referencing the past research outlined in the narrative. The low scoring students were persuaded by the suggestion to include oxygen as a focus variable and either made surface level connections (e.g., more is better) or attempted to demonstrate the benefit of adding oxygen as variable (see Table 26). For example, the interview elaborations of the low scoring seventh-grader suggested that the ubiquitousness of both gases (CO_2 & Oxygen) is cause for examining

Table 25

Examples of 5th Grade Responses on ERA and Interview for High and Low Scorers

Student	ERA Answer	Interview
Sophia (Low ERA)	Because they could have done more tests & got more info and I believe when they don't test something because they assume it they don't exactly prove it to be true so it's like stopping in the middle of a test. (2)	umm, maybe because they should have done both tests because that's what he's saying and like maybe they'd have different results if they did both tests and they'd have more like results if they did the test with the umm (visual?) information and one with the other thing that they did so they didn't have to like umm I don't know may-maybe got like they would get more precise information. (E)

Table 26

Examples of 7th Grade Responses on ERA and Interview for High and Low Scorers

Student	ERA Answer	Interview
Harper (High ERA)	Because mosquitos smell carbon dioxide so there's no point to use oxygen. (3)	well, I think if you go back to the story it says that they-ca like okay they use carbon dioxide to find them so oxygen wouldn't really contribute to the study because they don't use it to find food they use carbon dioxide. (E)
Elijah (Low ERA)	Because they only used carbon dioxide in their research. Because wouldn't it be important to see how mosquitos react to other gases. (2)	Because oxygen is all around us but so is carbon dioxide and if you add those two together-like if you had both of those in the same room it could really change the behavior of the mosquitos but if you just had oxygen maybe they wouldn't be able to like see-see or smell better and like see better or smell better or even see or smell worse in that study. (E)

mosquitos' response to oxygen, a point which appears to be self-refuting given that would mean oxygen was also present in the experimental tests. Perhaps the student viewed the lab setting as devoid of oxygen, in which case their position becomes more understandable. Regardless, their elaborations still fail to notice critical aspects of the study related to the selection and justification of focus variables and the evidence.

This item was also aimed at variable selection and justification. As noted previously, it was the only question comprised of multiple choice and constructed response. The decision to structure the question this way was to lessen the cognitive burden on students. A three or four was assigned to answers that agreed knowing how mosquitos respond to variables of size and or heat would provide important information or identifies the benefits in the additional granularity and provides a comparative example illustrating the advantages. A one or two was given to answers that either restated information in the question text or dismissed the additional variables as irrelevant.

On this question, high scoring students recognized the benefit of adding the variables and were able to provide examples. For example, the high scoring seventh-grader agreed adding the variables was beneficial and provided a comparative example in both their ERA answer and their interview elaborations. Interestingly, the example in their ERA answer only focuses on the amount of heat whereas their interview response incorporates an example considering both. Low scoring students selected the letter associated with the position that the additional variables were irrelevant and did not exhibit an understanding of the benefit to adding the additional variables in the written portion. However, both low scoring students changed their answers in the interviews to reflect the advantage of adding both variables (see Tables 27 and 28). Note how the seventh-grade student begins with a version of more equals better and through the sequence of discourse integrates both variables into their elaborations. Likewise, the interview elaborations of the fifth-grade student integrate the added benefit of size and heat through the process of discussion. Moreover, they were both able to articulate examples that incorporated heat and size. This suggests that through a process of discussion where students are prompted to reflect on their answer at strategic times, they can consider the connection between variables and their impact on evidence more fully.

Table 27

Examples of 5th Grade Responses on ERA and Interview for High and Low Scorer

Student	ERA Answer	Interview
Sophia (Low ERA)	Brian, b/c I believe you need to try everything like in all subjects in school you need to find everything before you answer or do a research paper. (2)	Like so like if you were to run an experiment and like just do 3 tests and there's another test you could do you probably would want to do the other test it might change your results. (E)

Table 28

Examples of 7th Grade Responses on ERA and Interview for High and Low Scorers

Student	ERA Answer	Interview
Elijah (High ERA)	I think it is the best one because how would a mosquito react to a rat giving off a little heat. Then see how a mosquito reacts to a horse giving off a lot of heat. (4)	I think b is the best choice because if let's say the mosquito's in the desert and the animal is either really big or has a lot of heat then it'd be able to ha-it'd be able to see it like visually…if they have heat vision so it'd be-so it'd have a better chance of finding food but if it was really small and it has less heat it'd be really ha-it'd be harder to find the blood of that animal. (E)
Harper (Low ERA)	C, the mosquitos only use heat to detect & blood (food) is blood for the mosquitos. (2)	Maybe it is beneficial to try new thing because I mean it would show you if it actually helped or not. *Researcher:* Is size important? H: umm-hmm *Researcher:* Is heat important? H: I think...well if they use heat to find their animals I think maybe if there's more heat they can maybe find them more without effort maybe if there's less they can't find them it might be the other way around it just depends on the test...and then as I said earlier the size maybe they're more attracted to like big animals because their visuals smaller animals are not as like pointed out to them. (C)

Interpretations and Conclusions. This item targeted the evidential interpretations and conclusions of the science narrative, and students reasoned about whether sample limitations were a legitimate concern and if they were, did they carry over to the evidence and the conclusions that could be drawn. Answers coded as intermediate or advancing demonstrated an understanding about the relationship between the sample and the evidence by referencing sensory abilities that may vary between diverse types or demonstrated the impact of a non-representative sample on the quality of evidence. Scores of one or two provided answers that either restated information found in the text or supported their answer by focusing on low level agreements such as the scientists should have tested more mosquitos because they would have more information.

Table 29 illustrates the lone high scoring student was persuaded by the argument that the evidence was limited. The basis for their agreement was 1) there are diverse types, and 2) their sensory abilities are likely different. In their interview elaborations, they cited previous knowledge from the news, science class, and their assumption distinct types exist as the foundation for their answer. When prompted for an example of how varied types of mosquitos

Table 29

Examples of 5th Grade Responses on ERA and Interview for High and Low Scorers

Student	ERA Answer	Interview
Isabella (Low ERA)	C, scientists are focusing on what instinct they use to find food. Not to see what size animal they prefer. (2)	Well because if they only change the size the size doesn't really matter cause but the body heat it-it says in the message that body heat was like one of the main things that mosquitos-cause if they can smell that and the-they can see the body heat they would want a lot of body heat *Researcher: oh so that would be different from what you circled here right* I: yeah *Researcher: so, are you changing your mind so that now you think possibly the amount of heat an animal puts off would be good information to know* I: yeah, yeah. *Researcher: can you give me an example of how you think* I: yeah, if you put in a small cow...you use the 2 fake cows 1 of them is really and the other is really big and the small one has very little body heat and the big one has a lot of body heat. *Researcher: Can I stop your just a second. Did you notice that in your example you're varying size and heat?* I: yeah...mmm-hmm *Researcher: So now size and heat?* I: yeah, a little bit...yeah... yeah, that is also (important?) *Researcher: Can you tell me what happened that made you change your mind?* I: Yeah... well if you look at it from different angles like you just read all the (couldn't make out) but see if changing the fake animal and the amount of heat it puts off would help to answer the question of how mosquitos find food, which I thought yeah that is true. (C)

suggested these could impact how mosquitos "process" information to find food. This suggests could change the results, the student provided sensory variations (hearing and seeing) and the student viewed sample characteristics and evidence as interrelated. Low scoring students either dismissed the limited sample position or supported it based on the idea that more experiments could have been done (see Table 30). The seventh-grader's response show they considered the limited sample claim and developed a rationale that, if true, would potentially undercut the assertion that flaws in the sample limited the evidence. It also implies the number of mosquitos in the study is acceptable, albeit never explicitly stated. They also note the additional complexity of experimenting with diverse types (sorting them) suggesting that even if mosquito populations were comprised of several types of varied abilities, scientists would not be able to study them individually. The low scoring fifth-grader, on the other hand, provides elaborations not

immediately relevant to the question and instead focuses on ideas of quantity. They reference the idea that a high number of tests results in the "right answer." Taken together, these responses reveal that low scoring students struggled to reason about how the experimental sample is related to evidence.

Table 30

Examples of 7th Grade Responses on ERA and Interview for High and Low Scorers

Student	ERA Answer	Interview
Elijah (High ERA)	Because there are a lot of types of mosquitos & each type would probably react differently than the other types of mosquitos. (3)	Because wi-because there's a lot of various different types of mosquitos around the world like even in a-in Indiana there's multiple types and each *Researcher: Do you know that there are different types?* E: I-both-both because you hear about all different types of mosquitos on the news and-and science class also so I somewhat know that and you just kind of assume that there's multiple...types I think that-I think that would *Researcher: You think different types would respond differently?* E: umm-hmm...yeah *Researcher: Can you give me an example of how you think a different type might respond differently?* E: Maybe that certain type has a different way of hearing or seeing so they have to process it differently and not-and not they find food and wa-like food and oxygen differently from other various types. (E)
Harper (Low ERA)	I think that mosquitos are pretty much mosquitos & that they're pretty much the same. (2)	well I think mosquitos are pretty much mosquitos and you're not I mean how you gonna like sort them out like I don't think that's really possible...but umm no because I think they all have like the same focus. (M)

This item also targeted evidential interpretations and conclusions. Here the crux of the debate for students to judge had to do with the issue of error with respect to the computer and determine whether it had any influence on the evidence and conclusions. Answers coded as a three or four made explicit connections between the accuracy of the computer and evidential quality or exhibited their understanding through detailing acceptable levels of accuracy (e.g., > 90%) within the context of the study. Similar to preceding questions, answers given a one or two either restated text found in the question or made claims that computer accuracy is irrelevant to the evidence.

Tables 31 and 32 contain data from high and low scoring students. High scoring students provided answers that highlighted the connection between the computer accuracy and the evidence. They correctly noted computer inaccuracies would result in the reduction of evidential

quality. With the assistance of interview prompts, the seventh-grader was also able provide elaborations that extended beyond the general issue of computer inaccuracy to focus on specific considerations such as the amount of CO_2 released, which could potentially have substantive effects on test outcomes. Additionally, while neither of the high scoring students offered an acceptable level of accuracy on their own, they did provide them when prompted (both suggested > 99%). Low scoring students did not reference the connection between computer accuracy and the quality of the evidence. Rather they dismissed the relevance altogether or focused on unrelated factors. For example, the seventh-grader took issue with whether the computer provided details on the general or specific route of mosquitos, while never revealing defining features of either one or how evidence comprised of specific routes was preferable. Even when prompted to review the portion of the science narrative containing this information, the student remained focus on the issue of specific flight paths and ignored the issue of computer accuracy and the evidence. The low scoring fifth-grader suggests the computer had no effect at all on the evidence. Thus, while some students recognized the connection between the computer accuracy and the evidence, no one demonstrated a more in-depth understanding by providing an acceptable rate of error or noting the complete elimination of such error is not possible. Further, none of the students referenced the importance of considering these types of issues early in the design process.

Table 31

Examples of 5th Grade Responses on ERA and Interview for High and Low Scorers

Student	ERA Answer	Interview
Sophia (Low ERA)	Because the scientists didn't do as many experiments as they could and limited their work. (2)	She was saying that like umm she thinks that the scientists need to like do all their tests for like finding food and like (couldn't make out?) not one individual...and like I don't know like I've been saying they need to just like look at all of the tests and do all of the tests so they can get uh right answer. (E)

Relationships. This item pair was designed to probe students' broader understanding about connections between distinct phases of scientific inquiry. Students had to consider if the change would impact the test outcomes and if so, to reflect on how they would be impacted.

Table 32

Examples of 7th Grade Responses on ERA and Interview for High and Low Scorers

Student	ERA Answer	Interview
Harper (High ERA)	Jackson, because if the computer isn't accurate then it won't be accurate information. (3)	yeah, Jackson is right because if-the-cause calculators and computers can mess up things a wrong-wrong sometimes and not everything is-you can't believe everything that's on a computer and when if-if it is messed up you get very wrong calculations and it could change the entire en-the entire like answer you could get from…from your study *Researcher: So, if the computer is not accurate-that could impact the accuracy of the scientists' conclusions?* umm-hmm because if the computer-like I'm pretty sure in the story the computer they typed in like how much oxygen if the computer puts the wrong amount of oxygen it could-well carbon dioxide...it could change the way the mosquitos act *Researcher: Do you have an idea how accurate the computer should be?* It would have to be fairly accurate *Researcher: So, when you say fairly, is that 95%* No, like 99 *Researcher: oh, it needs to be higher than 95%* 95 isn't very accurate. (E)
Charlotte (Low ERA)	Because the scientists didn't say if the if the computer tracked the mosquitos or if it just gave their general route. (2)	When the scientists were trying to track the mosquitos they didn't really give the I guess the most specific route which would have been valuable because then the scientists could see what like mosquitos patters were in order to CO2 like so they sense CO2 while they're were flying like would they turn towards or would they like try to find another route and they kind of just gave a general route so they didn't really like know if the mosquito was going to a specific place like multiple times or just like flying everywhere *Researcher: (showing student flight pattern pic on front page of narrative) So, you didn't think for example that on the 1st page that this generated image that's showing a flight path you didn't think that was enough information?* I guess I didn't think it was enough information umm because they only like I guess the general kind of direction he was going in-like-like what like the path he was taking. (E)

Answers scoring a three or four acknowledged change will occur and used a relevant example to support their position (e.g., environmental differences of natural environment) or provided an answer highlighting relevant issues such as the added complexity that would result from the change such as controlling variables. Answers given a one or two either restated information in the text or cited changes will take place without any other details.

Tables 33 and 34 contain the response by high and low scoring students. High scoring students recognized the connection between features of the experimental design such as location and the evidence. These students were also able to identify and articulate at least one area of the design that would become more difficult as a result. During their interviews, high scoring

students provided elaborations containing additional details about the about the difficulties that would ensue from the changes. For example, both students noted variables that would present measurement challenges in a natural environment. Low scoring students, on the other hand, did not. In fact, they (all 5th graders) suggested no changes whatsoever would result from altering the test environment.

Table 33

Examples of 5th Grade Responses on ERA and Interview for High and Low Scorers

Student	ERA Answer	Interview
Isabella (High ERA)	Because the computer could have glitched out and missed important evidence. (3)	Cause if the computer blacks out at some point and there was a really important evidence during that black out with the mosquitos and then it just turns back on they lost all that good evidence and they can't get it back because the computer blacked out *Researcher: So, you think how good your conclusions are have a lot to do with how accurate the computer is* yes *Researcher: how accurate do you think the computer has to be?* it should be at least 99.5% accurate...yeah *Researcher: That would mean in a hundred times the computer may be wrong ½ in those hundred. Is that acceptable?* yes...well no, not like no it shouldn't that can't it-it's not really acceptable because if it was wrong all those times then they're getting (wrong?) information *Researcher: I think I may have confused you. Let's say 99% instead of 99.5%. If you did 100 tests, the computer may be incorrect 1 time and right 99 times. Is that a good amount of accuracy?* well if you wanna get it right completely you have to have it all right like if want-if you're studying for a test and you're going for an A you're gonna wanna get all of it right not...you want an A+ 100%...so it should be 100% correct all the test should be correct. (E)
Ethan (Low ERA)	Because it has nothing to do with it. (1)	Because [student in mock debate] was doubting the quality of the evidence. (M)

This final item also targets students' broader understanding about the relationship between phases of inquiry. Here, students were asked to think about the relationship between variable selection and experimental design. Scores of a three or four were assigned to answers that noted changes would take place and related those changes to the design (e.g., would need to make a test for that) or also makes the connection between the new tests and increased outcomes/evidence. As with the other questions, scores of a one or two were assigned to answers

that either restated information in the text or agreed that changes would occur but provides no other details.

Table 11

Examples of 7th Grade Responses on ERA and Interview for High and Low Scorers

Student	ERA Answer	Interview
Charlotte (High ERA)	Yes, b/c the amount of heat & co2 levels would change tremendously, the visual sightings would vary greatly, and they wouldn't be able to control the senses they wanted mosquitos to use. (4)	umm I think change (would?) happen because if you change the amount of heat and carbon dioxide it could umm I guess make the mosquitos not as attracted to the object they were trying to attract them to or it could make them so attracted they couldn't really I guess like follow what the mosquitos were doing so I think umm I think the levels they did were pretty good but I don't think they should've changed them because the mosquito would have been either too much attracted or not attracted enough to the object du- due to not having its senses heightened or over-heightened *Researcher: Do you think doing this experiment in a natural environment makes the experiment more complicated, less, or the same as it was in the lab?* I think it would make it more complicated cause I think there would be other factors that would umm attract the mosquito instead of focusing on the object they were trying to attract the mosquito to. (E)

Note: No low scoring 7th grader on question 7

Similar to the previous question, high scoring students articulated the relationship between variables in a study and the design of tests. Specifically, they identified that if additional sense cues were discovered (e.g., hearing), scientists would need to design a test to investigate it. Further, they were also able to connect the additional tests to a higher quantity of outcomes (see Table 35 and 36). In the interview, the high scoring student provided elaborations supporting their ERA answer and connected mosquitos' sensory abilities to their behavior illustrating the relationship between the selection variables in the design of experimental tests and the generation of evidence. Low scoring students did not highlight the relationship between variables and test design. Rather, they either focused on surface level connections (e.g., changes here changes there) with no other details, provided an example of a sense ability already targeted in the study, or were not persuaded changes would occur. During the interviews, however, two of the low scoring students made the connection between the variables and the test design. In particular, one of the students changed their answer entirely and formulated a relevant example using a new sensory variable. While the student was unable to make the connection themselves,

the process of discussion was critical in developing their ideas and allowing them to consider the connection between adding variables and test design more fully.

Table 35

Examples of 5th Grade Responses on ERA and Interview for High and Low Scorers

Student	ERA Answer	Interview
Isabella (High ERA)	If they watched them in nature then it would be hard to track the mosquitos. Plus they would of got bit and have to itch all over. (3)	It'd be harder to track because if they take a little snooze and the other-they had 2 people and they were both watching them (one of them says can you watch them I need to fall asleep?) that person also falls asleep they're gonna have to go all over again and try to find them *Researcher: Can you think of anything else that would be hard to do that they did in the lab that would be really hard to do out in nature?* seeing how they tested the body heat of the animals like the body heat and the size the re-the sense of direction it's gonna be hard without any of that equipment (couldn't make out?)...and plus if they're out in nature they're gonna get bit by something. (E)
Ethan (Low ERA)	It would not because some questions might not be needed. (1)	I don't know. (N/A)

Table 36

Examples of 7th Grade Responses on ERA and Interview for High and Low Scorers

Student	ERA Answer	Interview
Elijah (High ERA)	I think it would influence the tests because if there were more senses the mosquitos had the more tests they would have to do. And if they had to do more tests there would be more outcomes from the tests that they would have to find. (4)	I think it would make a difference because if there was additional senses...at and like if they were all the other-by themselves it could really change the way mosquitos act because in nature there's always different things going on but in that lab there was only a certain amount of things they could use...but with more things to use like oxygen for example and then like maybe adding different animals you could change the way the mosquitos act. (E)
Michael (Low ERA)	yes, because everything affects something in different ways. (2)	Yeah because like everything is like it's affected in different ways and you would wanna like have a variation of like things that affected so (talking to himself but couldn't make out) you kinda wanna it-everything just like affect stuff in different ways so you'd wanna have (seems frustrated) if everything affects it...(stuff?) in different ways that's basically what I'm saying like then you would find out how stuff affects it-(almost whispering: if that makes-I don't know). (E)

Taken together, ERA results and student interviews reveal that whether reasoning about variables, an experimental test set, evidential interpretations, or the interrelated nature of discrete phases of inquiry, students from both grade levels struggled with these aspects of the conceptual framework in the context of the science narratives. At an aggregate level, for example, approximately 63% (337 out of 536) of all ERA responses were scored as a 1 or a 2 (no or beginning understanding) compared to just 36% (192 out of 536) as intermediate or advancing. When examining the cumulative distribution of these percentages across the samples, however, seventh-grade students were responsible for 75% of the intermediate or advancing codes, while fifth-graders provided 52% of answers coded as beginning but 92% of the no understanding category. This suggests that while students exhibited difficulty reasoning with and about evidential aspects of the framework generally, seventh-graders displayed more evidentiary knowledge compared to fifth-graders. Students in the fifth-grade sample consistently provided answers that could be grouped into three main categories: 1) simple judgments such as more equals better, 2) mirrored text found in the question, or 3) irrelevant observations. While seventh-graders scored higher generally, many of them scored in the upper coding levels (e.g., 3 or 4) on one question and then reverted to framing the issue as a simple comparison of quantity or providing simple answers on another. For example, Chloe rejects the idea that the scientists should have included oxygen as a focus variable on question three and justifies her answer by referring to a test outcome that demonstrates its irrelevancy. However, when reasoning about whether technologically-based error was a legitimate concern for the evidence in the mosquito narrative, she provides an answer that does not illustrate an understanding of the issue. These results suggest seventh-grade knowledge of scientific evidence is tentative and not well formed with respect to some facets and more developed on others.

Contextual Variables

Reading Ability. With respect to reading ability, the main differences were at the fifth-grade level. For example, there was a greater percentage of these students either at or below grade level reading (see Table 37) than the seventh-grade students. Data shows 1/3 of fifth-grade students were rated by their classroom teacher as below grade level reading compared to 0 for seventh-graders. Further, a full 2/3 of seventh-graders were above grade level readers compared to 37.5% of fifth-graders. Examination of fifth-grade performance differences between students

categorized as above and those at below grade level reading illustrate the way reading ability may have contributed to lower ERA scores (see Figure 7). These results may partly explain the higher difficulty indices for the fifth-grade sample in pilot data. Moreover, the increased difficulty of the science narratives may have contributed to students performing in ways that are not reflective of their understanding.

Table 37

Examples of 5th Grade Responses on ERA and Interview for High and Low Scorers

Student	ERA Answer	Interview
Isabella (Low ERA)	It would not influence the test, it would just be useless information. (1)	Because it'd be information that they didn't need even though having more information would be a good thing. If there's information that they didn't really need you didn't really need that. Having that would be if they didn't have any other senses that would just be information that they didn't really need. And if they did then that-they should also test with that information so it's kind of a both and both it would influence and it wouldn't influence *Researcher: Can you think of another sense that maybe mosquitos could have* umm…hmm (pause) I don't know if taste would be one like if they tasted something and that would've led them to food I guess *Researcher: How about I make a suggestion. What if the sense cue was hearing? Would that change the tests they did?* oh yeah (sure?)...yeah because if they heard something move or if there was a lot of motion movement that would've been-they would've heard that and they would've known oh that's food I want food so they would have found that *Researcher: What would the scientists have to do differently?* they would've had to test the hearing like have something I guess like very invisible like hard to see and then they would have had it like move around a little bit and see if the mosquitos would be able to find it. (C)

Task Differences. Analyses revealed a statistically significant effect of the task on total score, $F(1,62) = 42.04$, $p < .004$ at the $\alpha = .05$ level. These variances can also be seen by comparing the mean scores of the narratives. For example, there is less than a point difference between the fifth-graders mean scores on the mosquito and Acacia tree narratives, but the contrast is more than three points between the same narratives for seventh-graders (see Table 38). Though care was taken to structure the tasks similarly, the context and purposes of the narratives were different, which naturally led them to be comprised of distinct experimental methodologies and constraints. Further, research has demonstrated students' misconceptions about plants (e.g., Stavy & Wax, 1989). More specifically, students generally do not think of

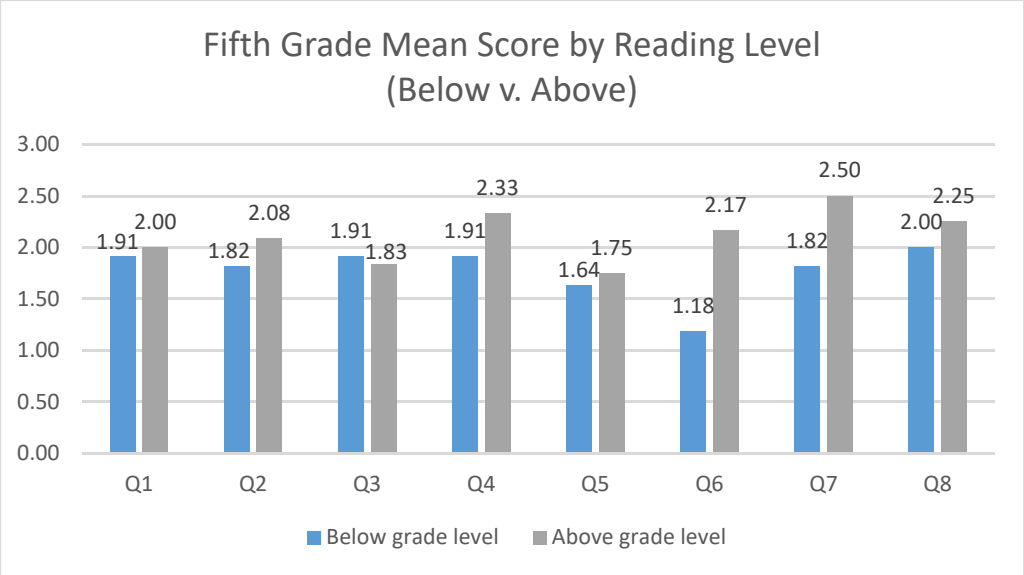


Figure 7

Fifth-grade Comparison of Reading Ability and Item Scores

Table 38

Reading Level Data

Grade	Above Grade	At Grade	Below Grade	Total
5	12	9	11	32
7	23	12	0	35

plants as living organisms that can respond and adapt to their environments. Another valid consideration is the absence of experiential knowledge of key aspects of the Acacia narrative. This includes the Acacia tree itself, the other animals (impalas, wild dogs, leopards) that contribute to its environment, or the processes and ways the tree changed to protect itself from predation. The same cannot be said of the mosquito narrative. For example, while it is doubtful many of the students had normative ideas about how mosquitos apprehend sensory data to locate food, aspects of the mosquito narrative contained several familiar elements such as what a mosquito is and that they rely on nutrients found in blood to stay alive. They were also familiar with the senses examined in the narrative (smell, touch, vision).

Assessment of Science Interest. To examine the effect of the assessment of science interest, a correlational analysis was computed to determine the relationship between each of the

questions to total ERA score. As shown in Table 39, the correlation between ERA total score and the statements science is interesting, I am good at science, and I liked the science story I was given was weak, thereby indicating the science interest questions played an insignificant role in the students' overall performance.

Table 39

Mean Differences by Task and Grade

Grade	Mosquito	Acacia Tree	Difference
Fifth	15.5625	14.7500	0.8125
Seventh	22.4118	19.0556	3.3562

While reading ability and differences between the science narratives appear to have influenced the variations in evidentiary knowledge both between and within the grade levels, teachers and the character of instruction students receive are two key facets in the development of robust notions of scientific evidence that have yet to be examined. The following sections present and discuss result from the teacher interviews and in-class observations.

Instructional Variables

Teacher Interviews. Semi-structured interviews were conducted with participating teachers (N=4) to gain insight into aspects of instruction related to scientific evidence. The questions addressed the following themes: 1) background and experience, 2) instructional methods, 3) instructional time and the nature of investigations, and 4) views on learning science and evidence. The results are presented and discussed below.

Background and Experience. Given the relationship between training and experience on instruction, the initial questions targeted teachers' educational backgrounds and experience teaching. Across the group, only one of them earned an undergraduate degree in science (biology). The other seventh-grade teacher completed an elementary education program and then received certification for middle school licensure later. Of the two elementary school teachers, one obtained an elementary education degree and the other earned a degree in kinesiology. All but one of the teachers earned a graduate degree. Seventh-grade teachers averaged 11.5 years of experience and fifth-grade teachers averaged 27 years (see Table 40).

Table 40

Correlation Table

		Science is interesting	I am good at science	Case was interesting	Total Score
Science is interesting	Pearson Correlation	1	.598**	.662**	.107
	Sig. (2-tailed)		.000	.000	.387
	N	67	67	67	67
I am good at science	Pearson Correlation	.598**	1	.262*	.132
	Sig. (2-tailed)	.000		.032	.286
	N	67	67	67	67
Case was interesting	Pearson Correlation	.662**	.262*	1	-.138
	Sig. (2-tailed)	.000	.032		.264
	N	67	67	67	67
Total Score	Pearson Correlation	.107	.132	-.138	1
	Sig. (2-tailed)	.387	.286	.264	
	N	67	67	67	67

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

Instructional Methods. The next series of questions focused on how teachers described their science instruction. Specifically, teachers were asked what they enjoyed about teaching science, how they would describe their science teaching, and to provide examples. Teachers unanimously highlighted the active and hands-on nature of science as major components of enjoyment. Moreover, they identified scientific investigations as an integral part of their instruction and promoted lively views of their classrooms. When describing examples of their science teaching, they all provided examples. Mrs. Samuels recapped an ecosystem activity where students looked up examples on iPads and then presented their findings back to the teacher. Mrs. Murray and Mrs. Keck (see Table 41) provided detailed responses about beginning of the year activities designed to introduce students to science process skills (e.g., observing and/or measuring qualities or quantities, sorting/classifying, inferring, predicting, etc.). Within these descriptions, students worked in problem-solution frameworks that require the application of key scientific practices such as collecting data and formulating evidence-based conclusions.

In Mrs. Murray's example, her seventh-graders collaboratively developed a list of outdoor games and identified relevant variables. They were also permitted to choose a game to investigate and were able to carry out a test of their ideas regarding variable change. This sequence allowed students to develop ways to examine how changes to particular variables

Table 41

Teacher background and experience

Name	Grade	Education	Experience
Murray	7	Biology, MS (Secondary Ed)	18
Carter	7	Elementary Education	5
Samuels	5	Elementary Education, MS (Education)	40
Keck	5	Kinesiology, MS (Education)	14

impacts their chosen game. Further, it allowed students to compare their initial ideas with experimental outcomes. The decision to introduce scientific practices by embedding them within a familiar topic (games) has the potential foster meaningful learning and transfer. Mrs. Keck's activity capitalized on the way science, as a method, can be used to approach problems and her innovative way of framing the problem is a creative way to generate student interest.

Instructional Time and the Nature of Investigations. The following questions were designed to gain insight into the time students spent in science instruction and scientific practice. These items were broken into 1) how much time is spent each week teaching science, 2) the frequency with which students participate in scientific investigations, 3) what a typical investigation looks like, 4) the duration of investigations (e.g., 1 class session or more), and 5) how often students reason with and about scientific evidence.

With respect to the amount of time spent each week teaching science, it is important to note the school corporation where one of the fifth and both seventh-grade classrooms were located defines elementary school as kindergarten to fifth-grade and middle school as grades six through eight. As such, students experience subject-specific teachers at the middle school level rather than one teacher for all subjects as is the case in elementary school. This likely results in students at the different grade level groupings receiving diverse amounts of weekly science instruction. According to the seventh-grade teachers, students spend an hour every day in science, while Mrs. Samuels teaches science content for three weeks of each nine-week rotation. During science rotations, students receive an hour day for a total of 15 hours. Taking the entirety of the rotation into consideration, this averages out to about 1 hour and 45 minutes a week. The other fifth-grade class was a part of district that structures grade five and up in a similar fashion as the middle school described above. Thus, Mrs. Keck's students receive about four hours of science instruction per week. In total, the seventh-grade teachers and Mrs. Keck reported similar

amounts of weekly science instruction and although Mrs. Samuels dedicated similar amounts of time when teaching a science rotation, the rotational schedule dictated she spend less overall time teaching science.

A key component of developing students' knowledge of scientific evidence is for them to have sustained opportunities to participate in investigations where they can reason with and about evidence. In the context of this question, a scientific investigation was defined as an activity that required students to apply the science process skills such as asking questions, making predictions, observing and/or measuring quantities or qualities, and developing evidence-based conclusions. With respect to how often students participated in scientific investigations, there were disparities once again between fifth and seventh-grade classrooms (see Figure 8). For example, Mrs. Murray projected half of her total science instruction per week was spent engaging in investigations, and Mrs. Carter reported an hour a week. The fifth-grade teachers

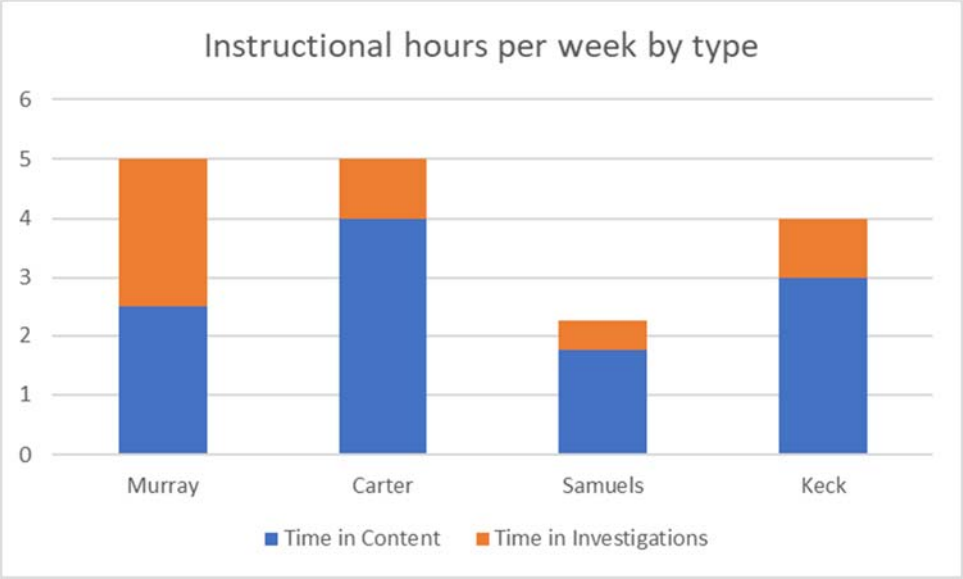


Figure 8
Distribution of Instructional Time

estimated their students spent an hour (Mrs. Keck) and less than an hour (Mrs. Samuels) per week working with investigations. Averaging these times, seventh-grade students spend more than twice the amount of time participating in scientific investigations than fifth-graders (1.75, .75 respectively). Over the course of a 36-week school year, this amounts to approximately 36

additional hours of investigative time for the seventh-grade classrooms. This estimates to be a considerable advantage for the older group.

The follow-up question probed how much class time was typically spent carrying out an investigation. The key idea being that the duration of the investigation can be a reliable indicator of how deep the teachers wade into aspects of investigations such as evidential features related to experimental design or evaluating aspects of outcomes. For example, an investigation spanning a single class meeting is likely to be limited in its ability to engage students with focal concepts such as operationalizing variables or productive discussions regarding sample characteristics and the way these aspects relate to evidence. Of course, this question includes an explicit recognition that teachers and their instruction are driven, in many cases, by the character and scale of content. Thus, a lesson or unit on the solar system presents investigational challenges that instruction on ecosystems or the water cycle does not.

Mrs. Samuels reported investigations that were contained within a single class. The remaining teachers reported varied lengths of time. Mrs. Keck, the other fifth-grade teacher, noted her class spent two to three sessions per investigation. However, this increase can be partly explained by the fact that her science classes meet for 50-minute periods, and the short duration of the class period may naturally cause investigations to spill over into other sessions. For comparison, Mrs. Carter and Mrs. Murray's classes meet for 75-minute periods. The seventh-grade teachers also reported variation in the amount of time spent on investigations, but further added that the duration was dependent on the type of investigation. For example, labs or closed inquiry investigations are completed in one or two class meetings. These investigations are marked by step-by-step instructions that lead to a pre-determined outcome. Students engaged in these activities approximately once a week. Full or true investigations require a minimum of three class periods. Mrs. Carter articulated the defining features of a true investigation as "…them [students] using their understanding of a topic to come up with their own question and then carrying out their own experiments." The seventh-grade teachers reported engaging students in this type of investigation once a quarter (3-4 times a year). For the most part, students in these classrooms typically participated in investigations that spanned one or two class meetings.

To understand the primary characteristics of investigative activities, teachers were asked to describe a typical investigation and to provide electronic examples. Fifth-grade teachers tended to describe what could be considered as a highly structured. Characteristics of these

investigations consist of following a series of steps to produce a pre-determined outcome and/or generating evidence-based statements from a single observation. More complex investigations, on the other hand, incorporate aspects such as developing questions, setting up experiments, collecting and analyzing data, and highlighting the relationship between these facets and scientific evidence. Note the comparison between Mrs. Keck and Mrs. Murray's descriptions in Table 42. Due to the emphasis on relatively straightforward observations, Mrs. Keck's investigation is best characterized as highly structured. The description of Mrs. Murray's mealworm activity is representative of a more complex investigation. Across these descriptions, there does not appear to be any dimensions of evidence from the framework present.

Table 42

Excerpts of teacher instruction examples

Teacher	Example
Mrs. Murray (7^{th})	…At the beginning of the year one of my favorite things that I do in introducing scientific processes skills, thinking, and methods I let the kids choose a game. We brain storm and we list all the games you can play outside on the board -whatever-when we list as many things as you can do outside then I let them choose whatever one they want to play and then they have to pick a variable to change. Does changing the size of the bottle increase your chances of throwing the ring on it. Does the distance that the cornhole things are away from each other affect how many that you get-ya know-umm, does using soccer balls with different pounds of pressure-this is all the things they've done this year-does filling a soccer ball with different amounts of pressure change the way or how far you can kick it. Umm, so they bring out a balance, they bring out a pressure gauge, they bring a bicycle pump, they're measuring the mass, they're measuring the volume-ya know-umm, does your athletic ability affect how well you can ride a hovercraft. You know so maybe you have athletes ride it or whatever, so they designed all these experiments. You know, what did you learn from it. You know, and then if we have time, change it-you know-after they find some conclusion make it work. And then I always tell them, this is the only class you can totally fail in whatever you're working on and still get an A+ cause it's about-you know-that you learn it.
Mrs. Keck (5^{th})	For an example one of the things we liked at the beginning of the year we kind of work on trying to follow the steps of the scientific method and we tell them that those can be interchangeable that scientists do it in different ways-one of the ones I told you I do is sewer lice (chuckles) and we like say like oh my gosh the New York Police department found these bugs living in the sewer umm-uh-an-we-so I give this whole big scenario we need to figure out what it is I show them the (sign?) that was sent with the hazardous sign on it to get them int-you know interested and then really what it is is we just-I just really want them to gather information just to write a hypothesis what we think-what they think it is and then umm you know follow the steps of the procedure umm and then to collect data in a table-in an organized fashion and umm and then be able to make claims based on the evidence that they found in the lab and really it's just reasons (can't make out) but you know that they go up and down (laughs) so that they're moving an-and then they like, wait a minute I think it's this so it's just interesting to see-oh I really think it's a lot like some kind of bug you know or whatever and then we really work on if you're claiming this, what's your evidence why do you think that and so we talk a lot too about opinions in science you can't use opinions in my class they have to be-you know-you have to have evidence to back it up and then again of course okay now what's your conclusion here's what you found out how would you share that with other scientists.

The final question in this group specifically addresses how often students reason with and about scientific evidence. Since teachers can engage students with tasks and activities that lie outside the boundaries of an investigation yet may still be designed to scaffold students' knowledge of scientific evidence, this question did not couple evidence with investigations so as not to unnecessarily constrain answers. Despite this, most teachers associated students working with evidence as a facet of investigative activities. Table 43 contains sample excerpts of teacher responses.

Table 43

Sample descriptions of investigations

Mrs. Keck (5th)	Mrs. Murray (7th)
I'll maybe show a picture and say can you tell me any claims and use some evidence from this photograph or we'll even go outside like during ecology and I'll say I want you to find 5 things and I want you to tell me what you know about it and you have to have evidence to back up what you what you know.	[Describing a mealworm investigation]: Put it in a petri dish and they have to make qualitative and quantitative observations about it. Then they have to draw inferences based on their observations. Then they have to ask questions about it. Then we classify their questions either as research or investigative. They're assigned to look up so many research ones, so they can find knowledge about the mealworm then they have to design an investigation. Like how does temperature affect the mealworm? So then we put it in the refrigerator, and they have it at room temperature, then we set it on the overhead on top of the light-but-put black paper over the light on the-on the overhead projector so the light doesn't bother them, it's black and then they just move around so much more when they're warmer, you know, so they take-data of their behavior-stuff like that.

The frequency with which seventh-grade students work with scientific evidence is about once a week, which aligns with their previous answers detailing how often students participate in investigations and their duration. Fifth-grade teachers' answers were less clear cut. These teachers were unable to specify how routinely students' reason with evidence. This suggests opportunities for students in these classrooms to work with evidence may be infrequent.

While thinking through their responses, the teachers also provided glimpses into how they defined evidence. Consider the responses provided by Mrs. Samuels (5th) and Mrs. Murray (7th) in Table 44. While these responses cannot be considered a comprehensive representation of Mrs. Samuels or Mrs. Murray's views, the notions of evidence students encounter through these examples is worthy of exploration. Both teachers underscore the importance of evidence and emphasize the evidence-claim relationship, which are important for students' science learning.

Table 44

Examples of teacher responses on how often students work with evidence

5th Grade Teachers		7th Grade Teachers	
Samuels	Keck	Murray	Carter
"…Things that I teach don't always apply themselves to experiments but yeah the research they get they always have to prove to me that they've what they've learned from it."	"(pause) I don't know…(laughs) I don't know."	"I'd say every time they do an experiment [investigation] they have to… So, after every single lab they have to reflect and make some claims about what they did and then use evidence from their data to support it."	"…maybe 5 to 6 times a month… it'd be like a lab."

Views on Learning Science and Evidence. The remaining questions explored teachers' notions in more detail. These questions targeted teachers' ideas about what they wanted students to know about science and what they wanted their students to learn about scientific evidence. With respect to the question of what important topics do you teach in science, all teachers underscored the importance of helping students to view science as both meaningful and practical. However, there were differences between fifth and seventh-grade teachers (see Table 45). For example, Mrs. Keck highlights curiosity and interest as important topics to teach, and advocates her students see science as a part of their individual identity. Rather than being a distant abstraction, the objective is to make science local and accessible. Mrs. Samuels also supports a view of science as personal and relevant but connects it to students through future career opportunities. The responses of Mrs. Carter and Murray, on the other hand, draw attention to the importance of teaching scientific processes and scientific thinking. Mrs. Carter stresses the methods and evidential foundations of science and the relationship between those foundations

Table 45

Examples of teachers' definitions of evidence

Mrs. Samuels	Mrs. Murray
I give them a sheet to write on but it's a photo like a girl stomping in a puddle and I what do you notice and then they have to use that visual evidence.	I have starter sentences for them and it-they-they have to make-depending on the lab I give them a number usually I average 3 – I claim that, you claim-you claim that, it's kind of an observation, what's your evidence that you proved that happened. I claim that umm zinc uhh produces a gas when mixed with hydrochloric acid. What's your evidence, when I put hydrochloric acid on the zinc, it fizzed.

and the knowledge about the world they provide. Mrs. Murray accentuates the importance of teaching students to think like scientists through the acquisition of knowledge about scientific practice. Further, she notes the value of then applying scientific thinking to other subjects.

The final question asked teachers to articulate what they wanted their students to learn about scientific evidence. This portion of the interview was designed to determine how the teachers' viewed evidence by defining what they wanted their students to learn about it. All responses contained remarks that conveyed the importance of evidence to science and instruction. For example, Mrs. Carter referred to it as a hallmark and both her and Mrs. Keck underlined the relationship between evidence and claims (see Table 46). Other notable themes include ideas about the objectiveness of scientific evidence, and the view that scientific knowledge is tentative and subject to change. When making the latter point, teachers used the word evidence, but they were referring to scientific knowledge. For example, Mrs. Murray notes both laws and theories can be provisional and lists novel discoveries ("new observations") and advances in technology as two revision triggering mechanisms.

Table 46

Teachers' views of important topics to teach in science

Teacher	Comments
Mrs. Keck	I want them to know that they're all scientists umm and then I want them just to be interested in science and to try to discover things on their own…and discover things and I try to tell them there's probably many things tha-that are not discovered out there it could be you you could be the one who finds some things.
Mrs. Samuels	I hope they get an interest in it and pursue a career because the future's going to be technology and uh there's um there's data out there I don't-I can't give the exact percentage it changes all the time but the jobs that these kids are going to have when they're older will first of all be many jobs and secondly may not have been invented yet. So, they have to be able to-to grasp those concepts whatever they need for their job learn those things.
Mrs. Carter	It's just the investigation of finding the truth whatever that may be and that people aren't just making up science. Like we're not just making this up there's years and years of research and development that backs the things we're teaching about and learning in class and so that's something that I feel kind of passionate about right now that it's important for them to know like there's a reason we're doing this…through scie-through the scientific process.
Mrs. Murray	Science processes and skills. And I can show you this (directs me to mini posters in room that list what appear to be practices from k-12 framework). Mostly the science processes and skills. If they learn the processes and skills, how to think like a scientist, how to ask questions, how to make observations, how to collect data, differentiate that data between qualitative and quantitative. Then you just take the content and apply all those skills with the content. And then I probably say making it applicable or integrating it with the other subjects.

The teacher interviews were designed to contribute valuable information about what aspects of evidentiary knowledge were integrated into these classrooms as well as key data about how these teachers viewed both scientific evidence and science more broadly. Overall, this was an experienced group of teachers that expressed enjoyment teaching science. Their descriptions portrayed active classrooms that included productive teaching strategies designed to support student learning. Examples include connecting science to areas and subjects outside the classroom, providing opportunities for students to engage in peer-to-peer discourse about science topics, and concerted efforts to highlight science process skills and scientific thinking. Additionally, all teachers identified scientific investigations as an integral part of their instruction. For example, seventh-grade teachers reported devoting up to half of their instructional time to investigations. While fifth-grade teachers committed less of their science instruction to investigative activities, they still highlighted their importance and were committed to affording students' opportunities to participate in them.

When talking about scientific evidence, the teachers tended to emphasize claim-evidence relationships. This came through in many of their comments where teachers talked about instructing students to be sure and connect their claims to the corresponding evidence. In these descriptions, there was no discussion of how evidence is construed of a complex web of scientific practices including experimental design, data collection procedures, or various features of analyses central to examining the scope and quality of evidence. Additionally, none of the teachers talked about evidence in a way that communicated its interconnected relationship to other phases of scientific inquiry.

The following section presents data from the in-class observations. A total of four observations were conducted. Additionally, teachers provided numerous electronic copies of activities/lessons they incorporate as part of their instruction. This information is also integrated in to observation section.

Classroom Observations. One observation (N=4) with each classroom was conducted to understand how instruction about scientific evidence was enacted in the classroom and to identify what aspects of evidence from my framework were addressed. During the observations, detailed notes were taken describing lesson activities and electronic copies of the observation activities were obtained. Once completed, field notes were written up in combination with the activity document, annotated and then interpreted through the lens of the conceptual framework

for thinking about scientific evidence. Teachers also provided additional lesson activities they defined as presenting students with quality opportunities to work with scientific evidence. These supplementary activities, like the observations, were analyzed to identify which evidential aspects from the conceptual framework were included. The descriptions presented below recount the instructional activities during the observations of the classrooms and which aspects of evidence students engaged with. In three of the observations students worked through an inquiry activity as a class. The remaining observation was comprised of students presenting the results of an experiment they carried out as part of a science project. The observations are organized by grade beginning with the elementary classrooms. Initially, two observations per classroom were scheduled, however, scheduling issues dictated only one per class.

Animal Adaptation. Class begins by calling attention to the overhead where the title, All the Living and Non-living Things in an Ecosystem Interact was displayed. Mrs. Samuels asked the class think about how to define adaptation. The teacher called on various tables around the room. Some students focused on the word ecosystem shown on the overhead in their responses and connected it to the water cycle and a plant unit completed previously. A table of four students suggested a set of animal habits (e.g., hibernation and geese flying south) as examples of adaptation. Mrs. Samuels applauded their thinking and then changed the overhead to display the results of the search terms animal adaptation and reads the definition provided by the top return. She informed students that adaptation can come in form of both physical and behavioral changes that have been produced by evolution.

Next Mrs. Samuel introduced mechanisms of adaption including changes in environment (climate change or natural disasters) and refers to how the long-neck of the giraffe permits them to eat leaves in a tall tree as a natural example of adaptation. She then directed students to work in groups on their iPads to discover their own examples of adaptation. Before releasing them to search in their groups, Mrs. Samuels modeled how students were to Google search animal adaptation to locate examples. After allowing student to work their own for 8-10 minutes, she then moved from group to group asking to see their findings and occasionally requested additional information about the specific adaptation of the student example. Once she worked through all groups, Mrs. Samuels told the class that they did a good job and asked them to get ready to transition to lunch. Total time for the lesson and activity was 30 minutes.

The evidence students encountered in this activity was produced from a Google search. I did not see a discussion of accuracy or considerations of quality. Further, I also did not observe the teacher provide directions to students about how to select reputable sources of information on the internet or the relationship between dependable resources and evidential quality. While the structure of the activity precluded examining aspects of evidence related to the sample or the quality of experimental design, there were unique opportunities to have productive discussions about the evidence returned from students' searches. For example, in broad strokes, the teacher could have outlined to the class that the adaptation evidence was the result of numerous scientific experiments and investigations occurring over many decades. This could have led to a rich discussion about the details of scientific experiments and the generation of evidence. Moreover, scaffolded discussions about how the presence of converging evidence influences both the quality and scope of the evidence would have made this activity even richer. Discussions such as these could have been used to engage students in conversations to help cement important ideas about the nature of scientific evidence. Outside of gathering examples of animal adaptation, I did not observe students engaging with any evidential aspects contained in the framework.

Ramps and Marbles. As students entered the room, Mrs. Keck directed their attention to the projector screen, which was displaying an activity based in physics. The teacher provided the focus question for the activity, and she asked students think about whether the height of a ramp influences the speed of a marble. She then outlined the procedure students are to go through and demonstrated each of the steps. Students were given a printout to complete that contains instructions and guides their progress through the activity. The central idea of the "investigation" was to examine how raising the height of a ramp influenced the speed of a marble over a stable distance. Prior to directing students to begin working in their table groups, Mrs. Keck asked students to identify the variables of the activity. After a brief discussion in which students were appeared confused trying to identify the independent, dependent, and constant variables, the teacher provided the information for them place in the gathering information section of their science sheet. Students were then asked to create a hypothesis about what they think will happen when the release height of the marble is raised.

The teacher directed students to begin working through the activity and to see her with any questions. Students began working in their table groups as the teacher walked from group to group. While she did ask both procedural and outcome-based questions (e.g., how did you work

the release of the marble and the start of the timer, what happened at the higher at the higher release point), Mrs. Keck spent the majority of the time making sure students remained focused and on track to complete the activity. After approximately 10 minutes, the teacher prompted students to make sure they have filled out the data table on their activity sheet and to begin work on their conclusions (see Table 47). The teacher provided the statement, "the marble rolled faster at the higher points" as an example to help students get started. She then switched the display on overhead to focus on the data and evidence portions of the activity sheet and detailed how these sections should look when students have filled them out. Students worked on filling out the times for each of the three trials across the three conditions (one textbook: low, two textbooks: higher, three textbooks: highest) and determining the average. Students then worked on constructing a simple evidence statement. After a few minutes, Mrs. Keck asked the students to reflect about what they learned and record it in the appropriate space on their sheet. The example provided was the conclusion listed above preceded by the words "Today I learned." Students were then asked to turn in their sheets and begin transitioning to math. Total time for the activity was 45 minutes.

Table 47

Teachers' views on students' learning about scientific evidence

Teacher	Comments
Mrs. Keck	I want them to know that it's I guess that umm evidence is important when umm making a claim about something and I use the example all the time in here when we first start talking about it you wanna claim that you're the best basketball player in this class (laughs) but what evidence do you have and I try to tell them how important evidence is and umm I want them to take away that facts are important that you can't really beat the facts if there are facts there you can't really if there's facts supporting something it's really hard to go against it
Mrs. Samuels	Oh, just about evidence itself that it changes…that there's a lot out there on the internet to see…because anymore you can find anything on youtube. You can find someone mixing chemicals and making a really cool explosion umm and that umm it can be extremely useful in whatever job they have in the future.
Mrs. Carter	It's a necessary part of the scientific process and that you cannot and should not make a claim about something if you don't have evidence to back up what you're claiming and that this is like the hallmark of what science is (begins laughing) is evidence. Yeah that anybody could look at the same thing and come up with the same conclusion if you have enough evidence let's say for instance like umm a graph of the world's temperature over time or a graph of carbon dioxide emissions over time in the atmosphere.
Mrs. Murray	That it can change…umm it's just a theory…it's not a law…it can change based on new observations…it can change based on umm the discovery of more technicalon-technologically advanced equipment…right…umm, that the evidence umm, has to be valid and true and coming from the data. It can be either quantitative-right it's numerical data or it's descriptive data but making the transition from here are these pieces and using that to construct the explanation.

This activity differs from Mrs. Samuel's. For example, it was investigative in nature and incorporated experimental procedures designed to explore relationships in physics. As such, its structure was more representative of empirical science. Students analysis of evidence was restrained to answering the question, what happened when you released the marble at different heights. I did not see a conversation about the utility of averaging the three trials across conditions, a comparison of times across groups, or dialogue about potential sources of error (e.g., imprecise timing between marble release and starting the timer). While there was a brief discussion of variables, it was largely driven by the teacher and students were not prompted to think about why those variables were important (either before or after the activity) or if other variables are worth considering (e.g., marble surface and contact surface). With respect to the framework, students were exposed to important aspects of evidence such as identifying and operationalizing variables. However, the observed discussion of these facets was brief. As noted, students were not prompted to think about the inclusion of specific variables. Moreover, how the variables were defined, measurement procedures, features of the design and the evidence from the trials were not topics of discussion.

Classifying Rocks Using a Key. Mrs. Murry welcomed the class to science and referred students to the set of papers at their lab tables. Pointing to the overhead, the teacher noted the topic of the activity while students distributed the packets to table members and began following along. Mrs. Murray communicated they will be working through a lab where they will learn about geological aspects of the earth. She goes on to explain the sequence of the lab. First, students are to read the investigation. The information provided in this section describes definitions of key terminology including rock texture and background material about the structural and observational differences of these varied types of rocks. Next, Mrs. Murry directed students' attention to the materials on each of the lab tables. Among these materials are different igneous, metamorphic, and sedimentary rocks that students will work to classify. The teacher identified the additional lab materials and reviewed proper safety procedures for working with chemicals such as hydrochloric acid. Mrs. Murray modeled for the class appropriate way to safely handle the chemicals by using safety goggles and disposable gloves. Mrs. Murry instructed students to begin working collaboratively through the activity and that she would be available to help work through any components students found confusing or uncertain about.

Students worked in groups, while Mrs. Murray walks from table to table to answering questions and providing guidance. One group asked her about rocks that have crystals and rather than provide them with the answer, the teacher offered question prompts that led students in the right direction. Mrs. Murray then provides suggestions to the class that contain additional sources to help with rock identification. A number of groups began asking procedural questions because some of the instructions and their corresponding outcomes do not match. For example, the second step provided two options if the answer to the question of the whether the rock has similar crystals (shape & color) is yes but contained no further directions if the crystals are diverse (see Table 48). Students noted that each of the other steps contained directions for yes or no, similar to step one below. There was also student debate surrounding the use of ambiguous color terms such as dark. Mrs. Murray reminded the class that the emphasis is not on the right answer but on process of the investigation. She then directed students to begin working on the analysis and conclusions section of their packet. After approximately 5-7 minutes, the teacher instructed students to move back to their original seats. Total time for the activity was 50 minutes.

Table 48

Evidence Portion of Ramps and Marbles Worksheet

COLLECT DATA AND MAKE OBSERVATIONS (Average – add the three trials and divide by 3)

	Trial 1	Trial 2	Trial 3	Average
One Textbook				
Two Textbooks				
Three Textbooks				

CONCLUSION

Evidence: My evidence_____

_____

_____

This activity presented students with valuable information in geological science and contained a phase with investigative elements that provided steps for students to follow and combined background information provided at the beginning of the lesson with observational clues to determine both the rock category and the specific type of rock (e.g., sedimentary, sandstone) from a set of group exemplars. As can be seen from the sample in table 48, the activity was structured. Similar to Mrs. Samuels animal adaptation activity, there were entry points for productive discussions about evidence that would have made this activity even richer. For example, instruction could have centered around how the evidence provided in the information portion of the activity was developed. In these discussions, students could have been introduced to evidential aspects such as sample representativeness, procedures for collecting and analyzing data, in addition to the way social practices of science contribute to formulations of evidence. Additionally, while students were working through the identification process, knowledge of scientific evidence could have been supported by underscoring how the evidence of the informational segment was the basis for the evidence encountered in the activity phase.

What Burns the Longest. Mrs. Carter called a group of students up to her desk and had a brief conversation with them. She reminded the class that they are viewing group presentations during class and asked students to take their seat. The presenting group loaded their science project, and Mrs. Carter set up the computer, so the project displayed on the overhead screen at the front of the class. Students began the presentation about their science project titled what burns the longest. Group members introduced themselves and their chosen topic. The driving question of the project was what burns the longest. They selected a range of materials to burn and timed how long it took for the item to be consumed to the point students could no longer hold the item safely with their tongs. Their hypothesis was that paper towels would burn the longest due to the fact its thin and dry. Other materials tested were newspaper, cardboard, pencil, tinfoil, a leaf, liquid soap, Styrofoam, plastic, and regular notebook paper. Students described their procedure for lighting each of the items and how they determined the amount of time it took for each of them to burn. The group then noted they tested each material twice, and then made connections between their project and the scientific processes of asking questions, developing a hypothesis, designing procedures for data collection, evaluating the evidence, and sharing their results.

Results of their experiment were displayed by material (see Table 49) in addition to digital photos of the experiment. Each trial was listed in seconds as well as an average time. The final slide was their conclusion, which listed some obstacles they encountered during the experiment and a reflection about what they would have done differently. Overall, students pitched the project as a success even though their hypothesis turned out to be wrong. Mrs. Carter thanked the students and invited the class to ask questions. Several students presented questions. For example, a question about the smell of burning specific materials was asked multiple times. Another student asked whether anything caught on fire accidentally during their experiment. Mrs. Carter asked the group to identify which evidence they used to formulate their conclusion. Students returned to the slides of their results and went through each material individually to demonstrate how their conclusion was formed. Mrs. Carter thanked the group and began to transition to other agenda items. Total time for the presentation was 35 minutes.

Table 49

Sample of Classifying Rocks Key

Key to Rock Classification	
1. Does the rock contain visible connecting crystals?	Yes: Go to Question 2 No: Go to Question 4
2. Are all the crystals the same color and shape?	Yes: The rock is a nonfoliated metamorphic rock (possibly marble or quartzite).
3. Are all the crystals in mixed “salt-and-pepper” pattern?	Yes: The rock if an intrusive Igneous rock (possibly granite or diortite). No: The rock is a foliated metamorphic rock (possibly schist or gneiss).

Compared to the other activities, this student project represented the closest example of a full investigation. Students developed an idea, formed a hypothesis, generated a design, developed data collection procedures, analyzed evidence and then integrated it into a conclusion. I did not observe a discussion about connections between the project topic and class content nor was there a discussion about where the ideas came from or how previous evidence they encountered through the research process informed the focus of the project. Likewise, no details were provided about how they chose variables or why each of the selected variables were important to the purposes of the study. Further, there was no observed discussion of the relationship between these considerations and the evidence. A further illustration of this can be seen in the evidential dimension of data collection errors. For example, many of the variables in

their project are available in a variety of sizes (e.g., newspaper, notebook paper, and paper towels), yet there was no reference to material size when explaining procedures or with respect to the variations in burn time across trials. Moreover, one of their photo slides displayed the use of tongs to grasp materials, and they were used in two separate places. On one object it was at the end and on the other it was in the middle. Given the students did not account for this in their project, they appear to have overlooked these features of evidence.

Across these observations students reasoned and worked with evidence, but there was no observed dialogue about evidential dimensions. The only consistently integrated aspect of evidence concerned the connective tissue between claims and evidence found in the interpretations/conclusions section. I did not observe students being afforded opportunities to develop understandings about other key features such as how the identification and justification of variables or examinations of error impacts and contributes to the scope and quality of evidence. Certainly, some amount of the why students are presented with forms of evidence described in the class observations can be explained by the need for heavily scaffolded activities. Students, especially at these ages, do not possess the requisite background knowledge to in engage in robust examinations of evidence. As noted, though, each of the above observations contained points of entry for key aspects of evidence to be introduced, explored, and then applied in service of students' knowledge development.

It is also worth noting that the above analysis is based on a single observation and therefore not able to make any definitive statements about the notions of evidence at play in these spaces. Students may encounter rich knowledge about the nature of evidence when participating in other activities and instruction. However, all 25 of the additional lessons contained notions of evidence that mirror those in the observations. To illustrate, in one of the more complex activities, students are presented with a set of small experiments to learn about macromolecules and complex compounds. The context of the activity is students are tasked with solving a crime. In the story, the victim has eaten at one of four restaurants, and students test fake stomach contents to determine where the victim ate their last meal. The tests conducted identify the presence of sugar, carbohydrates, lipids, and proteins, and each of these organic compounds is associated with one of the four restaurants. Detailed directions are provided for students to perform isolated tests on the four types of organic compounds to observe the proper

indicators prior to testing the fake stomach contents. Students then test the phony stomach contents and record their results in a table (see Table 50).

Table 50

Sample of Project Results

	Trial Times in Seconds
Newspaper	Trial 1~ 20.96 Trial 2~ 24 Avg~ 22.48
Cardboard	Trail 1~ 90 Trial 2~ 38 Avg~ 64
Pencil	Trail 1~ 22 Trial 2~ 17 Avg~ 19.5
Aluminum Foil	Didn't Burn Avg~0
Leaf	Trial 1~ Didn't Burn Trail 2~ 6 Avg~ 3

Outside of the column for students to record their results, there is no place in the activity for students to consider ideas about the evidential impact of conducting one trial or to think about potential sources of error in their data collection. The evidence portion of the activity directly follows the data table above (see Table 51). In this section, there appears to be no opportunity for students to consider important aspects of scientific evidence, even as they relate to this specific activity. Building on the issue of conducting a single trial already referenced, this could extend to discussions about evidential limits and connecting the two could help students develop understanding about the interrelated nature of evidence. Additionally, even though the activity details each step towards a predetermined outcome, there is no discussion of the evidence used to generate the indicators students relied on to determine which of the four organic compounds were present in their samples. While the process has been refined to such a degree that following a prescribed set of steps produces unambiguous evidence of sugar or a protein, there are aspects of this unseen history that can provide a valuable supplement to students developing notions of scientific evidence, especially when the instructional context dictates highly structured activities. In these cases, it would benefit students to be introduced to coordinated discussions that connect the evidence encountered in the activity with its evidential base.

Table 51

Data Table from Organic Compounds Lab

Investigation: Testing the "Mystery Stomach Contents"

Data

Solution Tested	**Indicator Used during test**	**Result**
Stomach contents	Benedicts	
Stomach contents	Lugol's	
Stomach contents	Brown bag	
Stomach contents	Biuret's	

Just as the observation portion cannot be used as the basis to formulate the claim that students only encounter circumscribed forms of evidence in the classrooms, the same is true for the detailed examination of the additional lessons teachers provided. Much more exposure to the form and character of instruction in the classroom is needed to generate a decisive analysis. However, taken together, the observations and supplementary documents represent a pattern where students and the activities they work through leave important evidential aspects unexamined. Without consistent opportunities to work and reason with rich notions of scientific evidence, students are unlikely to acquire deep understandings.

Table 52

Claims and Evidence Sample

Claims and Evidence

Remember a claim is what happened in the lab and the evidence is data to support or prove the claim to back it up!

Claim	**Evidence**
I claim that…	
I claim that…	

CHAPTER 4. DISCUSSION AND LIMITATIONS

Overall, these results show that these fifth and seventh-grade students exhibited difficulty reasoning with and about the varied aspects of evidence from the conceptual framework. While seventh-grade students did obtain higher ERA scores generally, their performance varied both across items and individuals. The most consistent response across all items was the application of a more equals better justification. This may be the result of a lack of knowledge about evidence itself. When considering the outcomes from the teacher interviews and classroom observations, the notions of evidence at play in these contexts appears relatively straightforward. Given the fact that little attention has been paid to the construct of evidence across the research fields, it is not surprising that teachers and their instructional materials do not incorporate robust notions of evidence. In many cases, teachers are not required to take courses aimed at unpacking scientific evidence as a part of their training. Without this valuable exposure, it unreasonable to expect teachers to develop this knowledge on their own.

When examining differences between the grade levels that emerged from the teacher interviews provides valuable insight into why seventh-grade performance was better. Overall, the seventh-grade teachers talked about science in ways that demonstrate they view it as a way of approaching the world to generate understanding. Fifth-grade teachers tended to talk about science in terms of developing student interest and its use for future employment. These differences will affect the versions of science and of evidence students encounter in the classroom. Further, seventh-grade students received more time on task. These students received daily science instruction and spent more time on investigations.

Another important consideration when evaluating the results is the extent to which students failed to transfer their knowledge. Given the significant task effect between the narratives, this appears to be a reasonable proposition. There are likely multiple explanations for this. As discussed previously, research into students' ideas about plants indicated they tend to not to view them as living things. This misconception could have led to diminished performance on the Acacia narrative. Student interest in the Acacia narrative also cannot be overlooked. While the assessment of science interest revealed no correlation between the three questions and total ERA score, there may still be differences that contributed to performance variations.

Additional analyses demonstrate these variations between the grades and within them can be explained, at least in part, by reading ability and key differences between the narratives. The issue of reading ability was especially relevant for fifth-graders. Although both narratives were experimental in nature and structured similarly, their differing purposes resulted in variations that played a role in the quality of responses. For example, question six on the mosquito narrative introduced notions of technological error and its corresponding effect on evidence, and students from both grade levels were able to identify how issues with the computer would negatively influence the evidence in this narrative. Due to the fact that the school corporation where the research was conducted is a 1:1 community, many students have likely had some previous experience of computer issues, and they were able to apply this knowledge to the ERA question. This underscores the influence that even some background knowledge and experience can have on working with complex ideas in science.

Another interesting performance related development is the extent to which the question type influenced scores. For example, fifth-graders performed at a higher level on question four than they did on other ERA items. As discussed previously, this item was the only one with a multiple-choice component, and fifth-graders appeared to benefit from the reduction in cognitive load. The second development is the performance differences between questions that asked students to evaluate causally plausible and causally implausible aspects of the narratives. Students performed much better when reasoning about causally plausible aspects. This may be due to the fact that students were unable to reason about underlying mechanism in these questions. For example, in question three on the mosquito narrative, one of the students in the mock debate suggest the inclusion of the implausible causal variable of oxygen. Since the narrative did not address this variable directly, students exhibited difficulty reasoning about why oxygen is causally implausible and reverted to a more equals better justification in their answers.

Finally, there were times during the interviews when fifth-grade students would provide responses that indicated a greater level of understanding than their ERA answer demonstrated. In these contexts, it appears that through the process of discussion and reflection, these students were able to generate higher quality answers. For example, Isabella, a fifth-grader, provided an answer on question eight of the ERA that was coded as no understanding (refer to student interviews for the entire example). The context of the question asked students to consider whether revisions to the experimental design would be required if a causally plausible variable

was discovered. In her ERA answer, she noted that it would not influence the tests and characterized the discovery as useless. During the interview, Isabella began to waiver in her original position. With some scaffolding, she was able to not only demonstrate an understanding of how a causally plausible variable would necessitate changes to the experimental design, but she also provided an example of how the scientists could test it. Examples like these suggest fifth-graders may have more evidentiary knowledge than indicated by their ERA performance.

This study was limited in several ways. First, the samples were relatively homogenous demographically and from a reading ability perspective, although the fifth-grade sample was more diverse in this regard. While the teachers assigned reading ability ratings, lack of access to standardized measures of performance is a limitation. The limited number of classroom observations and supplementary documents provided by teachers restricts the evidence produced by these instructional variables.

Conclusions

The combination of these results provides descriptive evidence that fifth and seventh-grade students had difficulty working with complex notions of evidence. Additional research is needed to further understand students' evidentiary knowledge and how it develops and can be supported. It is equally important to develop future projects that embed rich notions of evidence into curricula and design classroom instruction that fosters multifaceted views of scientific evidence. Following this, I want to partner with elementary science educators to design interventions that interface the dimensions of evidence contained in the framework with their existing science curricula to examine how to better develop and support students' knowledge of scientific evidence. This includes the development of lessons that contain introductory material about the nature of evidence to students as well as the construction of scaffolded activities designed to provide students opportunities to reason with and about scientific evidence.

REFERENCES

Aikenhead, G. S. (2005). Science-Based Occupations and the Science Curriculum: Concepts of Evidence (English). *Science Education, 89*(2), 242-275.

Amsel, E., & Brock, S. (1996). The Development of Evidence Evaluation Skills. *Cognitive Development, 11*(4), 523-550.

Anton, T. (2000). *Bold science: Seven Scientists Who Are Changing Our World*. New York, NY: W.H. Freeman.

Bacon, F. (1994). *Novum Organum* (P. Urbach & J. Gibson, Trans. P. Urbach & J. Gibson Eds. Sixth ed.). United States of America: Open Court.

Baofu, P. (2012). *The Future of Post-Human History: A Preface to a New Theory of Universality and Relativity*. United Kingdom: Cambridge Scholars

Blown, E., & Bryce, T. G. K. (2010). Conceptual Coherence Revealed in Multi-Modal Representations of Astronomy Knowledge. *International Journal of Science Education, 32*(1), 31-67.

Bod, R. (2014). *A New History of the Humanities: The Search for Principles and Patterns from Antiquity to the Present*. United States of America: Oxford University Press.

Boland, R. (1985). Phenomenology: A Preferred Approach to Research in Information Systems. In E. Mumford, R. A. Hirschheim, G. Fitzgerald, & A. T. Wood-Harper (Eds.), *Research Methods in Information Systems*. Amsterdam: North Holland.

Bonawitz, E. B., van Schijndel, T. J. P., Friel, D., & Schulz, L. (2012). Children Balance Theories and Evidence in Exploration, Explanation, and Learning. *Cognitive Psychology, 64*, 215-234.

Bouillion, L. M., & Gomez, L. M. (2001). Connecting school and community with science learning: real world problems and school-community partnerships as contextual scaffolds. *Journal of Research in Science Teaching, 38*(8), 878-898.

Brandon, R. N. (1996). *Concepts and Methods in Evolutionary Biology*. New York: Cambridge University Press.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (1999). *How People Learn: Brain, Mind, Experience, and School*. Washington, D. C.: National Academy Press.

Brewer, W. F., Chinn, C. A., & Samarapungavan, A. (1998). Explanation in Scientists and Children. *Minds and Machines, 8*(1), 119-136.

Brewer, W. F., & Lambert, B. L. (2001). The Theory-Ladenness of Observation and the Theory-Ladenness of the Rest of the Scientific Process. *Philosophy of Science, Supplement: Proceedings of the 2000 Biennial Meeting of the Philosophy of Science Association. Part I: Contributed Papers 68*(3), 176-186.

Brewer, W. F., & Samarapungavan, A. (1991). Children's Theories vs. Scientific Theories: Differences in Reasoning or Differences in Knowledge? In R. R. Hoffman & D. S. Palerno (Eds.), *Cognition and the Symbolic Processes: Applied and Ecological Perspectives*. New York: Lawrence Erlbaum Associates.

Bright-Paul, A., Jarrold, C., & Wright, D. B. (2008). Theory-of-Mind Development Influences Suggestibility and Source Monitoring. *Developmental Psychology, 44*(4), 1055-1068.

Brown, D. G. (2000). *Interactive Learning: Vignettes from America's Most Wired Campuses*: ERIC.

Carey, S. (2009). *The Origin of Concepts*. New York: Oxford University Press.

Carey, S., Evans, R., Honda, M., Jay, E., & Unger, C. (1989). 'An experiment is when you try it and see if it works': a study of grade 7 students' understanding of the construction of scientific knowledge. *International Journal of Science Education, 11*(5), 514-529.

Carey, S., & Spelke, E. (1994). Domain-specific knowledge and conceptual change. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificty in cognition and culture* (pp. 169-200). United States of America: Cambridge University Press.

CERN. (2015). About CERN. Retrieved from http://home.web.cern.ch/about

Cetina, K. K. (1999). *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, MA: Havard University Press.

Chalmers, A. F. (1999). *What is this thing called Science?* (3rd ed.). Indianapolis: Hackett Publishing.

Chappell, S. G. (2013). Plato on Knowledge in the *Theaetetus*. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Winter 2013 Edition)*: <http://plato.stanford.edu/archives/win2013/entries/plato-theaetetus/>.

Chi, M. T. H. (1997). Quantifying Qualitative Analyses of Verbal Data: A Practical Guide. *Journal of the Learning Sciences, 6*(3), 271-315.

Chinn, C. A., & Malhotra, B. A. (2002). Children's Responses to Anomalous Scientific Data: How Is Conceptual Change Impeded? *Journal of Educational Psychology, 94*(2), 327-343.

Chinn, C. A., O'Donnell, A. M., & Jinks, T. S. (2000). The Structure of Discourse in Collaborative Learning. *Journal of Experimental Education, 69*(1), 77-97.

Corot, C., Robert, P., Idée, J.-M., & Port, M. (2006). Recent advances in iron oxide nanocrystal technology for medical imaging. *Advanced Drug Delivery Reviews 58*, 1471-1504.

Corriveau, K. H., Pasquini, E. S., & Harris, P. L. (2005). "If it's in your mind, it's in your knowledge": Children's developing anatomy of identity. *Cognitive Development, 20*(2), 20p.

Council, N. R. (2007). *Taking Science to School: Learning and Teaching Science in Grades K-8*: The National Academies Press.

Craver, C., & Tabery, J. (2016). Mechanisms in Science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*: http://plato.stanford.edu/archives/spr2016/entries/science-mechanisms/.

Crupi, V., Tentori, K., & Gonzalez, M. (2007). On Bayesian Measures of Evidential Support: Theoretical and Empirical Issues. *Philosophy of Science, 74*(2), 229-252.

Descartes, R. (1989). *Descartes: Selected Philosophical Writings* (J. Cottingham, R. Stoothoff, & D. Murdoch, Trans.).

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the Norms of Scientific Argumentation in Classrooms. *Science Education, 84*(3), 287-312.

Duhem, P. (1954). *The Aim and Structure of Physical Theory*. United States of America: Princeton University Press.

Duncan, R. G., & Reiser, B. J. (2007). Reasoning Across Ontologically Distinct Levels: Students' Understandings of Molecular Genetics. *Journal of Research in Science Teaching, 44*(7), 938-959.

Ebel, R. L. (1965). *Measuring educational achievement*: Prentice-Hall.

Education, I. D. o. (2010). *Indiana's Academic Standards for Science*. Retrieved from https://learningconnection.doe.in.gov/Standards/PrintLibrary.aspx.

Fischer, K. W. (1980). A Theory of Cognitive Development: The Control and Construction of Hierarchies of Skills. *Psychological Review, 87*(6), 477-531.

Flavell, J. H. (2000). Development of children's knowledge about the mental world. *International Journal of Behavioral Development, 24*(1), 15-23.

Flavell, J. H., Flavell, E. R., Green, F. L., & Moses, L. J. (1990). Young children's understanding of fact beliefs versus value beliefs. *Child Development, 61*(4), 915-928.

Fodor, J. (1984). Observation Reconsidered. *Philosophy of Science, 51*, 23-43.

Ford, A. T., Goheen, J. R., Otieno, T. O., Bidner, L., Isbell, L. A., Palmer, T. M., . . . Pringle, R. M. (2014). Large carnivores make savanna tree communities less thorny. *Science, 346,* 346-349.

Franklin, A. (1986). *The Neglect of Experiment*. New York, NY: Cambridge University Press.

Franklin, A., & Perovic, S. (2015). Experiment in Physics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Summer 2015 Edition)*: http://plato.stanford.edu/archives/sum2015/entries/physics-experiment/.

Galison, P. (1987). *How Experiments End*. United States of America: The University of Chicago Press.

Gelman, R. (1996). Domain specificity in cognitive development: universals and nonuniversals. In M. Sabourin, F. Craik, & M. Robert (Eds.), *Advances in Psychological Science: Biological and cognitive aspects* (Vol. 2).

Gelman, R., & Brenneman, K. (2004). Science learning pathways for young children. *Early Childhood Research Quarterly, 19*(1), 150-158. doi:10.1016/j.ecresq.2004.01.009

Gelman, S. A. (2004). Psychological essentialism in children. *Trends in Cognitive Sciences, 8*(9), 6p.

Gelman, S. A., & Kremer, K. E. (1991). Understanding Natural Cause: Children's Explanations of How Objects and Their Properties Originate. *Child Development, 62*(2), 396-414.

Giere, R. N. (1984). *Understanding Scientific Reasoning* (2nd ed.). Texas: Holt, Rinehart, and Winston, Inc.

Glymour, C. (1980a). Bootstraps and Probabilities. *The Journal of Philosophy, LXXVII*(11), 691-699.

Glymour, C. (1980b). *Theory and Evidence*. Princeton, New Jersey: Princeton University Press.

Goldman, A. I. (1986). *Epistemology and Cognition*. Massachusetts: Harvard University Press.

Gopnik, A. (2012). Scientific Thinking in Young Children: Theoretical Advances, Empirical Research,and Policy Implications. *Science, 337*, 1623-1627.

Gopnik, A., & Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends In Cognitive Sciences, 8*(8), 371-377.

Greenberg, G. (2014). How New Ideas in Physics and Biology Influence Developmental Science. *Research in Human Development, 11*, 5-21.

Guajardo, N. R., & Turley-Ames, K. J. (2004). Preschoolers' Generation of Different Types of Counterfactual Statements and Theory of Mind Understanding. *Cognitive Development, 19*(1), 53-80.

Hacking, I. (1982). Experimentation and Scientific Realism. *Philosophical Topics, 13*, 81-87.

Hacking, I. (1983). *Representing and Intervening*. United States of America: Cambridge University Press.

Hee-Sun, L., Liu, O. L., & Linn, M. C. (2011). Validating Measurement of Knowledge Integration in Science Using Multiple-Choice and Explanation Items. *Applied Measurement in Education, 24*(2), 115-136. doi:10.1080/08957347.2011.554604

Heilbron, J. L. (2003). *Ernest Rutherford: And the Explosion of Atoms* New York: Oxford University Press.

Hempel, C. G. (1965). *Aspects of Scientific Explanation: And Other Essays in the Philosophy of Science*. New York: The Free Press.

Hoffert, M. I., Caldeira, K., Benford, G., Criswell, D. R., Green, C., Herzog, H., . . . Wigley, T. M. L. (2002). Advanced Technology Paths to Global Climate Stability: Energy for a Greenhouse Planet. *Science, 298*.

Hoffman, R. (2007). What Might Philosophy of Science Look like If Chemists Built It? . *Synthese, 155*(3), 321-336.

Holger, A. (2013). Theoretical Terms in Science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Summer 2013 Edition)*: http://plato.stanford.edu/archives/sum2013/entries/theoretical-terms-science/.

Jackson, S., Krajcik, J., & Soloway, E. (2000). Model-It: A design retrospective. In M. J. Jacobson & R. B. Kozma (Eds.), *Innovations in Science and Mathematics Education: Advanced Designs for the Technologies of Learning:* . Hillsdale, NJ: Erlbaum.

Jeong, H., Songer, N. B., & Lee, S.-Y. (2007). Evidentiary Competence: Sixth Graders' Understanding for Gathering and Interpreting Evidence in Scientific Investigations. *Research in Science Education, 37*(1), 75-97.

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

Johnson, R. B., & Christensen, L. (2014). *Educational Research: Quantitative, Qualitative, and Mixed Approaches* (5th ed.). Los Angeles, CA: Sage Publications.

Joyce, J. M. (2005). How Probabilities Reflect Evidence. *Philosophical Perspectives, 19*(1), 153-178.

Kanari, Z., & Millar, R. (2004). Reasoning from Data: How Students Collect and Interpret Data in Science Investigations. *Journal of Research in Science Teaching, 41*(7), 748-769.

Kelly, G. J., & Chen, C. (1999). The Sound of Music: Constructing Science as Sociocultural Practices through Oral and Written Discourse. *Journal of Research in Science Teaching, 36*(8), 883-915.

Kelly, T. (2008). Evidence: Fundamental Concepts and the Phenomenal Conception. *Philosophy Compass, 3*(55), 933-955.

Kelly, T. (2014). Evidence. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2014 Edition)*: http://plato.stanford.edu/archives/fall2014/entries/evidence/.

Khishfe, R., & Abd-El-Khalick, F. (2002). Influence of Explicit and Reflective versus Implicit Inquiry-Oriented Instruction on Sixth Graders’ Views of Nature of Science. *Journal of Research in Science Teaching, 39*(7), 551-578.

Klahr, D., & Li, J. (2005). Cognitive Research and Elementary Science Instruction: From the Laboratory, to the Classroom, and Back. *Journal of Science Education & Technology, 14*(2), 217-238.

Koerber, S., Sodian, B., Thoermer, C., & Nett, U. (2005). Scientific Reasoning in Young Children: Preschoolers' Ability to Evaluate Covariation Evidence. *Swiss Journal of Psychology, 64*(3), 141-152.

Koslowski, B. (1996). *Theory and Evidence: The Development of Scientific Reasoning*. United States: Massachusetts Institute of Technology.

Kuhn, D. (2000). Metacognitive Development. *Current Directions in Psychological Science, 9*(5), 178-181.

Kuhn, D., Amsel, E., & O'Loughlin, M. (1988). *The development of scientific thinking skills*. Orlando, FL: Academic Press.

Kuhn, D., Iordanou, K., Pease, M., & Wirkala, C. (2008). Beyond Control of Variables: What Needs to Develop to Achieve Skilled Scientific Thinking? *Cognitive Development, 23*(4), 435-451.

Kuhn, D., & Pearsall, S. (2000). Developmental Origins of Scientific Thinking. *Journal of Cognition & Development, 1*(1), 113-129.

Kuhn, D., & Udell, W. (2007). Coordinating own and other perspectives in argument. *Thinking & Reasoning, 13*(2), 15p. doi:10.1080/13546780600625447

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions* (3rd ed.). Chicago: University of Chicago Press.

Kuhn, T. S. (2003). *The Copernican Revolution: Planetary Astronomy in the Development of Western Thought*. United States of America: Harvard University Press.

Lakatos, I. (1970). Falsification and the Methodology of Scientific Research Programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the Growth of Knowledge: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965*. United States of America: Cambridge University Press.

Lane, J. D., Wellman, H. M., & Evans, E. M. (2010). Children's Understanding of Ordinary and Extraordinary Minds. *Child Development, 81*(5), 15p.

Latour, B., & Woolgar, S. (1986). *Laboratory Life: the Social Construction of Scientific Facts*. Princeton, NJ: Princeton University Press.

Laudan, L. (1990). Demystifying Underdetermination. In C. W. Savage (Ed.), *Scientific Theories* (Vol. 14). Minneapolis: University of Minnesota Press.

Laudan, L. (1996). *Beyond Positivism and Relativism: Theory, Method, and Evidence*. United States: Westview Press.

Leach, J. (1999). Students' understanding of the co-ordination of theory and evidence in science. *International Journal of Science Education, 21*(8), 789-806.

Lederman, N. G. (1992). Students' and Teachers' Conceptions of the Nature of Science: A Review of the Research. *Journal of Research in Science Teaching, 29*(4), 331-359.

Lederman, N. G., Abd-El-Khalick, F., Bell, R. L., & Schwartz, R. S. (2002). Views of Nature of Science Questionnaire: Toward Valid and Meaningful Assessment of Learners' Conceptions of Nature of Science. *Journal of Research in Science Teaching, 39*(6), 497-521.

Legare, C. H. (2012). Exploring Explanation: Explaining Inconsistent Evidence Informs Exploratory, Hypothesis-Testing Behavior in Young Children. *Child Development, 83*(1), 173-185.

Lehrer, R., Schauble, L., & Lucas, D. (2008). Supporting Development of the Epistemology of Inquiry. *Cognitive Development, 23*(4), 512-529.

Linn, M. C., Clark, D. B., & Slotta, J. D. (2003). WISE Design for Knowledge Integration. *Science Education*(87), 517-538. doi:10.1002/sce.10086

Lloyd, E. A. (2010). Confirmation and Robustness of Climate Models. *Philosophy of Science, 77*(5), 971-984.

Longino, H. (2015). The Social Dimensions of Scientific Knowledge. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Spring 2015 Edition)*: http://plato.stanford.edu/archives/spr2015/entries/scientific-knowledge-social/.

Longino, H. E. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, NJ: Princeton University Press.

Lord, F. M. (1952). *A Theory of Test Scores (Psychometric Monograph No. 7)*. New York: Psychometric Society.

Lubben, F., & Millar, R. (1996). Children's ideas about the reliability of experimental data. *International Journal of Science Education, 18*(8), 955-968.

Masnick, A. M., & Klahr, D. (2003). Error Matters: An Initial Exploration of Elementary School Children's Understanding of Experimental Error. *Journal of Cognition and Development, 4*(1), 67-98.

Masnick, A. M., & Morris, B. J. (2008). Investigating the Development of Data Evaluation: The Role of Data Characteristics. *Child Development, 79*(4), 1032-1048.

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. United States of America: The University of Chicago Press.

Mayo, D. G. (2005). Evidence as Passing Severe Tests: Highly Probable versus Highly Probed Hypotheses. In P. Achinstein (Ed.), *Scientific Evidence: Philosophical Theories & Applications* (pp. 95-128): The Johns Hopkins University Press.

Mayr, E. (2004). *What Makes Biology Unique? Considerations on the autonomy of a scientific discipline*. New York: Cambridge University Press.

McNeill, K. L. (2011). Elementary Students' Views of Explanation, Argumentation, and Evidence, and Their Abilities to Construct Arguments over the School Year. *Journal of Research in Science Teaching, 48*(7), 793-823.

McNeill, K. L., & Krajcik, J. (2007). Inquiry and Scientific Explanations: Helping Students Use Evidence and Reasoning. In J. Luft, R. Bell, & J. Gess-Newsome (Eds.), *Science as inquiry in the secondary setting* (pp. 121-134). Arlington, VA: National Science Teachers Association Press.

Metz, K. E. (2004). Children's Understanding of Scientific Inquiry: Their Conceptualization of Uncertainty in Investigations of Their Own Design. *Cognition & Instruction, 22*(2), 219-290.

Metz, K. E. (2008). Narrowing the Gulf between the Practices of Science and the Elementary School Science Classroom. *The Elementary School Journal, 109*(2), 138-161.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. United States of America: The National Academies Press.

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards*. Washington, DC: The National Academies Press.

National Research Council. (2015). *Enhancing the Effectiveness of Team Science*. Washington, DC: The National Academies Press.

Nersessian, N. J. (2006). Model-Based Reasoning in Distributed Cognitive Systems. *Philosophy of Science, 73*(5), 699-709.

Nersessian, N. J. (2008). *Creating Scientific Concepts*. Cambridge, Massachussets: MIT Press.

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

O'Neil, D. K., Astington, J. W., & Flavell, J. H. (1992). Young Children's Understanding of the Role That Sensory Experiences Play in Knowledge Acquisition. *Child Development, 63*(2), 474-490.

O'Neill, D. K., & Gopnik, A. (1991). Young children's ability to identify the sources of their beliefs. *Developmental Psychology, 27*(3), 390-397.

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the Quality of Argumentation in School Science. *Journal of Research in Science Teaching, 41*, 994-1020.

Peabody, J. W., Luck, J., Glassman, P., & et al. (2004). Measuring the quality of physician practice by using clinical vignettes: A prospective validation study. *Annals of Internal Medicine, 141*(10), 771-780. doi:10.7326/0003-4819-141-10-200411160-00008

Penner, D. E., Giles, N. D., Lehrer, R., & Schauble, L. (1997). Building Functional Models: Designing an Elbow. *Journal of Research in Science Teaching, 34*(2), 125-143.

Penner, D. E., & Klahr, D. (1996). When to trust the data: further investigations of system error in a scientific reasoning task. *Memory & Cognition, 24*(5), 655-668.

Piekny, J., Grube, D., & Maehler, C. (2014). The Development of Experimentation and Evidence Evaluation Skills at Preschool Age. *International Journal of Science Education, 36*(2), 334-354.

Popper, K. R. (1992). *The logic of scientific discovery*. London ; New York: Routledge.

Quine, W. V. (1951). Two Dogmas of Empiricism. *The Philosophical Review, 60*(1), 20-43.

Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., . . . Edelson, D. (2004). A Scaffolding Design Framework for Software to Support Science Inquiry. *The Journal of the Learning Sciences, 13*(3), 337-386.

Radder, H. (2009a). Science, Technology and the Science-Technology Relationship. In D. M. Gabbay, A. Meijers, & J. Woods (Eds.), *Philosophy of Technology and Engineering Sciences*. Oxford, UK: Elsevier.

Radder, H. (2009b). Toward a More Developed Philosophy of Scientific Experimentation. In H. Radder (Ed.), *The Philosophy of Scientific Experimentation*. Pittsburgh, PA: University of Pittsburgh Press.

Raghavan, K., & Glaser, R. (1995). Model-Based Analysis and Reasoning in Science: The MARS Curriculum. *Science Education, 79*(1), 37-61.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to Psychometric Theory*. New York: Routledge.

Reiser, B. J., Tabak, I., & Sandoval, W. A. (2001). BGuILE: Strategic and Conceptual Scaffolds for Scientific Inquiry in Biology Classrooms. In S. M. Carver & D. Klahr (Eds.), *Cognition and Instruction: Twenty-five years of progress*. Mahwah, NJ: Erlbaum.

Rheinberger, H.-J. (1997). *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Stanford: Stanford University Press.

Rodriguez, M. C. (2002). Choosing an Item Format. In G. Tindal & T. M. Haladyna (Eds.), *Large-Scale Assessment Programs for All Students: Validity, Technical Adequacy, and Implementation* (pp. 182-198). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Roth, W.-M. (2005). *Talking Science: Language and Learning in Science Classrooms*. United States: Rowman and Littlefield.

Russell, B. (1912). *The Problems of Philosophy*. Oxford: Oxford University Press.

Samarapungavan, A. (1992). Children's judgments in theory choice tasks: Scientific rationality in childhood. *Cognition, 45*, 32.

Samarapungavan, A., Mantzicopoulos, P., & Patrick, H. (2008). Learning Science Through Inquiry in Kindergarten. *Science Education, 92*(5), 868-908.

Samarapungavan, A., & Wiers, R. W. (1997). Children's thoughts on the origin of species: A study of Explanatory Coherence. *Cognitive Science, 21*(2), 147-177.

Sandoval, W. A. (1998). *Inquire to explain: Structuring inquiry around explanation construction in a technology-supported biology curriculum.* P.h.D. Dissertation. Northwestern University.

Sandoval, W. A., & Cam, A. (2010). Elementary Children's Judgments of the Epistemic Status of Sources of Justification. *Science Education, 95*(3), 383-408.

Sandoval, W. A., & Reiser, B. J. (2004). Explanation-Driven Inquiry: Integrating Conceptual and Epistemic Scaffolds for Scientific Inquiry. *Science Education, 88*, 345-372.

Sandoval, W. A., Sodian, B., Koerber, S., & Wong, J. (2014). Developing Children's Early Competencies to Engage With Science. *Educational Psychologist, 49*(2), 139-152.

Schauble, L. (1996). The Development of Scientific Reasoning in Knowledge-Rich Contexts. *Developmental Psychology, 32*(1), 102-119.

Schauble, L., Glaser, R., Duschl, R. A., Schulze, S., & John, J. (1995). Students' Understanding of the Objectives and Procedures of Experimentation in the Science Classroom. *Journal of the Learning Sciences, 4*(2), 131-166.

Schindler, S. (2011). Bogen and Woodward's data-phenomena distinction, forms of theory-ladenness, and the reliability of data *Synthese, 182*, 39-55.

Schulz, L. E., & Gopnik, A. (2004). Causal Learning Across Domains. *Developmental Psychology, 40*(2), 162-176.

Siegler, R. S. (2000). The Rebirth of Children's Learning. *Child Development, 71*(1), 26-35.

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young Children's Differentiation of Hypothetical Beliefs from Evidence. *Child Development, 62*(4), 753-766.

Somerville, J. (1941). Umbrellaology, or, Methodology in Social Science. *Philosophy of Science, 8*(4), 557-566.

Songer, N. B. (2006). BioKIDS: An Animated Conversation On the Development of Complex Reasoning in Science. In R. K. Sawyer (Ed.), *Cambridge Handbook of the Learning Sciences*. New York: Cambridge University Press.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science, 10*(1), 8p.

Staley, K. W. (2004). *The Evidence for the Top Quark: Objectivity and Bias in Collaborative Experimentation*. New York: Cambridge University Press.

Stanford, K. (2013). Underdetermination of Scientific Theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Winter 2013 Edition)*: http://plato.stanford.edu/archives/win2013/entries/scientific-underdetermination/.

Stavy, R. (1991). Children's Ideas about Matter. *School Science and Mathematics, 91*(6), 240-244.

Stecher, B., Le, V.-N., Hamilton, L., Ryan, G., Robyn, A., & Lockwood, J. (2006). Using structured classroom vignettes to measure instructional practices in mathematics. *Educational Evaluation and Policy Analysis, 28*(2), 101-130.

Strand-Cary, M., & Klahr, D. (2008). Developing Elementary Science Skills: Instructional Effectiveness and Path Independence. *Cognitive Development, 23*(4), 488-511.

Stratford, S. J., Krajcik, J., & Soloway, E. (1998). Secondary Students' Dynamic Modeling Processes: Analyzing, Reasoning About, Synthesizing, and Testing Models of Stream Ecosystems. *Journal of Science Education & Technology, 7*(3), 215-234.

Tao, P.-K., & Gunstone, R. F. (1999). Conceptual change in science through collaborative learning at the computer. *2001, 21*(1), 39-57.

Tullos, A., & Woolley, J. D. (2009). The Development of Children's Ability to Use Evidence to Infer Reality Status (English). *Child development, 80*(1), 101-114.

van Bruegel, F., Riffell, J., Fairhall, A., & Dickinson, M. H. (2015). Mosquitoes Use Vision to Associate Odor Plumes with Thermal Targets. *Current Biology, 25*, 1-7.

Varma, K., & Linn, M. C. (2012). Using interactive technology to support students' understanding of the greenhouse effect and global warming. *Journal of Science Education and Technology, 21*(4), 453-464.

Veal, W. R. (2002). Content Specific Vignettes as Tools for Research and Teaching. *Electronic Journal of Science Education, 6*(4).

Vosniadou, S., & Brewer, W. F. (1992). Mental Models of the Earth: A Study of Conceptual Change in Childhood. *Cognitive Psychology, 24*, 535-585.

Weber, M. (2012). Experiment in Biology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Winter 2014 Edition)*: http://plato.stanford.edu/archives/win2014/entries/biology-experiment/.

Wellman, H. M., & Lagattuta, K. H. (2004). Theory of Mind for Learning and Teaching: The Nature and Role of Explanation. *Cognitive Development, 19*(4), 479-497.

White, B. Y. (1993). ThinkerTools: Causal Models, Conceptual Change, and Science Education. *Cognition and Instruction, 10*(1), 1-100.

Wills, J. M., & Samarapungavan, A. (2017). Rethinking Notions of Evidence in Research on Scientific Reasoning and Science Learning. *Manuscript in Preparation.*

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. doi:10.1002/sce.20259

Woodward, J. (2011). Mechanisms Revisited. *Synthese, 183*, 409-427.

Wu, H., Krajcik, J., & Soloway, E. (2002). Promoting conceptual understanding of chemical representations: Students' use of a visualization tool in the classroom. *Journal of Research in Science Teaching, 38*, 821-842.

Zimmerman, C., & Glaser, R. (2001). *Testing Positive Versus Negative Claims: A Preliminary Investigation of the Role of Cover Story on the Assessment of Experimental Design Skills (CSE Technical Report 554)*. Retrieved from Los Angeles, CA:

Ziv, M., & Frye, D. (2004). Children's Understanding of Teaching: The Role of Knowledge and Belief. *Cognitive Development, 19*(4), 457-477.

Zohar, A. (1995). Reasoning about interactions between variables. *Journal of Research in Science Teaching, 32*(10), 1039-1063.

APPENDIX A. EVIDENTIARY REASONING ASSESSMENT

Research Case 1: Do Acacia Trees Defend Themselves?

A team of Ecologists wants to understand why some of the Acacia trees in Africa have long thorns and some have short ones. Ecologists know that the Acacia tree naturally grows a short thorn. They also know plants respond to factors in their environment. Past studies have shown that if plants grow in the shadow of another plant, it will grow a stem high enough to reach the sunlight. Plants have also been shown to produce chemicals to make their leaves taste bad when their survival is threatened by a plant-eating animal or insect. This team wondered if the different sized thorns of the Acacia trees was a response to a threat in their environment. They wanted to test the hypothesis that the Acacia trees grew longer thorns as a way to defend themselves.

Before testing their ideas, they needed to be sure the trees with the long thorns were not a new kind of Acacia tree. They took DNA samples of both types and the DNA tests revealed the trees were the same. The picture on the left is shows the Acacia tree with the short thorns. The image on the right shows the Acacia tree with the long thorns.

Ecosystems are made up of very complex groups of living things that share a location. The ecosystem the trees lived in is filled with animals like impalas, leopards, and wild dogs. It also had a number of different plants and insects. Since the team could not accurately recreate the ecosystem, the scientists needed to carry out their study where the trees lived. They could then study other living things that share the Acacia tree's natural environment. This helped them to understand why only some of the trees had longer thorns. The scientists created several areas

for them to observe the ecosystem. These areas allowed the team to remain out of view. This was important since past studies have shown that animals behave differently when they are being watched.

To see how factors in the ecosystem affect the thorn length of the Acacia trees, the team conducted four tests in the Acacia trees natural environment. Over a five-year period, the scientists generated hundreds of detailed notes about the ecosystem and their tests. The tests and their results are described below:

Test 1: Observing the Ecosystem. The scientists spent a lot of time studying the environments of the trees with long thorns and the ones with short thorns.

Results of test 1: The Acacia trees with the long thorns were only found in open areas. The trees with the short thorns were found in wooded areas. The scientists also discovered that plant-eating Impalas spent most of their time in the open areas and they would feed on the Acacia leaves. Other animals that lived in the area like leopards and wild dogs spent their time in the wooded areas.

Test 2: **Impalas and Open Areas.** To see whether impalas preferred the open areas because they could easily see the leopards and wild dogs and not some other reason, they cleaned out part of a wooded area and turned into an open one.

Results of test 2: Impalas began spending time in the cleared area.

Test 3: **Do Impalas have a Preference**. The scientists created an eating space for the impalas. They pulled the long and short thorns off of their branches. They put the long thorns on the branches that originally had short ones and put the short thorns on the branches that originally had long ones. They also placed unchanged branches from both trees in the eating space.

Results of test 3: Impalas showed a preference for leaves on the branches with short thorns.

Test 4: Remove the Plant-Eating Impala. The scientists blocked off sections of open areas where the Acacia trees with long thorns were found. This prevented the impalas from eating the leaves of the tree.

Results of test 4: Over time, the large thorns surrounding the leaves began to get smaller.

After looking at the data from the hundreds of detailed notes and the results of the four tests, the reason why some of the Acacia trees grew long thorns became clear. Acacia trees with

long thorns were only found in open areas where the plant-eating impalas also spent their time. When the impalas were not able to feed on the Acacia trees, the long thorns began to return to regular size. The Acacia trees grew longer thorns as a way to defend themselves.

Question: Some people have asked why the scientists looked at plant eating animals that lived in the same area as the trees. The other people said the scientists should have looked at how much sun the trees received. They said sunlight could tell us why there were different sized thorns. Should the scientists have looked at other factors or were they right to focus on the other animals?

Explain your answer in a few sentences below.

_____

_____

_____

_____

_____

_____

The scientists made **three hypotheses**:

1) Impalas made choices about where they would eat based on how easily they could see the leopards and wild dogs.
2) Impalas liked the leaves of the trees with short thorns because they were easier to eat and not because they liked the taste better.
3) The trees with long thorns were found in the open areas because they had a higher risk of being eaten by impalas.

After talking with each other, the scientists designed a series of tests. They took place over five years in the areas where the Acacia trees lived.

In test 1**,** the scientists looked at detailed pictures of the area. They also spent a lot of time in the area to learn about where the animals lived.

Results of test 1: Impalas spent most of their time in open areas like meadows. The leopards and wild dogs spent their time in the wooded areas.

In test 2, the scientists thought about other reasons why the impalas may like the open areas. They cleaned a part of a wooded area so that it would look like an open area. The new area

did not have any of the other plants that lived in the open areas, though. Then the scientists watched to see if the impalas would begin to use the area.

Results of test 2: Impalas used the newly cleaned out area. They did not care that the area did not have any of the other plants found in open areas.

In test 3, the scientists switched the leaves of the trees to see if the impalas liked one better than another. They put leaves from branches with long thorns on branches with the short thorns and the other way around. The branches were then offered to a group of impalas as food.

Results of test 3: Impalas ate the leaves from the branches with short thorns. These were the leaves that originally came from the trees with large thorns.

Test 4: In the fourth test, the scientists blocked off sections of open areas where the trees with long thorns were found. The fencing stopped the Impalas from eating the leaves of the tree.

Results of test 4: Over time, the large thorns surrounding the leaves began to get smaller. The scientists thought about the results of all the tests. They decided that the evidence showed that the trees grew longer thorns as a way to protect their leaves from being eaten. So plants do defend themselves against threats to their survival.

During science, another 5th grade class read the same story you just read about the Acacia trees. Their teacher asked the class to pair up and talk about what the scientists did. Below are some examples of the class discussions. After reading samples of the student discussions, circle who you agree with most and explain your choice. Remember to do your BEST.

Serena and Jaden focused their conversation on the decisions the scientists made to test whether the longer thorns were a way for the Acacia tree to defend itself.

Question 1: Serena questions the number of tests the scientists did. She says that test four was the only experiment needed to show that the longer thorns were a survival response of the Acacia trees. Jaden thinks that all of the tests are important because they each provide unique information about how the Acacia trees respond to factors in their environment.

Do you agree with Serena or Jaden? Explain **why** you agree with Serena or Jaden.

Question 2: Jaden thinks the scientists should have done a test where the Acacia trees with short thorns were placed in open areas with the impalas to see if thorn length would change. Serena said that doing this test was not necessary because tests 3 & 4 show that thorn length was a response to environmental threats?

Do you agree with Jaden or Serena? Explain **why** you agree with Jaden or Serena.

Kevin and Rachel focused their conversation on the different aspects of the ecosystem the scientists decided to focus on in their study.

Question 3: Kevin questioned why the scientists chose to focus on the impalas as important to explaining why some of the Acacia trees had longer thorns. He thinks the scientists should also look at how the leopards and wild dogs influences the types of thorns the Acacia trees grow. Rachel thinks the scientists had good reasons to only focus on the impalas.

Do you agree with Kevin or Rachel? Explain **why** you agree with Kevin or Rachel.

Question 4: Rachel questioned why the scientists chose to ignore the make-up of the soil and the amount of sunlight the trees received. She thinks the scientists should have examined the soil and the amount of sunlight received for trees with both types of thorns. Kevin asked Rachel how investigating the soil and the amount of sunlight the trees receive helps to answer the question of whether the Acacia trees grew longer thorns as a way to defend themselves.

Rachel asks her classmates for help answering Kevin's question. After reading their responses, please circle the letter of the response you agree with.

a. Only examining the soil of the trees with the long thorns would help to answer the question. It would show whether the tree was getting what it needed to grow properly.
b. Examining the soil of the trees and the amount of sunlight they received would help to answer the question of whether the Acacia tree grew longer thorns as a way to defend itself because soil make-up and sunlight influence plant health and growth.
c. Examining the soil or the amount of sunlight they received would not help to answer the question of whether the Acacia trees grew longer thorns as a way to defend themselves. Adding these tests would only show if the trees lived in a healthy environment.
d. Only examining the amount of sunlight the trees received would help answer the question. It would show the amount of sunlight the trees received and that has an effect on growth.

Explain **why** your choice is the best one.

Alicia and Michael focused their discussion on the evidence the scientists used to decide that the Acacia trees grew longer thorns as a way to defend themselves.

Question 5: Alicia questions the scientists' conclusion that the Acacia trees grew longer thorns as a response to the plant-eating impalas in their environment. Since the scientists didn't examine whether other plant-eating organisms like insects or birds also fed on the Acacia trees leaves, she

thinks the scientists' evidence is limited. Michael thinks the evidence from the study is NOT limited.

Do you agree with Alicia or Michael? Explain **why** you agree with Alicia or Michael.

Question 6: Michael questions the evidence from the four tests. He thinks the scientists should have reported how many Acacia trees of each thorn size were in the study. Without this information, Michael has doubts about the quality of the evidence. Alicia thinks the number of trees with long and short thorns have nothing to do with the quality of the evidence.

Do you agree with Michael or Alicia? Explain **why** you agree with Do you agree with Michael or Alicia.

The final two questions were presented to the class by their science teacher. After reading the questions, write how you would respond to the teacher and why.

Question 7: The teacher asked the class to imagine that the scientists decided to plant some Acacia trees at a local zoo that had some impalas, leopards, and wild dogs instead of observing the trees in their natural environment. The teacher then asked the class to think carefully about whether changing the experimental tests in this way would influence the results of the study. Do you think changing the tests would or would not influence the results? Be sure to support your answer.

Question 8: The teacher asked the class to imagine that the scientists thought there were other factors in addition to the impalas that contributed to the Acacia trees growing longer thorns. The teacher then asked the class to think carefully about whether this information would change the tests the scientists decided to do.

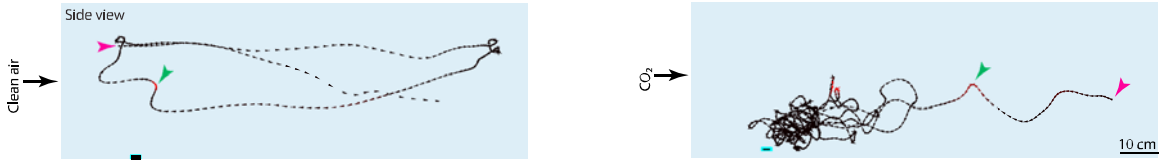
Do you think the addition of other factors would or would not change the tests the scientists decided to do? Be sure to support your answer.

Research Case 2: Why Do Mosquitos Bite?

A team of biologists wants to understand how mosquitos discover possible food sources. Biologists know that animals and humans give off body heat. They also breathe out an invisible gas called carbon dioxide. From past studies, scientists have learned that mosquitos can smell carbon dioxide. Mosquitos can also use their sense of touch to detect heat. Scientists' believe that sensing carbon dioxide and heat helps mosquitos to find their food sources. This team wondered if mosquitos also used their sense of sight to locate potential targets. They want to test the

hypothesis that mosquitos combine information from their senses of smell, touch and sight to find their food.

It is hard to test these ideas in nature because humans and animals give off smell, heat, and visual clues to mosquitos all at the same time. If mosquitos only use one type of clue such as carbon dioxide or body heat, it is impossible to tell which clue mosquitos use in the wild. To test their ideas, the scientists needed to control and be able to consistently change the type of sense information given to the mosquitos. They could then study how the mosquito behaved when each type of sense clue (smell, heat, or visual information) was present on its own as well as together. The scientists decided to use 100 mosquitos in a specially designed indoor lab. In their lab, the scientists could control which type of sense information was given to mosquitos. They could give each type of sense clue, such as carbon dioxide, heat, or visual information on its own or in combination with another clue. They could also track mosquito behavior better in the lab. It is not easy to follow mosquitos with just our eyes because they are small and move fast. The scientists used a special computer with a video camera to detect and record mosquito flight paths. A mosquito's flight path was defined as where the mosquito went during a test. The picture on the left shows a computer image of a mosquito's flight path when there was no carbon dioxide. The one on the right shows how the flight path of a mosquito when it smelled carbon dioxide.



To see how carbon dioxide, heat, and visual information affect mosquito behavior, the team ran four tests in their lab and recorded mosquito behavior for each test. The computer recorded thousands of mosquito flight paths for each test. The tests and their results are described below:

Test 1: Carbon Dioxide. To see how mosquitos behaved when only carbon dioxide was present, the scientists released carbon dioxide gas into the lab for a while and then turned it off.

Results of Test 1: While the carbon dioxide was present, the mosquitos flew all over the lab room. When the carbon dioxide gas was gone, the mosquitos went back to the walls and ceiling.

Test 2: Combined Heat and Visual: No carbon dioxide released in this test. To see how mosquitos behaved with only heat and visual information present, the scientists put two fake cows that they built in the room. One fake cow gave off heat and the other fake cow did not.

Results of Test 2: Even though both of the fake cows were in the room, the mosquitos did not move around much.

Test 3: Combined Visual and Smell Clues. The scientists put the fake cow that did not release any heat into the room. Then they released the carbon dioxide.

Results of Test 3: The mosquitos began to fly all over the lab room again. They flew to the fake cow that did not release heat, but when they got close they moved away and flew back to the walls and ceiling.

Test 4: Combined Smell, Visual and Heat Clues. The scientists put the fake cow that let off heat into the room and then released carbon dioxide.

Results of Test 4: The mosquitos left the walls and ceiling and began flying towards the fake cow that put off heat. The mosquitos then flew close to the fake cow and landed on it. After looking at the data from thousands of flight paths the computer recorded from the four tests, the pattern for how mosquitos use sense information to find food find food became clear. When mosquitos smell carbon dioxide, they begin to search for food. During the search mosquitos use visual cues to locate a potential food source. When a potential food source is located, mosquitos fly close enough to sense body heat. If the object puts off heat, mosquitos will land. This is how mosquitos find food.

During science, another 5th grade class read the same story you just read about how mosquitos find food. Their teacher asked the class to pair up and talk about what the scientists did. Below are some examples of the class discussions. After reading samples of the student discussions, circle who you agree with most and explain your choice. Remember to do your BEST.

Michele and Howard focused their conversation on the decisions the scientists made to test how mosquitos find food.

Question 1: Michele questions the number of tests the scientists did. She says that test four was the only experiment needed to show that mosquitos use a combination of senses to locate food

and bite. Howard thinks that all of the tests are important because they each provide unique information about how mosquitos find food.

Do you agree with Michele or Howard? Explain **why** you agree with Michele or Howard.

Question 2: Howard thinks the scientists should have done a test where the mosquitos were only given visual information to see if they use it to find food. Michele said that doing this test was not necessary because tests 3 & 4 show that mosquitos use visual information to find food?

Do you agree with Howard or Michele? Explain **why** you agree with Howard or Michele.

Brian and Jordan focused their conversation on the different sense information the scientists decided to focus on in their study.

Question 3: Brian questioned why the scientists chose to focus on carbon dioxide, the gas that people and animals breathe out, as important to explaining how mosquitos find food. He thinks the scientists should also look at how oxygen, the gas people and animals breathe in, influences mosquitos' search for food. Jordan thinks the scientists had good reasons to only focus on carbon dioxide.

Do you agree with Brian or Jordan? Explain **why** you agree with Brian or Jordan.

Question 4: Jordan questioned why the scientists chose to focus only on one size of fake animal and whether or not it gave off heat. She thinks the scientists should have varied the size of fake animals and the amount of heat they gave off. Brian asked Jordan how changing the size of the fake animals or the amount of heat they gave off helps to answer the question of how mosquitos use sense information to find food.

Jordan asks her classmates for help answering Brian's question. After reading their responses, please circle the letter of the response you agree with.

a. Only changing the size of the fake animals would help to answer the question of how mosquitos find food. It would show if mosquitos preferred large or small animals.
b. Changing the size of the fake animals and the amount of heat they put out would help to answer the question. Changing the size would show if mosquitos preferred large or small animals. And changing the amount of heat would show if they have a preference for animals that put out a certain amount of heat.
c. Changing the size of the fake animals and the amount of heat they put out would not help to answer the question of how mosquitos find food. Adding these tests would only show if mosquitos had a size preference and how sensitive mosquitos are to heat information.

d. Only changing the amount of heat the fake animals put out would help to answer the question of how mosquitos find food. It would show if they have a preference for animals that put out a certain amount of heat.

Explain **why** your choice is the best one.

Olivia and Jackson focused their discussion on the evidence the scientists used to decide that mosquitos rely on a combination of sense information to find food.

Question 5: Olivia questions the scientists' conclusion that mosquitos rely on a combination of smell, visual, and heat information to find food. Since the scientists didn't experiment with different types of mosquitos, she thinks their evidence is limited to only the mosquitos used in the study. Jackson thinks the evidence from the study is NOT limited.

Do you agree with Olivia or Jackson? Explain **why** you agree with Olivia or Jackson.

Question 6: Jackson questions the evidence from the four tests. He thinks the scientists should have reported how accurate the computer was at recording the mosquitos. Without this information, Jackson has doubts about the quality of the evidence. Olivia thinks the accuracy of the computer doesn't have anything to do with the quality of the evidence.

Do you agree with Jackson or Olivia? Explain **why** you agree with Jackson or Olivia.

The final two questions were presented to the class by their science teacher. After reading the questions, write how you would respond to the teacher and why.

Question 7: The teacher asked the class to imagine that the scientists decided to watch the mosquitos in nature instead of using a lab with a computer to record them. The teacher then asked the class to think carefully about whether changing the experimental tests in this way would influence the results of the study.

Do you think changing the tests would or would not influence the results? Be sure to support your answer.

Question 8: The teacher asked the class to imagine that the scientists thought there were other sense cues in addition to smell, heat, and visual information that mosquitos relied on to find food. The teacher then asked the class to think carefully about whether this information would change the tests the scientists decided to do.

Do you think the addition of other sense cues would or would not influence the tests the scientists decided to do? Be sure to support your answer.

APPENDIX B. ASSESSMENT OF SCIENCE INTEREST

1. **Science is interesting.**

   1— strongly disagree

   2—disagree

   3—do not know

   4—agree

   5—strongly agree

2. **I am good at science.**

   1— strongly disagree

   2—disagree

   3—do not know

   4—agree

   5—strongly agree

3. **I liked the research case I completed.**

1— strongly disagree

2—disagree

3—do not know

4—agree

5—strongly agree

APPENDIX C. ERA ITEM SCORING: ACACIA TASK

Questions 1 & 2: This question set targets constructs in the quality of design & data collection procedures section of the conceptual framework for thinking about scientific evidence. Question 1 asks students to consider the value of the experimental tests as a set. Question 2 asks students to think about the added value of adding another test to the study that targets a variable of the study isolation.

Question 1: Serena questions the number of tests the scientists did. She says that test four was the only experiment needed to show that the longer thorns were a survival response of the Acacia trees. Jaden thinks that all the tests are important because they each provide unique information about the Acacia trees and their environment.

Do you agree with Serena or Jaden? Explain **why** you agree with Serena or Jaden.

Score	Description
0	No response, I don't know, or an irrelevant answer.
1	Response simply restates information given in the text and/or does not contain evidence of an understanding about the way each test in the study is needed to generate a picture of why some trees had longer thorns. For example, I agree that the scientists only needed to do test four or I think they needed all of them.
2	Answer contains a beginning understanding of the selection & design of the experimental tests. Student focuses on simple rules (e.g., values test set based on a simple rule: more (tests, research, etc.) = more information = better) or if you didn't do all the tests you wouldn't have enough information.
3	Response reflects a developing understanding. References particulars from one or more tests (e.g., information from test 1) to illustrate value but does not provide any other details about how the quality of information from the test set would be impacted by only conducting one test.
4	Answer demonstrates an advancing understanding about the role the experimental tests played in developing an explanation of why some trees have longer thorns. Incorporates greater level of detail about relevant issues such as: connects information gained in the test set to understanding why some of the trees had longer thorns or focuses on the details of a specific test outcome and identifies its importance; makes a comparison between the information acquired by the test set with the single suggested study.

Question 2: Jaden thinks the scientists should have done a test where the Acacia trees with short thorns were placed in open areas with the impalas to see if thorn length would change. Serena said that doing this test was not necessary because tests 3 & 4 show that thorn length was a response to environmental threats?

Do you agree with Jaden or Serena? Explain **why** you agree with Jaden or Serena.

Score	Description
0	No response, I don't know, or an irrelevant answer.
1	Response simply restates information given in the text and/or does not contain evidence of an understanding of the selection & design of the experimental tests. For example, I agree/disagree that the scientists should have done a test on the trees with short thorns.
2	Answer contains a beginning understanding of the selection & design of the experimental tests. Student focuses on simple rules (more=better) or low level agreement/critiques like the scientists should do that test because they did other tests with only certain things (e.g., roped off wooded areas).
3	Response reflects a developing understanding about the selection & design of the experimental tests. For example, student references aspects of a specific test that illustrates the Acacia trees responded to threats by growing longer thorns (e.g., test 4 showed that longer thorns began to shorten) or notes that the suggested test is the reverse of test 4 but does not provide any other details.
4	Answer demonstrates an advancing understanding about the selection & design of the experimental tests. Incorporates greater level of detail about relevant issues such as: proposes a plausible reason to revise the design (e.g., other environmental differences btw the open & wooded areas exist (e.g., soil differences, etc.)) to include the suggested test; goes further than recognizing that the suggested test is a reversal of test 4 and highlights that scientists don't just do tests to do them – tests are selected based on their ability to contribute important information.

Questions 3 & 4: This question set targets constructs in the variable selection and operationalization section of the conceptual framework for thinking about scientific evidence. Question 3 asks students to think about the studies focus on a specific variable at the exclusion of a similar type of variable. Question 4 proposes the addition of 2 variables and asks students to consider value of this change in determining whether the Acacia trees defend themselves.

Question 3: Kevin questioned why the scientists chose to focus on the impalas as an important part of why some of the Acacia trees had longer thorns. He thinks the scientists should also look at how the leopards and wild dogs influence the types of thorns the Acacia trees grow. Rachel thinks the scientists had good reasons to only focus on the impalas.

Do you agree with Kevin or Rachel? Explain **why** you agree with Kevin or Rachel.

Score	Description
0	No response, I don't know, or an irrelevant answer.
1	Response simply restates information given in the text and/or does not contain evidence of an understanding of the studies variables or how they relate to the purpose of the study. For example, I agree/disagree that the scientists should have also looked at the influence of the wild dogs and leopards on the trees.
2	Answer contains a beginning understanding about why specific variables were chosen. Student focuses on simple rules (more=better) or low level agreement/critiques; for example, the scientists did tests with the impalas so they should have tested the leopards and wild dogs or they should've tested other animals too.
3	Response reflects a developing understanding about why specific variables were chosen. For example, references a specific test outcome that shows the importance of the impalas (e.g., the impalas are the ones that eat the leaves) but does not provide any other details.
4	Answer demonstrates an advancing understanding about why specific variables were chosen. Incorporates greater level of detail about relevant issues such as: details how past research and test outcomes support focusing on the impalas (e.g., longer thorns began to get smaller when impalas were removed); identifies issues like if wild dogs and leopards were important, they would be feeding on the trees; highlights indirect influence of predators (e.g., their presence in wooded areas causes impalas to feed in open areas).

Question 4: Rachel questioned why the scientists chose to ignore the soil the trees with short and long thorns lived in and the amount of sunlight they received. She thinks the scientists should have examined the soil and the amount of sunlight received for trees with both types of thorns. Kevin asked Rachel how investigating the soil and the amount of sunlight the trees receive helps to answer the question of whether the Acacia trees grew longer thorns as a way to defend themselves.

Rachel asks her classmates for help answering Kevin's question. After reading their responses, please circle the letter of the response you agree with.

a. Only examining the soil of the trees with the long thorns would help to answer the question. It would show whether the tree was getting what it needed to grow properly.
b. Examining the soil of the trees and the amount of sunlight they received would help to answer the question of whether the Acacia tree grew longer thorns as a way to defend itself because soil make-up and sunlight influence plant health and growth.
c. Examining the soil or the amount of sunlight they received would not help to answer the question of whether the Acacia trees grew longer thorns as a way to defend themselves. Adding these tests would only show if the trees lived in a healthy environment.
d. Only examining the amount of sunlight the trees received would help answer the question. It would show the amount of sunlight the trees received and that has an effect on growth.

Explain **why** your choice is the best one.

Score	Description
0	No response, I don't know, or an irrelevant answer.
1	Response simply restates information given in the text and/or does not contain evidence of an understanding of why the variables are targeted or how they connect to the purposes of the study. For example, selects [A, B, or D] and says that the scientists should've examined these things or simply says the scientists shouldn't have examined one or both.
2	Answer contains a beginning understanding about variable selection. For example, selects [A, B, C, or D] and attempts to demonstrate the value of examining soil nutrients and sunlight by referring to a simple heuristic (more = better); selects C and dismisses the suggested variables as irrelevant.
3	Response reflects a developing understanding about the selection of variables and how they contribute information to the answering the original question. For example, agrees that knowing whether differences exist in soil make-up or sun exposure would provide important data (e.g., sunlight/soil make-up influence growth; potential impact) but does not provide any other details.
4	Answer demonstrates an advancing understanding about the selection of variables and how they contribute information to the answering the original question. Incorporates greater level of detail about relevant issues such as: identifies the benefit in the granularity of examining soil make-up and sun exposure (provides comparative example or additional detail about its benefit); identifies that the additional variables would contribute an increase in knowledge about thorn length (there is or is not a relationship).

Questions 5 & 6: This pair of questions engages students about constructs in the interpretations / conclusions section of the conceptual framework for thinking about scientific evidence. Question 5 presents students with a claim that the evidence is incomplete because the scientists didn't examine other variables that could help explain the phenomena. Question 6 asks students to consider a claim that the quality of the evidence is reduced based on the experimental sample.

Question 5: Alicia questions the scientists' conclusion that the Acacia trees grew longer thorns as a response to the plant-eating impalas in their environment. Since the scientists didn't examine whether other plant-eating organisms like insects or birds also fed on the Acacia trees leaves, she thinks the scientists' evidence is incomplete. Michael thinks the evidence from the study is NOT incomplete.

Do you agree with Alicia or Michael? Explain **why** you agree with Alicia or Michael.

Score	Description
0	No response, I don't know, or an irrelevant answer.
1	Response simply restates information given in the text and/or does not contain evidence of an understanding of how the focus variables are related to the completeness of evidence. For example, I agree/disagree that the scientists should have examined whether other organisms fed on the trees.
2	Answer contains a beginning understanding of the relationship between the experimental variables and the evidence. For example, the scientists should have looked at other plant-eating organisms; justifies answer based on simple heuristic: more (tests, research, etc.) = more information = better; dismisses the claim that examining other plant-eating organisms is connected to the evidence.
3	Response reflects a developing understanding about the relationship between the focus variables and the evidence. For example, answer references the results of test 1 when the scientists spent a lot of time observing the ecosystem and/or singles out the impalas as the causal force behind the longer thorns (e.g., they were the only ones eating the leaves) but doesn't provide any other details.
4	Answer demonstrates an advancing understanding about the relationship between the focus variables and the evidence. Incorporates greater level of detail about relevant issues such as: criticizes the plausibility that insects & birds were potentially eating the leaves because their small size would not be influenced by thorn length; references important aspects of the study to refute the suggestion that other organisms could be eating the leaves such as the lengthy duration of the study & connects that to the reduced possibility that birds or insects would've been missed; questions why birds or plant-eating insects would only be eating leaves of trees in open areas.

Question 6: Michael questions the evidence from the four tests. He thinks the scientists should have reported how many Acacia trees of each thorn size were in the study. Without this information, Michael has doubts about the quality of the evidence. Alicia thinks the number of trees with long and short thorns have nothing to do with the quality of the evidence.

Do you agree with Michael or Alicia? Explain **why** you agree with Do you agree with Michael or Alicia.

Score	Description
0	No response, I don't know, or an irrelevant answer.
1	Response restates information given in the text and/or does not contain evidence of an understanding of how the experimental sample is related to the quality of evidence. For example, the scientists should have reported how many trees were in the study.
2	Answer contains a beginning understanding about relationship between the sample size and the quality of evidence. For example, the scientists should have included how many trees of each they looked but justifies answer based on simple heuristic: more (tests, research, etc.) information = better; dismisses the claim that the number of trees with each type of thorn is relevant to the quality of the evidence.
3	Response reflects a developing understanding about the relationship between sample characteristics and the evidence. For example, answer references the way sample size is related to evidence generally (e.g., the more trees of each the better) but doesn't provide any other details.
4	Answer demonstrates an advancing understanding of experimental sample and how it can impact the quality of evidence. Incorporates greater level of detail about relevant issues such as: impact of a non-representative sample on the quality of evidence (e.g., if you have a small number that you are looking at, then maybe your results aren't as good as if you have a lot); gives example of how small sample (e.g., 4 total trees) makes results much less compelling than one with 20 of each; assigns an acceptable number of trees (e.g., if sample > than 20 of each); also notes the importance of equality (e.g., 2 long thorn vs 15 short thorn or vice versa).

Questions 7 & 8: The final set of questions asks students to consider the interrelatedness between the distinct phases of scientific inquiry. Question 7 presents students with a scenario where the design of the experimental tests has been altered and then asks students to think about whether this would change the results. Question 8 provides students with a scenario where the selection of variables had changed and asks students to reason about whether it would influence the experimental tests.

Question 7: The teacher asked the class to imagine that the scientists decided to plant some Acacia trees at a local zoo that had some impalas, leopards, and wild dogs instead of observing the trees in their natural environment. The teacher then asked the class to think carefully about whether changing the experimental tests in this way would influence the results of the study.

Do you think changing the tests would or would not influence the results? Be sure to support your answer.

Score	Description
0	No response, I don't know, or an irrelevant answer.
1	Response simply restates information given in the text and/or does not demonstrate an understanding of how changes in the design of a test impact the evidence. For example, I think it will/will not change.
2	Answer contains a beginning understanding about the relationship between the test design and their outcomes. For example, student cites that change will occur but relies on a simple rule (changes here = changes there); notes a potential change but it is irrelevant.
3	Response reflects a developing understanding about the relationship between the experimental design and evidence in the study. Answer acknowledges that change will occur and uses a relevant example to support position (e.g., environmental differences of zoo) but doesn't provide any other details.
4	Answer demonstrates an advancing understanding of the relationship between the design of experiments and the evidence. Incorporates greater level of detail about relevant issues such as: complexity related to variable control (e.g., impalas, wild dogs, leopards & their interactions); the difficulty with recreating an ecosystem (as stated in the article); details the difficulty/complexity of executing study in zoo and connects it to accuracy of results.

Question 8: The teacher asked the class to imagine that the scientists thought there were other factors in addition to the impalas that contributed to the Acacia trees growing longer thorns. The teacher then asked the class to think carefully about whether this information would change the tests the scientists decided to do and how the tests would change.

Do you think the addition of other factors **would** or **would not** change the tests the scientists decided to do? Be sure to support your answer.

Score	Description
0	No response, I don't know, or an irrelevant answer.
1	Response simply restates information given in the text and/or does not demonstrate an understanding of how changes in the selection of focus variables impacts the design of the tests. For example, I think it will/will not change.
2	Answer reflects a beginning understanding that changing or adding variables will result in changes to the experimental tests. For example, student cites that change will occur but is relying on a simple rule (changes here = changes there) or references potential changes but they are irrelevant.
3	Response reflects a developing understanding about the relationship between the selection of focus variables and experimental tests. For example, I think they will change because adding other factors would require additional tests – student may even suggest a hypothetical factor but doesn't provide any additional details.
4	Answer demonstrates an advancing understanding of the relationship between the focus variables and the experimental tests. Incorporates greater level of detail about relevant issues such as: connects additional variables to the need for new tests and connects them to increased outcomes/evidence; student suggests a hypothetical variable and demonstrates impact on test set; identifies the need for additional controls.

APPENDIX D. ERA ITEM SCORING: MOSQUITO TASK

Questions 1 & 2: This question set targets constructs in the quality of design & data collection procedures section of the conceptual framework for thinking about scientific evidence. Question 1 asks students to consider the value of the experimental tests as a set. Question 2 asks students to consider adding a test where one of the variables in the study is examined in isolation of the others.

Question 1: Michele questions the number of tests the scientists did. She says that test four was the only experiment needed to show that mosquitos use a combination of senses to locate food and bite. Howard thinks that all the tests are important because each of them adds valuable information about how mosquitos find food.

Do you agree with Michele or Howard? Explain **why** you agree with Michele or Howard.

Score	Description
0	No response, I don't know, or an irrelevant answer.
1	Response simply restates information given in the text and/or does not contain evidence of an understanding about the way each test in the study is needed to generate a picture of how mosquitos find food. For example, I agree that the scientists only needed to do test four or I think they needed all of them.
2	Answer contains a beginning understanding of the selection & design of the experimental tests. Student focuses on simple rules (e.g., values test set based on a simple rule: more (tests, research, etc.) = more information = better) or if you didn't do all the tests you wouldn't have enough information.
3	Response reflects a developing understanding. References details from any one or more tests (e.g., impact of CO_2 on behavior) but does not provide any other details about how the quality of information from the test set would be impacted by only conducting one test.
4	Answer demonstrates an advancing understanding about the role experimental tests play in developing the explanation of how mosquitos find food. Incorporates greater level of detail about relevant issues such as: how a test or multiple tests contribute to understanding how mosquitos make use of sense data; references how the single suggested test would limit the amount of information about how sensory information is utilized (e.g., lack of controls).

Question 2: Howard thinks the scientists should have done a test where the mosquitos were only given visual information to see if they use it to find food. Michele said that doing this test was not necessary because tests 3 & 4 show that mosquitos use visual information to find food?

Do you agree with Howard or Michele? Explain **why** you agree with Howard or Michele.

Score	Description
0	No response, I don't know, or an irrelevant answer.
1	Response simply restates information given in the text and/or does not contain evidence of an understanding of the selection & design of the experimental tests. For example, I agree/disagree that the scientists should have done a test on just visual information.
2	Answer contains a beginning understanding of the selection & design of the experimental tests. Student focuses on simple rules (more=better) or low level agreement/critiques like the scientists should do that test because they did other tests with only certain things (e.g., CO_2).
3	Response reflects a developing understanding about the selection & design of the experimental tests. For example, student references aspects of a specific test that illustrates mosquitos use visual information to find food (e.g., tests 2, 3, & 4 show that mosquitos use visual information in their search for food) or values the suggested test by pointing out the tests where vision was used also included other variables but does not provide any other details.
4	Answer demonstrates an advancing understanding about the selection & design of the experimental tests. Incorporates greater level of detail about relevant issues such as: outcome that showed CO_2 was search trigger; outcome showing presence of visual information and mosquito behavior; suggests a plausible reason to revise the design (e.g., inter-mosquito sense differentiation or test to determine the extent to which eyesight is relied upon/simple landing could be directed solely by other sense information (e.g., bats)).

Questions 3 & 4: This question set targets constructs in the variable selection and operationalization section of the conceptual framework for thinking about scientific evidence. Question 3 asks students to think about the studies focus on a specific variable at the exclusion of a similar type of variable. Question 4 proposes an additional layer of variation between 2 of the studies variables and asks students to consider value of this change to learning how mosquitos find food.

Question 3: Brian questioned why the scientists chose to focus on carbon dioxide, the gas that people and animals breathe out, as important to explaining how mosquitos find food. He thinks the scientists should also look at how oxygen, the gas people and animals breathe in, influences mosquitos' search for food. Jordan thinks the scientists had good reasons to only focus on carbon dioxide.

Do you agree with Brian or Jordan? Explain **why** you agree with Brian or Jordan.

Score	Description
0	No response, I don't know, or an irrelevant answer.
1	Response simply restates information given in the text and/or does not contain evidence of an understanding of the studies variables or how they relate to the purpose of the study. For example, I agree/disagree that the scientists should have also looked at the influence of oxygen (or other gases) on how mosquitos find food.
2	Answer contains a beginning understanding about why specific variables were chosen. Student focuses on simple rules (more=better) or low level agreement/critiques; for example, the scientists tested carbon dioxide so they should have tested oxygen as well or they should've tested other gases too.
3	Response reflects a developing understanding about why specific variables were chosen. For example, references specific test outcome that shows the importance of CO_2 (mosquitos response to CO_2 or their actions when it was absent) but does not provide any other details.
4	Answer demonstrates an advancing understanding about why specific variables were chosen. Incorporates greater level of detail about relevant issues such as: details how past research and test outcomes support focusing on CO_2 (e.g., mosquitos' reaction with/out); identifies issues like if oxygen was important, mosquitos would be interested in plants; highlights the ubiquitous nature of oxygen; details the way CO_2 contributes location information in a way oxygen does not.

Question 4: Jordan questioned why the scientists chose to focus only on one size of fake animal and whether or not it gave off heat. She thinks the scientists should have varied the size of fake animals and the amount of heat they gave off. Brian asked Jordan how changing the size of the fake animals or the amount of heat they gave off helps to answer the question of how mosquitos use sense information to find food.

Jordan asks her classmates for help answering Brian's question. After reading their responses, please circle the letter of the response you agree with.

a. Only changing the size of the fake animals would help to answer the question of how mosquitos find food. It would show if mosquitos preferred large or small animals.
b. Changing the size of the fake animals and the amount of heat they put out would help to answer the question. Changing the size would show if mosquitos preferred large or small animals. And changing the amount of heat would show if they have a preference for animals that put out a certain amount of heat.
c. Changing the size of the fake animals and the amount of heat they put out would not help to answer the question of how mosquitos find food. Adding these tests would only show if mosquitos had a size preference and how sensitive mosquitos are to heat information.
d. Only changing the amount of heat the fake animals put out would help to answer the question of how mosquitos find food. It would show if they have a preference for animals that put out a certain amount of heat.

Explain **why** your choice is the best one.

Score	Description
0	No response, I don't know, or an irrelevant answer.
1	Response simply restates information given in the text and/or does not contain evidence of an understanding of why the variables are targeted or how they connect to the purposes of the study. For example, selects [A, B, C, or D] and says that the scientists should have varied these things or simply says the scientists shouldn't vary one or both.
2	Answer contains a beginning understanding about variable selection. For example, selects [A, B, or D] and attempts to demonstrate the value of varying size and/or heat by referring to a simple heuristic (more = better); selects C and dismisses the suggested variables as irrelevant.
3	Response reflects a developing understanding about the selection of variables and how they contribute information to the answering the original question. For example, agrees that knowing how mosquitos respond to either size and/or heat variations would provide important data about how mosquitos' find food but does not provide any other details.
4	Answer demonstrates an advancing understanding about the selection of variables and how they contribute information to the answering the original question. Incorporates greater level of detail about relevant issues such as: identifies the benefit in the granularity of varying heat and size (provides comparative example or additional detail about its benefit); identifies that the additional variables would contribute an increase in knowledge about how mosquitos find food (there is or is not a relationship).

Questions 5 & 6: This pair of questions engages students about constructs in the interpretations / conclusions section of the conceptual framework for thinking about scientific evidence. Question 5 presents students with a claim about the limits of the reported evidence based on the studies sample. Question 6 asks students to consider a claim reducing the impact of the evidence based on the accuracy/precision of experimental tools.

Question 5: Olivia questions the scientists' conclusion that mosquitos rely on a combination of smell, visual, and heat information to find food. Since the scientists didn't experiment with different types of mosquitos, she thinks their evidence is limited to only the mosquitos used in the study. Jackson thinks the evidence from the study is NOT limited.

Do you agree with Olivia or Jackson? Explain **why** you agree with Olivia or Jackson.

Score	Description
0	No response, I don't know, or an irrelevant answer.
1	Response simply restates information given in the text and/or does not contain evidence of an understanding of how the experimental sample is related to the evidence. For example, I agree/disagree that the scientists should have experimented with different types of mosquitos.
2	Answer contains a beginning understanding about the relationship between the sample and the quality of evidence. For example, the scientists should have looked at different types of mosquitos; justifies answer based on a simple heuristic: more (tests, research, etc.) = more information = better; dismisses the claim that testing different types is connected to the evidence.
3	Response reflects a developing understanding about the relationship between the sample characteristics and the evidence. For example, answer references the way sample size is related to evidence generally (e.g., the more mosquitos they examine that better) but doesn't provide any other details; references general differences that may obtain from examining different types (e.g., different types of mosquitos could be attracted to different things); cites the # (100) of mosquitos used in the study as acceptable size/number to draw conclusions.
4	Answer demonstrates an advancing understanding of experimental sample and how it can impact the quality of evidence. Incorporates greater level of detail about relevant issues such as: impact of a non-representative sample on the quality of evidence (e.g., if you have a small number that you are looking at, then your results will not be as good as if you have a lot); isolates a specific piece of sensory information and demonstrates how it could vary across types (e.g., different types of mosquitos may respond to CO_2, heat, or visual information differently); argues that without knowing the types that were used in the study (there actually is no information about type in the story), no definitive answer can be generated.

Question 6: Jackson questions the evidence from the four tests. He thinks the scientists should have reported how accurate the computer was at recording the mosquitos. Without this information, Jackson has doubts about the quality of the evidence. Olivia thinks the accuracy of the computer doesn't have anything to do with the quality of the evidence.

Do you agree with Jackson or Olivia? Explain **why** you agree with Jackson or Olivia.

Score	Description
0	No response, I don't know, or an irrelevant answer.
1	Response simply restates information given in the text and/or does not demonstrate an understanding of how the accuracy/precision of experimental tools are related to the quality of evidence. For example, the scientists should've reported the accuracy of the computer.
2	Answer contains a beginning understanding of the relationship between the accuracy of the experimental tools and the quality of evidence. For example, dismisses claims about the accuracy (e.g., doesn't matter because the scientists were watching); refers to the real tendency for technology to be inaccurate but does not connect it to results; applies simple rule more = better.
3	Response reflects a developing understanding about the relationship between the tools used in an experiment and the quality of evidence. For example, the scientists should have reported how accurate the computer was because the results of the tests depend on it or notes that if the computer malfunctioned, the results are also affected but does not provide any other details.
4	Answer demonstrates an advancing understanding about the relationship between the accuracy of experimental tools and quality of evidence. Incorporates greater level of detail about relevant issues such as: assigns an acceptable value of accuracy (e.g., if > 90%, scientists shouldn't worry about it); references considerations of experimental error; connects tool accuracy to the design process (e.g., the scientists should have looked at how accurate the computer program when they were thinking about their experiments).

Questions 7 & 8: The final set of questions asks students to consider the interrelatedness between the distinct phases of scientific inquiry. Question 7 presents students with a scenario where the design of the experimental tests has been altered and then asks students to think about whether this would change the results. Question 8 provides students with a scenario where the selection of variables had changed and asks students to reason about whether it would influence the experimental tests.

Question 7: The teacher asked the class to imagine that the scientists decided to watch the mosquitos in nature instead of using a lab with a computer to record them. The teacher then asked the class to think carefully about whether changing the experimental tests in this way would influence the results of the study.

Do you think changing the tests would or would not influence the results? Be sure to support your answer.

Score	Description
0	No response, I don't know, or an irrelevant answer.
1	Response simply restates information given in the text and/or does not demonstrate an understanding of how changes in the test design impacts the outcomes. For example, I think it will/will not change.
2	Answer contains a beginning understanding about the relationship between the test design and their outcomes. For example, student cites that change will occur but is relying on a simple rule (changes here = changes there) or references potential changes but they are irrelevant.
3	Response reflects a developing understanding about the relationship between the experimental design and evidence in the study. Answer acknowledges that change will occur and uses a relevant example to support position (e.g., environmental differences of natural environment) but doesn't provide any other details.
4	Answer demonstrates an advancing understanding of the relationship between the design of experiments and the evidence. Incorporates greater level of detail about relevant issues such as: complexity related to variable control (e.g., carbon dioxide levels, visual information, heat; sample); how could the scientists visually track individual mosquitos – connects difficulty/complexity of executing study in nature to accuracy of results.

Question 8: The teacher asked the class to imagine that the scientists thought there were other sense cues in addition to smell, heat, and visual information that mosquitos relied on to find food. The teacher then asked the class to think carefully about whether this information would change the tests that the scientists did in the study and how the tests would change.

Do you think the addition of other sense cues **would** or **would not** influence the tests that the scientists did? Be sure to support your answer.

Score	Description
0	No response, I don't know, or an irrelevant answer.
1	Response simply restates information given in the text and/or does not demonstrate an understanding of how changes in the selection of focus variables impacts the design of the tests. For example, I think it will/will not change.
2	Answer reflects a beginning understanding that changing or adding variables will result in changes to the experimental tests. For example, student cites that change will occur but is relying on a simple rule (changes here = changes there) or references potential changes but they are irrelevant.
3	Response reflects a developing understanding about the relationship between the selection of focus variables and experimental tests. For example, I think they will change because adding sense cues would require additional tests – student may even suggest a hypothetical sense cue but doesn't provide any other details.
4	Answer demonstrates an advancing understanding of the relationship between the focus variables and the experimental tests. Incorporates greater level of detail about relevant issues such as: connects additional variables to the need for new tests and connects them to increased outcomes/evidence; student suggests a hypothetical variable and demonstrates impact on test set; identifies the need for additional controls.

APPENDIX E. STUDENT INTERVIEWS BY CASE

Study 1 – Acacia Task

Hello, (say student's first name). My name is Jamison Wills; I am from Purdue University. I am going to ask you some questions about your answers to the stories you read a few days ago. I will be recording (show them the recorder) what we talk about to help me remember what you said.

[**Remind student that the questions I am asking are not a test & will not affect their grades in any way**]

Question 1: Serena questions the number of tests the scientists did. She says that test four was the only experiment needed to show that the longer thorns were a survival response of the Acacia trees. Jaden thinks that all of the tests are important because they each provide unique information about how the Acacia trees respond to factors in their environment.

You answered **(show student their response**)

Can you tell me more about why you agreed with [Serena or Jaden]?

Once student finishes, ask them if there is anything else?

Question 2: Jaden thinks the scientists should have done a test where the Acacia trees with short thorns were placed in open areas with the impalas to see if thorn length would change. Serena said that doing this test was not necessary because tests 3 & 4 show that thorn length was a response to environmental threats?

You answered **(show student their response**)

Can you tell me more about why you agreed with [Serena or Jaden]?

Once student finishes, ask them if there is anything else?

Question 3: Kevin questioned why the scientists chose to focus on the impalas as important to explaining why some of the Acacia trees had longer thorns. He thinks the scientists should also look at how the leopards and wild dogs influences the types of thorns the Acacia trees grow. Rachel thinks the scientists had good reasons to only focus on the impalas.

You answered **(show student their response**)

Can you tell me more about why you agreed with [Kevin or Rachel]?

Once student finishes, ask them if there is anything else?

Question 4: Rachel questioned why the scientists chose to ignore the make-up of the soil and the amount of sunlight the trees received. She thinks the scientists should have examined the soil and the amount of sunlight received for trees with both types of thorns. Kevin asked Rachel how investigating the soil and the amount of sunlight the trees receive helps to answer the question of whether the Acacia trees grew longer thorns as a way to defend themselves.

You selected (a-b-c-d-e-f, **show student their response**), can you tell me why this is the best choice?

Once student finishes, ask them if there is anything else?

Question 5: Alicia questions the scientists' conclusion that the Acacia trees grew longer thorns as a response to the plant-eating impalas in their environment. Since the scientists didn't examine whether other plant-eating organisms like insects or birds also fed on the Acacia trees leaves, she thinks the scientists' evidence is incomplete. Michael thinks the evidence from the study is NOT incomplete.

You answered **(show student their response**)

Can you tell me more about why you agreed with [Alicia or Michael]?

Once student finishes, ask them if there is anything else?

Question 6: Michael questions the evidence from the four tests. He thinks the scientists should have reported how many Acacia trees of each thorn size were in the study. Without this information, Michael has doubts about the quality of the evidence. Alicia thinks the number of trees with long and short thorns have nothing to do with the quality of the evidence.

You answered **(show student their response**)

Can you tell me more about why you agreed with [Alicia or Michael]?

Once student finishes, ask them if there is anything else?

Question 7: The teacher asked the class to imagine that the scientists decided to plant some Acacia trees at a local zoo that had some impalas, leopards, and wild dogs instead of observing the trees in their natural environment. The teacher then asked the class to think carefully about whether changing the experimental tests in this way would influence the results of the study.

Can you tell me more about why your answer is the best?

Once student finishes, ask them if there is anything else?

Question 8: The teacher asked the class to imagine that the scientists thought there were other factors in addition to the impalas that contributed to the Acacia trees growing longer thorns. The

teacher then asked the class to think carefully about whether this information would change the tests the scientists decided to do.

Can you tell me more about your thoughts on this question?

Once student finishes, ask them if there is anything else?

Study 2 – Mosquito Task

Hello, (say student's first name). My name is Jamison Wills; I am from Purdue University. I am going to ask you some questions about your answers to the stories you read a few days ago. I will be recording (show them the recorder) what we talk about to help me remember what you said.

[**Remind student that the questions I am asking are not a test & will not affect their grades in any way**]

Question 1: Michele questions the number of tests the scientists did. She says that test four was the only experiment needed to show that mosquitos use a combination of senses to locate food and bite. Howard thinks that all of the tests are important because they each provide unique information about how mosquitos find food.

You answered (**show student their response**)

Can you tell me more about why you agreed with [Michele or Howard]?

Once student finishes, ask them if there is anything else?

Question 2: Howard thinks the scientists should have done a test where the mosquitos were only given visual information to see if they use it to find food. Michele said that doing this test was not necessary because tests 3 & 4 show that mosquitos use visual information to find food?

You answered (**show student their response**)

Can you tell me more about why you agreed with [Michele or Howard]?

Once student finishes, ask them if there is anything else?

Question 3: Brian questioned why the scientists chose to focus on carbon dioxide, the gas that people and animals breathe out, as important to explaining how mosquitos find food. He thinks the scientists should also look at how oxygen, the gas people and animals breathe in, influences mosquitos' search for food. Jordan thinks the scientists had good reasons to only focus on carbon dioxide.

You answered (**show student their response**)

Can you tell me more about why you agreed with [Brian or Jordan]?

Once student finishes, ask them if there is anything else?

Question 4: Jordan questioned why the scientists chose to focus only on one size of fake animal and whether or not it gave off heat. She thinks the scientists should have varied the size of fake animals and the amount of heat they gave off. Brian asked Jordan how changing the size of the fake animals or the amount of heat they gave off helps to answer the question of how mosquitos use sense information to find food.

You selected (a-b-c-d-e-f, **show student their response**), can you tell me why this is the best choice?

Once student finishes, ask them if there is anything else?

Question 5: Olivia questions the scientists' conclusion that mosquitos rely on a combination of smell, visual, and heat information to find food. Since the scientists didn't experiment with different types of mosquitos, she thinks their evidence is limited to only the mosquitos used in the study. Jackson thinks the evidence from the study is NOT limited.

You answered (**show student their response**)

Can you tell me more about why you agreed with [Olivia or Jackson]?

Once student finishes, ask them if there is anything else?

Question 6: Jackson questions the evidence from the four tests. He thinks the scientists should have reported how accurate the computer was at recording the mosquitos. Without this information, Jackson has doubts about the quality of the evidence. Olivia thinks the accuracy of the computer doesn't have anything to do with the quality of the evidence.

You answered (**show student their response**)

Can you tell me more about why you agreed with [Olivia or Jackson]?

Once student finishes, ask them if there is anything else?

Question 7: The teacher asked the class to imagine that the scientists decided to watch the mosquitos in nature instead of using a lab with a computer to record them. The teacher then asked the class to think carefully about whether changing the experimental tests in this way would influence the results of the study.

Can you tell me more about why your answer is the best?

Once student finishes, ask them if there is anything else?

Question 8: The teacher asked the class to imagine that the scientists thought there were other sense cues in addition to smell, heat, and visual information that mosquitos relied on to find

food. The teacher then asked the class to think carefully about whether this information would change the tests the scientists decided to do.

Can you tell me more about your thoughts on this question?

Once student finishes, ask them if there is anything else?

APPENDIX F. ERA TEACHER INTERVIEWS

Teacher Interview

Name:

School:

1. Educational background:
2. Experience (Years teaching/years teaching science):
3. How much time is spent each week on science instruction:
4. What do you like about teaching science:
5. What are some of the important topics do you cover in science:
6. How would you describe your science teaching:
7. Can you give me an example:
8. What do you want your students to learn about science:
9. How often do students conduct investigations (per week/month):
10. How much class time does a typical investigation require:
11. Can you describe an investigation:
12. How often do your students work with evidence in activities or investigations:
13. Can you provide an example of an investigation you think does a really nice job of presenting students with scientific evidence (obtain detailed task descriptions/lessons):
14. What do you want your students to learn about scientific evidence during science:

APPENDIX G. ERA OBSERVATION

Teacher: _____ Grade: _____ Start time: _____ End time: _____

Planning, Design, and Collection		
Question Generation	Based on what is known and are shaped by potential/anticipated evidence and in turn delineate what will count as evidence	☐
Variable Selection and Operationalization	Relevant variables are identified/selected and justified Are variables: Continuous/categorical What is the sampling interval /range/ frequency	☐
Quality of design & data collection procedures	Is the design appropriate for the purposes of the study? Does it target the variables in an unconfounded way? Are the methods of data collection appropriate and trusted? Technical precision and sensitivity of measurement tools/devices: Do they have acceptable accuracy and sensitivity for measuring the variables of interest and are they used properly Sampling: Are the data collected in an unbiased way, representative of the population, and of sufficient range Are there diverse kinds/sources of relevant data collected? Are there appropriate models for aggregating and analyzing primary data that guide collection? Accounting for potential sources of error in data collection	☐
Analysis, Interpretation, & Explanation		
Analyses of Data	Do examinations of data meet accepted standards Descriptive statistics vs more complex analyses Examinations of error How are anomalies (e.g., outliers) resolved Graphical representations to organize data	☐
Interpretations / Conclusions	Are claims supported by evidence? Are the results consistent with past research? Alternative explanations explicitly addressed? Free from bias/conflicts of interest? Were limits discussed?	☐
Social Factors		
	Scientific evidence and its communication relies on: Expertise/training (researcher) Reporting of results to community Peer-review of work (proposal, publication) Expert feedback and evaluation Journal quality	☐

Notes: