

© 2017

Ronald W. Rinehart

ALL RIGHTS RESERVED

PROMOTING STUDENTS' EPISTEMIC COGNITION AND CONCEPTUAL
LEARNING THROUGH THE DESIGN OF SCIENCE LEARNING ENVIRONMENTS

By

RONALD WAYNE RINEHART

A dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Education

Written under the direction of

Clark A. Chinn & Ravit Golan Duncan

And approved by

New Brunswick, New Jersey

OCTOBER, 2017

ABSTRACT OF THE DISSERTATION

Promoting students' epistemic cognition and conceptual learning through the design of science learning environments

By RONALD WAYNE RINEHART

Dissertation Directors:

Clark A. Chinn & Ravit Golan Duncan

When scientists develop knowledge about the world, they engage in a variety of complex epistemic processes (Allchin, 2011; Hardwig, 1985). They evaluate scientific models and evidence (Giere, 2004) and evaluate not only their own claims but the claims of others (Chinn, Rinehart, & Buckland, 2014) through the use of argumentation (Thagard, 2000). School science often omits the authentic epistemic practices of scientists, producing a false characterization of their work (Allchin, 2004; Chinn & Malhotra, 2002; Duschl, 1988). Science classrooms tend to be epistemically sterile environments (Goldberg, 2013) focused on unproblematic accounts of science (Allchin, 2004; Duschl, 1990). Recent calls for reform argue that there is a need for learning environment designs where students grapple with opposing perspectives and uncertainty like that found in the world outside of school (Britt, Richter, & Router, 2014). This research addresses these concerns in three parts.

Chapter 2 presents a design case discussing four key design principles for engaging students with models and evidence in environments that embrace uncertainty and multiple, sometimes conflicting, perspectives. These decisions involve: identifying phenomena for students to investigate, designing for student engagement with modeling, developing evidence for use during modeling, and fostering productive disciplinary engagement (Engle & Conant, 2002).

Chapter 3 examines how students use, evaluate, and re-evaluate evidence over time and how their ideas about one piece of evidence impact their ideas about other evidence. I present the results of a three-day model-based inquiry lesson with seventh-grade students who investigated the possibility that some humans might be genetically resistant to HIV. Existing frameworks for evaluating student reasoning do not include evidence re-evaluation or the combination of pieces of evidence to construct a new body of evidence. I argue that normative accounts of good reasoning in science classes could be improved by taking both of these practices into account.

Chapter 4 presents the results of a three-day modeling activity in which 7th grade life science students developed models of inheritance in response to multiple evidence sets. Students developed models that: were consistent with evidence, were internally consistent, increased in their use of causal mechanisms, and increased in their consistency with normative explanations of inheritance. Students' abilities to correctly make predictions about novel inheritance problems significantly increased over time.

ACKNOWLEDGEMENTS

First I would like to thank my wife Rebecca Weaver Rinehart for her tireless efforts. I would like to thank my daughters, Sofia Avery Rinehart and Alyssa Jane Rinehart, for the love and happiness they have brought to my life as I completed my graduate studies. I would like to thank my parents Janette S. Parmley and Ronald A. Rinehart. I would like to thank Dr. Clark Chinn for his mentorship, guidance, and patience. I would like to thank Dr. Ravit Golan Duncan for her drive and passion. I would like to thank Dr. Drew Gitomer and Dr. Rick Duschl for their insights. I would like to thank Trudy Atkins for her enthusiastic support of my work. I would like to thank Lars Sorensen for being himself no matter what happens, and my colleague Luke Buckland for his guidance through my early readings of the philosophical work on epistemology that undergirds much of my present research. I would like to thank the many graduate students in both the PRACCIS and Epistemic Cognition research groups for their friendship. I would like to thank Dr. Manu Kapur, Dr. Frank Fischer, Dr. Gale Sinatra, Dr. Bill Sandoval and Dr. Rainer Bromme for their support and guidance at the variety of graduate student mentoring events through which I came to know them. I would like to thank Janet and Jon Weaver for all of their support and assistance. Finally, for their enthusiasm for cheesy, weird and obscure board games I would like to thank the crew of the Flip The Table podcast: Moderator Chris Michaud, “Flip” Florey, Chris Barter, and Jared Hunnefeld. Flip The Table was the best podcast of my Ph.D. era, and made my many long commutes bearable, and so I thank you gentleman. Finally, Chapter 2 was written by me, Ronald W. Rinehart, and published as Rinehart, R. W., Duncan, R. G., Chinn, C. A., Atkins, T. A., & DiBenedetti, J. (2016). Critical design decisions for successful model-based inquiry in science classrooms. *International Journal of Designs for Learning*, 7(2), 17-40.

TABLE OF CONTENTS

ABSTRACT OF THE DISSERTATION	ii
ACKNOWLEDGEMENTS	iv
LIST OF TABLES	x
LIST OF FIGURES	xii
Chapter 1: Introduction	1
1.1 Statement of the Problem	2
1.2 Epistemic Practices of Science	3
1.3 Students Inside and Outside the STEM Pipeline	4
1.4 The Epistemically Sterile Classroom	7
1.5 Considering Epistemic Messiness	8
1.6 Overview of the Three Papers	10
1.7 References	14
Chapter 2: Critical Design Decisions for Successful Model-Based Inquiry in Science Classrooms	16
Abstract	17

2.1 Introduction	18
2.1.1 Brief Unit and Lesson Design Description	22
2.2 Design Challenge 1: Choosing Phenomena	23
2.2.1 Design Challenge 1 Example: The HIV Lessons	31
2.3 Design Challenge 2: Developing Models	33
2.4 Design Challenge 3: Developing Evidence	42
2.4.1 Guideline 1: Designers should take into account the variety of evidence features that can be varied along two continua: (a) complexity and (b) quality.....	45
2.5 Design Challenge 4: Productive Disciplinary Engagement	59
2.6 Conclusion	73
2.7 Acknowledgements	73
2.8 References	74
Chapter 3: The Body of Evidence: The Role of Epistemic Cognition and Evidence	
Evaluation in Science Classes	77
Abstract	78
3.1 Introduction	79

3.1.1 Evaluating and Re-evaluating Evidence	82
3.1.2 Bodies of Evidence	83
3.1.3 Criteria for Evidence	85
3.1.4 Insiders, Outsiders and the STEM Pipeline	86
3.1.5 The Role of Multiple Documents Coordination in Model-Based Inquiry Environments	89
3.1.6 The Present Study	91
3.2 Method	102
3.2.1 Coding	103
3.3 Results	109
3.4 Discussion	128
3.4.1 Epistemic Cognition and Evidence Evaluation	128
3.4.2 Evidence Re-evaluation	130
3.4.3 Responding to New Evidence and Changing Beliefs	132
3.4.4 Developing a Body of Evidence	133
3.4.5 The Body of Evidence and Argument Complexity	135

3.5 Conclusion and Implications	137
3.6 Acknowledgments	139
3.7 References	140
Chapter 4: Using Evidence to Develop and Refine Models of Inheritance	145
Abstract	146
4.1 Introduction	147
4.1.1 Designing for Learning from Anomalous Data	148
4.1.2 Designing for Productive Success	150
4.1.3 Evidence to Model Connections	152
4.1.4 Developing Internally Consistent Models	153
4.1.5 Linking the “Seen” and the “Unseen”	154
4.1.6 The Present Study	157
4.2 Methods	158
4.2.1 Timeline and Noteworthy Activities	159
4.2.2 Data Collection and Analysis	169
4.2.3 Coding	169

4.3 Results	174
4.3.1 Changes in Students' Rules: Consistency with Evidence	175
4.3.2 Changes in Students' Rules: Internal Coherence	177
4.3.3 Changes in Students' Rules: Genotype to Phenotype Connections	179
4.3.4 Changes in Students' Rules: Genes Occurring in Pairs	182
4.3.5 Prediction Correctness	184
4.4 Discussion and Implications	186
4.5 Acknowledgments	191
4.6 References	192
4.7 Appendix A	194
Chapter 5: Conclusion	196
5.1 Introduction	197
5.2 Findings and Implications	198
5.3 Future Research	203
5.4 References	205

LIST OF TABLES

Table 2.1 Guidelines for Choosing Phenomena for Scientific Modeling Activities.	25
Table 2.2 Guidelines for Developing Models for Scientific Modeling Activities.....	35
Table 2.3 Guidelines for Developing Evidence for Scientific Modeling Activities.....	44
Table 2.4 Guidelines for Generating Productive Disciplinary Engagement with Scientific Modeling Activities	61
Table 3.1 A brief summary of the three day HIV lesson	94
Table 3.2 Students' evaluations of the quality of evidence taken from their Day 3 final essays.....	115
Table 3.3 An analysis of one student's evaluation and re-evaluation of evidence	122
Table 3.4 Students' development of bodies of evidence in their final essay	127
Table 4.1 A brief summary of the first five days of the genetics lesson.....	165
Table 4.2 Is the student's rule consistent with the available evidence?.....	170
Table 4.3 Does the student's rule contradict other rules within the rule set?	171
Table 4.4 Does the rule make a connection between the proposed genotype (genes) of individuals and their traits?	172

Table 4.5 Does the student's rule reflect the notion of 2-allele combinations (e.g. a pair of genes) for each trait?	173
Table 4.6 Does the student make a correct prediction?	174
Table 4.7 Are the rules consistent with the evidence?	176
Table 4.8 Are students' models consistent with the evidence?	177
Table 4.9 Does the rule contradict other rules?	178
Table 4.10 Does the student's model have internal contradictions?	179
Table 4.11 Does the rule connect the genotype to the phenotype?	181
Table 4.12 Does the student's model connect genotype to phenotype?	182
Table 4.13 How many genes are described by the rule?	183
Table 4.14 Does the student's model make use of the concept of alleles?	184
Table 4.15 Students' points earned on pedigree predictions.	185

LIST OF FIGURES

Figure 2.1. PRACCIS Elements.....	22
Figure 2.2. The “Attack-and-destroy” model	38
Figure 2.3. The “Keep-it-out” model	39
Figure 2.4. A heuristic for the combinations of evidence quality and evidence complexity	46
Figure 2.5. Evidence 1.....	49
Figure 2.6. Evidence 2	51
Figure 2.7. Evidence 3	53
Figure 2.8. Evidence 4	55
Figure 2.9. Discussion stems used in an earlier iteration of the project	63
Figure 2.10. The MEL Matrix for HIV Lesson 1	66
Figure 2.11. A typical example of student work using the MEL 1.0.....	68
Figure 2.12. The three basic combinations of relevance and diagnosticity	72
Figure 3.1. A MEL Matrix	96
Figure 3.2. The Berland & McNeill (2010) argument structure.....	134

Figure 3.3. The upper level anchor for the Osborne et al. (2016) learning progression for argumentation	134
Figure 3.4. A proposed alternative model of argumentation	135
Figure 4.1. Evidence Set 1.....	161
Figure 4.2. A prediction question from the first day of Lesson 3	162
Figure 4.3. Evidence Set 2.....	163
Figure 4.4. Evidence Set 3.....	164
Figure 4.5. A single student’s initial and intermediate rule sets	180

Chapter 1: Introduction

1.1 Statement of the Problem

This dissertation examines the impacts of several learning environment designs that aim to promote growth in students' use of sophisticated scientific practices (i.e., modeling, argumentation, evidence evaluation) and conceptual knowledge through an approach that embraces the range of situations in which people use science both as scientists as well as outsiders (laypersons) making use of science for matters of personal or societal import. Typical K-12 schools, particularly secondary schools, teach science in a manner that is consistent with the aim of delivering "science-ready students to colleges and universities" (Feinstein, Allen, & Jenkins, 2013, p. 314); in other words all students, regardless of their interests or career objectives, are educated in the "STEM pipeline." The fundamental problem with the STEM pipeline approach to science education is that it may not fully meet the everyday knowledge needs of the majority of students in science (Britt, Richter, & Rouet, 2014; Duschl, 1988; Feinstein et al., 2013).

Collecting data to analyze, revising models of scientific phenomena, and communicating findings to a research community are all worthwhile pipeline-oriented activities that are consistent with the vision of the Next Generation Science Standards (NGSS Lead States, 2013) and the general zeitgeist of the model-based inquiry movement in science (Windschitl, Thompson, & Braaten, 2008). However, laypeople rarely, if ever, collect and analyze their own data and publicly revise or add to theoretical commitments in a research community. Instead laypeople are often faced with whom to trust and what to believe when making decisions about personal or societal issues related to science (Bromme, Kienhues, & Porsch, 2010). The skills needed for seeking multiple sources of evidence, evaluating them, and integrating them into a coherent mental model are central to the reasoning practices of laypeople confronted with making decisions

about matters of science that intersect with their lives (Britt et al., 2014; Bromme, Thomm, & Wolf, 2015; Stadtler & Bromme, 2013). Similarly, scientists also need to construct mental models that integrate evidence and theory (Nersessian, 1992).

In this dissertation I examine instructional interventions designed to promote elements of both pipeline-oriented science and sophisticated reasoning about science for the layperson. First I begin with a brief account of the interaction between epistemic processes of science and model-based inquiry. Then I examine the rationale for an expanded vision of science education, one that includes more attention to the reasoning of the layperson. Then I briefly and non-exhaustively examine some of the epistemological underpinnings of school science and propose an alternative perspective that embraces the epistemically messy real world of science outside of school. Finally I end with an examination of the research questions addressed in the three papers in this dissertation.

1.2 Epistemic Practices of Science

When scientists develop knowledge about the world, they engage in complex epistemic processes (Allchin, 1999, 2004, 2011; Brush, 1974; Goldberg, 2013; Hardwig, 1985, 1991; Shapin, 1994). For example, scientists evaluate scientific models and evidence, often evidence provided by other scientists via the process of peer reviewed publishing, and they evaluate not only their own claims but the claims of others (Chinn, Buckland, & Samarapungavan, 2011; Chinn & Rinehart, 2016; Chinn, Rinehart, & Buckland, 2014; Giere, 2004). Scientists also use argumentation as a reliable epistemic process to generate knowledge about the world (Thagard, 2000). School science, on the other hand, often omits the authentic epistemic practices of scientists, producing a false characterization of their work (Allchin, 2004; Chinn & Malhotra, 2002; Duschl, 1988),

while focusing on the acquisition of the vocabulary of science, the memorization of diagrams and formulae, and traditional “cookbook” style labs. The omission of authentic epistemic practices from science classrooms is troubling. For example, scientific modeling, a core epistemic practice of scientists (Giere, 2004), is rarely encountered in the typical school science curriculum (Windschitl et al., 2008).

Scientific models, which are also referred to as explanatory or conceptual models, are used to explain a range of phenomena (Giere, 2004). Explanatory models are not attempts to reproduce a phenomenon, or represent it at a different scale as is the case with physical models; they are abstractions that are used to describe and explain certain aspects of the world under certain conditions (Giere, 2004). Models may be well known and widely applicable, such as the double helix model of DNA, or highly localized, such as a hydrographic model of a particular stream network. Valid models can be used to generate predictions, hypotheses, and generalizations about particular phenomena. Scientific models are developed, evaluated, and refined in light of empirical evidence; thus, evaluating the quality of evidence, and coordinating the relationships (e.g., support, contradict, irrelevant) of evidence to models through scientific argumentation is integral to the practice of science. Evidence can also stand in relation to other pieces of evidence, a relationship that is largely underexplored in the work on modeling and argumentation.

1.3 Students Inside and Outside the STEM Pipeline

Modeling, evidence evaluation, evidence integration, and argumentation are all practices that scientists use in the construction of scientific knowledge. Using these authentic practices to learn the content of science would represent a productive shift in school science, a shift embraced by recent reforms like the Next Generation Science

Standards (NGSS Lead States, 2013). However, not all students, especially those in K-12 or non-science majors in college, are aiming toward a career as an expert in a scientific field; it is therefore imperative that science education meet the needs of these learners, too (Duschl, 1988; Feinstein et al., 2013).

Students and adults who are in the position of making sense of science from the perspective of non-scientists (laypeople) do not engage in first-hand evidence collection or in the construction and revision of scientific models. Instead the layperson is often confronted with making sense of secondhand evidence. They are faced with decisions about who and what to believe. This questions a core epistemological assumption that “one’s own knowledge is better than knowledge attained from others” (Bromme et al., 2010, p. 164). The typical science classroom, and even reform-based classrooms, places a heavy emphasis on experimentation to collect and analyze data, an approach that reflects this kind of commitment. Instead the reasoning of the layperson is often characterized by epistemic dependence on others for their knowledge (Bromme et al., 2010); particularly on the knowledge of experts when it comes to complex scientific topics.

Expertise is a scarce commodity, and everyone is a layperson in all domains in which they are not an expert (Bromme et al., 2010; Feinstein et al., 2013). For example, an expert biologist is not expert in physics, computer science, health care policy, economics or any other discipline in which the bar for expertise requires years of dedicated study and practice. Research on the *division of cognitive labor* is based on the premise that specialized knowledge and labor are unevenly distributed through society and take the form of disciplines that reflect such specialization (Bromme et al., 2010). Given that expertise is scarce and the division of cognitive labor permeates society,

everyone is in the position of being a layperson with respect to most domains of knowledge. Scholarship on the division of cognitive labor suggests that there are specific skills one needs when dealing with information as a layperson. For example, evaluating source characteristics of information (e.g., bias, competence, authorship, date and venue of publication) is important for making decisions about whom to trust, including decisions that involve science and medicine (Bromme et al., 2010; Scharrer, Stadtler, & Bromme, 2014). Oftentimes it is the case that scientists and laypeople alike need to make sense of phenomena based on multiple sources of information; a task that has proven challenging for students in a variety of studies (Braasch, Bråten, Strømsø, Anmarkrud, & Ferguson, 2013; Britt & Anglinskis, 2002; Wiley, Goldman, Graesser, Sanchez, Ash, & Hemmerich, 2009).

Beyond gaining facility with the knowledge-making practices of scientists (e.g., evidence evaluation, argumentation, and modeling), learning in science classrooms should also include a host of other reliable processes for producing knowledge and making decisions, because these are the skills that laypeople need to make sense of a science information rich world. These processes include (a) evaluating the claims of experts, especially when experts disagree (Allchin, 2011; Bromme et al., 2015; Collins, 2014); (b) gaining knowledge about how scientific findings are communicated both within the community of scientists as well as with the general public (Allchin, 2011; Jiménez-Aleixandre & Federico-Agraso, 2009; Feinstein, 2011; Yarden, 2009); and finally, (c) learning to consider under what conditions scientific consensus is trustworthy or untrustworthy (Allchin, 2011; Bromme et al., 2015; Collins, 2014); and (d) evaluating and integrating evidence from a variety of sources (Britt et al., 2014). Eventually

students leave formal schooling and become the next generation of lay adults; it is therefore important that their science classes equip them with the skills that lay adults need to navigate a world defined by the division of cognitive labor. In short, students should develop facility with the reliable practices of laypeople needing to make sense of science in a science rich world (Feinstein, 2011; Feinstein et al., 2013; Phillips & Norris, 2009).

1.4 The Epistemically Sterile Classroom

Scientists use complex epistemic practices to generate knowledge about the world. Laypeople have a need to navigate a world full of rich science content (Feinstein, 2011; Feinstein et al., 2013). Designing learning environments that teach the content of science by having students engage in the knowledge production practices of an expert (i.e., a scientist) while simultaneously learning the evaluation practices needed by the layperson is a challenging task. Compounding this already difficult task is the epistemically sterile nature of the typical science classroom (Goldberg, 2013), to which I turn next.

In the epistemically sterile classroom, students are rarely confronted with real dilemmas about who and what to believe and why it should be believed (Goldberg, 2013). The typical science classroom reflects “final form” science (Duschl, 1990). It presents science as a compendium of facts about which there is a high degree of certainty, so much so that laypeople conflate theory and fact and often hold scientific facts in higher regard than theory (Duschl, 1990). Moreover, this presentation of science as facts fails to portray active professional science as engaging with uncertainty rather than certainty. Even the activities that have the most uncertainty, so-called labs and

experiments, have predictable and certain outcomes that can often be discovered with a quick internet search. If a high school student wants to know what happens when they mix two chemicals, they can in all likelihood quickly search for it and find a video showing the outcome and explanation for why it happened. These kinds of activities have no uncertainty, no epistemic messiness and are highly sterile.

However, the world outside the classroom is far from epistemically sterile; it is characterized by uncertainty. Epistemically sterile classrooms are unlikely to prepare students for a world full of complex and conflicting claims and evidence (Britt et al., 2014; Goldberg, 2013). Similar to Chinn and Malhotra (2002) I argue that more effective inquiry oriented instruction will promote student reasoning about authentic evidence, which is very different than inauthentic school science tasks (Rinehart, Duncan, & Chinn, 2014).

1.5 Considering Epistemic Messiness

Science teachers are faced with a dilemma. What is the right balance in the classroom between certain and sterile knowledge and uncertain and messy knowledge? On the one hand, there are concerns that students may develop incorrect conceptions if they are confronted with conflicting claims and a mixture of high and low quality evidence. This is not a speculative concern; during professional development sessions for this study, as well as during the implementation phase of the project, teachers expressed on numerous occasions concerns about “students getting the wrong idea” or “what if they don’t get the right idea?” On the other hand, if students learn the practices of science in an epistemically sterile environment, then how can they gain the ability to critically evaluate a wider range of evidence? Teachers in these studies were also aware that there

is a gap between what students experience in the classroom and what they encounter in the world outside the classroom. Bridging the gap between the science classroom and evidence and claims in the real world requires pedagogy focused on evaluative epistemic practices, with particular attention to skills and epistemic dispositions that enable the layperson to competently evaluate secondhand evidence and stitch together multiple pieces of evidence to reach an informed decision (Britt et al., 2014).

To counteract this pervasive epistemic sterility, I recommend designing lessons that embrace uncertainty by expanding the range of evidence considered in the classroom to encompass the full range of evidence students find outside the classroom. This is commensurate with multiple calls from the research community to include readings beyond textbook style expository writing (Phillips & Norris, 2009) and to incorporate more evidence that reflects “conflict, opposing perspectives and uncertainty” (Britt et al., 2014, p. 119).

In the *epistemically messy science* classroom, students encounter evidence of variable quality (i.e., from bad to good evidence), as well as a range of models from incomplete and inaccurate to complete and accurate, and they make determinations about who and what to believe, even in the presence of uncertainty. Students engage with the epistemic practices of scientists while grappling with meaningful sense making of phenomena in the world through the lens of science. Here I refer to both *epistemic* and *science* in an attempt to place science squarely in the business of producing knowledge. This is not, and has not historically been, the *only* role that science has played (Shapin, 1994), however for the purposes of this dissertation I will work with science as an epistemic system, one among many for producing knowledge (Goldman, 1999).

The Promoting Reasoning And Conceptual Change In Science (PRACCIS) project is a research and development endeavor funded by the National Science Foundation. PRACCIS lessons often leverage some epistemic messiness while at the same time promoting epistemically sophisticated practices. The aim of PRACCIS is to promote conceptual learning of science content through authentic engagement in the sophisticated epistemic practices of science. This approach is based on the general principles of model-based inquiry (Windschitl et al., 2008). The data for this dissertation are drawn from a subset of lessons I designed as part of the larger PRACCIS project.

1.6 Overview of the Three Papers

Broadly speaking, this dissertation addresses three areas: (a) instructional design considerations for science classes, particularly designs that can foster sophisticated epistemic practices for students in the science pipeline as well as students who are becoming competent outsiders; (b) how students evaluate, re-evaluate and engage in written evidence based argumentation in an epistemically messy environment; and (c) how students use evidence over time to revise models. There are three research projects which will be presented in detail in Chapters 2, 3, and 4 of this dissertation.

Chapter 2 describes a lesson design framework for generating model-based science curricula for secondary science students. This chapter was published in July 2016 in an *International Journal of Designs for Learning* special issue about K-12 classroom implementation design cases (Rinehart, Duncan, Chinn, Atkins, & DiBenedetti, 2016). The aim of this method of instruction is to increase and refine students' use of sophisticated epistemic practices like evidence evaluation, argumentation and evidence to

model coordination. This chapter outlines four major design challenges, and principles to meet those challenges, for model-based inquiry including:

1. Choosing a phenomenon as a context for inquiry
2. Developing models for students to use
3. Developing evidence for students to evaluate
4. Developing scaffolds for productive disciplinary engagement

Chapter 3 examines the epistemic practices of students engaged with learning about the possibility that humans can be genetically resistant to HIV with a particular focus on how students reason about evidence across a range of types of evidence. In this lesson, lasting about three days, middle school life science students were presented with four pieces of evidence of variable quality. Students selected the claim they thought was best, either genetic resistance to HIV exists or not, and supported their selection with an evidence-based argument. Students made this selection three times: (a) after being introduced to the problem but before seeing any evidence; (b) after seeing the first two pieces of evidence; and (c) after they had seen all four pieces of evidence. Students were actively engaged in evidence evaluation, evidence-to-model coordination, and written and verbal argumentation throughout the lesson. At the end of the lesson students wrote an extended argument about the models and evidence. My analysis concentrated on students' evidence evaluation and re-evaluation, conceptual links between pieces of evidence, evidence-to-model coordination, and written final arguments. Research questions for this study included:

1. What are students' implicit criteria for evidence evaluation?

2. Do students adjust their evaluation of the quality of the evidence with exposure to new evidence?
3. Do students shift their model selection with exposure to new evidence?
4. Do students recognize the opportunity to construct an integrated body of evidence, and if so, what criteria do they use in its construction?
5. Does constructing an integrated body of evidence lead to increases in argument complexity?

Chapter 4 examines middle school life science students' use of evidence and their subsequent refinement of models of inheritance over the span of several days. In this study I investigated how seventh-grade students developed new understandings of the biological mechanisms that govern gene-trait inheritance patterns by modeling the "rules" of inheritance, using three sets of scientific evidence in the form of family trees (pedigrees). This study examines how middle-school students engaged in model development, model revision and evidence-model coordination while developing their knowledge of genetics and inheritance. My research questions included:

1. Are students' rules consistent with the available evidence and does consistency change as the complexity of the evidence increases over time?
2. Are students' models internally coherent (i.e., do they contain rules that contradict one another?) and does the degree of coherence change as the complexity of the evidence increases over time?
3. To what extent do student models change over time with respect to their ability to develop scientifically accurate causal accounts of inheritance?

4. Can students make useful predictions about novel inheritance problems based on their rules?

There are two themes that cut across these studies. First, the role of evidence in science classrooms is explored in some detail. Chapter 2 examines design considerations for developing lessons that make use of evidence for modeling and argumentation. Chapter 3 empirically explores a lesson designed with these principles. Chapter 4 explores how students used a patchwork of evidence to revise their own model of how genetic inheritance operates. Epistemic messiness, the second theme, is intertwined with the considerations of evidence. Chapter 2 provides guidelines for developing evidence of variable quality, relevance and diagnosticity. Chapter 3 explores how students evaluate evidence, and re-evaluate evidence in light of new evidence, with an emphasis on secondhand evidence. Chapter 4 makes use of epistemic messiness by allowing students to grapple with their own flawed models of inheritance and revise them over time by considering new evidence.

1.7 References

- Allchin, D. (1999). Do we see through a social microscope?: Credibility as a vicarious selector. *Philosophy of Science (Proceedings)*, 60, S287-S298.
- Allchin, D. (2004). Should the sociology of science be rated x?. *Science Education*, 88(6), 934-946.
- Allchin, D. (2011). Evaluating knowledge of the nature of (whole) science. *Science Education*, 95(3), 518-542. doi: 10.1002/sce.20432
- Braasch, J. L., Bråten, I., Strømsø, H. I., Anmarkrud, Ø., & Ferguson, L. E. (2013). Promoting secondary school students' evaluation of source features of multiple documents. *Contemporary Educational Psychology*, 38(3), 180-195.
- Britt, M. A., & Aglinskias, C. (2002). Improving students' ability to identify and use source information. *Cognition and Instruction*, 20(4), 485-522.
- Britt, M. A., Richter, T., & Rouet, J. F. (2014). Scientific literacy: The role of goal-directed reading and evaluation in understanding scientific information. *Educational Psychologist*, 49(2), 104-122.
- Bromme, R., Kienhues, D., & Porsch, T. (2010). Who knows what and who can we believe? Epistemological beliefs are beliefs about knowledge (mostly) to be attained from others. In L. D. Bendixen & F. C. Feucht (Eds.), *Personal epistemology in the classroom: Theory, research, and implications for practice* (pp. 163-193). Cambridge: Cambridge University Press.
- Bromme, R., Thomm, E., & Wolf, V. (2015). From understanding to deference: Laypeoples' and medical students' views on conflicts within medicine. *International Journal of Science Education, Part B*, 5(1), 68-91.
- Brush, S. G. (1974). Should the history of science be rated x? The way scientists behave (according to historians) might not be a good model for students. *Science*, 183(4130), 1164-1172.
- Chinn, C. A., Buckland, L. A., & Samarapungavan, A. L. A. (2011). Expanding the dimensions of epistemic cognition: Arguments from philosophy and psychology. *Educational Psychologist*, 46(3), 141-167. doi: 10.1080/00461520.2011.587722
- Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education*, 86(2), 175-218.
- Chinn, C. A., & Rinehart, R. W. (2016). Epistemic cognition and philosophy: Developing a new framework for epistemic cognition. In W. A. Sandoval, J.A. Greene, & I. Braten (Eds.), *Handbook of epistemic cognition* (pp. 460-478). New York: Routledge.
- Chinn, C. A., Rinehart, R. W., & Buckland, L. A. (2014). Epistemic cognition and evaluating information: Applying the AIR model of epistemic cognition. In D. Rapp & J. Braasch (Eds.), *Processing inaccurate information* (pp. 425-453). Cambridge, MA: MIT Press.
- Collins, H. (2014). *Are we all scientific experts now?*. John Wiley & Sons.
- Duschl, R. A. (1988). Abandoning the scientific legacy of science education. *Science Education*, 72(1), 51-62.
- Duschl, R. A. (1990). *Restructuring science education: The importance of theories and their development*. Teachers College Press.

- Feinstein, N. (2011). Salvaging science literacy. *Science Education*, 95(1), 168-185.
- Feinstein, N. W., Allen, S., & Jenkins, E. (2013). Outside the pipeline: Reimagining science education for nonscientists. *Science*, 340(6130), 314-317.
- Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science*, 71(5), 742-752.
- Goldberg, S. C. (2013). Epistemic dependence in testimonial belief, in the classroom and beyond. *Journal of Philosophy of Education*, 47(2), 168-186.
- Goldman, A. I. (1999). *Knowledge in a social world* (Vol. 281). Oxford: Clarendon Press.
- Hardwig, J. (1985). Epistemic dependence. *The Journal of Philosophy*, 82(7), 335-349.
- Hardwig, J. (1991). The role of trust in knowledge. *The Journal of Philosophy*, 88(12), 693-708.
- Jiménez-Aleixandre, M. P., & Federico-Agraso, M. (2009). Justification and persuasion about cloning: Arguments in Hwang's paper and journalistic reported versions. *Research in Science Education*, 39(3), 331-347.
- Nersessian, N. J. (1992). How do scientists think? Capturing the dynamics of conceptual change in science. *Cognitive Models of Science*, 15, 3-44.
- NGSS Lead States. 2013. *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.
- Phillips, L. M., & Norris, S. P. (2009). Bridging the gap between the language of science and the language of school science through the use of adapted primary literature. *Research in Science Education*, 39(3), 313-319.
- Rinehart, R. W., Duncan, R. G., & Chinn, C. A. (2014). A scaffolding suite to support evidence-based modeling and argumentation. *Science Scope*, 38(4): 70-77.
- Rinehart, R. W., Duncan, R. G., Chinn, C. A., Atkins, T. A., & DiBenedetti, J. (2016). Critical design decisions for successful model-based inquiry in science classrooms. *International Journal of Designs for Learning*, 7(2).
- Scharrer, L., Stadtler, M., & Bromme, R. (2014). You'd better ask an expert: Mitigating the comprehensibility effect on laypeople's decisions about science-based knowledge claims. *Applied Cognitive Psychology*, 28(4), 465-471.
- Shapin, S. (1994). *A social history of truth: Civility and science in seventeenth-century England*. University of Chicago Press.
- Stadtler, M., & Bromme, R. (2013). Multiple document comprehension: An approach to public understanding of science. *Cognition and Instruction*, 31(2), 122-129.
- Thagard, P. (2000). *How scientists explain disease*. Princeton, NJ: Princeton University Press.
- Wiley, J., Goldman, S. R., Graesser, A. C., Sanchez, C. A., Ash, I. K., & Hemmerich, J. A. (2009). Source evaluation, comprehension, and learning in Internet science inquiry tasks. *American Educational Research Journal*, 46(4), 1060-1106.
- 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.
- Yarden, A. (2009). Reading scientific texts: Adapting primary literature for promoting scientific literacy. *Research in Science Education*, 39(3), 307-311.

Published as:

Rinehart, R. W., Duncan, R. G., Chinn, C. A., Atkins, T. A., & DiBenedetti, J. (2016). Critical design decisions for successful model-based inquiry in science classrooms. *International Journal of Designs for Learning*, 7(2), 17-40.

Chapter 2: Critical Design Decisions for Successful Model-Based Inquiry in Science Classrooms

Abstract

Current science education reforms and the new standards (Next Generation Science Standards [NGSS], 2013) advocate that K-12 students gain proficiency in the knowledge-generating practices of scientists. These practices include argumentation, modeling, and coordinating evidence with theories and models. Practice-based instruction is very different from traditional methods. Creating practice-rich instructional materials presents substantive challenges even for experienced educational designers because of the unlimited choice of potential phenomena to study and the inherent difficulties of developing the associated models and evidence. In this design case we will discuss some of the affordances, constraints and tradeoffs associated with making decisions about four key design principles of engaging students with evidence-based scientific modeling. The first set of decisions involves identifying the focus phenomenon. The second set of decisions regards how to represent the focus phenomenon as an explanatory scientific model and how to design for student engagement with modeling. The third set of decisions involves selecting and developing the evidence students will use to evaluate models. The final set of design decisions pertains to developing supporting activities that foster disciplinary engagement (Engle & Conant, 2002) during modeling. We developed a variety of approaches that address these four design challenges and present them in the context of a unit we developed for a middle school life science course focusing on genetics and inheritance. This design case illustrates how a group of designers, including university researchers, teachers, and school administrators, arrived at collective design decisions bearing on these four problems.

2.1 Introduction

Traditional approaches to science instruction have often embraced science in its "final form" which "consists of solved problems and theories to be transmitted" (Duschl, Schweingruber, & Shouse, 2007, p. 254). This form of science often lacks the social epistemic practices embraced by scientists that are integral to the production of knowledge. What are needed are scaffolds that introduce students to the practices of science (Grandy & Duschl, 2007). Recent reforms in science education (i.e., the Next Generation Science Standards [NGSS]) in the United States have embraced this approach by positioning students to be the constructors of their own knowledge through authentic scientific practices like those described in the NGSS.

Here we describe our approach to scaffolding student involvement in developing life science knowledge using some of the authentic practices of science. These scientific practices include (a) argumentation as a process by which students and scientists alike arrive at reasoned judgments (Fischer et al., 2014); (b) coordinating evidence with theories and models (Windschitl, Thompson, & Braaten, 2008a, 2008b), particularly in cases where there are competing theories and models supported by evidence of variable quality; as well as (c) evaluations of the quality of the evidence and models themselves. This combination of evaluating evidence, coordinating evidence and models, and arriving at evidence-based judgments that are communicated through argumentation, forms the core of our instructional approach and embodies many of the scientific practices embraced by the NGSS. We will refer to this pedagogical approach interchangeably as modeling or model-based inquiry.

Explanatory models are causal and purposeful abstractions developed by scientists to explain a range of phenomena; their use is central to the natural sciences

(Giere, 2004; Kitcher, 1993). Well known examples of explanatory models include the Bohr model of the atom, the Standard Model of particle physics, the double helix model of DNA, and the Copernican heliocentric model of the solar system. Explanatory models are abstractions in that they do not seek to replicate the actual phenomenon but rather are used to describe and explain certain elements of the phenomenon and make predictions about it (Giere, 2004; Kitcher, 1993).

Additionally, scientific models contain purposeful simplifications. Scientists choose to include some details and leave out others. Models used for pedagogical purposes also contain purposeful simplifications. For example, models of photosynthesis, like those used by middle school science students, are often simple representations of carbon dioxide and water being converted into oxygen and sugar in the presence of light. As students progress through biology, additional elements are added to the model like the light-dependent and light-independent reactions. Models, as used traditionally in schools, are given to students in a finished form with little justification, no evidence, and they rarely, if ever, are revised by the students themselves. These models are often poorly understood by students and persistent alternative conceptions represent significant impediments to meaningful understanding (Private Universe Project, 1995). This method of instruction is not epistemologically authentic (Chinn & Malhotra, 2002) and is not compatible with modeling or the NGSS.

Epistemologically authentic practices used by scientists include evaluating the quality of evidence, developing new lines of inquiry, evaluating the utility of conceptual models, and generating evidence based arguments (Chinn & Malhotra, 2002). These practices contrast with approaches to learning that are particular to "school science" but

not authentic to actual science practices, like carrying out well-defined experimental procedures with well-known results (i.e., the so-called "cookbook lab") and memorizing terms and definitions to be repeated on tests (Chinn & Malhotra, 2002). Reading in the epistemologically authentic science classroom would be different as well. At present most textbooks are purely expository, contrasting sharply with primary scientific literature which has an argumentative structure characterized by claims, reasons, evidence, qualifiers, and so on (Phillips & Norris, 2009).

Model-based inquiry is very different from traditional instructional methods. It is clear that extensive design efforts will be needed over the next decade to develop additional instructional materials that are consistent with the NGSS. To a large extent this burden will fall on teachers, most of whom currently do not have the knowledge or capacity to engage in this effort. The primary purpose of this paper is twofold: (a) to illustrate learning environment design challenges associated with science practices-rich designs; and (b) to present a framework for resolving those challenges grounded in examples from a six-month long middle school life science curriculum. It is our hope that other learning environment designers can benefit specifically from three elements of this paper: (a) the framework of design challenges; (b) strategies to solve these challenges; and (c) selected designs which represent our solutions to these challenges. The lesson designs described here represent the collaborative effort of a university-based research team, middle school science teachers, and school administration working as part of a National Science Foundation (NSF) funded research project titled Promoting Reasoning and Conceptual Change in Science (PRACCIS).

The PRACCIS project ran in two large phases during the 2011-2012 and 2012-2013 school years as well as a smaller project in 2013-2014. The project ran for five to six months during each of the two larger implementations. Many of the design challenges and solutions presented in this article represent a blend of insights from the research literature as well as practical wisdom derived from our experiences working together as teachers and researchers. On this point it is worth mentioning that every design decision comes with associated potential for success or failure, and while we cannot address all of the potential pitfalls, or successes, this design case is a distillation of what we feel are some of the most important considerations we have encountered.

The PRACCIS project lesson and unit designs make use of a variety of instructional scaffolds and include elements of evidence based argumentation, reading and writing in the discipline of life science, hands-on science experiences, and technology elements like animations and simulations. Unlike some research in science education that aims at developing a particular piece of software or hardware, there is no single unifying technology product for PRACCIS but rather the thoughtful integration of tools and techniques, described later in this design case, that are already accessible to most science teachers. We feel that this is a strength of our approach.

In this design case we present two lessons that make use of epistemologically authentic methods of instruction centered on model and evidence evaluation that are consistent with the NGSS. We first briefly introduce two lessons that embody the outcome of the design process, with the intent of giving the reader an idea of the aim of this particular design process. Next we develop a framework for the challenges involved in creating learning environments that embrace the scientific practices and disciplinary

core ideas outlined in the NGSS. The framework addresses four major challenges that learning environment designers face: (a) selecting appropriate scientific phenomena, (b) designing models, (c) developing evidence, and (d) developing scaffolds (e.g., disciplinary discussions, epistemic criteria, and student-generated written arguments) that foster disciplinary engagement during modeling. Lessons and units developed for PRACCIS typically include six major elements as shown in Figure 2.1. Each PRACCIS element presents the designer with particular challenges. Each element and associated design challenge is presented in greater detail later.

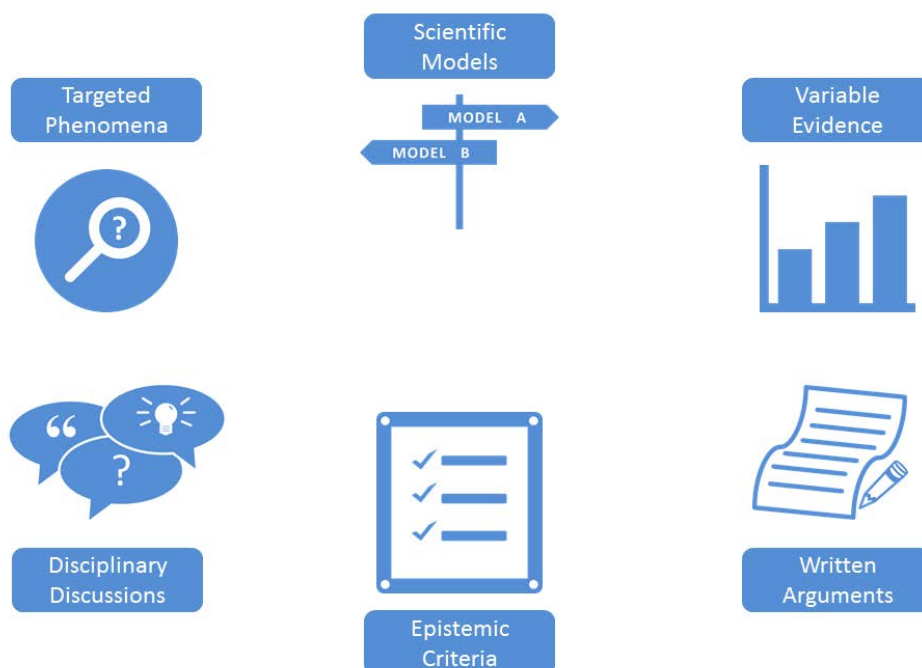


Figure 2.1. PRACCIS Elements.. Most lessons and units developed for the PRACCIS project include these six major elements. Each element presents a suite of challenges that we had to consider during our design process.

2.1.1 Brief Unit and Lesson Design Description

Our research group has developed many middle school level life science lessons on topics like cells, inheritance, genetics, and evolution. Here we will describe two

lessons from our genetics and inheritance unit, which is about three to four weeks in length. The aim of the unit is two-pronged: (a) to engage students in the authentic practices of science, as described in the NGSS, like modeling and evidence based argumentation; and (b) to help students develop competence in understanding the mechanisms of inheritance, specifically the role of alleles, parental contributions to offspring's traits, and the concept that distinct genes code for specific proteins that perform particular functions in the body (i.e., NGSS-DCI: LS3.A; NGSS, 2013).

Throughout this paper we will ground our discussion of design challenges in two lessons in which students engaged in modeling about the possible existence and mechanism of genetically based HIV resistance in humans. Lesson one introduces students to HIV (i.e., we do not assume that students know what the virus is or how it works) and the possibility that genetically based resistance to HIV might exist. This lesson is focused on helping students develop their evidence evaluation and argumentation skills and serves as preparation for lesson two, which engages with the biology content at a deeper level. Lesson two revolves around the cellular and molecular mechanism underlying HIV resistance. Given the space limitations of a single article we will mostly focus on lesson one and we will only discuss the models, and not the evidence, from lesson two. This is because the models from lesson two do a better job of illustrating key design decisions.

2.2 Design Challenge 1: Choosing Phenomena

When students engage in the practice of modeling, they invariably engage with it in the context of a particular phenomenon. In some cases a model may explain a single phenomenon, for example the inheritance pattern of albinism, a relatively common

genetic condition. In other cases a model may be more generalized and explain a class of phenomena, for example, a general model of recessive inheritance patterns can explain the occurrence of albinism, sickle cell anemia, attached earlobes, hitchhikers thumb, and many other traits. Often such generalized models come about after multiple models of individual phenomena are compared to reveal patterns that hold across the distinct examples. In fact, the choice of the initial phenomenon to study can impede or facilitate discovery of the underlying mechanism. Consider the discoveries of Gregor Mendel: his choice of the pea plant and the specific traits he followed allowed him to develop a model of inheritance, where others, choosing more complex traits and organisms, had failed (Berg & Singer, 1998). Therefore choosing a phenomenon to investigate is a critical and influential step in science inquiry. Here we argue that the same is true for science learning.

In this section we provide guidelines, as shown in Table 2.1, that have directed our work for developing modeling lessons and units for use by science students. The guidelines are derived from a blend of our own experiences as a design team as well as the published work of others. Work on "driving questions" informed our ideas regarding how to choose phenomena for modeling (Kanter & Konstantopoulos, 2010; Krajcik et al., 1998) although there are differences with our approach. Below we describe our thinking about selecting phenomena and questions that relate to a particular topic or standard.

Table 2.1	
<i>Guidelines for Choosing Phenomena for Scientific Modeling Activities</i>	
<u>Design Challenge: Choosing Phenomena</u>	<u>Principles</u>
Guideline 1: The phenomenon should be accessible to students and well understood by scientists, but the mechanism that drives the phenomenon should be unfamiliar to students.	<p>1a. We recommend that designers choose a phenomenon that is familiar or understandable to students, but the mechanism should be unfamiliar to them (Falk & Brodksy, 2014).</p> <p>1b. To the extent possible, the designers should choose phenomena that are meaningful and relevant to students.</p> <p>1c. It can be advantageous for a designer to choose mysterious, counterintuitive, and non-obvious phenomena, which can enhance engagement (Hidi & Baird, 1986).</p> <p>1d. Mechanisms relevant to the phenomenon are more accessible if they have real world analogues that students are familiar with, especially macro-scale analogues.</p>
Guideline 2: Modeling should promote mechanistic understandings of phenomena.	<p>2a. Developing mechanistic models of phenomena is the primary aim of much of the work scientists do, and modeling activities should be consistent with this central feature of scientific work (Giere, 2004).</p> <p>2b. Many phenomena have multiple underlying mechanisms that causally intersect to produce them. It is often advantageous for students to explore multiple instantiations of the model.</p>
Guideline 3: There should be a significant base of evidence that supports the existence of the phenomenon and underlying mechanisms.	<p>3a. Models of candidate phenomena should be grounded in a significant amount of evidence.</p> <p>3b. Designers should carefully develop evidence so that it is accessible to students.</p>

We wanted to give students a chance to explore the role of mutations in human health and continue discussions about the topic of how genes and proteins produce a variety of traits (students had been studying genes and inheritance for about three weeks at that point). We chose the phenomenon of "HIV resistance" and developed

two questions around the phenomenon, first "Does resistance exist and is it genetic?" and second "How does HIV resistance work?" These two questions were explored sequentially. Students first explored two models and four pieces of evidence regarding the existence and genetic basis for HIV resistance. Then they engaged in the second lesson that included two new models and four new pieces of evidence devoted to understanding how the HIV resistance, established in the first lesson, works. In this way we have three levels of design decisions at work: (a) the topic of interest (i.e., the relationship between genes, mutations, and proteins); (b) the phenomenon of interest (i.e., HIV resistance); and (c) the driving questions (i.e., does HIV resistance exist and is it genetic? how does HIV resistance work?).

We will describe this in some detail later but a contrast here might be helpful. While staying on the topic of genes, proteins, and mutations we originally considered looking not at HIV resistance but rather at a range of other phenomena like allergies, obesity, and genetic diseases like Duchenne Muscular Dystrophy. We did not make much progress on allergies as a phenomenon or Duchenne Muscular Dystrophy. The phenomenon of obesity in mice was strongly considered and a lesson was partially developed surrounding that phenomenon but in the end we went with HIV resistance. Our reasons for selecting HIV instead of the obese mice are described in detail later.

2.2.1 Guideline 1: The phenomenon should be accessible and well understood by scientists but the mechanism that drives the phenomenon should be unfamiliar to students.

For any disciplinary core idea in the Framework for K-12 Science Education (National Research Council [NRC], 2012) there are many candidate phenomena that could be used to teach that idea. However, not all phenomena are equally compelling and

accessible for students. A phenomenon that is entirely novel and unfamiliar to students can be problematic, as students may not have any productive initial ideas to inform their early models and initial exploration. For example, the evolution of antibiotic resistant bacteria may be a compelling and cutting edge problem in medicine and highly relevant to the core idea of natural selection, yet, students who know little about bacteria or antibiotics will not have a productive starting point in exploring this phenomenon. This is not to say that one should never use this phenomenon in teaching evolution, but rather that as an initial entry point it is probably not the best option.

On the other hand, a phenomenon may be familiar and accessible but not compelling to students because it does not intersect with their lives in meaningful and relevant ways. For example, the evolution of Darwin's finches is a seminal phenomenon for scientists, however, students may not get nearly as excited about the beaks of little brown birds. Finding the right balance between familiar and perplexing is challenging. In their work on fostering student engagement Pitts and Edelson (2004, 2006) found that a mixture of motivations drove students' interest. They examined students' engagement during two modeling units; one focused on removing the sea lamprey (i.e., an ecologically disruptive invasive species) from the Great Lakes, and another lesson focused on finding out why some finches died and others did not on the Galapagos Islands. Researchers initially thought that either the role of the student (i.e., being asked to take on the role of a scientist) or the goal (i.e., finding out how to get rid of lampreys or explain differential mortality in finches) would be primary drivers for a student's engagement over a several week timespan as students engaged in extended inquiry activities (Pitts & Edelson, 2004, 2006). What they found was that while the role and goal

were salient for a few students, others were motivated by more situational factors like a particular lab exercise they completed or by considerations of receiving a grade for their work.

In this case what seemed interesting and curious to teachers and education researchers, namely adopting the role of being a scientist with the goal of solving problems and providing explanations, may not have been motivating for students (Pitts & Edelson, 2006). Fortunately, the opposite can hold true as well. Students can get invested in phenomena presented as mysteries even when the actual story seems rather dull, like a "made-up" letter from the Great Lakes Fishery Commission outlining the project with the invasive sea lampreys (Pitts & Edelson, 2006, p. 546).

Returning to the evolution of finches mentioned earlier, positioning this as a mystery of "what happened to the finches?" could generate enough puzzlement and curiosity even in students who do not find the organism or its problems particularly fascinating. Such an approach was successfully used in a software-based investigation of the finch population on one of the Galapagos Islands (Reiser et al., 2001).

Two additional constraints worth emphasizing relate to the compelling nature of phenomena. First, phenomena cannot be compelling and unexplained. The designers must know and understand the underlying mechanism involved. Thus phenomena on the cutting edge of science may not be resolved enough to serve as worthwhile cases for investigation by students. Second, as alluded to in the finch evolution example, we recommend that the phenomena be perplexing, puzzling, or counterintuitive in order to generate a need to know about the underlying mechanism (Hidi & Baird, 1986). Learning is goal-directed, and without a need to know, students are unlikely to expend the mental

effort involved in figuring out complex models and phenomena (Edelson, 2002). Thus a phenomenon needs to be known to designers with a balanced mix of familiar, accessible, and puzzling to students.

We distinguish between familiarity and accessibility to underscore that the accessibility of the phenomenon is not solely about familiarity with the phenomenon itself. It is the underlying mechanism, that students are expected to uncover, which needs to be accessible. That is, students should be able to reason about and conceptualize this mechanism; it does not mean they should know it (or be taught it) before engaging in modeling the phenomenon in question.

2.2.2 Guideline 2: Modeling should promote mechanistic understandings of phenomena.

Developing explanatory models of phenomena is central to the work of scientists in many fields (Giere, 2004). These kinds of models generally employ a mechanistic understanding of a phenomenon (i.e., the phenomenon is produced through a network of causal relations between components of the model). Scientists often work with multiple models across many scales of a phenomenon (Kitcher, 1993).

Consider a case where students are learning about genetics with the following learning goal: understanding the relationship between a gene, a protein's structure and function, and the resulting trait. The mechanism here is that genes are instructions for making the proteins necessary for normal cell and body function. If we want students to develop a model that links genes to proteins and traits, they need to explore multiple instantiations of the model.

Investigating several examples of relevant phenomena can help students generate a model which they can apply to other examples. Here too there are design decisions to

be made. The overall set of phenomena that students investigate or explain needs to reflect the explanatory scope of the model. These phenomena should include relevant nuance and distinctions that are important in the general model. In genetics this entails exploring both normal and abnormal traits, an array of protein functions that are affected, and both beneficial and harmful consequences of mutations. No single lesson can capture the full range of considerations here, which would be explored at the level of an entire unit on genetics. In the design case we describe in this article students are exploring how a single mutation can be beneficial to an organism, but in earlier lessons students explored other phenomena related to the central story of genes, proteins, and traits.

2.2.3 Guideline 3: There should be a significant base of evidence that supports the existence of the phenomenon and underlying mechanisms.

The identification of a puzzling, accessible, and known phenomenon is only the start of the process. Next, one must find evidence that students can use or generate in order to build or evaluate explanatory models. We discuss design decisions associated with evidence below. However, at this point we wish to stress that a good phenomenon with little evidence, or evidence that is not accessible to students, is not a workable design. At times we have identified a great phenomenon but upon closer inspection of the existing body of evidence it became clear that to understand the evidence, (even in adapted form) students would need knowledge above and beyond what was required by the target concept.

For example, we originally had plans to develop a third HIV resistance lesson that would focus on the origin and spread of the resistance mutation. At the time we were developing the lesson the science was not settled, which violated our first guideline that the phenomenon be well documented. Moreover, while we found a lot of studies that

could serve as evidence, it was the case that many of the methods used in these studies were well beyond what we felt could be productively adapted to a middle school classroom, given our time constraints on the project. It is possible a longer lesson could make productive use of this phenomenon, but at that time in our project it was not logistically feasible and we decided to move on. The role of evidence in modeling is a complex topic and will be addressed in greater detail in Design Challenge 3.

Lastly, in terms of beginning the search for phenomena, there are several resources we have found useful. The Next Generation Science Standards (NGSS, 2013) offer some suggestions regarding phenomena that can be used to teach the core ideas, and thus are a useful starting point in selecting a target phenomenon. Researching the scientific developments that led to the generation of the target model also often yielded interesting and productive phenomena for consideration. In addition, our large team included several domain experts who were familiar with a large array of phenomena; having deep knowledge of the domain is a critical characteristic of a team that can readily identify multiple candidate phenomena as well as relevant evidence.

2.2.1 Design Challenge 1 Example: The HIV Lessons

The HIV lessons were developed to help students understand the role that mutations and genes play in the production of proteins. The HIV resistance story has several compelling features that led us to choose this phenomenon. First, the mechanism by which resistance actually works has a macro-world analogue: the protein molecule on the surface of cells that the virus uses as an anchor is missing in HIV resistant individuals. Reasoning about anchors and their role in enabling an object to "dock" is not new to students. A macro-world analogue is an important consideration when students

are working with unobservable phenomena. Second, understanding how disease impacts human lives, and the role that genetics plays in how our bodies respond to disease, can be meaningful and relevant for students. Third, the actual mechanism is unknown to students and fairly esoteric and mysterious (at least initially), but students do have familiarity with the general idea of resistance to infections. Finally, students can understand the evidence that can be brought to bear in evaluating the models. Thus the HIV resistance phenomenon meets the proposed criteria for a productive choice for the design.

HIV resistance, however, was not the initial phenomenon of interest and our team's decision can shed light on navigating how to select a phenomenon. We initially identified research on links between obesity and genetics as a potential phenomenon of interest. On the one hand, there are numerous high quality studies about the interactions between genes, proteins, and diet. On the other hand, many of the most controlled studies have been conducted on laboratory animals, particularly mice. The role of some genes and the proteins they produce are relatively well documented in animals, especially control animals like knock-out mice (i.e., populations of mice that are identical except that they have been engineered so that they don't produce a particular protein). In many of these experiments the animals are tightly controlled for exercise, diet, and so on, so that researchers can isolate the role of the protein. However, obesity in humans is considerably more complicated, so making the connection from a model laboratory organism to human populations might be problematic for students. Additionally, we felt that understanding why lab mice are fat is not as meaningful as understanding how disease resistance works, particularly a disease with the cultural significance of HIV. Moreover, the idea that lab mice can be engineered to be obese does not seem as

counterintuitive, and fails to provoke a sense of inquiry or wonder compared to investigating how a relatively unknown population of humans can resist a deadly pandemic like HIV. During discussion in our professional development sessions both researchers and teachers shared the same concerns about the two phenomena and the consensus was that HIV would be a better phenomenon for reasons elaborated on above.

Once a suitable phenomenon has been selected, the learning environment designer is tasked with deciding how to represent the phenomenon in a way that is consistent with model-based inquiry. In short, the designer will need to develop a coherent set of models (i.e., Design Challenge 2) and evidence (i.e., Design Challenge 3) based on the phenomenon that engages students in ways that promote productive disciplinary engagement (i.e., Design Challenge 4). We will discuss each of these design challenges next.

2.3 Design Challenge 2: Developing Models

Scientists use models to describe, explain, and predict phenomena that are under investigation; successful models can point the way toward new investigations that previously had not been considered. For example, scientists have developed and refined, over the span of many decades, many models of the particulate nature of matter (e.g., the plum pudding model of the atom, the Bohr model of the atom, the standard model of particle physics). Each model of the particulate nature of matter opened up new avenues of inquiry leading to revisions of older models. Developing and revising models is central to science and is a challenging practice for scientists.

Developing models for students to use is challenging as well, in ways that are the same for scientists (i.e., students still try to describe, explain, and predict with models),

and in ways unique to learners or novices (i.e., students lack the years of training and experience and the deep disciplinary background knowledge of professional scientists). In the lessons we describe here students are provided with models. There are numerous pedagogical factors to consider, like how many models students should consider? (i.e., is just one model sufficient or should there be multiple competing models?). If competing models are used, how plausible should the alternative (i.e., incorrect) models be? Keeping in mind that students do not have the background knowledge of professional scientists, what is the right level of complexity? (i.e., what level of detail needs to be included and what can be left out?). Table 2.2 summarizes the guidelines and principles, discussed in more detail below, for resolving the challenges our team faced when developing modeling activities for student use.

Table 2.2

Guidelines for Developing Models for Scientific Modeling Activities

<u>Design Challenge: Developing Models</u>	<u>Principles</u>
Guideline 1: We recommend that models generated by a designer are, at least initially, comprehensible, plausible, compelling, and of comparable quality.	<p>1a. We recommend that designers develop models such that students cannot use surface features of the models to rule out, or embrace, a particular model before seeing any evidence.</p> <p>1b. We recommend that designers avoid models that are already well understood by students because the alternative models are implausible even before the activity begins.</p> <p>1c. When possible designers should choose incorrect models that reflect misconceptions that have been identified in the research literature (Pfundt & Duit, 1998).</p>
Guideline 2: We recommend that designers choose a developmentally appropriate modeling task from a range of tasks that represents a progression of different levels of sophistication.	<p>2a. Designers can choose from four basic core modeling tasks, and these can be combined in novel ways. These tasks are arranged from least to most difficult below:</p> <ul style="list-style-type: none"> i. Select a model and justify the selection ii. Rule out a model and justify its exclusion iii. Revise a model and justify the revision iv. Generate a model and justify its development <p>2b. The selection of a modeling activity (e.g., generating models) should reflect a consideration of what aspects of the phenomenon a student needs to come to know and the means (e.g., making models) by which they come to know it.</p>

2.3.1 Guideline 1: We recommend that models generated by a designer are, at least initially, comprehensible, plausible, compelling, and of comparable quality.

We favor engaging students in modeling tasks that involve comparing and evaluating multiple models. In science there is often more than one viable explanation, or model, for a phenomenon, and much of the work of the scientific community is centered on figuring out which explanation or model, among a field of competing alternatives, is the best. Therefore many of our instructional activities involve a multiplicity of models. This imposes a challenge in that a designer needs to create multiple models for students to use.

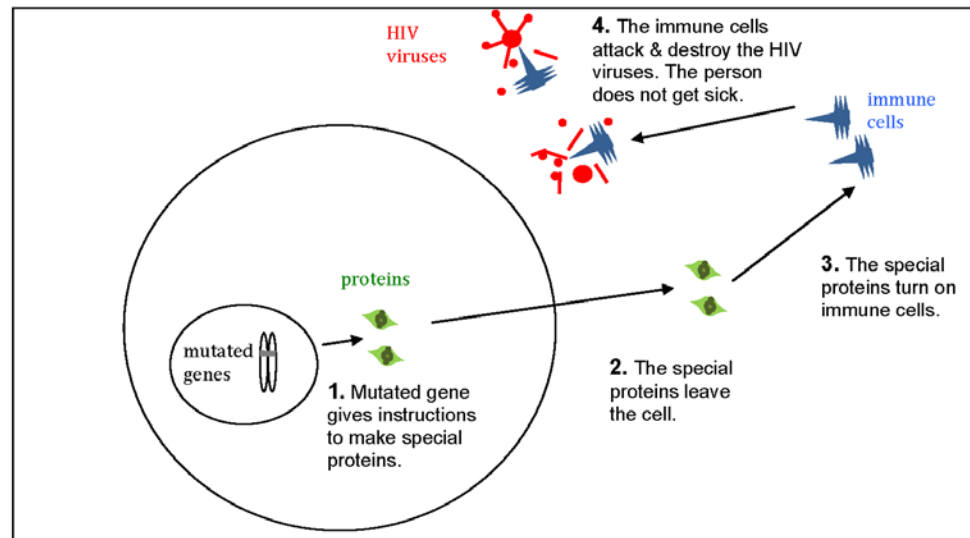
One of the key challenges designers face in creating modeling activities is developing two or more plausible models that are compelling and comparable in quality for the students to consider. Students will spontaneously use surface features of the models to make decisions about which model is better. Therefore it is up to the designer to develop models that require students to engage with the evidence before coming to a decision about which model is better.

While we are largely focused on describing the first of two HIV lessons throughout this paper, we would like to take a brief aside into what we feel is a very informative comparison of the models we used in the second lesson that will highlight some key features of this guideline. The reason for this is that the two models of HIV Lesson 1 are not particularly detailed. In HIV Lesson 1 students assess two competing claims: (a) Genetic resistance to HIV does not exist, and (b) Genetic resistance to HIV does exist. These models are intentionally simple and lack detail so that students can focus first on evaluating evidence and writing arguments.

In the second HIV lesson we introduce models that are more complex and include some of the mechanism of the resistance. The two models in brief are the "keep-it-out" model, which posits that a mutated gene fails to make a cell membrane protein (specifically an anchor protein) that the HIV virus uses to infect a cell, and the "attack-and-destroy" model, which posits that a mutated gene generates a protein that stimulates the immune system in a way that enables it to destroy the virus. In this case one of the models is correct (for the curious reader it is the "keep-it-out" model in which the anchor protein is missing) and the alternative model is incorrect, but both models have some initial plausibility for middle-school students. Figures 2.2 and 2.3 show both models of the second HIV lesson.

Model 1: THE ATTACK-AND-DESTROY MODEL

Resistant people have a mutated gene that keeps them from getting sick.



Sick people have a gene that does not make the special protein.

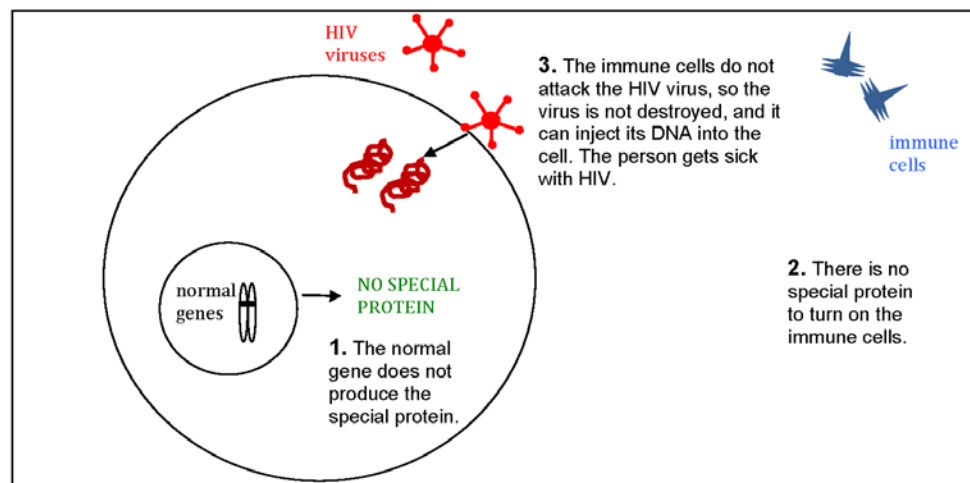
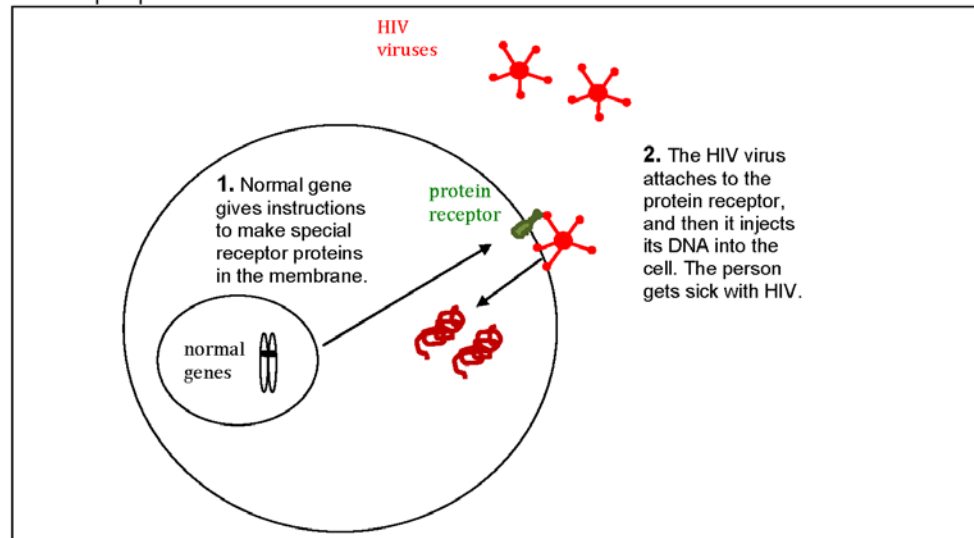


Figure 2.2. The “Attack-and-destroy” model shows that people resist HIV because of a mutated gene that makes a special protein that activates the immune system to fight the HIV.

Model 2: THE KEEP-IT-OUT MODEL

Sick people have a normal gene that makes a receptor protein that the HIV virus uses to infect people.



Resistant people have a mutated gene that keeps them from getting sick.

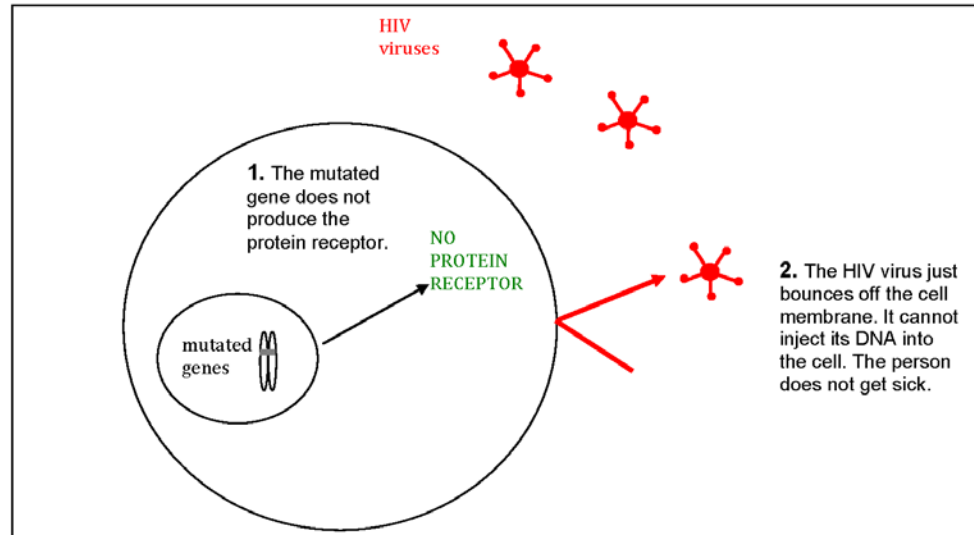


Figure 2.3. The “Keep-it-out” model shows that people resist HIV because of a mutated gene that fails to make a protein receptor that HIV needs to enter the cell.

While the phenomenon of disease resistance is well known, the correct model for HIV resistance is not. Well known models are not a particularly good choice for modeling

activities. The reason for this is that the alternative models are implausible before the activity even begins. For example, it is doubtful that middle school or high school students would carefully consider the details of evidence bearing on two models of the solar system, a heliocentric model and a geocentric model, because they know in advance which model is correct. A better approach to developing models is to develop multiple competing alternatives that have similar initial plausibility. The HIV models are a good example because students find them both equally compelling and plausible at the start.

In our designs we strive to make surface features like the number of steps in the model, how many words are used to describe the model, the amount of technical science language, and the layout and presence of images, as similar as possible, so that students are not favoring one model over another due to these superficial features. Given equal plausibility and similar structure, students focus on the relative merits of the models, evidence, and the relationship between them in order to arrive at an informed decision about which model is best.

Finally, it is advisable that when possible, lesson designers use modeling as an opportunity to address common student misconceptions. A common misconception about mutations is that they typically add a new function to the body (Nehm & Ha, 2011). Using the HIV models from above, some students think that seemingly positive mutations (e.g., resisting HIV Type 1) must involve adding a new function (i.e., ability to attack and destroy) and they do not consider that a beneficial mutation might remove a function.

2.3.2 Guideline 2: We recommend that designers choose a developmentally appropriate modeling task from a range of tasks that represents a progression of different levels of sophistication.

There are four basic categories of modeling activities that designers can choose from: (a) selecting a model from two or more competing alternatives based on evidence, (b) ruling out a model (eliminating it) from a field of competitors based on evidence, (c) revising an existing model and justifying the revision based on evidence, and (c) generating a new model and justifying its various components based on evidence.

Selecting a model is typically one of the least complex activities because students do not have the additional cognitive demands of ruling out a model, revising a model, or generating a model themselves. Ruling out a model is more cognitively demanding than selecting a model because it requires refuting a model by identifying the elements in the model that are inconsistent or incorrect. Revising and generating models are more demanding still, because they require revising an existing model by spotting and resolving incongruities in the proposed mechanism(s) or representing the mechanism(s) in a causal form from scratch.

We do believe that these four kinds of designs represent a progression toward higher levels of sophistication, but we recognize that it is possible to increase or decrease a particular activity's complexity, and subsequent demands on students' cognition, by manipulating a variety of relevant variables (e.g., how many models are present, what is the model complexity, how much evidence is needed to make a determination about the validity of a model, and so on). Designers can mix and match these four activities, for example, a lesson might have students first select a model from several slightly flawed or

simplified alternative models, and then engage in model revision as they gather and evaluate new evidence related to the model.

Beyond considerations of complexity, there are two primary criteria that a learning environment designer needs to consider when developing a model-based inquiry lesson or unit. First, what practices do we want students to develop facility with, and second, what elements of a phenomenon do students need to come to understand? For example, if a learning designer wants to focus on evidence-to-model relations (i.e., does a piece of evidence support, contradict, or lack relevance to a model) for students without much prior modeling experience, it might be better to focus on select-a-model activities or rule out a model activities. If the phenomenon of interest has a number of complex steps, then a focus on model revision might be better because through the model revision process students will develop a deeper understanding of the mechanisms involved (e.g., the steps in photosynthesis).

2.4 Design Challenge 3: Developing Evidence

Evidence plays a central role in the modeling practices of scientists and it also plays a central role in our lesson and unit designs. Students and scientists alike use evidence to make sense of models and arguments and to evaluate their plausibility and correctness. Considerable effort is expended by scientists to produce evidence. The scientific community, through academic publishing and conferences, expends even more effort in making sense of evidence and how it connects with the various arguments and models in a given scientific field. For example, establishing the bacterial cause of ulcers involved numerous empirical studies that were initially rejected by the majority of medical professionals working on the problem (Thagard, 2000). It wasn't until after

further empirical studies were conducted and examined in detail over a span of many years that the community finally came to accept an explanation that involved bacteria as a primary cause of stomach ulcers (Thagard, 2000). Similar to scientists and medical professionals, students also need time and social processes (e.g., evidence-based argumentation) to engage in the deep sensemaking process of examining evidence and its relationship to various explanatory models, if they are to gain facility in evidence evaluation practices. Table 2.3 summarizes the guidelines and principles, discussed in more detail below, for resolving the challenges our team faced when developing evidence for use by students engaged in model-based inquiry.

Table 2.3

Guidelines for Developing Evidence for Scientific Modeling Activities

<u>Design Challenge: Developing Evidence</u>	<u>Principles</u>
Guideline 1: Designers should take into account the variety of evidence features that can be varied along two continua: (1) complexity and (2) quality.	<p>1a. Evidence exists along two continua: (1) simple to complex evidence and (2) high quality to low quality.</p> <p>1b. Designers can foster students' evidence evaluation skills by designing evidence that exists along the full range of both the complexity and quality continua.</p>
Guideline 2: We recommend that designers create evidence that represents the authentic range of sources that can be encountered when learning about the phenomenon both inside and outside the classroom.	<p>2a. Designers can help students develop facility with evaluating evidence in different media by making sure that their evidence comes in a variety of formats including video, audio, text, simulations, charts, tables, and graphs.</p> <p>2b. Evidence exists along a continuum of fairly impartial to highly biased. Designers can encourage growth in students' sourcing skills by making sure that the sources of evidence span this continuum.</p>
Guideline 3: Evidence should often, but not always, contain data.	<p>3a. Authentic scientific evidence often contains data and analysis; the evidence students use should reflect this. The research on Adapted Primary Literature (APL) provides some grounding for designers looking to adapt primary sources for use by students (Yarden, 2009).</p> <p>3b. Much of the evidence we use in everyday reasoning does not contain data. Developing a complete toolkit of evidence evaluation skills requires students to encounter everyday evidence as well as scientific evidence.</p> <p>3c. Data can include qualitative evaluations by experts and non-experts.</p>

2.4.1 Guideline 1: Designers should take into account the variety of evidence features that can be varied along two continua: (a) complexity and (b) quality.

Here we will argue that there are at least two important continua that designers should consider when developing evidence. The continua are: (a) complexity, and (b) quality. We operationalize complexity as the features of evidence that place cognitive demands on students as they work toward understanding and using the evidence during modeling activities. These include, but are not limited to: reading level, use of specialized scientific terms, generating research questions, designing studies, and handling data by collecting, interpreting, and drawing conclusions from it. We operationalize the second continuum, quality, as the internal features of evidence that can be assessed against criteria for good evidence. For example, evidence quality criteria might include: the completeness of the data, the appropriateness of methods employed in the study, and the expertise and biases of the investigators. Numerous other evidence quality features can be considered as well.

The complexity and quality of evidence can interact in a variety of ways, as seen in Figure 2.4. We offer Figure 2.4 as a guideline to think about the relative strengths and weaknesses of each of the four major categories of evidence. The labels "Low Quality/High Quality" and "Simple/Complex" only indicate the extremes of each continuum. We do not want to suggest that there are only four kinds of evidence; rather we recognize that both evidence complexity and evidence quality exist along continua, and thinking about interactions between these two continua can give the designer a rough heuristic for considering important characteristics of the evidence.

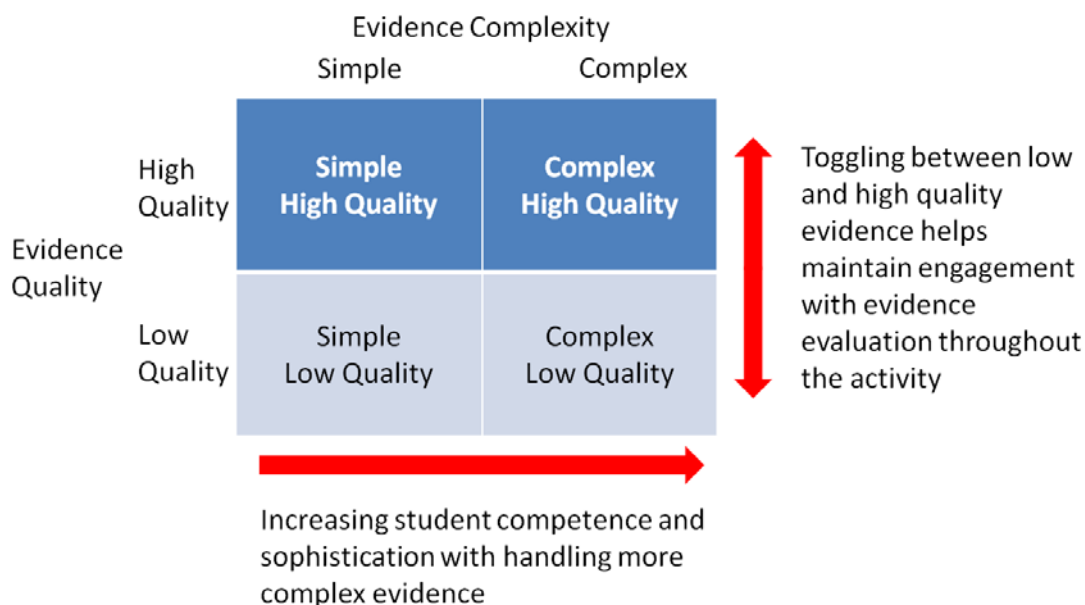


Figure 2.4. A heuristic for the combinations of evidence quality and evidence complexity

2.4.1.1 Complexity continuum.

Evidence can increase, or decrease, in complexity along a number of dimensions. To be clear, by complexity we mean complex for students to understand and use in modeling activities. Evidence that is generally simpler for students to understand and use has the following characteristics: it has a reading level that is at or below the students' level, it uses few specialized scientific terms, and it generates few demands on students like data collection, data interpretation, and drawing conclusions. More complicated evidence places more demands on the student (e.g., includes graphs or complex data tables), has a higher reading level, and uses more specialized science terms.

The complexity of evidence can be manipulated along these dimensions in ways that fit the pedagogical aims of the designer. For example, one may wish for students to gain facility with drawing conclusions and design evidence that requires students to engage with the evidence in this way. Similarly designers can manipulate the reading level complexity and use of scientific terms in ways that scaffold student work toward

promoting greater facility with reading scientific texts. There are numerous other ways that evidence complexity can be manipulated, more than can be addressed in this paper. Here we have highlighted some of the major ways that evidence complexity can be adjusted, with the aim of providing suitable challenges that offer students the opportunity to grow by engaging in the authentic practices of scientists and building their own knowledge.

2.4.1.2 Quality continuum.

Designers may also wish to scaffold students' thinking about evidence quality. We believe that promoting evidence quality evaluation is a worthy aim of science instruction and can be accomplished by manipulating different evidence quality parameters. For example, the designer may include data that are incomplete or contain anomalies in an effort to help students extend their thinking about how to deal with problematic data sources. The methods might include procedures that students are unfamiliar with or contain flaws that can only be identified with deeper content knowledge of the domain.

2.4.1.3 Evidence 1: simple and high quality evidence.

There are times when use of simple, high quality evidence is warranted and other times when it is not. On the one hand, simple high quality evidence provides students with an easy to understand exemplar of what strong evidence looks like, a benchmark against which to compare other evidence. On the other hand it does not provide for a very rich discussion about the merits of authentic scientific evidence, which oftentimes is much more mixed in terms of its quality.

Figure 2.5 shows the first piece of evidence that students consider in HIV Lesson 1. In evidence 1 students learn about the Feline Immunodeficiency Virus (FIV), which is

a virus that attacks the immune system in house cats in a way that is similar to how HIV attacks the immune system in humans. It is observed that house cats contract FIV easily. Dr. O'Brien gathered blood samples from thousands of large wild cats from around the world. After analyzing the samples, Dr. O'Brien concluded that wild cats are genetically resistant to FIV, and house cats are not genetically resistant to FIV.

Evidence 1 – FIV Video

Video Summary: The following is a summary of the video about FIV in cats.

Introduction: FIV stands for Feline Immunodeficiency Virus. FIV is a virus that attacks the immune system in house cats in a way that is similar to how HIV attacks the immune system in humans.



FIV was first observed in house cats, also called domestic cats. Dr. Stephen O'Brien noticed that house cats could get FIV very easily, and he was worried that FIV would spread from house cats to the large wild cats like cheetahs, lions, and pumas. Many of these species of large wild cats are endangered and could become extinct. Dr. O'Brien was afraid that many of these endangered species could die out if they were exposed to FIV.

Method: Dr. O'Brien gathered blood samples from thousands of large wild cats from around the world. He analyzed these samples. He used well known, reliable techniques for analyzing the blood for the presence of the virus.

Results: Most large wild cats like cheetahs, lions, and pumas already had FIV in their blood. However, they were not negatively affected by it because they possessed a genetic mutation that makes them resistant to the disease. Even though large wild cats get the virus, they do not become sick. Unlike wild cats, house cats do not have this genetic mutation and are not resistant to the disease. When house cats get infected with FIV, they often become very sick and can die.

Conclusion: From the blood samples of thousands of wild cats and house cats, Dr. O'Brien concluded that wild cats are genetically resistant to FIV, and house cats are not genetically resistant to FIV.

- 3A. Most wild cats who get FIV become sick and can die. True False
- 3B. House cats do not get the FIV resistant gene. True False
4. Geeta and Jose are arguing about this evidence. Circle the one you agree with the most.
- A. Geeta thinks cats are mammals like humans and research on cats is useful for understanding HIV.
 - B. Jose thinks cats are different from humans and research on cats is not useful for understanding HIV.
 - C. I don't agree with either of them.

Explain your choice for your answer to question 4.

Figure 2.5. Evidence 1, a summary of a video interviewing a well-respected geneticist discussing FIV in cats, is an example of simple high quality evidence.

Evidence 1 is a fairly simple piece of evidence because students have some familiarity with it (i.e., they are aware that animals can be sick), and the methods used in

the study are not described in great detail (i.e. the actual blood work methods are fairly complex, but that has been glossed over here for the middle school audience). It is seemingly high quality evidence because Dr. Stephen O'Brien is a well-regarded geneticist with a long track record of publishing studies on this topic.

We did introduce, via the questions at the end of lesson, a new concept that students may or may not have spontaneously considered with regard to evidence quality, and that is the validity of animal models. In this case, FIV is really quite different from HIV; however we chose to leave that topic open for student discussion and further consideration so that students could engage in the practices of scientists, like arguing about the validity of animal model evidence.

2.4.1.4 Evidence 2: simple and low quality evidence.

A designer might be inclined to provide students with only high quality evidence that supports the correct model lest students make mistakes, such as choosing the wrong model. Similarly a designer might be afraid that during evidence evaluation activities students might mistakenly form the belief that what is normatively weak evidence, especially simple low quality evidence, is in fact strong evidence.

Avoiding low quality evidence is a mistake because it, along with higher quality evidence, represents the epistemologically authentic range of evidence that people encounter in everyday life. Classrooms should not be epistemically sterile environments where only good evidence and models exist, rather a productive science classroom will provide students with the opportunity to develop heuristics of what is good and bad evidence and what makes some models better than others.

Evidence 2, as shown in Figure 2.6, is a simple low quality piece of evidence. This evidence is a report produced by a journalist after interviewing several subjects. The subjects are all experienced health care professionals working in a clinic that specializes in treating HIV positive patients. This evidence supports the incorrect model (i.e., that HIV resistance does not exist) because several of the clinic staff say they have never encountered an HIV resistant person.

Evidence 2 – Greater Area Health Clinic

Interview Report:

It is common for people with HIV to be treated in health clinics. A journalist interested in whether some people are genetically resistant to HIV interviewed the nurses and doctors at the Greater Area Health Clinic.

The journalist interviewed fifteen different nurses and doctors at this health clinic. Here are a few things the interviewees said:

Dr. Gutierrez: “It used to be, back in the 1980s, people would come in with HIV and there was very little that we could do to help. In the 1990s we developed medicine that attacked HIV in the blood stream. This reduced the infection but it didn’t cure it. People taking the medicine people live longer than people who don’t take the medicine.”

Nurse Singh: “I have worked in the labor and delivery ward for twenty-seven years. It used to be that if a pregnant woman came in and she had HIV, the baby would usually get the disease too. Now we can give mothers some medicine that reduces the chance the baby will get it. If the mothers don’t get the medicine, the babies will still usually get the disease.”

Dr. Morse: “With my patients I try to stress the point that everyone can get HIV. You can get it from injecting drugs with contaminated needles or having sex with someone who has the disease.”

Lab Assistant Feld: “I have worked in the blood lab for about five years. We check patients’ blood for HIV. The test is about 99% accurate. I have never met anyone who is resistant to HIV. We have had some patients who thought they were resistant because they injected drugs for a long time and didn’t get it. But within a few years they eventually got HIV.”

5. How do you rate the quality of this piece of evidence (0, 1, or 2)?

Give reasons for your rating.

Figure 2.6. Evidence 2, a report including statements made by a number of medical professionals, is an example of simple low quality evidence.

Some students tend to think of this as higher quality evidence because it involves medical professionals. However, once they encounter other pieces of evidence that are better, especially Evidence 4 discussed later, many students change the valence of their evaluation of Evidence 2 (i.e., the simple low quality evidence) and tend to think of it as weaker evidence because of the biased sample of individuals who visit an AIDS clinic (HIV resistant individuals are not likely to go there). Thus, facility with evaluating evidence quality relies on exposure to a variety of evidence of both low and high complexity and quality.

2.4.1.5 Evidence 3: complex and low quality evidence.

Similar to our reasons for why simple low quality evidence is worth student consideration, it is also good for students to consider evidence that, on its surface, has the trappings of complexity, like Evidence 3 shown in Figure 2.7. It is well established that novices tend to focus on surface features and fail to see the deeper connections that experts see (Chi, Feltovich, & Glaser, 1981). In this case students are presented with data that seem to allude to resistance having an inherited component.

Evidence 3 – SIV

Introduction: Monkeys can be infected by SIV (Simian Immunodeficiency Virus). SIV is similar to HIV, the virus found in humans. Some monkeys seem to be resistant to SIV even when exposed to the virus. Resistant monkeys have SIV in their blood, but they do not develop AIDS. Monkeys that are not resistant to SIV develop AIDS and get sick.

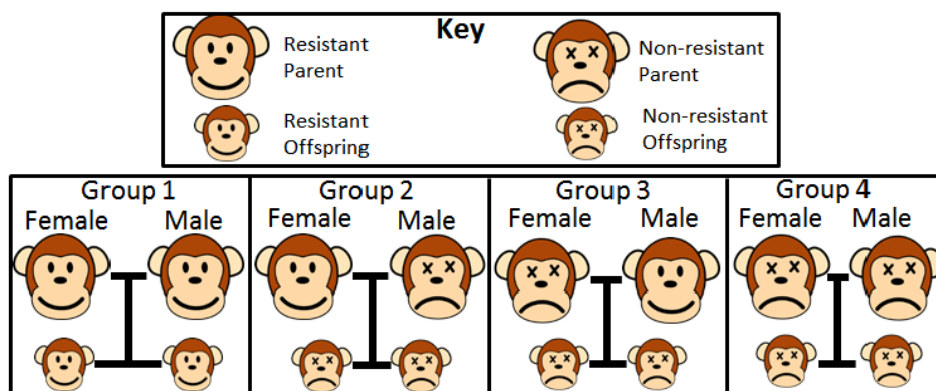
Method and Results: Scientists did four breeding experiments with eight parent monkeys. The groups were completely separated so that they did not have contact with monkeys outside of their group. All monkeys were tested for SIV resistance using high-quality blood tests.

Group 1: A resistant mother and resistant father have resistant offspring

Group 2: A resistant mother and non-resistant father have non-resistant offspring

Group 3: A non-resistant mother and resistant father have non-resistant offspring

Group 4: A non-resistant mother and non-resistant father have non-resistant offspring



10a. Is SIV resistance in monkeys genetic? Circle your answer.

- A. No it is not genetic.
- B. Yes it is genetic and resistance is a dominant trait.
- C. Yes it is genetic and resistance is a recessive trait.

10b. Explain why it is or is not genetic based on the results of this study. Give reasons for your answer.

Figure 2.7. Evidence 3, the results of an experiment using SIV in monkeys, is an example of complex low quality evidence.

In evidence 3 students learn that monkeys can be infected by the Simian Immunodeficiency Virus (SIV). SIV is similar to HIV, the virus found in humans. Scientists did four breeding experiments with eight parent monkeys. All monkeys were

tested for SIV resistance using high-quality blood tests. The only resistant offspring came from a pair of two resistant parents.

This evidence is more complex than either of the other two simple pieces of evidence (FIV and Health Clinic Interview) because it contains actual data in the form of four different family pedigrees for resistance/non-resistance to SIV and necessitates some additional processing to make sense of it and draw a conclusion.

In the case of evidence 3, the SIV study, the data are actually inconclusive. The pedigrees do not fully establish whether the trait is dominant or recessive and fail to establish that SIV resistance is genetically based. Moreover the study has a very small sample size, which decreases the quality of this evidence. This evidence also gives students a chance to revisit the issue of the utility of animal models in understanding human disease. In this case, SIV is actually a close relative of HIV, unlike FIV, which is highlighted in the first piece of evidence. Students also have a chance to discuss issues related to sample size as well as use their knowledge of pedigrees (gained in a prior lesson) to puzzle out the phenomenon of potential SIV resistance.

2.4.1.6 Evidence 4: complex and high quality evidence.

Evidence that is both complex and high quality provides students with the opportunity to develop sophisticated practices in two ways. First, designers scaffold students toward handling more complex evidence. Second, higher quality evidence presents an important contrast with lower quality evidence. This contrast affords students opportunities to engage in important discussions about evidence quality that would not be possible without contrasting high and low quality evidence.

Evidence 4, as shown in Figure 2.8, describes how Dr. Paxton and his team of researchers studied a group of 25 people who had been exposed to HIV many times. Despite many exposures, the people in the study were HIV negative. Their white blood cells were exposed to different levels of HIV in a test tube. All 25 peoples' white blood cells showed some resistance, with some being resistant to very high levels of HIV. This evidence strongly supports the correct model, that HIV resistance does exist.

Evidence 4 – Dr. Paxton's Study

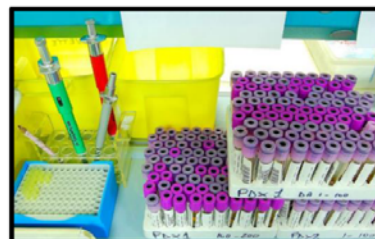
Introduction: During the 1990s Dr. Paxton heard that there were some people who had been exposed to HIV, but didn't develop AIDS. He wanted to see if their immune system cells would be resistant to HIV if they were exposed to it again. People who have unprotected sex or inject illegal drugs are more likely to get HIV, so they decided to study these people.



Method: Dr. Paxton and his team of researchers studied a group of 25 people who had been exposed to HIV many times. Despite many exposures, the people in the study were HIV negative, which means that there was no HIV in their blood.

The researchers used white blood cells taken from these 25 people. The white blood cells were exposed to different levels of HIV in a test tube.

Results: All 25 peoples' white blood cells showed some resistance. Some people had immune system cells that were resistant to very high levels of HIV in the test tube.



11. What conclusion do you draw from this study? Explain your answer.

Figure 2.8. Evidence 4, an example of adapted primary literature, is a simplified version of the methods and results of a study carried out by Paxton and colleagues (1996).

The Dr. Paxton Study is more complex than evidence 1, the FIV study, and evidence 2, the interview with health clinic staff, because it includes more detailed method and results sections and asks students to draw their own conclusions. The evidence is higher quality because it (a) involves a larger sample size than the previous pieces of evidence students have seen in this lesson, (b) it directly uses humans as test subjects, and (c) it uses established medical science procedures for "in vitro" experiments with white blood cells.

2.4.2 Guideline 2: We recommend that designers create evidence that represents the authentic range of sources that can be encountered when learning about the phenomenon both inside and outside the classroom.

Real world evidence also comes in a variety of formats (e.g., text, video, animations, simulations, tables, charts, graphs), from a variety of sources (e.g., first hand observations, second hand accounts of empirical work like work published in scientific journals, popular science texts), and spans the full range of quality from low to high (i.e., some evidence comes from competent investigators with robust methods and other evidence comes from less competent sources). Assessing the quality of evidence also affords students the opportunity to evaluate the role of bias in scientific evidence. The act of gathering or presenting evidence is often purposefully aimed at solving a problem or bringing clarity to a situation, and as such the bias of those involved in the collection of evidence is important to assess.

2.4.3 Guideline 3: Evidence should often, but not always, contain data.

Data play a central role in authentic scientific evidence. However, laypeople (non-scientists or even scientists outside of their own domain) rarely engage with primary literature (Bromme, Kienhues, & Porsche, 2010). It is often the case that laypeople make

sense of scientific phenomena, like the latest discoveries of the New Horizons Probe to Pluto or the latest particle discoveries at the Large Hadron Collider, based on secondary sources as reported in popular media outlets. It is usually the case that the original published articles are beyond the expertise of the average layperson. Even in the case of health care decisions, the layperson is often faced with reasoning about phenomena with only secondary sources or anecdotes, like a doctor's account of what he or she personally feels works with his or her patients, to guide them. We feel it is important to capture the range of everyday evidence, which usually lacks data, while still engaging students in reasoning about data in the way that scientists do. Consequently we argue that some evidence, but not all evidence, should contain data. Reasoning about evidence that lacks data is just as useful a life skill as reasoning about evidence with data.

We have a three-pronged approach to developing evidence with variable levels of data inclusion. The first, broadly speaking, is encompassed by developing Adapted Primary Literature (APL) sources of evidence (Yarden, 2009). The second involves developing evidence that is more consistent with a Journalistic Reported Versions (JRV) approach to evidence. The third and final prong involves the typical kinds of anecdotal evidence encountered in daily life. To briefly distinguish between the three options we can say that APL- style evidence includes data, JRV-style evidence frequently points to another source that has data, and anecdotes typically use low-quality data (often qualitative in nature) that are not gathered systematically. We will describe each style in greater detail next.

Adapted Primary Literature (APL) involves the designer transforming a piece of primary literature, like an article in *Science* or *Nature*, into a succinct and comprehensible

piece of evidence. APL style evidence often mirrors the typical style of a published peer-reviewed scientific article in that it contains an introduction, methods, results and conclusion. We have found that problematizing one or more of these four structural elements (e.g., a slightly flawed methods section, a conclusion that isn't quite supported by the evidence and so on) can make for rich discussions about evidence quality.

Consider the following example. It is common to teach students that large sample sizes make the findings of a study more robust and smaller sample sizes are problematic. A sample size of one could in fact be highly problematic in some contexts but in the context of medical studies, particularly case studies, a sample of one can yield very important findings. One piece of APL evidence we have developed is based on an important medical case study (Allers et al., 2011) involving the "Berlin Patient" who is the first known human being to be cured of HIV by leveraging knowledge about the mechanism of genetically based resistance to HIV. The "problem" with this study is that it rests on a single patient, however in the eyes of scientists this was a highly influential finding. Grappling with the tension between large and small sample size studies, gives students the opportunity to discuss the relative strengths and weaknesses of various authentic investigative techniques employed by scientists, in a way that would not be possible if students did not have a range of evidence of variable quality to consider.

Journalistic Reported Version (JRV) evidence often makes use of a primary source, similar to APL, but as is typically consistent with journalistic conventions, actual data and statistics are not part of the evidence itself but are rather referred to with some sort of in-text citation. We often use JRV-style evidence because it is an important part of the authentic range of evidence that students encounter outside of school. For example,

evidence 1 (the FIV video) is a typical JRV piece of evidence, albeit in video form (note: we present a written summary here because it was used in classes as well, as a reference document for students so that they didn't need to watch the video more than once). The video is a short narrative about an individual scientist's concerns about a possible connection between FIV and HIV. No data are presented in the video but the scientist, Dr. Stephen O'Brien, does refer to past empirical research he has conducted on the topic.

Finally, anecdotal evidence is common in everyday life. Evidence 2, an interview with several medical professionals, represents the typical type of anecdotal evidence people encounter as they attempt to make sense of their world, through the lens of past personal experiences or insights gleaned from their educational and professional backgrounds.

We argue that using all three types of evidence provides students with the opportunity to engage with the full range of evidence one can encounter. While we do not specifically label evidence for students as any one of these three types, we think that contrasting different styles of evidence provides learners the chance to discuss what role data, or lack of data, plays in evidence evaluation and modeling activities.

2.5 Design Challenge 4: Productive Disciplinary Engagement

One of the aims of reform-oriented science instruction is to move students into the position of being constructors of their own knowledge through the authentic practices of scientists. We take productive disciplinary engagement to be deep student involvement in problem solving while engaging with the epistemic (Pluta, Chinn, & Duncan, 2011) and social norms of the knowledge production processes used by scientists (Engle & Conant,

2002). Engle and Conant (2002, p. 399) recommend four principles for fostering productive disciplinary engagement including:

1. "problematizing subject matter"
2. "giving students authority to address such problems"
3. "holding students accountable to others and to shared disciplinary norms"
4. "providing students with relevant resources"

In general we agree that all four principles are important and we will elaborate on how our lesson and unit designs have instantiated these. So far in this paper we have described several ways of selecting phenomena for modeling as well as structuring models and evidence to promote "problematizing of subject matter." The next set of guidelines, as shown in Table 2.4, draws on a blend of our experiences as a team and primary literature that is relevant to learning in science classrooms. We have found these principles useful in guiding the development of our learning environments where we aim to promote productive disciplinary engagement during modeling, with particular emphasis on the remaining three principles from Engle and Conant (2002).

Table 2.4

Guidelines for Generating Productive Disciplinary Engagement with Scientific Modeling Activities

<u>Design Challenge: Generating Productive Disciplinary Engagement</u>	<u>Principles</u>
<p>Guideline 1: Student autonomy and accountability can be promoted through adoption of the norms of science like disciplinary talk (Engle & Conant, 2002) and epistemic criteria (Pluta et al., 2011).</p>	<p>1a. Learning environment designers can promote autonomy by putting students in the role of decision makers and problem solvers.</p> <p>1b. We recommend that designers guide students toward developing discussion stems that foster disciplinary talk (Michaels, Connor, Resnick, L. B. 2008).</p> <p>1c. Learning environment designers can encourage the use and adoption of disciplinary scientific practices by focusing students' attention on the use of epistemic criteria (Pluta et al., 2011).</p>
<p>Guideline 2: To foster deep cognitive processing, inquiry should be structured with scaffolds that promote quality of evidence evaluation and help students develop systematic relations between evidence and models.</p>	<p>2a. Designers are encouraged to incorporate scaffolds that promote systematic examination of the relationship between evidence and models (Rinehart, Duncan, & Chinn, 2014; Lombardi, D., Sibley, B., & Carroll, K., 2013; Toth, Suthers, & Lesgold, 2002; Suthers & Hundhausen, 2003).</p> <p>2b. Designers can incorporate scaffolds that promote model and evidence quality evaluation (Authors, 2014).</p>
<p>Guideline 3: Designers should take into account the variety of evidence-to-model relations that can be varied along two continua: (1) relevancy and (2) diagnosticity.</p>	<p>3a. Evidence exists along two continua: (1) low relevance to high relevance and (2) low diagnosticity to high diagnosticity.</p> <p>3b. Students' evidence-to-model relation skills can be fostered when they encounter evidence that exists along the full range of both the relevancy and diagnosticity continua.</p>
<p>Guideline 4: To foster productive disciplinary engagement, the designer should consider incorporating into their lessons designs that engage students in the socio-epistemic practices of science.</p>	<p>4a. Argumentation is a central socio-epistemic practice of science (Erduran, Simon, & Osborne, 2004). Written argumentation activities can be designed to enhance the authenticity of modeling in science classes and promote deep processing of evidence and models.</p> <p>4b. We encourage designers to develop assessments and activities that effectively capture students' facility with the scientific practices and content of the modeling activities.</p>

2.5.1 Guideline 1: Student autonomy and accountability can be promoted through adoption of the norms of science like disciplinary talk and epistemic criteria.

Our use of discussion stems to promote disciplinary talk is inspired by work on Accountable Talk™ (Michaels et al., 2008) and Guided Questioning (King, 1992). The aim of Accountable Talk™ is to develop a community of practice that is grounded in respectful, yet critical, discussions about evidence, claims, knowledge, and reasons. Our use of discussion stems is also rooted in the work of Guided Questioning where students are provided with general questions that are "content free" to guide their discussions (King, 1992). We built our discussion stems with Accountable Talk™ and Guided Questioning in mind, although our instantiation is particular to our project and is not a direct implementation of either.

In the first year of our project we used extensive lists of discussion stems with the aim of promoting sophisticated disciplinary talk (see Figure 2.9). Feedback from teachers, as well as our own observations in class, indicated that this approach was problematic. The lists were too lengthy, too specific, and were difficult for students to use because of the additional cognitive load imposed by tracking which discussion stems should be in use for a particular activity. Moreover, those lists were generated by the research team rather than by the teachers or students, and we have reason to believe based on teacher feedback that student "buy-in" was low. In the following year we changed our approach.

Discussion STEMS

General STEMS	Evidence Understanding and Evaluation
Listening and sharing ideas with the whole group I don't know what you mean by _____. Could you explain _____ more? What do you think _____? I want to add to what (name) said about _____. To expand on what (name) said _____, _____, what do you think?	Purpose Why did they _____? What was the purpose of _____?
Giving reasons and developing arguments I think _____ because _____. _____ because _____. Why do you (agree/disagree/think) _____? I agree with _____ because _____.	Method The most important steps in the method were _____. In this study, they _____. Why did they (do) _____? What did they do after _____? After they _____, they _____. They were careful to _____.
Challenging and thinking carefully about issues I disagree with _____ because _____. An argument on the other side is _____. What about the argument that _____? I still have questions about _____. A question I have is _____. An example of _____ is _____. This reminds me of _____. I understand _____. I'm confused by _____.	Results What were (the results) _____? This (graph/table/photograph) shows _____. What does (graph/table/photograph) mean? Why are _____ and _____ the same / different?
	Conclusion The (conclusion) is _____.
	Evaluating the evidence They could have made the study better if they had _____. What if they had done _____ rather than _____? Why is this study/evidence (good/bad) _____? A problem with this study is _____. What are the (problems/good points) of this evidence? What are your reasons for rating this study 0, 1, 2, 3? What criteria does this evidence (meet/not meet)? This study is (0, 1, 2, 3, bad, good) because _____. We can, can't believe the conclusion because _____.

Figure 2.9. Discussion stems used in an earlier iteration of the project

In the second year of the project we included in our designs very short lessons in which students generated discussion stems that they used to structure their own conversations. Having students develop the criteria themselves we believed would lead to greater "buy-in" as well as get students comfortable with taking on the autonomy of being problem solvers. We developed a very short 15 minute activity in which students had the opportunity to develop their own discussion stems. In this activity, which was a preparatory activity that students participated in before engaging with the modeling lessons about HIV, students were placed in the role of a city council in Christchurch, New Zealand. They viewed a few PowerPoint slides containing information about the major 2011 Christchurch Earthquake that destroyed many of the buildings in the city. As the city council, they were asked to consider if the new replacement buildings should be

constructed of wood or stone? The aim of the lesson was not to develop a lot of content knowledge about earthquakes, but rather to provide an opportunity to use an accessible topic (i.e., buildings being destroyed by earthquakes) to foster disciplinary norms for argumentation.

Students were asked to guide their discussion using stems that they themselves had developed. To do this, students generated three lists of stems: (a) giving reasons, (b) asking for reasons, and (c) disagreeing with the reasons of others. Examples of these include: (a) I think that ____ is better because of ____, (b) What is another reason that you think ____ is better, and (c) I disagree with ____ because of _____. The activity promoted autonomy by giving students the opportunity to act as decision makers. It promoted disciplinary norms like asking for reasons, giving reasons, and making it "ok" to disagree with one another, as well as establishing disciplinary talk by using student-generated discussion stems to guide their conversation. No systematic investigation into the impacts of the stems has been undertaken at this time, but teacher feedback indicated that students were not overwhelmed as had been the case the previous year.





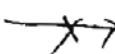






In addition to developing social norms, students also generate epistemic criteria for use in modeling activities. Epistemic criteria guide scientists and students in their evaluation of scientific processes and products (Pluta et al., 2011), and for the purposes of model-based inquiry classrooms we can distinguish at least three types of criteria: (a) model criteria, (b) evidence criteria, and (c) argumentation criteria. Past research has shown that students are surprisingly adept at generating and refining lists of criteria that match the sophisticated criteria used by practicing scientists (Pluta et al., 2011). Our own designs make use of explicit aggregated class lists (i.e., lists that pull together

contributions from different groups of students within a class) of student-generated epistemic criteria of the three types mentioned earlier. Example criteria might include items like "Good evidence should usually have a large sample size," "Good arguments should have reasons," or finally, "A good model will include clearly labeled steps." A more detailed treatment of students' use of model criteria has previously been published (Pluta et al., 2011).

2.5.2 Guideline 2: To foster deep cognitive processing, inquiry should be structured with scaffolds that promote quality of evidence evaluation and help students develop systematic relations between evidence and models.

Engaging in the practices of modeling can be cognitively demanding and designers should take this into account. Research has shown that even undergraduate college students find modeling challenging (Windschitl et al., 2008a). To offload some of the simultaneous cognitive demands imposed by modeling we have developed a suite of scaffolds and graphical organizers, based on the work of Suthers and colleagues (Suthers & Hundhausen, 2003; Toth et al., 2002) called the Model Evidence Link (MEL) matrix (Chinn, Duschl, Duncan, Buckland, & Pluta, 2008; Rinehart et al., 2014). The MEL matrix is designed to facilitate systematic model and evidence evaluation. We feel that it meets the fourth criterion set forth by Engle and Conant (2002), that students should be provided with the resources needed to be effective problem solvers. This is also commensurate with the scaffolding framework by Quintana and colleagues (Quintana et al., 2004), which suggests that making disciplinary strategies explicit in the tools and artifacts students use is beneficial for novices because it makes the expert practices salient. A sample MEL matrix is shown in Figure 2.10.

Arrows Diagram

Evidence Goodness Rating	Model 1: Genetic resistance to HIV does <u>not</u> exist.	Model 2: Genetic resistance to HIV does exist.
1. FIV Video  <div style="border: 1px solid black; padding: 2px; display: inline-block;">2</div>		
2. Greater Area Health Clinic: Interview Report <div style="border: 1px solid black; padding: 2px; display: inline-block;">1</div>		
3. SIV Study  <div style="border: 1px solid black; padding: 2px; display: inline-block;">2</div>		
4. Paxton Study  <div style="border: 1px solid black; padding: 2px; display: inline-block;">2</div>		

12. For **all** the pieces of evidence make sure to rate them (0, 1, or 2) and draw an arrow for how the evidence relates to each model.

13. Which model is better? Circle your selection.

Model 1: Genetic resistance to HIV does not exist.

Model 2: Genetic resistance to HIV does exist.






Support	
Strongly Support	
Contradict	
Strongly Contradict	
Irrelevant	

Figure 2.10. The MEL Matrix for HIV Lesson 1. including the arrows diagram, evidence quality boxes, and student model selection boxes.

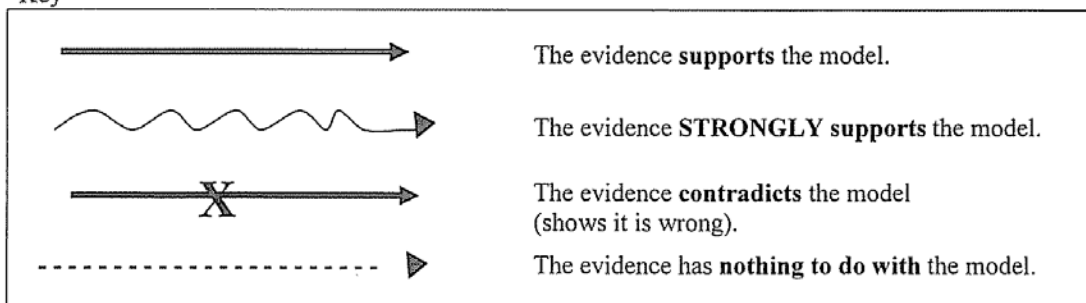
Across the top (i.e., the columns) of the MEL matrix are the various models under consideration (two in this case) and across the side of the chart (i.e., the rows) is each piece of evidence with a brief reminder (picture or label). Students complete the table by filling in the evidence-to-model connection arrows, of which there are five kinds: (a) strongly support, (b) support, (c) irrelevant, (d) contradict, and (e) strongly contradict. The arrows show the connection between the evidence and the model. Within the

evidence boxes (i.e., the rows) there is another box to display a numerical rating of evidence quality ranging from 0 (i.e., evidence that is so bad it shouldn't be considered evidence) up to 3 (i.e., excellent high quality evidence). We recommend that designers develop evidence so that there is a range of relationships between evidence and models. When there is a range of relationships (from strongly support to strongly contradict) represented across the full body of evidence that they consider, students have the opportunity to engage in disciplinary talk about what makes a piece of evidence support, or even strongly support, a model and perhaps contradict another model.

The MEL shown in Figure 2.10 is a highly refined product that has been through several major rounds of revision. Our earliest attempts at using the MEL (i.e., MEL 1.0) can be seen in Figure 2.11. The MEL 1.0 varied from the MEL 2.0 in several ways. First, and probably most noticeable, is the tangle of justification arrows (i.e., the crisscrossing mass of arrows). It is also worth noting that there were only four arrow types (strongly support, support, irrelevant, and contradict) and there were no evidence rating boxes. The MEL 2.0 introduced an arrow type, the strongly contradict arrow. With the revised MEL we hoped that students would be able to have finer grained networks of justification. For example, a student could now make a statement like "evidence 1 supports model A and evidence two strongly contradicts model A." The idea was that finer distinctions would give students grounds to be more discerning about evidence features (i.e., attending to why one study might support a model while another piece of evidence strongly contradicts a model).

Draw one arrow from each evidence box to each model. You will draw six total arrows.

Key



Evidence #1. Scientists have found that many small mammals—including mice, rats, squirrels, and rabbits—have the Red Fever virus in their bodies.

Evidence #2. Scientists found that many workers who shared a cafeteria at their office got the Red Fever virus.

Evidence #3. Scientists have found that in areas where there are many mosquitoes, such as near swamps or marshes, there are many more people who get Red Fever disease than in areas where there are few mosquitoes.

Model A Mosquito model

1. Many small mammals carry the Red Fever virus even though they do not get sick.
2. Mosquitoes bite small animals as well as humans.
3. When mosquitoes bite these small mammals, they get the virus inside them.
4. Those mosquitoes bite people, and the people get sick.

Model B Feces model

1. Many small mammals carry the Red Fever virus even though they do not get sick.
2. Some of these small mammals get into indoor areas where people keep food, and they leave feces that contain the virus.
3. The feces with the virus touches the food that people eat.
4. People eat the food and get sick.

Figure 2.11. A typical example of student work using the MEL 1.0 for a modeling activity about the cause of a disease. The HIV lesson described throughout this design case did not make use of the MEL 1.0 so we had to use a representation from another activity. For our purposes here the key features are the elements of the MEL (i.e., the lack of evidence rating boxes, the free form arrows, fewer linking arrow choices, and so on) rather than the evidence and models.

Second, the early MELs were useable with smaller evidence sets, perhaps three or four pieces of evidence, and most appropriate when only one or two models were being considered. Later designs introduced more evidence and the "connect the arrow to the models" method became unwieldy. Both teachers and researchers found the tangle of arrows a bit difficult to navigate. For the MEL 2.0 we shifted from the tangle of arrows to a table format to enhance readability while still maintaining the metaphor of "connecting evidence to models" that the arrows represented.

Finally, and most significantly, we added evidence rating boxes. Our decision to include this in the design revolved around our desire to promote student comprehension and consideration about the quality of the evidence. Students rated evidence on a numeric scale with a range of 0–3, where 0 is very low quality evidence that is so bad it probably should not be considered worthwhile evidence and probably does not merit a justification arrow, and a 3 would be considered very high quality evidence. We also tried a narrower range of 0–2, but felt that 0–3 was more successful. The aim of reducing the range was to try to encourage students to give really bad evidence a rating of zero, because in previous studies we noticed considerable student resistance to giving lower quality ratings to bad evidence. However, students often times just alternated between giving evidence a 1 or a 2 and still resisted giving evidence a 0. To provide support for using the evidence quality ratings effectively teachers worked with students to develop class level criteria lists for what counted as high quality evidence. These lists were refined over time, typically on an interval of four to six weeks.

2.5.3 Guideline 3: Designers should take into account the variety of evidence-to-model relations that can be varied along two continua: (a) relevancy and (b) diagnosticity.

Beyond considerations of how each piece of evidence relates (e.g., support, contradict, etc.) to each model under consideration, there are two additional parameters of interest that designers should consider when developing evidence to be used with models. The parameters are: (a) relevancy and (b) diagnosticity. We place them here in the section on disciplinary engagement, rather than in the developing evidence section, because relevancy and diagnosticity surface only when evidence is considered in relation to models, as discussed in the previous guideline. To be clear, evidence cannot be relevant or irrelevant, nor diagnostic or non-diagnostic, without considering the model to which it applies (or fails to apply). Moreover, engaging in discussion about the relevance and diagnosticity of evidence as it relates to the models in question pulls students into deeper engagement with the disciplinary norms of science.

Rarely does a single piece of evidence relate to all of the elements of a given model. For example, a simple model of disease resistance might still contain many elements, like the role of proteins produced by the genes in a cell, the role of antibodies, and the location of these entities within or between cells. Oftentimes it is the case that evidence connects to just one, or a few, elements of a model. For example, evidence 4 in Figure 2.8, "The Paxton Study," is relevant to one part of the model students worked with, namely the existence of HIV resistance. However the same piece of evidence is silent on the second element of the model, that HIV resistance is genetic. So in the case of the two models discussed above, "The Paxton Study" is relevant to part, but not all, of the model.

The second parameter, diagnosticity, is intimately related to, but not the same as relevance. Diagnosticity rests on the learners' ability to distinguish differential levels of support or contradiction for two or more models. Again consider the case of the "The Paxton Study." It is highly diagnostic between the two models in terms of the existence of HIV resistance (it exists). Based on this evidence the learner can support one model (that resistance exists) and reject the alternate model (it does not exist). This is unlike some of the other pieces of evidence that may be perceived as having lower relevance and subsequently lower diagnosticity. For example, the FIV video might be thought of as irrelevant because FIV and HIV are very different diseases and it might be the case that the findings from feline animal models do not map well onto investigations with humans. Engaging students in considerations of the diagnosticity and relevance of evidence, as it relates to the models in question, is a highly authentic epistemic practice of scientists and worthy of consideration when designing modeling lessons.

Relevance and diagnosticity interact in ways that can be complex for the lesson designer. It is the case that both relevance and diagnosticity exist on a continuum of possibilities. With that in mind we offer Figure 2.12 as a guide to thinking about the relative strengths and weaknesses of each of the four major categories of evidence. The labels "Low Relevance/High Relevance" and "Low Diagnosticity/High Diagnosticity" only indicate the extremes of each continuum. We do not want to suggest that there are only three kinds of relationships; rather we recognize that both relevance and diagnosticity exist along two continua. We provide Figure 2.12 as a rough heuristic that designers can use for thinking about the relationships between the evidence and models they develop. While it is certainly the case that scientists hope to develop studies that aim

for high relevance and high diagnosticity, not all studies achieve this. To simulate the authentic range of evidence found in real science we encourage designers to consider manipulating both the diagnosticity and the relevance of the evidence they design.

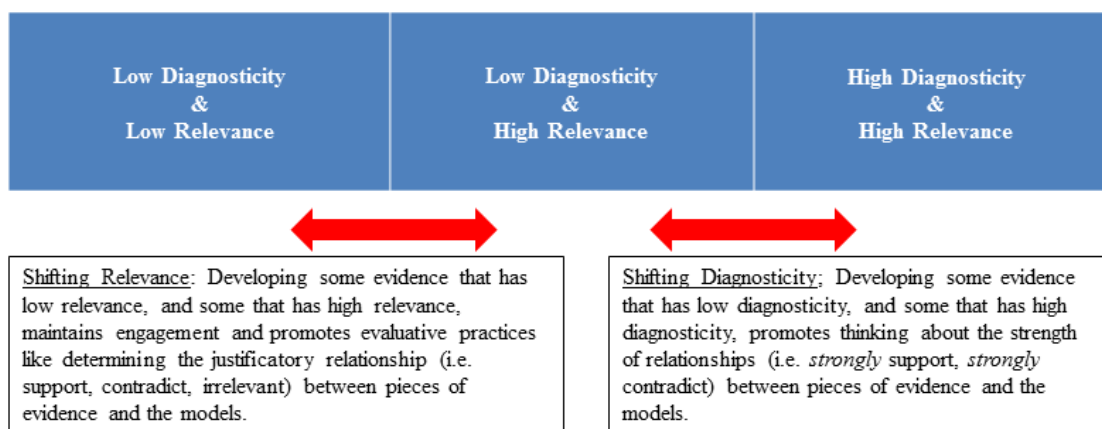


Figure 2.12. The three basic combinations of relevance and diagnosticity, showing the two major design decisions: (a) *Shifting Relevance* and (b) *Shifting Diagnosticity*.

2.5.4 Guideline 4: To foster productive disciplinary engagement, the designer should consider incorporating into their lessons designs that engage students in the socio-epistemic practices of science.

At the conclusion of a lesson (keeping in mind lessons sometimes stretch across several days), students are offered a final opportunity to revise their MEL matrix and write a final argument in support of the model they favor. The chance to revise is important because as students are exposed to more evidence their evaluation of the quality of evidence can change. For example what once may have seemed like good evidence may not seem so strong after seeing other evidence that is even better. Once revisions are completed students write a final argument, leveraging their argument criteria, based on the evidence they have worked with. This final epistemic product is authentic to science in that they are making a case for (and/or against) a model that attempts to explain a phenomenon or class of phenomena. The culminating activity of the

final argument and revised MEL Matrix affords teachers the opportunity to assess the content and practices of what students have learned in a setting that is more epistemologically authentic than, for example, a multiple choice or fill in the blank type assessment.

2.6 Conclusion

The Next Generation Science Standards necessitate a serious shift in the way we engage in classroom practices, and as such require a move away from epistemologically inauthentic practices, such as "cookbook" labs, and toward the epistemic and social practices that scientists actually use, like scientific modeling and argumentation. Many of the requirements to generate new reform-oriented classroom materials will fall on the shoulders of teachers and science administrators. In this paper we have outlined what we feel are the four major challenges faced by reform-oriented designers in creating modeling and argumentation activities: (a) choosing a phenomenon, (b) developing models, (c) developing evidence, and (d) generating productive disciplinary engagement. Within each challenge we provide guidelines as heuristics aimed at illustrating the variety of parameters one must consider. Our own designs are presented as one among many potentially productive paths toward addressing these challenges.

2.7 Acknowledgements

We would like to thank the many teachers, administrators, and research assistants who have had a hand in shaping, refining and contributing to the course of the learning environment designs we have presented here. This material is based upon work supported by the National Science Foundation under Grant No. 1008634. Any opinions, findings,

and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

2.8 References

- Allers, K., Hütter, G., Hofmann, J., Loddenkemper, C., Rieger, K., Thiel, E., & Schneider, T. (2011). Evidence for the cure of HIV infection by CCR5 Δ 32/ Δ 32 stem cell transplantation. *Blood*, 117(10), 2791-2799.
- Berg, P., & Singer, M. (1998). Inspired choices. *Science*, 282(5390), 873-874.
- Bromme, R., Kienhues, D., & Porsch, T. (2010). Who knows what and who can we believe? Epistemological beliefs are beliefs about knowledge (mostly) to be attained from others. In L. D. Bendixen & F. C. Feucht (Eds.), *Personal epistemology in the classroom: Theory, research, and implications for practice* (pp. 163-194). Cambridge University Press.
- Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5(2), 121-152.
- Chinn, C. A., Duschl, R. A., Duncan, R. G., Buckland, L. A., & Pluta, W. J. (2008, June). A microgenetic classroom study of learning to reason scientifically through modeling and argumentation. In ICLS'08: Proceedings of the 8th International Conference for the Learning Sciences, (Vol. 3, pp. 14-15). International Society of the Learning Sciences.
- Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education*, 86(2), 175-218.
- Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K-8*. National Academies Press.
- Edelson, D. C. (2002). Design research: What we learn when we engage in design. *The Journal of the Learning Sciences*, 11(1), 105-121.
- Engle, R. A., & Conant, F. R. (2002). Guiding principles for fostering productive disciplinary engagement: Explaining an emergent argument in a community of learners classroom. *Cognition and Instruction*, 20(4), 399-483.
- Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education*, 88(6), 915-933.
- Falk, A., & Brodsky, L. (2014). Scientific explanations and arguments: Accessible experiences through exploratory arguments. *Science Scope*, 37(5), 4-9.
- Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., ... & Strijbos, J. W. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research*, 2(3), 28-45.
- Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science*, 71(5), 742-752.
- Grandy, R., & Duschl, R. A. (2007). Reconsidering the character and role of inquiry in school science: Analysis of a conference. *Science & Education*, 16(2), 141-166.
- Hidi, S., & Baird, W. (1986). Interestingness—A neglected variable in discourse processing. *Cognitive Science*, 10(2), 179-194.

- Kanter, D. E. and Konstantopoulos, S. (2010). The impact of a project-based science curriculum on minority student achievement, attitudes, and careers: The effects of teacher content and pedagogical content knowledge and inquiry-based practices. *Science Education*, 94(5), 855-887. doi: 10.1002/sce.20391
- King, A. (1992). Facilitating elaborative learning through guided student-generated questioning. *Educational Psychologist*, 27(1), 111-126.
- Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions*. New York: Oxford University Press.
- Krajcik, J., Blumenfeld, P. C., Marx, R. W., Bass, K. M., Fredricks, J., & Soloway, E. (1998). Inquiry in project-based science classrooms: Initial attempts by middle school students. *Journal of the Learning Sciences*, 7(3-4), 313-350.
- Lombardi, D., Sibley, B., & Carroll, K. (2013). What's the alternative? Using model-evidence link diagrams to weigh alternative models in argumentation. *The Science Teacher*, 80(5), 50-55.
- Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in Philosophy and Education*, 27(4), 283-297.
- National Research Council (NRC). (2012). *A framework for k-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.
- Nehm, R. H., & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching*, 48(3), 237-256.
- Next Generation Science Standards (NGSS) Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.
- Paxton, W. A., Martin, S. R., Tse, D., O'Brien, T. R., Skurnick, J., VanDevanter, N. L., ... & Koup, R. A. (1996). Relative resistance to HIV-1 infection of CD4 lymphocytes from persons who remain uninfected despite multiple high-risk sexual exposures. *Nature Medicine*, 2(4), 412-417.
- Pfundt, H., & Duit, R. (1988). Students; alternative frameworks and science education bibliography. Retrieved from ERIC database (ED315266).
- Phillips, L. M., & Norris, S. P. (2009). Bridging the gap between the language of science and the language of school science through the use of adapted primary literature. *Research in Science Education*, 39(3), 313-319.
- Pitts, V. M., & Edelson, D. C. (2004, June). Role, goal, and activity: A framework for characterizing participation and engagement in project-based learning environments. In Proceedings of the 6th International Conference on Learning Sciences (pp. 420-426). International Society of the Learning Sciences.
- Pitts, V. M., & Edelson, D. C. (2006, June). The role-goal-activity framework revisited: Examining student buy-in in a project-based learning environment. In Proceedings of the 7th International Conference on Learning Sciences (pp. 544-549). International Society of the Learning Sciences.
- Pluta, W. J., Chinn, C. A., & Duncan, R. G. (2011). Learners' epistemic criteria for good scientific models. *Journal of Research in Science Teaching*, 48(5), 486-511.
- Private Universe Project. (1995). *The private universe teacher workshop series [DVD]*. South Burlington, VT: The Annenberg/CPB Math and Science Collection.

- Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., ... & Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. *The Journal of the Learning Sciences*, 13(3), 337-386.
- Reiser, B. J., Tabak, I., Sandoval, W. A., Smith, B. K., Steinmuller, F., & Leone, A. J. (2001). BGuILE: Strategic and conceptual scaffolds for scientific inquiry in biology classrooms. In M. S. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 263-305). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Rinehart, R. W., Duncan, R. G., & Chinn, C. A. (2014). A scaffolding suite to support evidence-based modeling and argumentation. *Science Scope*, 38(4), 70-77.
- Suthers, D. D., & Hundhausen, C. D. (2003). An experimental study of the effects of representational guidance on collaborative learning processes. *The Journal of the Learning Sciences*, 12(2), 183-218.
- Thagard, P. (2000). *How scientists explain disease*. Princeton University Press.
- Toth, E. E., Suthers, D. D., & Lesgold, A. M. (2002). "Mapping to know": The effects of representational guidance and reflective assessment on scientific inquiry. *Science Education* 86(2), 264-286. doi: 10.1002/ sce.10004
- Windschitl, M., Thompson, J., & Braaten, M. (2008a). Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. *Science Education*, 92(5), 941-967.
- Windschitl, M., Thompson, J., & Braaten, M. (2008b). How novice science teachers appropriate epistemic discourses around model- based inquiry for use in classrooms. *Cognition and Instruction*, 26(3), 310-378.
- Yarden, A. (2009). Reading scientific texts: Adapting primary literature for promoting scientific literacy. *Research in Science Education*, 39(3), 307-311.

**Chapter 3: The Body of Evidence: The Role of Epistemic Cognition and Evidence
Evaluation in Science Classes**

Abstract

Promoting students' use of sophisticated epistemic practices has become a central feature of classroom-based interventions designed to scaffold students' reasoning about scientific phenomena in model-based inquiry environments. Prior research has focused largely on structural elements of the argumentative frame or has examined how students use pieces of evidence in isolation. My research examined students' use, evaluation, and re-evaluation of evidence over time with exposure to new evidence. The intervention was designed to scaffold middle school (ages 12-13) science students' reasoning about evidence, argumentative practices, and epistemic cognition. I present the results of a three-day model-based inquiry lesson in which students investigated the possibility that some humans might be genetically resistant to Human Immunodeficiency Virus (HIV). Written work from students ($N = 88$) was coded for evidence evaluation based on students' implicit criteria, evidence-to-model coordination, and model selections. Students significantly shifted their evaluations (re-evaluated) pieces of evidence over time as they encountered more evidence. Moreover, students engaged in the development of supra-evidence structures, a body of evidence, combining two or more pieces of evidence into a coherent whole that motivated their beliefs about different models. Developing a body of evidence led to increases in argument complexity. Existing frameworks for evaluating student reasoning do not include (a) evidence re-evaluation and (b) combining pieces of evidence to construct a new body of evidence. I argue that normative accounts of good reasoning in science classes could be improved by taking both of these practices into account. Further, instruction designed to promote sophisticated practices for evaluating and handling evidence can build on students' latent reasoning capacities.

3.1 Introduction

Promoting students' use of sophisticated epistemic practices has become a central feature of classroom-based interventions designed to scaffold students' reasoning about scientific phenomena in model-based inquiry environments (Windschitl, Thompson, & Braaten, 2008a). A broad program of research by the science education community has embraced greater inclusion of the epistemic practices of science (e.g., argumentation, communicating findings, and so on) in the classroom (Driver, Newton, & Osborne, 2000; Duschl, 1990, 2008). The trend in recent research has been to develop interventions to support students' use of scientific epistemic practices such as evaluating evidence and using evidence to support or refute particular claims (Fischer et al., 2014) and revise explanatory models (Windschitl, 2008). Prior research on students' use of evidence has tended to focus on the structural elements of the argumentative frame that students use to motivate their claims in science (Berland & McNeill, 2010; Chinn & Brewer, 2001; Garcia-Mila, Gilabert, Erduran, & Felton, 2013; Linn, Clark, & Slotta, 2003; Osborne, Erduran, & Simon, 2004; Schwarz, Neuman, Gil, & Ilya, 2003) or has examined how students use pieces of evidence in isolation (Chinn & Brewer, 2001). However, we know less about how students reason about evidence in light of other evidence. This paper provides a fine-grained account of how students evaluate, re-evaluate, and make use of evidence during model-based inquiry activities (Windschitl et al., 2008a) in science class while investigating an authentic life and health science topic, the possibility that humans can be genetically resistant to the Human Immunodeficiency Virus (HIV).

The purposes of this paper are fivefold. First, I offer a detailed account regarding students' epistemic practices for grappling with multiple, sometimes conflicting, pieces of evidence in a model-based inquiry environment. Second, I describe how students'

reasoning about evidence changes in light of new evidence. Reasoning about evidence has ties to both epistemic cognition and domain specific knowledge. Epistemic cognition, as conceptualized in the AIR model, is taken to be the suite of cognitions that are used to guide a person's aims and values for developing knowledge, processes used to achieve those aims, and the ideals used to evaluate the merit of one's knowledge generating practices (Chinn, Rinehart, & Buckland, 2014; Chinn & Rinehart, 2016). Evaluating evidence is an epistemic process; this study examines the implicit criteria that students use to evaluate evidence. Third, I present an analysis of how students construct an integrated *body of evidence*, multiple pieces of evidence conceptually linked together to support a claim, and how constructing *bodies of evidence* influences student reasoning and impacts some forms of argument complexity. Fourth, I argue that the kind of science learning environment described here could represent a productive synthesis of document-based learning techniques that are appropriate when firsthand data collection and analysis are not a possibility in the classroom. Finally, I argue that productive modifications can be made to existing learning progressions for argumentation by including insights from research on how people reason about multiple text documents while trying to resolve conflicting claims in a model-based inquiry environment.

Effective science instruction and curriculum design can provide opportunities for learners to engage in authentic inquiry (Minner, Levy, & Century, 2010). Model-based inquiry has as its goal for students to develop “defensible explanations of the way the natural world works” by generating and revising scientific models (Windschitl et al., 2008a, p. 2). This has been described by Windschitl (2008) as being composed of four conversations about: (a) organizing what we know and would like to know; (b) generating

hypotheses from models; (c) seeking evidence to test those hypotheses; (d) constructing arguments. Common practices in support of these core principles of scientific inquiry include reading background research and defining variables that are to be measured, recorded and interpreted with the aim of making sense of, constructing and revising scientific models. Model-based inquiry is a welcome alternative to typical school accounts of The Scientific Method (TSM). It is more strongly grounded in the epistemic practices of authentic science (Windschitl et al., 2008a).

The practices of inquiry in the science classroom could be productively expanded in ways that might better meet the needs of students who are faced with making decisions about problems where firsthand data collection in the science classroom is not a possibility (Palincsar & Magnusson, 2001). There are occasions in science class when it can be productive to elevate the role of text-based evidence in ways that still support the inquiry practices of students. In cases where there are multiple pieces of conflicting evidence, students' practices around evidence evaluation take on increased importance (Wiley, Goldman, Graesser, Sanchez, Ash & Hemmerich, 2009). The ways students in a model-based inquiry classroom make sense of multiple and conflicting pieces of evidence are underexplored (Britt, Richter, & Rouet, 2014). In the research presented in this paper I will describe not only how evidence evaluation impacts students' reasoning and written argumentation and will consider the role of re-evaluation of evidence in light of new evidence. Moreover, I will address a gap in the literature about the implicit evidence criteria students use to evaluate evidence and how notions about criteria and evidence re-evaluation can be integrated into learning progressions for argumentation.

3.1.1 Evaluating and Re-evaluating Evidence

Prior research has shown that students struggle with evidence evaluation tasks. Phillips and Norris (1999) had 91 high school seniors read multiple texts from popular science magazines. They found that many students simply deferred to the authorities in the text. They wrote that it was rare for students to “challenge the authority of the reports or authors” (Phillips & Norris, 1999, p. 325). Wiley and colleagues (2009) had undergraduate students evaluate a variety of web-based pieces of evidence about the causes of volcanic eruptions (Wiley et al., 2009). Students were put into two conditions; one tasked with writing an argument and another condition for writing a descriptive essay. The authors found that “the argument writing task did not improve the ability to discriminate between reliable and unreliable sources” (p. 1084) and that students infrequently justified their evaluation of the sources of evidence (Wiley et al., 2009).

Scientists and medical experts often re-evaluate evidence in light of new evidence, revising beliefs about evidence previously seen. For example, Tenopir and colleagues (2005) found that astronomers read on average 228 articles per year, one-quarter of which were rereadings of articles previously read (Tenopir, King, Boyce, Grayson, & Paulson, 2005). In his work on how scientists reason about and explain disease, Thagard (2000) found that part of the development of the bacterial theory of ulcers relied on deeper reconsideration of evidence that had previously been dismissed (wrongly) by panels of scientists and medical experts. In short, evidence re-evaluation plays an important role in the reasoning practices of scientists and medical experts. Students evaluating multiple pieces of evidence over time might afford them the opportunity to develop and calibrate a sense for what counts as good evidence. Re-

evaluating older evidence in light of new evidence provides the opportunities for this calibration to occur.

3.1.2 Bodies of Evidence

Scientists reason not only about pieces of evidence in isolation from one another but also about evidence in the context of other evidence (Sober, 2008). In their account of the differences between various forms of creationism and evolutionary theory Chinn and Buckland (2011) noted that multiple converging lines of evidence are an important source of conviction about the correctness of evolutionary theory, and that it would be irrational to engage in wholesale belief abandonment when minor bits of anomalous data are found. Ault (1998) argued that using multiple lines of independent converging evidence is one attribute of excellent reasoning. Ault's principles were used by Kelly and Takao (2002) in their evaluation of undergraduate students' arguments about oceanography topics. They found that some students were able to productively impose constraints on the range of possible interpretations of a phenomenon by using multiple converging lines of evidence to reduce ambiguity. Reducing the range of possible explanations is part of productive science.

Converging lines of evidence can be thought of as independent investigations that tend to support (or rule out) the same theory. Converging lines of evidence privilege evidence-to-model coordination. Although related, a body of evidence is different. A body of evidence typically includes converging lines of evidence, but there are evidence-to-evidence links that become important as well. Evidence-to-evidence links have received little attention compared with evidence-to-model coordination.

To develop a body of evidence, students need to see at least one conceptual link between two pieces of evidence. For example, if students think that two studies share a similar investigative technique, such as sampling blood for the presence of a virus, they might integrate the two pieces of evidence in their own internalized conceptual model. This integrated conceptual structure can be thought of as a body of evidence, two or more pieces of evidence whose conceptual links create a unified body of evidence, whose weight is considered greater than the individual pieces that make it up. The body of evidence can then be used to make informed judgments about what to believe.

The AIR model of epistemic cognition (Chinn, Rinehart, & Buckland, 2014; Chinn & Rinehart, 2016) posits that people bring Aims and values, Ideals, and Reliable processes to bear on epistemic goals like knowledge production, truth attainment and avoiding false beliefs. Developing a body of evidence and using it as a guide can, within this framework, be thought of as one process (among many potential processes) for attaining an epistemic aim like selecting a theory to believe or refining a scientific model. Similarly, the sub-processes of developing a body of evidence, evaluating evidence and seeing connections between pieces of evidence, are epistemic processes as well.

Chinn and Buckland (2011) argued that bodies of evidence constrain the range of theories to be considered, and that one function of a body of evidence can be to rule out some theories, as in the case of young earth creationism and intelligent design. A body of evidence is something that scientists create and use; it is an open question if middle-school students in life science can productively engage with this practice.

Examining how students develop a body of evidence, and the impact it has on the structure of their arguments, represents an alternative approach to analyzing

argumentation. Examining student-constructed bodies of evidence necessarily involves a detailed focus on the role of evidence and evidence-to-model relations. Different approaches to examining argumentation have emphasized certain elements of the Toulmin Argument Pattern over other elements (Sampson, 2008). For example, Garcia et al. (2013) examined the complexity of students' verbal arguments with a typology of eleven kinds of arguments, which were various combinations of claims, data, warrants, backing, and rebuttals. For Garcia et al. (2013) a structurally complex argument included the presence of these five major Toulmin elements. Schwarz and colleagues (2003) examined 5th grade students' verbal argumentation about the topic of permitting or forbidding animal experimentation. Their focus also included an emphasis on structure, where the least sophisticated argument contained a claim with no other elements and the most sophisticated form was a "two-sided" argument in which students had considered pros and cons for their position and articulated rebuttals. The present study also attempts to examine argument structure. Since bodies of evidence has not been examined before it is open question about the impact they have on student argumentation.

3.1.3 Criteria for Evidence

In addition to epistemic processes, ideals can also play an important role in model-based inquiry (Chinn & Rinehart, 2016). Ideals are taken to be criteria that can be used to evaluate epistemic aims, processes, and products. Ideals are manifold and can apply to many elements of the scientific endeavor including theory evaluation, methods, results reporting and even what counts as good (or bad) evidence (Chinn & Rinehart, 2016). Pluta and colleagues (2011) found that middle-school science students could, at the class level and without significant prior training, produce a list of criteria for what

counts as a good scientific model that was commensurate with what philosophers and historians of science have found in their work. This paper is in part concerned with the evaluative criteria that students bring to bear on evidence as they engage in coordinating evidence to generate an argument in light of competing claims.

The coordination of multiple pieces of evidence to generate a valid scientific argument is a complex task (Kuhn, 1992). Doing this well relies in part on successfully evaluating the quality of evidence to sift the good from the bad. Toulmin (1958) described this as the “...field dependence of our standards,” (p. 33) which I take to be epistemic ideals, or criteria, used to evaluate evidence. The field dependent features of Toulmin’s framework have not been well addressed by argumentation researchers in favor of the more field-invariant features like claims, data, and so on (Sampson & Clark, 2008). Given the situated nature of cognition, it is likely the case that evidence criteria are sensitive to situational factors (Chinn, Buckland, & Samarapungavan, 2011) and have ties to disciplinary knowledge (Elby & Hammer, 2001). This paper is one attempt at exploring students’ implicit criteria for evidence quality related to human genetic resistance to HIV in a model-based inquiry environment.

3.1.4 Insiders, Outsiders and the STEM Pipeline

One of the many aims of school science is to provide students with the knowledge, skills and experience to make sense of and use science in everyday life (Feinstein, Allen, & Jenkins, 2013). Feinstein (2011; et al., 2013) argued that science classes often contain (at least) two kinds of students; those who are inside the pipeline and those who are outside the pipeline. Students in the pipeline aim for a career in the

sciences. Students outside the pipeline are those for whom a career in science is not an aim.

Feinstein et al. further argued that science education should aim to develop competent outsiders— “Nonscientists who can access and make sense of science relevant to their lives” (Feinstein et al., 2013, p. 314). He further clarified that “They remain anchored outside of science, reaching in for bits and pieces that enrich their understanding of their own lives” (Feinstein, 2011, p. 180). Much of the reasoning science outsiders engage in revolves around consulting a variety of sources of evidence (e.g., print media, videos, internet sources, local experts and so on) to arrive at informed judgments about what to believe (e.g., is climate change is occurring?) or how to act (e.g., what is the best course of treatment for a medical condition?) (Feinstein et al., 2013). For the layperson, skills that revolve around assessing the reliability of sources and integrating knowledge from a variety of sources are central to everyday knowledge and decision making (Bromme, Kienhues, & Porsch, 2010).

Training within a STEM pipeline may fail to equip students for productive reasoning from the standpoint of the layperson. Keselman and colleagues (2015) explored the relationship between training in biology and beliefs about common health misconception statements. They found that upper division biology students were “unequivocally better than non-science majors on only one statement” (Keselman, Hundal, Chentsova-Dutton, Bibi, & Edelman, 2015, p. 174). Although biology majors were more likely to employ systems- and cell-level thinking, this had a minimal impact on their responses to the health misconceptions. This is some evidence that biology majors function (mostly) as laypeople about health topics.

Moreover, even scientists are laypeople in domains in which they are not experts (see Bromme et al., 2010; Bromme, Thomm, & Wolf, 2015; Thomm & Bromme, 2012, for a more in-depth treatment). Bromme and colleagues have suggested that a division of cognitive labor permeates society. They write that “Most knowledge claims are based on specialized knowledge provided by specialized experts and the knowledge is organized into disciplines, reflecting such specialization” (Bromme et al., 2010, p. 167). The result is that we are all laypeople in any domain, field, or topic in which we are not experts, and as such rely on the testimony of others for our knowledge. What this means is that training in the use of lay-reasoning practices (i.e., coordinating multiple documents, sourcing, evidence evaluation, evidence integration and so on) could be valuable for those who are in a STEM pipeline as well.

Testimony, learning from the words and actions of others, is how we learn about most of the world (Lackey, 2008; 2011; Origgi, 2012). It would be impossible to engage in firsthand verification of the totality of knowledge a person possesses. Instead, knowledge production is characterized and driven by the division of cognitive labor communicated through testimony. Even communities of experts exchange vast amounts of information based on testimony. Hardwig (1985, 1991) investigated the inner epistemological workings of the group of physicists responsible for the discovery of the charm quark. He found that sub-communities within this group of physicists reliably exchange information through testimonial channels. Recent work on how communities of mathematicians operate reveals similar results. Weber, Inglis, & Mejia-Ramos (2014) showed that mathematicians rarely fully interrogate the proof of a mathematical concept, as it is too time consuming, writing “We argue that mathematicians frequently believe

mathematical assertions are true on the testimony of others and that the perceived authority of the mathematician advancing a claim influences which testimony mathematicians choose to accept" (Weber et al., 2014, p. 43).

Savvy navigation of the webs and chains of testimony requires a suite of skills that are often not well represented in typical science courses, which often make heavy use of lectures, textbooks written in an expository rather than argumentative style (Yarden, 2009) and cookbook labs (Windschitl et al., 2008a), to the detriment of those inside and outside the pipeline. Students inside the pipeline would benefit from a better understanding of how information flows within a professional community of scientists. Students both inside and outside the pipeline could benefit from learning about how to reason about a patchwork of evidence from a variety of sources (Britt, Richter, & Rouet; 2014).

3.1.5 The Role of Multiple Documents Coordination in Model-Based Inquiry Environments

In model-based inquiry, reading is conceived of as a “supporting activity” (Windschitl et al., 2008a). However, reading evidence, and reasoning about that evidence, is a central focus of both scientific (Phillips & Norris, 2009; Tenopir et al., 2004; Tenopir et al., 2005, Tenopir, King, Edwards, & Wu, 2009) and lay practice (Britt et al., 2014; Bromme & Goldman, 2014). Scientists, medical professionals, and engineers spend a significant amount of their work time evaluating and re-evaluating secondhand evidence (Tenopir et al., 2005; 2009), and the most productive scientists spend the most time reading evidence (Tenopir et al., 2004). Given that evaluating secondhand evidence is an authentic inquiry practice extensively used by scientists, and that it figures largely in lay

information-seeking about science topics as well, it deserves greater attention in the science classroom.

Insights from multiple documents research, which explores how people process information from disparate sources, may be useful here. Prior research in both the domains of history and science has shown that students rarely spontaneously use source information (e.g., author, date, publication type, and so on) when reading from multiple documents (Britt & Angliskas, 2002; Wiley et al., 2009; Wineburg, 1991). Studies that examine source recall and corroboration are numerous and show a pattern that in general, student recall of source information and corroboration is sparse in unscaffolded learning environments (Britt & Angliskas, 2002, Chinn & Rinehart, 2016; Wiley et al., 2009). However, Wiley et al. (2009) showed that scaffolding could increase students' use of evidence and prompt them to pay attention to evidence quality. Cerdan & Vidal-Abarca (2008) showed that making sense of multiple documents and writing about them could lead to enhanced understanding of causal processes in biology. Argumentative essays in particular can promote deep processing of science texts when combined with evidence evaluation (Anmarkrud, Braten, & Stromso, 2014). These studies, however, did not encourage students to re-evaluate evidence and did not elaborate on the implicit criteria that students use to evaluate evidence beyond what counts as a good source and how students conceptually link together individual pieces of evidence to develop a body of evidence. At present, several important connections between multiple documents processing in science and scientific argumentation are under-described. I investigated these previously underexplored areas at the nexus of science education and the use of multiple documents in inquiry environments.

Adding a multiple documents component to science instruction takes into account the three major considerations mentioned previously: (a) most of what we know comes from testimony; (b) the division of cognitive labor permeates society; and (c) the lived experience of most students, who eventually become non-scientist adults, exists outside the science pipeline. Science lessons that make use of multiple documents can benefit students within the pipeline as well. Part of a scientist's work is the evaluation, and re-evaluation, of secondhand evidence. Coordinating firsthand evidence with a patchwork of secondhand evidence could be productive for students (Hapgood, Magnusson, & Palincsar 2004).

3.1.6 The Present Study

The Promoting Reasoning And Conceptual Change In Science (PRACCIS) project represents an instructional approach that fosters conceptual learning while helping students develop sophisticated epistemic practices in middle grades (age 12-13) life science classes. Students engaged with PRACCIS materials for about five months of their school year. The PRACCIS project included the development of numerous instructional units on a variety of biology topics such as cell organelles, genetics, and evolution. Units included a suite of instructional scaffolds to promote students' engagement with the practices and epistemology of science in a model-based inquiry environment (Rinehart, Duncan, & Chinn, 2014). These scaffolds included public criteria lists, evidence rating scales, and the MEL (Model Evidence Link) Matrix, which will be described in more detail later (Rinehart et al., 2014; Rinehart, Duncan, Chinn, Atkins, & DiBenedetti, 2016). PRACCIS lessons, and teacher professional development materials, were designed

to further science education for students in the pipeline and for those students who are presently outside the pipeline.

The lesson described in this study occurred near the end of a three-week unit on genetics and inheritance. During the genetics unit, students developed and revised their own models of inheritance in light of evidence, learned standard genetics and inheritance terminology (e.g., heterozygous, homozygous, etc.), completed several lab activities related to chromosome pairs, and learned basic monohybrid Punnett squares. For the end of the unit I developed two lessons for students to investigate genetic resistance to HIV. In the first HIV lesson, which was the basis for this study, students considered whether or not genetic HIV resistance could exist. In the second lesson, students used evidence to choose between two competing models of the mechanism of genetic resistance to HIV.

3.1.6.1 Timeline and Noteworthy Activities.

A description of the important activities that made up the first HIV lesson, and the sequence in which students experienced them, is presented in Table 3.1. Briefly, students began Day 1 of the lesson with a review of “How to make good scientific arguments.” While students had engaged in written argumentation in previous PRACCIS lessons, the teachers in the study expressed a desire to have more focused attention on written student argumentation and thought that a review of some of the key elements of a written argument would be useful. The PRACCIS research team agreed that such a review could prove useful and developed a short (approximately 25 minute) review of scientific argumentation. The lesson provided students with four criteria for a good argument (see Table 3.1 for details), and then students, as part of a class discussion, critiqued eight arguments on a variety of topics. Students then engaged in an individual written analysis

of an argument about a topic (i.e., the role of the nucleus in the cell) that they had previously engaged with as a model-based inquiry lesson. Next teachers presented a very brief PowerPoint composed of seven slides that introduced the HIV topic. Students learned the distinction between HIV and AIDS and how the disease is transmitted. The presentation ended with a problem: scientists and doctors had heard that a small number of people seemed to be genetically resistant to HIV, and the students were tasked with figuring out whether this was true. Day 1 concluded with students selecting their initial position on the topic, whether genetic resistance to HIV does or does not exist.



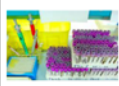
On Day 2 students engaged with Evidence 1 and Evidence 2, evaluated the quality of the evidence, completed a MEL matrix, selected which model they thought was best, and wrote an argument in favor of their chosen model. Evidence 1 was a short video clip in which Dr. Steven O'Brien, a geneticist, was interviewed about his research on cats and Feline Immunodeficiency Virus (FIV). Students were also provided with a short (1 page) text that summarized the key points of the video. The finding from this research was that wild cats tend to be resistant to the FIV virus and do not develop AIDS, unlike house cats, which are not resistant and do develop AIDS. Evidence 2 was a written interview with medical professionals at a health clinic treating HIV patients. In short, the clinic staff asserted that HIV can be contracted by anyone, and that they had never had a patient who was resistant to HIV. Students worked in pairs as they evaluated evidence and discussed their evidence-to-model coordination. Students periodically completed individual questions in their own written work packet, typically after a pair discussion. At the end of Evidence 1 and 2 students wrote a response to the prompts described in Table 3.1. To conclude Day 2 of the lesson, students completed a MEL Matrix.

Table 3.1			
<i>A brief summary of the three day HIV lesson</i>			
<u>Activity</u>	<u>Sequence</u>	<u>Summary</u>	<u>Noteworthy Details</u>
How to make good scientific arguments	Day 1	Students examined 8 arguments and evaluated them based on 4 criteria. The key point was that they are open to critique for knowledge building, they tend to be more understandable, and are more persuasive.	The 4 criteria for good arguments were: It tells what your position is. It tells what the evidence for your position is. It is accurate. It explains why the evidence supports your position.
Evaluating arguments	Day 1	Students evaluated a written argument drawn from a prior model-based inquiry lesson about the nucleus in a cell.	The argument had multiple pieces of evidence in support of the claim that the nucleus gives instructions for proteins. Students listed 4 good points and 1 bad point for the essay.
What is HIV	Day 1	The teacher presented a short 7 slide PowerPoint that contained basic declarative information about HIV/AIDS, how prevalent it is, and how people contract the virus.	The PowerPoint problematized HIV by introducing the idea that there were rumors and anecdotes that some humans might be resistant to HIV, and that scientists were interested in finding out if this was true or not.
Initial model selection*	Day 1	Students chose an initial model after seeing the PowerPoint, but before seeing any evidence.	Model 1: Genetic resistance to HIV does not exist Model 2: Genetic resistance to HIV does exist.
Evidence 1 Cats and FIV*	Day 2	A 3 minute video and summary text about research on Feline Immunodeficiency Virus (FIV). The conclusion was that some cat species are resistant to FIV and some are not.	After seeing Evidence 1 students responded to this prompt: “Geeta and Jose are arguing about this evidence. Circle the one you agree with the most. A. Geeta thinks cats are mammals like humans and research on cats is useful for understanding HIV. B. Jose thinks cats are different from humans and research on cats is not useful for understanding HIV. C. I don’t agree with either of them. Explain your choice for your answer.”

<u>Activity</u>	<u>Sequence</u>	<u>Summary</u>	<u>Noteworthy Details</u>
Evidence 2 Health clinic interview*	Day 2	A 1 page description of several medical professionals' experience with treating HIV patients. Two of them asserted that anyone can get HIV if exposed.	Students responded to the following prompt: "How do you rate the quality of this piece of evidence (0, 1, or 2)? Give reasons for your rating." <i>Author's note: 0 is bad evidence, 1 is o.k. evidence, and 2 is good evidence.</i>
Intermediate model selection*	Day 2	Students completed a MEL Matrix, selected the model they thought was best, and wrote an argument in support of their model based on Evidence 1 and 2.	Students responded to the following prompt: "Which model do you think is best and why? Be sure to give reasons for your answer."
Evidence 3 Monkeys and SIV	Day 3	A 1 page description of a breeding experiment with monkeys. Some monkeys were resistant to Simian Immunodeficiency Virus (SIV) and some were not.	Students responded to the following prompt: "Is SIV resistance in monkeys genetic? Circle your answer. A. No it is not genetic. B. Yes it is genetic and resistance is a dominant trait. C. Yes it is genetic and resistance is a recessive trait. Explain why it is or is not genetic based on the results of this study. Give reasons for your answer."
Evidence 4 Dr. Paxton and HIV	Day 3	A 1 page description of an experiment on the white blood cells of 25 humans who had been repeatedly exposed to HIV and were still HIV negative.	Students responded to the following prompt: "What conclusion do you draw from this study? Explain your answer."
Reflection on evidence	Day 3	Students reflected on the utility of all four pieces of evidence.	Students responded to the following prompt: "Which evidence is most useful for helping you decide between the models? Explain why."
Final model selection*	Day 3	Students completed a MEL Matrix, selected the model they thought was best, and wrote an argument based on all four pieces of evidence.	Students responded to the following prompt: "Write an argument to support your model. Write to someone who may not agree with you. Give detailed reasons for your answer."

* Indicates items that were coded and analyzed

The MEL Matrix is a scaffold that has been adapted by the members of PRACCIS from work on graphical organizers (Suthers & Hundhausen, 2003; Toth, Suthers, & Lesgold, 2002) to help students keep track of their thoughts about evidence quality and evidence-to-model relations as shown in Figure 3.1. The MEL Matrix contains several important elements, including (a) a place for students to rate evidence quality; and (b) arrow boxes where students systematically connect each piece of evidence (shown in the rows) with each model (shown in the columns). For the evidence quality rating students used a three point scale, where 0 represented bad evidence, 1 represented evidence with both good and bad qualities, and 2 represented good evidence. For the arrow boxes, students could choose from five different evidence-to-model connections including (a)

Evidence Goodness Rating	Model 1 Genetic resistance to HIV does <u>not</u> exist.	Model 2 Genetic resistance to HIV in does exist.
1. FIV Video  <input type="checkbox"/>		
2. Greater Area Health Clinic: <u>Interview Report</u> <input type="checkbox"/>		
3. SIV Study  <input type="checkbox"/>		
4. Paxton Study  <input type="checkbox"/>		

12. For all the pieces of evidence make sure to rate them (0, 1, or 2) and draw an arrow for how the evidence relates to each model.

13. Which model is better? Circle your selection.

Model 1: Genetic resistance to HIV does not exist.

Model 2: Genetic resistance to HIV does exist.

Support	→
Strongly Support	⇒
Contradict	↯
Strongly Contradict	⇸
Irrelevant	→

Figure 3.1. A MEL Matrix showing evidence rating boxes next to the name of the evidence within each row, the arrow types (e.g., support, contradict etc.), and the model selection boxes (e.g., Model 1: Genetic Resistance...)

strongly support; (b) support; (c) irrelevant; (d) contradict; and (e) strongly contradict. Below the arrow diagram are model selections boxes where students indicated which model they thought was best. After completing the MEL Matrix and model selection, students wrote an argument in favor of the model they thought was best supported by the evidence. After evaluating Evidence 3 and 4 students wrote a response to the prompts described in Table 3.1. Students then completed a second MEL Matrix (the first was completed on Day 2), selected a model, and wrote a brief essay about which piece of evidence was most useful in helping them pick the model they thought was best. The lesson concluded with students writing an evidence-based argument in support of the model they thought was best. Students were instructed to write their essay to someone who disagreed with them.

3.1.6.2 Lesson Design Considerations.

The lesson was designed to scaffold students' epistemic cognition and conceptual learning in life science through attention to their: (a) reasoning about evidence by evaluating and re-evaluating it in light of new evidence; (b) reasoning with evidence by coordinating with other evidence and coordinating evidence with models; and (c) written arguments making connections between evidence and models. Students engaged with scaffolds while facing competing claims and conflicting evidence of variable quality.

An exclusive focus on expository texts in science classrooms might be responsible for a pattern of results across many studies showing that science students often cannot discern the differences between claims, reasons, and evidence (Goldman & Bisanz, 2002). Moreover, expository texts likely play a very small role in laypersons' reasoning about scientific issues (Phillips & Norris, 2009; Yarden, 2009). This has led to

numerous calls for the inclusion of a greater variety of text types to represent a more authentic range of what students are likely to encounter outside the science classroom (Britt et al., 2014; Goldman & Bisanz, 2002; Phillips & Norris, 2009; Yarden, 2009).

The design of the evidence for this lesson was a critical element of the investigation. The topic, HIV resistance in humans, did not lend itself well to authentic hands-on exploration in science classrooms. Therefore, I designed the evidence to be commensurate with research on how to improve the quality of texts, in this case written pieces of evidence, in the science classroom. To be clear, this was not an investigation of reading, but rather an inquiry into how students reason about text-based evidence. The lesson described here used two pieces of evidence that were Adapted Primary Literature (APL) and two pieces of evidence that was a Journalistic Reported Version (JRV). APL texts are derived from empirical scientific reports and still contain some of the organizational features of science texts, and science students are their intended audience (Yarden, 2009). Rather than expository texts that invite little critique, APL style articles generally contain information about the aims, methods and results of research whose quality can be evaluated by students. JRV texts do not include these organizational features and are aimed at a more general audience (Yarden, 2009). The JRV pieces of evidence were included to provide some contrast with the APL articles. Moreover, JRV style evidence is what students will more commonly encounter outside of school. Using a blend of both JRV (Evidence 1 and 2) and APL (Evidence 3 and 4) texts was an attempt to meet the call for more diverse texts in science classrooms that can promote critical stances toward evidence and models and drive the need for evidence evaluation, evidence-to-model coordination, and argumentation in ways that expository texts do not.

Evidence 1 (cats/FIV) and 2 (health clinic interview) were designed with several considerations in mind (see Table 3.1 for a description). First, they were developed to have low diagnosticity, meaning that they did not strongly support, or strongly rule out, competing models. The reason for this was to maintain an ongoing sense of investigation and inquiry. Had the most diagnostic evidence come first, the need for continued inquiry would have diminished. Second, each piece of evidence was intended to support one particular model, so that students would have some evidence to cite for the evidence-based argument that they wrote at the end of Day 2. Finally, I used evidence that was low in diagnosticity to find out if students would shift their model selections in response to new (and more diagnostic) evidence. In short, would students be change their model selections based on evidence.

The prompts for both pieces of evidence (see Table 3.1 for details) were designed to generate discussion about the quality of the evidence. For Evidence 1, students were asked if they agreed with Geeta or Jose, who had opposing views on the utility of research on cats and FIV for resolving issues related to HIV resistance in humans. The prompt was designed to get students discussing the utility of experimental organisms and the relatedness of the various immunodeficiency viruses in non-human animals. This discussion was authentic to the field of HIV research, given that HIV is a strain of SIV that infects humans, and provided an opportunity for students to engage with discussions about domain specific problems in biology. The prompt for Evidence 2 was a simple open ended prompt that asked students to rate the evidence on a 3-point scale (bad, o.k. and good) and then write their reasons for their evaluation. The intent was to keep students focused on evaluating evidence quality during Day 2 in order to elicit their implicit

criteria for evidence quality. Additionally, it provided an initial evaluation against which I could investigate students' later re-evaluation of this same evidence. Students evaluated Evidence 1 and 2 on three occasions (a) first after seeing each piece of evidence; (b) next when they completed a MEL Matrix and wrote their evidence based argument at the end of Day 2; and (c) at the end of day three when they completed a MEL matrix, were asked to write about which piece of evidence was most useful in helping them pick a model, and then completed their final evidence based argument. This sequence of activities provided students with multiple opportunities to re-evaluate evidence in light of other evidence.

Evidence 3 (monkeys/SIV) was designed to serve as a possible conceptual link between the cats/FIV evidence and the humans/HIV evidence. Like Evidence 1 and 2 it was not strongly diagnostic, but similar to Evidence 1 it suggested that another kind of mammal (monkeys) are regularly infected by an immunodeficiency virus and that some form of resistance exists and can be passed down from parent to offspring. The prompt at the end of Evidence 3 asked students to analyze the results shown in the monkey family pedigrees and identify the pattern of resistance (dominant or recessive). The aim was to have students use some of their genetics knowledge gained from earlier lessons (i.e., pedigrees, inheritance, recessive/dominant traits) while reasoning about the possibilities of the HIV resistance possibly being inherited.

Evidence 4 (Paxton study) was designed to be the most diagnostic piece of evidence. The results of the research on the white blood cells of people exposed to HIV clearly show that these cells are resistant to high levels of the virus. This was saved for

last because it strongly suggests that humans can at least be resistant, although whether or not that resistance is genetic is not addressed by this evidence.

Three of the pieces of evidence were designed in a way that could subtly promote students' development and use of a body of evidence. Evidence 1 (cats/FIV) in particular was designed to complement two other pieces of evidence, namely Evidence 3 (monkeys/SIV) and Evidence 4 (humans/HIV). There are two common themes on which one could build a body of evidence that is shared across all three pieces of evidence. The first theme is that there is a range of immunodeficiency viruses (FIV, SIV, and HIV) that infect mammals, and one might infer that SIV is potentially more similar to HIV than FIV, given the taxonomic relationships of the host organisms (monkey are more similar to humans than are cats). The second theme is a plausible progression of mammals (cats, then monkeys) with increasing taxonomic similarity to humans. Seeing a plausible connection of taxonomic and viral similarities was designed with the intention that some evidence that may have seemed like it had low relevance, namely cats/FIV and monkeys/SIV, might together be seen as more relevant to figuring out if humans can be resistant to HIV. As described in detail in the results section, some students did in fact see these connections and used them in their reasoning. It is important to note that students were asked to figure out if humans could be resistant to HIV, and if so, whether resistance is genetic. There was direct evidence from the Paxton study (Evidence 4) that humans can have resistance to HIV, however the link to genetic resistance was underdetermined in that study. Evidence 1 mentioned that wild cats are resistant to FIV and that this resistance has been passed down for many generations. Evidence 3 showed, in the form of pedigrees, that some pattern of heritability for SIV resistance does seem to exist. Some

students recognized the connections between these pieces of evidence and used this line of reasoning in constructing a body of evidence.

The evidence and the tasks around the evidence (the prompts, peer discussions, MEL Matrix and written arguments) were all designed to promote students' use of various evidence evaluation, re-evaluation, and integration strategies. My analyses concentrated on students' evidence evaluation, their conceptual links between pieces of evidence that could form a body of evidence, their ability to coordinate evidence with models, and their written arguments. Research questions for this study included:

1. What are students' implicit criteria for evidence evaluation?
2. Do students adjust their evaluation of the quality of the evidence with exposure to new evidence?
3. Do students shift their model selection with exposure to new evidence?
4. Do students recognize the opportunity to construct an integrated body of evidence, and if so, what criteria do they use in its construction?
5. Does constructing an integrated body of evidence lead to increases in argument complexity?

3.2 Method

The data for this study are drawn from a five-month classroom based intervention designed to increase the sophistication of middle school (approximately age 12) life science students' epistemic practices and conceptual understanding. The instructional intervention took place in a middle-class suburban middle school (grades 6 and 7) in the United States. Students eligible for free and reduced lunch made up 14% of the total population. Demographically the school's students were 61% Caucasian, 28% Asian, 6%

Hispanic, and 5% African-American. The research presented here was conducted in one life science classroom with 88 seventh-grade students.

3.2.1 Coding

Students' ($N = 88$) essays were coded for several epistemic practices including evidence quality evaluation (Research Questions 1 and 2), evidence-to-model coordination, responsiveness to new evidence as indicated by changing the model selected (Research Question 3), constructing a body of evidence (Research Question 4), and argument complexity related to the body of evidence (Research Question 5). As described in Table 3.1, students recorded their model selections and wrote responses to multiple prompts in their student packet on all three days. Not all prompts were coded because not all pertained to the research questions addressed in this paper. Four written items were coded, and are marked with an asterisk in Table 3.1: (a) students' initial evaluations of Evidence 1 and 2; (b) students' essays in support of their chosen model on the second day after reading Evidence 1 and 2; (c) students' essays in support of their chosen model on the third day after reading all four pieces of evidence; and (d) students' responses to a prompt about which evidence was most useful in helping them decide between Models 1 and 2. The coding used here is derived from the simplified Toulmin model, used by other science education researchers (McNeill & Krajick, 2009), that combines the warrant and backing into a *reasons* category so that claims, reasons, evidence and rebuttals are analyzed for. Similar to other researchers (Garcia-Mila et al., 2013) I did not examine the use of qualifiers.

To address Research Question 1, students' implicit criteria for evidence, their final essays were coded for their evaluation of each piece of evidence (i.e., good or bad) and

the reason attached to that evaluation (e.g., this evidence is bad because it doesn't talk about humans or HIV). Two coders overlapped on 25% of the items in each of the categories mentioned we coded with a reliability of 82% for reasons codes, 92% agreement on the quality of evidence, and 95% agreement on what pieces of evidence were cited. Differences were resolved through discussion and a single coder coded the remaining items.

The reasons were organized into nine major categories including (a) Taxonomic similarities; (b) Viral similarities; (c) Heritability; (d) Processes; (e) Results; (f) Communicative features; (g) Diagnosticity; (h) Source characteristics; and (i) Role of medicine. Four of the categories (a, b, c, and i) were particular to the phenomenon being investigated. These categories were coded to give a clearer picture of how students reason about biological phenomena and give some insight into the topic specific criteria that are relevant to students' evaluations of the evidence. For example, the category *Taxonomic similarities* captured students' thoughts about whether the experimental organisms (cats and monkeys) are enough like humans to bear on the question of human resistance to HIV. The other five categories (d through h) are more generalizable across tasks that involve the coordination of evidence and models. A more detailed description and examples of the codes are contained in Table 3.2.

To address Research Question 2, whether students would adjust their evaluation of the quality of the evidence with exposure to new evidence, students' evaluations of the quality of Evidence 1 and 2 were coded at three time points: (a) when they first encountered the evidence and responded to prompts about it (see Table 3.1 for the exact prompts); (b) when they wrote their intermediate argument essay at the end of Day 2; and

(c) when they wrote their final essay at the end of Day 3, after seeing all of the evidence. The same coding scheme that was used for Research Question 1 was used here as well; see Table 3.2 for examples. Two coders overlapped on 25% of the items for each of the three time points mentioned above with reliability of 82.5% for Evidence 2 evaluation and 97.2% for Evidence 1 evaluation. Differences were resolved through discussion and a single coder coded the remaining items.

Students selected the model they thought was best, either genetic resistance to HIV exists or not, three times: (a) after being introduced to the problem but before seeing any evidence; (b) after seeing the first two pieces of evidence; and (c) after they had seen all four pieces of evidence. On each occasion, students circled a box on their worksheet to indicate their model selection and, following their second and third model selections, wrote an essay about the model they selected. Students' selections were recorded at all three time points so that they could be analyzed to address Research Question 3.

Students' final written arguments were coded for the creation of a body of evidence to address Research Question 4. The creation of a body of evidence occurred when students saw similar features across pieces of evidence and combined them into new supra-evidence structures that had their own valence (good/bad) and relationship (support/contradict/etc.) to the models. The evidence in the intervention was designed to foster the potential for this development by first introducing a more distant mammalian evolutionary cousin (cats) and their immunodeficiency virus (FIV), which has similarities and differences when compared to HIV, and then later introducing a closer evolutionary cousin (monkeys) and immunodeficiency virus (SIV) that is more similar to HIV. Students seemed to respond to three different broad sets of relationships in the evidence.

The first relationship, “pattern of resistance across species,” conjoined taxonomic relationships between cats, monkeys, and humans, all three of which are mammals, to the apparent resistance of some members of each species to their immunodeficiency virus. Second, some students saw a connection between FIV, SIV, and HIV (which were discussed in Evidences 1, 3, and 4) and came to think of them as a broad category of immunodeficiency viruses. This was coded as “pattern of resistance to immune deficiency viruses.” The third relationship that students made note of was “passing on a resistance gene to offspring.” This code captured students’ thinking that the resistance that occurs in cats, monkeys, and humans must be genetic.

To address Research Question 5, students’ final essays were analyzed for argument complexity. This required several steps. First, the codes from Research Question 1 were used to tally the number of reasons a student gave in the argument. Second, the number of pieces of evidence the student cited was also counted. Then each student received a score that combined the number of pieces of evidence cited with the number of reasons given. For example, a student who cited three pieces of evidence and provided seven reasons received a score of ten. Other argument features like the presence of qualifying statements or rebuttals were not included in this score, only the number of pieces of evidence cited and the number of reasons given. This sum was used as a dependent variable. This method has some similarities to the method used by Schwarz et al. (2003) in their analysis of verbal argument used with triads debating whether experimentation on animals is permissible or not. With the focus of this paper on the role of evidence I included a summation of the amount of evidence cited in addition to the

number of reasons as an approximation of argument complexity. It is worth noting that argument complexity is not the same as argument quality.

In general the research community has tended to value argumentative products that are more structurally complex in terms of the Toulmin argument components that are included. Garcia-Mila et al.'s (2013) framework identified 11 argument structures of increasing complexity, where a claim with no evidence (data) was the lowest level and the highest level included a claim with data, warrants, backing, and rebuttals to counterclaims. In their analysis they collapsed repeated elements down to a single instance. On this point they wrote "...we collapsed all the structures according to the types of elements rather than to the number of elements in the same category. For instance, CDDD, CDD, and/or CD were considered in the category CD. That is, the repeated elements in each structure were not taken into consideration" (Garcia-Mila et al., 2013, p. 508). For example, providing a claim with three pieces of data was treated the same as a claim with a single piece of data. The same would apply to providing warrants where multiple warrants would be treated as a single instance. While this move may have been appropriate given the aims of their study, such an analysis would obscure the kind of complexity and nuance presented in this study, given the particular focus on the role of evidence. Therefore in computing complexity for this study repeated elements were not collapsed, so that a student who cited several pieces of evidence and provided multiple reasons was distinguishable from a student who provided a single piece of evidence and a single reason (the two would be collapsed together in the Garcia-Mila et al. (2013) framework).

Schwarz et al. (2003) also used a structural approach for analyzing argument complexity and categorized arguments into four types. The least sophisticated type of argument included "...statements unsupported by any reason" and the most complex included what they referred to as a two-sided argument in which "the individual or group undertook an analysis of the pros and cons to solve the issue" (Schwarz et al., 2003, p. 229). Their analysis also included four other dimensions including the "soundness of the argument," "the overall number of reasons," "the number of reasons supporting counterarguments" and the "quality of reasons" (Schwarz et al., 2003, p. 230-232).

This analysis has some similarities, as each essay was categorized into one of four basic argument structures: (a) students who developed a body of evidence *and* used a rebuttal of a counterargument in their essay; (b) students who just developed a body of evidence; (c) students who just used a rebuttal to a counterargument; and (d) students who did not develop a body of evidence or use a rebuttal to a counterargument. The four categories are exclusive, no student could be in more than one category. The four categories of argument type were treated as an independent variable. The presence of a rebuttal to a counterargument was coded for when a student used evidence and reasons to discuss the model they did not choose; typically students gave reasons why the evidence they viewed as supporting the alternative model was somehow insufficient or of low quality. As a reminder, students were directed to write their argument to someone who disagreed with them, affording students a chance to develop a rebuttal against the alternative claim. Differences in argument complexity, the score of the total number of reasons and pieces of evidence cited, were analyzed for between the four groups.

3.3 Results

Research Question 1: *What are students' implicit criteria for evidence evaluation?*

Students used numerous criteria to evaluate evidence, which I organized into nine major categories, as shown in Table 3.2 and as described in the coding methods section. Table 3.2 contains the codes, a description of each code, examples from students' final essays, and the frequency of each code as it occurred in relation to each piece of evidence.

The first code in the table, *animals are similar to humans*, is in the category *Taxonomic similarities*. This code was given when a student made a statement that animals (cats or monkeys) are similar to humans, and occurred only in the context of Evidence 1 (cats/FIV) and Evidence 3 (monkeys/SIV). Taxonomic similarities played a role in student reasoning about these pieces of evidence. This was planned for in the design of the evidence and the prompts, as noted earlier. Some students felt that the taxonomic relationship of mammals made the three species similar enough to be useful for deciding which model was correct, and these students typically wrote positively about these pieces of evidence and used them to support the model they chose. Some students denied the plausibility of this connection and tended to downgrade their evaluation of this evidence. Understanding the role of experimental organisms, and their relationship to one another, was a major element for some of the bodies of evidence that students constructed, and will be discussed in more detail.

Other codes occurred across all four pieces of evidence. For example, under the category *Results* is the *Cats/monkeys/humans are resistant* code, which was used in relation to all four pieces of evidence. Like the *animals are similar to humans* code, the *cats/monkeys/humans are resistant* code also played a role in some students' construction of a body of evidence.

Viral similarities, the second major code category, captured students' reasons regarding similarities or differences between FIV, SIV, and HIV. The connection between FIV and HIV was mentioned most often, with 27 reasons given, and the connection between SIV and HIV was mentioned 21 times, often in conjunction with one another. Although the evidence did not contain a lot of details about the structure and function of the viruses, some students constructed a body of evidence based on the similarities of immunodeficiency viruses.

The third major category, *Heritability*, arose mostly in response to Evidence 1 and 3. Students who thought these two pieces of evidence were good tended use them to support the second part of the model, that resistance can be inherited. It is worth noting that there was no evidence about humans that showed heritability of resistance, so some students who selected Model 2, that genetic resistance exists, found themselves in the position of using the animal evidence to support this part of the model.

The fourth category, *Processes*, is derived from work in epistemic cognition that highlights the role of processes in the development of epistemic products (e.g., knowledge, scientific models, and so on). It was an open question if students would make note of processes and distinguish between reliable and unreliable ones. In their final essays students frequently commented on Evidence 4, the Paxton study, as one that used reliable processes. This was an APL piece of evidence derived from one of the early studies on HIV resistance, arguably one of the first to highlight that such resistance is possible, which has been cited thousands of times. In their final essays, students often mentioned that it was a blood test of a relevant group (humans) and a relevant virus (HIV) and that the levels of HIV were manipulated to show that the white blood cells

resisted even high levels of the virus. The results provide some support for the claim that APL style evidence provides students with an opportunity to comment on processes and that, at least within the scope of this study, they recognized and frequently commented on the highest quality piece of evidence.

Evidence 2 was the most common piece of evidence to be evaluated as having a poor process, with 82% (14 of 17 responses) of the *unreliable process* codes being applied to that evidence. The reasons given were diverse, ranging from the small sample of doctors, the limited geographic location (just one health clinic), and the accuracy of the blood test (although it was 99% accurate, some students felt that it was not accurate enough and that someone with resistance may have slipped through). The most interesting critique offered by a few students was that because this was a health clinic for patients with HIV, a person who was resistant to HIV would have no need to attend such a clinic and would likely be unknown to the medical professionals there. Evidence 1 and 3 did not generate much writing about reliable processes for the final essay.

Each study contained data about resistance, and/or lack of resistance, to immunodeficiency viruses. This was coded for in the *Results* category. Evidence 1 and 3 contained similar information that some animals are resistant and some are not. Evidence 2 strongly asserted that humans cannot be resistant. Evidence 4 showed that white blood cells from some humans show resistance to high levels of HIV. The distinctions here are important because of the prevalence of these codes, with results codes accounting for more than 40% of all the codes assigned. Many students used Evidence 4 as the most conclusive piece of evidence; it was cited 80 times as support for the most commonly selected model (80 out of 88 students selected Model 2, that resistance to HIV exists and

is genetic). The results of Evidence 1 and 3 were also frequently cited, with 52 and 50 reasons given respectively, accounting for 11% and 10% of all codes. The results of these two studies were more indirectly related to the models than the results of Evidence 4, and their use typically elicited additional justifications. Commonly students cited Evidence 1 and Evidence 3 and argued that they were relevant because they were both about mammals and humans are mammals. In this, the experimental organisms lent some additional plausibility for many students to the idea that humans might be resistant.

Communicative features played a small role in students' thinking about the evidence; just over 1% of the total number of codes occurred in this category. This code was used when students cited features of a study such as its inclusion of a lot of details or its readability. Prior work by Pluta et al. (2011) suggested that students are sensitive to these elements in models, but they were less prevalent in the written essays in this study. If asked whether ease of understanding was important, it would not have been surprising if many students had affirmed this idea, but it did not play a large role in their essays. Given that no students made a negative comment about the communicative features of the evidence in the final essay, such as its understandability, and only a few positive comments were offered, it is reasonable to think that students felt the evidence was accessible.

The seventh major category, *Diagnosticity*, targets students' thinking about diagnosticity, relevance and irrelevance. The three codes represent a rough progression from *irrelevant and not useful*, to *relevant and useful*, to *diagnostic*. The *irrelevant* code was used 9 out of a total of 10 times in reference to Evidence 2. These evaluations tended to focus on the fact that most of the doctors talked about issues related to HIV, but not the

possibility of genetic resistance. Four medical staff were interviewed and only one, the lab assistant, spoke about the possibility of genetic resistance (he denied it was possible). It was possible for a piece of evidence to be coded as both relevant and irrelevant; a few students made a distinction between the lab assistant and the rest of the medical staff, which showed detailed attention to the evidence.

The study most cited as *relevant* was the cats/FIV evidence. This was a bit surprising given the more distant links between humans and cats, and FIV and HIV. One interpretation is that students were expending some additional effort to convince the reader of their essay (they were instructed to write to someone who disagreed with them) that despite perhaps surface level irrelevance (i.e., wrong species, wrong virus), the cats/FIV evidence was relevant. The body of evidence codes that will be discussed later support this interpretation.

Diagnosticity is hierarchically above the *relevance* code. To move from a relevance code to a diagnostic code, the student needed to state that the evidence they were discussing was involved in their decision about which model was better. This was most commonly attributed to Evidence 4. As explained in the Lesson Design Considerations section, the most diagnostic piece of evidence was intentionally presented last.

The *Source* category of codes was included given the prevalence of studies that have investigated how students engage in sourcing when considering a variety of secondhand evidence. Source information did not figure strongly into the design of this evidence. It is worth noting that Evidence 2, the health clinic, received the most positive source evaluations (4) whereas the Dr. Paxton study (Evidence 4) received no such codes.

This is probably for two reasons. First, Evidence 2 contained the most source information. Second, the sheer volume of codes for Evidence 4 pertaining to *Process*, 50 reasons given, and *Results*, 80 reasons given, showed that students were focused on methods and results of the studies rather than their sources.

The ninth major category, the *Role of medicine*, mostly related to Evidence 2. In that evidence a medical professional mentioned that medicines can be given to pregnant and nursing mothers to cut the transmission rate of HIV from mother to child. A few students interpreted this as resistance, indicated by the *medicine conveys resistance* code, and used it to support Model 2. It would seem that these students were focused on the resistance portion of the claim and not the genetic portion of the claim. One student argued that this was resistance but not genetic resistance. For most students this information was not a major factor in their reasoning.

Finally, there were a few reasons offered that did not fit well into the existing categories and therefore received an *other* code. There were very few other codes, less than 1% of the total. The *no reason* code was used when a student cited a piece of evidence but provided no reasoning about the evidence. This occurred with just over 1% of the pieces of evidence cited, about half of which were attributable to a single student's essay.

As an overview of students' criteria for evaluating the evidence, students targeted *Processes*, *Results*, *Heritability*, *Taxonomic similarities*, and *Viral similarities* for most of their reasoning in their final essays. Results and processes, particularly with respect to Evidence 4, were deeply involved in the final essays of students, but many students realized that Evidence 4 was not fully conclusive as it did not contain information about

the heritability component. To address this gap, students needed to marshal additional evidence. Evidence 1 and 3, as can be seen in Table 3.2 below, often filled this gap with a combination of taxonomic similarities, viral similarities and heritability reasons ascribed to both pieces of evidence.

Table 3.2						
<i>Students' evaluations of the quality of evidence taken from their Day 3 final essays</i>						
<u>Code Category</u>	<u>Code Description</u>	<u>E1</u>	<u>E2</u>	<u>E3</u>	<u>E4</u>	<u>Example</u>
<i>1. Taxonomic Similarities</i>						
Animals are similar to humans (g)	The student states that animals, cats, and/or monkeys are similar to humans	10	-	12	-	"monkeys are considered the closest animal species to humans..."
Animals are not similar to humans (b)	The student states that animals, cats, and/or monkeys are not similar to humans	4	-	4	-	"monkeys and cats are NOT humans, even if they are mammals"
<i>2. Viral Similarities</i>						
FIV/SIV is similar to HIV (g)	The student states that FIV and/or SIV is similar to HIV	27	-	21	-	"FIV is similar to HIV in humans with AIDS like symptoms"
FIV/SIV is not similar to HIV (b)	The student states that FIV and/or SIV is not similar to HIV	2	-	2	-	"Maybe there is a big factor or difference between SIV, FIV and HIV that we are missing"

<u>Code Category</u>	<u>Code Description</u>	<u>E1</u>	<u>E2</u>	<u>E3</u>	<u>E4</u>	<u>Example</u>
3. Heritability						
Resistance is passed down (g)	The student states that resistance to the viruses is passed down	6	-	30	-	"...one resistant parent and one non-resistant parent. They had offspring and both were non-resistant. After seeing this I realized resistance is genetic it's just recessive"
Resistance is not passed down (g)	The student states that resistance is not passed down	-	-	7	-	"The monkeys can pass down the SIV but can't pass the resistance"
Disease is passed down (g)	The student states that the disease is passed down from parents to offspring	-	-	4	-	"In evidence 3 the monkeys can pass down the SIV"
Disease is not passed down (g)	The student states that the disease is not passed down from parents to offspring	-	2	-	-	"In evidence two it tells about how doctors & nurses had patients that have HIV or AIDS, & how they kept it from passing it to their offspring"
4. Process						
Reliable process (g)	The student states that a reliable process was used to obtain the evidence	3	1	-	50	"Dr. Paxton actually exposed the blood of some people who were possibly resistant to high levels of HIV, but all of them were still resistant"
Unreliable process (b)	The student states that an unreliable process was used to obtain the evidence	-	14	1	2	"people may believe that according to Evidence 2 the scientist had never found anyone resistant. However, maybe that doctor only tested people in 1 area"

<u>Code Category</u>	<u>Code Description</u>	<u>E1</u>	<u>E2</u>	<u>E3</u>	<u>E4</u>	<u>Example</u>
<i>5. Results</i>						
Cats, monkeys, humans are resistant (g)	The student states that some cats, monkeys, or humans resist their virus	57	7	52	80	“the SIV study shows that resistant monkeys that are SIV resistant have the virus in their blood already”
Cats, monkeys, humans are resistant (b)	The student states that some cats, monkeys, or humans resist their virus	2	-	3	-	“... animals are different from humans so there might be a difference that animals can be resistant but not humans”
Cats, monkeys, humans are not resistant (g)	The student states some cats, monkeys, or humans don’t resist the virus	11	11	8	2	“[evidence 2] because it is saying that everyone can get it but you can’t get rid of it.”
Humans are not resistant (b)	The student states that some or all humans do not resist HIV	-	9	-	1	“this evidence [E2] shows a group of doctors saying people can’t be resistant. The reason I go against this evidence is because it doesn’t have many doctors being interviewed so it’s not a wide study”
<i>6. Communicative Features</i>						
Communicative features (g)	The student indicates that the communicative features (i.e., ease of understanding) are good	1	2	1	3	“it gave a lot of detail of how the white blood cell being resistant to the disease...”
Medicine conveys resistance (g)	The student states that medicine can make people resistant to HIV	-	4	-	-	“in evidence 2, it says that if the mother takes medicine to not get the disease, the baby probably won’t get it either. That shows that people can be resistant”
Medicine is not resistance (b)	The student states that medicine is not the same thing as resistance	-	1	-	1	“pregnant women can take medicine if they have HIV and the medicine will reduce the babies chance of getting HIV but the evidence never says anything about families and generations being resistant”

[illegible]

Research Question 2: *Do students adjust their evaluation of the quality of the evidence with exposure to new evidence?*

The coding scheme used for Research Question 1 (Table 3.2) was used for the evaluation prompts for Evidence 1 and 2 (time point 1), for the model selection essay students wrote after seeing those two pieces of evidence (time point 2), and for the final essay (time point 3). Based on the coded data I analyzed students' evaluations of the quality of Evidence 1 (cats/FIV) and Evidence 2 (health clinic) at these three time points with two Cochran's Q tests.

The first analysis examined changes in students' evaluations of the quality of Evidence 1. Initially, 67% of students ($N = 88$) rated Evidence 1 as good evidence, but when students wrote an argument about their second model selection, this dipped to 48% who adopted a positive view of this evidence. In the final model selection essay, 65% believed it was good evidence. A Cochran's Q test showed a significant difference (Cochran's Q, $df = 2$, $Q = 9.418$, $p = 0.009$); the percentage change indicated a dip at time 2 in their confidence about the quality of Evidence 1. This was likely driven by several factors. First, some students who initially thought this might be good evidence re-evaluated it after seeing Evidence 2; some students pointed out that the second evidence was about humans, not cats, and thus more relevant. This may have caused part of the dip. It was possible that students would maintain this stance through the final essay, but that was not the case. Instead students re-evaluated their stance on Evidence 1 and 2 in the final essay. It is possible that this was in part driven by seeing more evidence about immunodeficiency virus resistant mammals (Evidence 3 and Evidence 4). An example of

this kind of reasoning can be seen in Table 3.3 and will be explained in more detail below.

I conducted a similar analysis of Evidence 2 (i.e., health clinic interview). Initially, 41% of students adopted a favorable view of Evidence 2, rising to 58% when they wrote an argument about their second model selection. In the final model selection essay, however, only 22% believed it was good evidence. A Cochran's Q test showed a significant result (Cochran's Q , $df = 2$, $Q = 25.633$, $p < 0.001$); the percentage change indicated an increase at time 2 and a decrease at time 3 in students' confidence about the quality of Evidence 2. The Cochran's Q analyses for both Evidence 1 and Evidence 2 show that students re-evaluated old evidence in light of new evidence, which was articulated in their written argumentation and motivated their model selection.

With respect to both pieces of evidence, one interpretation of the pattern of results is that the dip for Evidence 1, and bump up for Evidence 2 at time 2 occurred partly because after seeing Evidence 2 (humans/HIV), students believed it was more relevant than Evidence 1 (cats/FIV). A student essay, Stephanie's (pseudonym), that is elaborated in Table 3.3 shows an example of how this series of reasoning moves occurred over time with exposure to new evidence. Initially after reading Evidence 1 (cats/FIV) and without seeing any other evidence, Stephanie offered a fairly positive evaluation of Evidence 1. After seeing Evidence 2, which spoke much more directly to the HIV resistance issue, she viewed Evidence 1 more negatively and believed that humans could not be genetically resistant to HIV. In the final essay, she shifted to believing that humans can be genetically resistant to HIV. On Day 2, she viewed Evidence 2 (the health clinic) as good evidence based specifically on the testimony of Lab Assistant Feld, who tested blood

samples and had never met anyone resistant to HIV. On Day 3 she no longer believed Evidence 2 was good, for two reasons. First, she stated that the evidence was bad because people who were resistant to HIV would not develop the disease and would have no need to go to a health clinic for treatment. Her second reason was that the evidence was an interview and not a study or experiment aimed at finding people who were resistant. Her critique likely stemmed from having read Evidence 4, which aimed to investigate people who were possibly resistant. Evidence 1, which she previously thought was good, and then bad, was viewed as a good piece of evidence again because of Evidence 3. Evidence 1 and 3, when considered in concert by this student, seemed to support the idea that heritable resistance to immunodeficiency viruses exists for other species, and was used as part of her justification for why humans can be genetically resistant to HIV, a change from her essay on the previous day.

Table 3.3

An analysis of Stephanie's evaluation and re-evaluation of evidence

<u>Time</u>	<u>Evidence</u>	<u>Quality</u>	<u>Student Reasoning</u>	<u>Comment</u>
T1	E1	Good	"FIV and HIV are basically the same thing but they affect different animals, therefore studying cats and their FIV resistant gene would be beneficial to us. Furthermore, both cats and humans are mammals which makes us genetically similar"	This was the initial evaluation (occurred at the start of Day 2) that the student made after watching the video and reading Evidence 1.
T1	E2	Good	"It does point to humans not being HIV resistant, and it also seems very accurate."	This was the initial evaluation (occurred on Day 2) after reading Evidence 2.
T2	E1	Bad	"Evidence 1 talks about wild cats being resistant to FIV. However, they are cats and they are immune to FIV which is entirely different (yet admittedly similar) to HIV."	At Time 2 (the first essay that referred to two pieces of evidence) the student believed that humans are not resistant to HIV and discounted Evidence 1.
T2	E2	Good	"I believe model 1 is better because of evidence 2...Specifically the lab assistant Feld says he has never met someone who was HIV resistant...this points to HIV resistance not existing."	At Time 2 the student thought that Evidence 2 was good and supported the claim that HIV resistance does not exist (Model 1).
T3	E1	Good	"Evidence 1 which is a video on how wild cats are resist to FIV, and Evidence 3 which is a study on how SIV resistance is passed on down in families support Model 2 as they state how other species can be resistant to immunodeficiency viruses."	For the final essay the student changed to Model 2 (resistance exists). They re-evaluated Evidence 1, re-establishing why it is good evidence. This is in part because of the connection to Evidence 3.
T3	E2	Bad	"The small part of evidence 2 that relates to HIV resistance specifically, the Lab assistants interview, does state he has never met anyone genetically resistant. However one that was genetically resistant to HIV wouldn't need to get treated for HIV at a clinic. Furthermore it was a simple interview, not a specific study or experiment to find someone resistant to HIV, thus decreasing the chance of actually finding HIV resistance and therefore not very believable."	Again, at Time 3 the student re-evaluated the evidence, in this case downgrading the quality of Evidence 2, a piece of evidence they previously thought was good and that they had used to motivate their selection of the alternate model on the previous day.

Research Question 3: *Do students shift their model selection with exposure to new evidence?*

Students selected a model at three time points: (a) after the lesson introduction and PowerPoint but before seeing any evidence; (b) after seeing two pieces of evidence; and (c) after seeing all four pieces of evidence. Students initially found both models compelling. On Day 1, before seeing any evidence, 42 students chose the “no resistance” model, 41 chose the “resistance” model, and 5 students did not select a model. On Day 2, after seeing Evidence 1 and 2, 42 students chose “no resistance” and 46 chose the “resistance” model. On Day 3, when students had seen all four pieces of evidence, there was a significant shift, with only 8 students picking the incorrect model (no resistance) and 80 students picking the correct model (resistance). Applying a McNemar’s test of Day 2 compared to Day 3 model selections ($N = 88$) revealed a statistically significant shift ($p < .001$). This shows that students’ model selections were made in response to the changing landscape of evidence. A common sequence of model selections is shown in Stephanie’s essay in Table 3.3.

It was possible that students who had selected Model 1 (the incorrect model) at the first two time points would stick with their model in the face of evidential challenges. There was little evidence on Day 3 to support Model 1, and only one piece of evidence overall (Evidence 2) that supported it. However, sticking with a model that is not well supported by the evidence was not necessarily an unsophisticated stance to take in this context. The two part claim that (a) humans can be resistant to HIV and that (b) the resistance is genetic was underdetermined by the four pieces of evidence. Evidence 4, the Paxton study, strongly supported part (a) that humans can be resistant to HIV; but part (b)

that HIV resistance is genetic, was at best indirectly supported by Evidence 1 and 3. Some of the students recognized this weakness in the evidence and wrote about it in their essays. Olivia's (pseudonym) essay below is one such case.

In short, she argued that Evidence 2 clearly showed that everyone can get HIV, and Evidence 1 and 3 were largely irrelevant because the species and the viruses were not similar to humans and HIV. She believed that resistance exists, but that there was little evidence that the resistance is genetic. She went one step further and provided her own explanation for how resistance could happen through an alternative mechanism. She believed that the subjects in the Paxton study developed an immunity to HIV through their repeated exposures, similar to a vaccine.

Olivia's Final Essay:

I think that Model A, genetic resistance to HIV, does not exist, is the best. First off, Evidence 2 stated that everyone can get HIV, so therefore no one can be resistant and therefore supports my model. Further, though Ev. 1 and 3 support genetic resistance, SIV and FIV are NOT the same as HIV, no matter [how] similar they may be (unless scientists renamed the same disease to fit the animal species). Also, monkeys and cats are NOT humans, even if they are mammals. Moreover, in Ev. 4, the study was on people who were exposed to HIV frequently. However, I feel that the people who were exposed to HIV frequently were slowly building up immunity to the virus, therefore they became resistant. I don't think it was genetics that made them resistant – it was more like a HIV “vaccine” that made them resistant. So therefore genetic resistance to HIV does not exist.

Olivia's essay shows that even students who did not switch to the correct model could still be responsive to evidence and reason about that evidence in sophisticated ways. Moreover, her attention to the underdetermined part of the claim, and attempt to explain it, highlights that some students recognize that even simple models can occur in pieces and that each piece of the model needs evidential support.

Research Question 4: *Do students recognize the opportunity to construct an integrated body of evidence, and if so, what criteria do they use in its construction?*

Thirty of 88 students (34%) engaged in a practice I call *developing a body of evidence* in their final written arguments. This occurred when students saw similar features across pieces of evidence and combined them into new supra-evidence structures that had their own valence (good/bad) and relationship (support/contradict/etc.) to the models. This most commonly occurred between Cats/FIV (Evidence 1) and Monkeys/SIV (Evidence 3), with 25 out of the 30 students combining these evidences. The evidence in the intervention was designed to foster the potential for this development by first introducing a more distant mammalian evolutionary cousin (cats) and their immunodeficiency virus (FIV), which has similarities and differences when compared to HIV, and then later introducing a closer evolutionary cousin (monkeys) and immunodeficiency virus (SIV) that is more similar to HIV.

Students seemed to respond to three different broad sets of relationships in the evidence. The first relationship conjoined taxonomic relationships between cats, monkeys, and humans, all three of which are mammals, to the apparent resistance of some members of each species to their respective immunodeficiency virus. For the code *pattern of resistance across species*, I found that 21 students used this kind of reasoning to partially justify their belief that humans can be resistant to HIV. As an example, one student said “...evidences 1 and 3 showed animals other than humans that have a resistance...” Second, some students saw a connection between FIV, SIV, and HIV (which were discussed in Evidences 1, 3, and 4) and came to think of them as a broad category of immunodeficiency viruses. The code *pattern of resistance to immune*

deficiency viruses occurred in 12 students' essays. As an example, one student said "I believe that there is a resistance to HIV that is genetic. I think this because other animals with other immune deficiency viruses such as monkeys and wild felines (evidence 1/3) have built up a resistance..." The third relationship that students made note of was *passing on a resistance gene to offspring*, which occurred in five students' essays. This code captured students' thinking that the resistance that occurs in cats, monkeys, and humans must be genetic. Students typically drew on Evidence 1 and 3. With regard to the cats evidence (Evidence 1), students tended to state that parents passed resistance down to offspring, while for the monkeys evidence (Evidence 3), students commented on several family pedigrees of monkeys, some of whom were resistant to SIV and some of whom were not. As an example, one student said "Furthermore, as to the genetic resistance, both evidence 1 and 3 show that it does exist. Evidence 1 states that the mutation occurred in the wild cats' ancestor and the resistance was passed on from generation to generation...Evidence 3, the SIV study, showed that the resistance was recessive but was passed on through genetics." The codes, frequencies, and student examples of these practices that resulted in the development of a body of evidence are included in Table 3.4.

Table 3.4		
<i>Students' development of bodies of evidence in their final essay</i>		
<u>Reason</u>	<u>Frequency</u>	<u>Examples, excerpts from students' final essays</u>
Pattern of resistance across species	21	"...in evidence 1 and 3 they both show mammals (like humans) that are at least a little resistant to the HIV. That is why I believe that there is a way to have genetic resistance."
Pattern of resistance to immune deficiency viruses	15	"[Evidence 1 and Evidence 3] support Model 2 as they state how other species can be resistant to immunodeficiency viruses" "Because both FIV & SIV are similar to HIV so the same conclusions should apply to HIV"
Passing on a resistance gene to offspring	5	"Another evidence that were relevant were the evidence 1 and 3. In evidence 1 wild cats had a resistance gene in them which stopped them from having FIV. This had happened because their ancestors developed a resistance. In evidence 3 group two and group three had a parent or both parents with resistance..." ^a
<p><i>Note.</i> Students could develop more than one body of evidence interpretation, so the frequency is the number of bodies of evidence, not the number of students.</p> <p>^a "group two and group three" refers to pedigree charts of monkeys passing on, or failing to pass on, SIV resistance to their offspring, as shown in Evidence 3.</p>		

Research Question 5: *Does constructing an integrated body of evidence lead to increases in argument complexity?*

For the final essay, all of the argument elements related to evidence evaluation and evidence-to-model coordination were summed for each student to provide an approximation of argument complexity. For example, if a student cited three pieces of evidence and gave one reason for each piece of evidence's quality then that student received a score of six (i.e., three pieces of evidence and three reasons). Scores ranged from 0 to 19 ($M = 8.74$, $SD = 3.87$). Essays were also categorized into four types. The four basic argument structures identified here include: (a) students who used a body of

evidence and a rebuttal to a counterargument in their essay ($n = 10$); (b) students who just developed a body of evidence ($n = 20$); (c) students who just used a rebuttal to a counterargument ($n = 18$); and finally (d) students who did neither ($n = 40$).

A one-way between subjects ANOVA was performed to compare argument complexity scores across the four categories of argument structure. A Levene's test for the equality of error variance did not show a significant result ($F(3,84) = 0.474, p = 0.702$), meaning the variance of the data was suitable for this type of ANOVA. There was a significant effect of argument structure on the argument complexity produced ($F(3,84) = 15.06, p < 0.001$). A post-hoc Tukey HSD comparison showed that the mean argument complexity score for the body of evidence plus counterargument category ($M = 12.5, SD = 3.89$) and the body of evidence alone category ($M = 11.45, SD = 2.74$) were significantly different from students' written arguments that included only a counterargument ($M = 8.1, SD = 3.52$) and those that did not include a counterargument or body of evidence ($M = 6.75, SD = 3.02$). The post-hoc test revealed that there were no significant differences between the body of evidence plus counterargument and the body of evidence argument structures ($p = 0.764$). The post-hoc test also revealed that there were no significant differences between the argument component complexity of written arguments that included a counterargument and those that did not ($p = 0.621$).

3.4 Discussion

3.4.1 Epistemic Cognition and Evidence Evaluation

Coordinating multiple pieces of evidence is an important practice in science (Sober, 2008). Doing this well relies in part on successfully evaluating the quality of evidence to sift the good from the bad. Research Question 1 was aimed at finding out

students' implicit criteria for evidence evaluation. As pointed out by Sampson and Clark (2008) much of the research using the Toulmin Argumentation Pattern has focused on "the field invariant features of an argument..." (p. 452). This study attempted to extend our knowledge of students' notions about evidence quality with some particular attention to field-variant elements. The results here show that students regularly made use of their own implicit criteria for evidence quality. They used fine-grained criteria like taxonomic similarities (or dissimilarities) among host organisms or similarities (or differences) between different families of viruses. Students also evaluated the methods and processes used to generate the evidence. The role of experimental organisms (i.e., domestic and wild cats, monkeys) and their taxonomic relationships to humans were an important component of students' reasoning in this study.

Theoretical work on epistemic cognition suggests that there could be many ideals (criteria) for what counts as good evidence (Chinn et al., 2011; 2014; Chinn & Rinehart, 2016). This particular facet of epistemic cognition has not been empirically explored in a model-based inquiry environment. An implication of this work is that it affirms claims advanced in the AIR model of epistemic cognition that some of the cognitive processes used by students were situated in the task and specific for the topic (heritability of immunodeficiency virus resistance) and domain (biology). Further, it is the case that medical studies often make use of experimental organisms, and reasoning about them could be a significant factor in how people process evidence that makes use of non-human subjects. Future research on the role of experimental organisms and the agents and objects (e.g., pathogens, drugs, new biotechnologies) they interact with might further reveal how students and laypeople reason about biology, health, and medicine topics.

The research presented here made use of Adapted Primary Literature (APL) and Journalistic Reported Versions (JRV) sources and was, at least in some respects, similar to research done on multiple documents coordination. A limitation of the research done in the multiple documents tradition is a relative lack of attention to evidence criteria, other than those dealing with source information and corroboration, which are largely used in field-invariant ways. This research attempted to contribute to multiple documents research by highlighting some important considerations of evidence that students reason about in a more fine-grained way. A limitation of the research presented here is related to its strength; not all of the evidence quality considerations that students put forward are widely applicable.

3.4.2 Evidence Re-evaluation

Evidence re-evaluation plays an important role in the research reading habits (Tenopir et al., 2005) and reasoning practices (Thagard, 2000) of scientists. Laypeople often encounter evidence related to scientific issues, and providing students with multiple pieces of evidence of variable quality can provide them with opportunities to develop a sense for what counts as good evidence, or importantly what counts as bad evidence. In addressing Research Question 2, students adjusting their evaluation of the quality of evidence with exposure to new evidence, I show that students did in fact make such moves; they frequently re-evaluated the quality of evidence. This happened for at least two reasons. First, students began to calibrate their sense of what counts as good evidence as shown in Stephanie's work. Like Stephanie, many students rated Evidence 2 highly at first, only to revise that that estimation later when they came in contact with higher quality evidence, specifically Evidence 4. This contrasts with findings that suggest that

students rarely challenge the authority of what they read (Phillips & Norris, 1999). This study did not explore the reasons why students felt empowered to challenge the evidence, it merely established that they did challenge it. There are several reasons why this might have occurred based on the design of the study. First, students were introduced to argument critique, this may have helped establish a general norm that critique in the classroom is acceptable. More specifically students were asked to rate, or evaluate evidence, and justify this in writing or in discussion with a peer. Again, this helped to establish a general classroom norm that critique is part of science.

Second, as they encountered new evidence they began to see conceptual connections between pieces of evidence, and these conceptual connections in some cases strengthened or weakened students' evaluation of the quality of evidence previously seen. For some students in this study, connections between Evidence 1 (cats/FIV) and Evidence 3 (monkeys/SIV) provided the opportunity to make these kinds of links.

Research on students' written argumentation has suggested that writing an argument does not lead to improvements in students' abilities to differentiate reliable and unreliable sources (Wiley et al., 2009). Others have suggested that students infrequently use evidence that is appropriate or sufficient for supporting a claim (McNeill & Krajcik, 2007). In this study, students responded to evidence in a situated fashion. When seeing the first two pieces of evidence students approached both with a critical stance, but in the end tended to favor one piece of evidence over another. After seeing Evidence 3 and 4, students made several moves. First, they tended to downgrade Evidence 2 (health clinic) because stronger evidence (Evidence 4, Paxton study) contradicted the message from Evidence 2. Rather than just going on the say-so of the medical professionals in the

interview, students strongly preferred the empirical results presented in Evidence 4. This is commensurate with Sandoval and Cam's (2010) findings that students tend to prefer data over the word of an authoritative figure. This is interesting in this case because the medical professionals in question had relevant experience with HIV patients and could be considered experts, not just authoritative figures, on the topic. One possible difference here is that students were asked to select evidence to support or contradict a model, rather than gather and generate their own data to use as evidence to support a model.

3.4.3 Responding to New Evidence and Changing Beliefs

Zimmerman (2000) asserted that "The ability to consider alternative hypotheses is an important skill, as evidence may relate to competing hypotheses" (Zimmerman, 2000, p. 114). Some prior research has shown that students engage in a variety of psychological stances toward evidence, particularly anomalous evidence, to limit the need for belief change when faced with belief inconsistent information (Chinn & Brewer, 1993). On this basis one might predict that once a piece of evidence is viewed as good evidence which supports a favored model, a student would be unlikely to change their evaluation of the quality of the evidence or their favored model. Other empirical research has documented that students will entrench to some degree and defend a preferred position; Garcia-Mila et al. (2013) found that having a goal to persuade each other limited the quality of students' argumentation, while having a goal of reaching consensus produced better arguments. This factor was not explored in this study. However, the final essay prompt did ask students to engage in persuasion by writing their final argument to someone who may disagree with them. I found students to be flexible in their positions, with many of them switching their model selection in response to new evidence. Most of the students

selected the model that was best supported by the available evidence, even though approximately half of them had expressed a belief in the alternative model just one day before. One possible explanation is that verbal argumentation might enhance this “solidify and defend” factor but written argumentation might not. Another possible explanation is that during the PRACCIS project students engaged with a number of lessons where there were competing models, and switching between models was not uncommon. Given that this lesson occurred midway through the PRACCIS intervention, it is possible that students had developed a norm that changing one’s mind in response to evidence was acceptable.

3.4.4 Developing a Body of Evidence

Little prior research has examined how students develop and use networks of evidence. The study described here intentionally included pieces of evidence that had the potential for students to link together into a body of evidence. Research Question 4 was aimed at finding out if students could recognize the opportunity to develop a body of evidence, and if so, to discover what criteria they would use. In this study many students developed a body of evidence and used it to justify the selection of a model.

Typically analyses of how students handle evidence consider that evidence in an isolated fashion. Chinn and Brewer’s (2001) analyses, while highly detailed, did not examine how students coordinate a body of evidence. Recent learning progressions for argumentation also tend to envision the argumentative product in such way that pieces of evidence are isolated from one another. Figure 3.2 is a representation of such an argument structure (Berland & McNeill, 2010). The focus of this structure is on supporting claims with evidence and rebutting alternative claims with evidence.

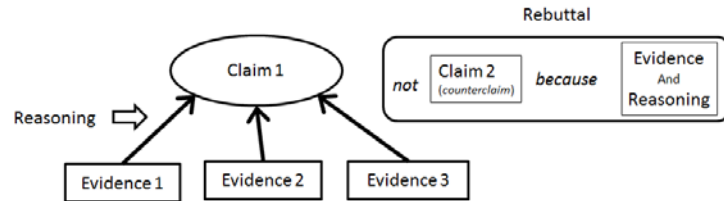


Figure 3.2. The Berland & McNeill (2010) argument structure. This is a faithful representation.

A more recent learning progression for argument structure, as seen in Figure 3.3, conceives of a structurally complex argument as one where multiple claims and their associated warrants are rebutted (Osborne et al., 2016).

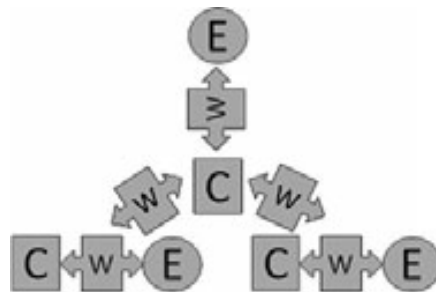


Figure 3.3. The upper level anchor for the Osborne et al. (2016) learning progression for argumentation.

Osborne et al. stated that “The limitations of the current study that are worthy of attention in future work are further elaboration of the sublevels of the map and additional investigation of the highest level of the map” (Osborne et al., 2016, p. 841). In this case I would argue that this progression could be productively expanded to include a deeper account of the role of evidence, specifically (a) criteria for good evidence; and (b) the development of bodies of evidence as elements that populate some of the sublevels of the learning progression.

Figure 3.4 shows how some students in this study came to see conceptual links, represented as horizontal arrows, between pieces of evidence. Seeing conceptual links

between pieces of evidence can cause students to develop a mental model of integrated evidence. The student has selected Model 2 as the best model based on Evidence 1, 3 and 4 supporting it. In this instance Evidence 1 and 3 are conceptually linked to develop a body of evidence that is used to justify (partly) the student's selection of this model. Note that this is not a complete representation of the activities in this study. A more complete representation would include how students reasoned about and rebutted Model 1. This representation is intended to highlight the structural elements detailed in this paper (i.e., body of evidence and criteria for good (or bad) evidence; it is intentionally simplified.

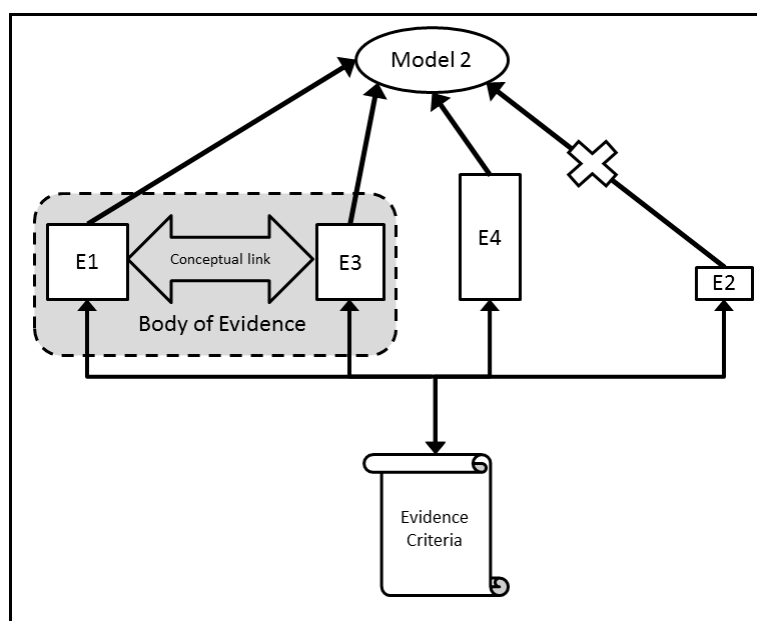


Figure 3.4. A proposed alternative model of argumentation that shows conceptual links forming a body of evidence and the role of evidence criteria.

3.4.5 The Body of Evidence and Argument Complexity

Argument complexity has been conceived of in a variety of ways, sometimes with little overlap (Sampson & Clark, 2008). It is not my position that complexity entails quality. However, promoting argument complexity has been of interest to many researchers (Garcia-Mila et al., 2013; Kelly & Takao, 2002; Sandoval & Millwood, 2005;

Schwarz et al., 2003). Some researchers have focused in large part on the role of counterarguments and rebuttals in generating argument complexity. For example, a young-earth creationist's short rebuttal that evolution is false because it contradicts scripture is not particularly sophisticated. The mere inclusion of any given argument element does not lead to sophistication, however the exclusion of many elements would likely create a less sophisticated product; many schemes for argument complexity acknowledge this (Garcia-Mila et al., 2013; Schwarz et al., 2003).

In this study I was concerned with argument complexity and the role that constructing a body of evidence plays in generating complexity. Prior work has been focused on the role of rebuttals and constructing counterarguments; these have often been positioned as some of the highest level performances. This is probably because rebuttals tend to generate the discourse we look for in verbal argumentation, and responding to counterarguments can produce epistemic products for evaluation. This is worthwhile of course, particularly in a school setting. However, it may be the case the role of evidence is underrepresented. Given the number of articles scientists and other professionals read it would seem that reasoning about evidence, especially evidence gathered by other epistemic peers participating in a community, is an important element of the scientific process, one that has not received a lot of attention to date. This is not to say that evidence has not played a role, it clearly does otherwise the claims, arguments and rebuttals described in other research would represent mere sophistry. The claim here is that constructing a body of evidence can generate additional argument complexity in ways that have not been previously documented. I make the case that students who saw

connections between pieces of evidence tended to cite and provide more reasons about that evidence, and that the role of evidence is worth unpacking in greater detail.

3.5 Conclusion and Implications

Students evaluated and reevaluated evidence, guided by their own implicit criteria, and motivated their written arguments by using evidence, particularly looking for lines of converging evidence, developing a body of evidence, or focusing on the strongest empirical study, as a means of justifying their stance on whether or not humans can be resistant to HIV.

Based on the results it is clear that students shifted their critical evaluation of the quality of the evidence over time. Critical evaluation of science texts is an important part of scientific literacy. “For students to be scientifically literate, they must not only remember what science texts say, but also take a critical stance toward those texts” (Phillips & Norris, 1999). Critical evaluations were in part driven by students’ implicit ideals for what counts as good evidence. The coding of the students’ written work reveals in some detail the kinds of criteria that seventh-grade students bring to the classroom. A limitation of this study is that it focused on the criteria that students bring with them to the classroom which is a subset of the criteria that scientists use and that many of their criteria are very domain or even topic specific.

It seems to be the case that very small scaffolds can promote a great deal of high quality reasoning. For example, in Evidence 1 the students were asked what their stance was on the value of various cat species as potentially useful experimental organisms for exploring immunodeficiency virus differences. This single question drove students to engage with disciplinary and topical questions about the similarities between humans and

cats and about the potential complex interplay of host-pathogen relations, suggesting that short, carefully designed questions can drive student thinking in ways that promote evaluation of evidence and sometimes the integration of evidence resulting in students developing a body of evidence.

Hands-on style inquiry will not be possible at all times and for all topics taught in a science classroom. The approach presented here is an alternative to learning about science through expository texts, as is often the case when field or laboratory work is not possible, and instead presents a way for teachers to still engage their students in core scientific practices like evaluating evidence and developing arguments in support of a claim or model. An approach using APL and JRV can provide the opportunity to grapple with complex issues in a scientific way. Moreover, this approach is commensurate with calls for increasing students' contact with uncertainty and conflicting perspectives (Allchin 2011, Britt, Richter, & Rouet, 2014, Goldberg, 2013).

Developing the conceptual links that fuse together multiple pieces of evidence into a coherent body of evidence is a sophisticated task. It was surprising to find that so many students engaged in this kind of high level cognitive work. This skill was not explicitly taught to students but the potential for it was embedded in the context of the evidence and claims the students evaluated. This suggests that this is a skill that could be taught more explicitly with additional scaffolds.

Existing frameworks for evaluating student reasoning (Berland & McNeill, 2010; Chinn & Brewer, 2001; Osborne, Erduran, & Simon, 2004; Osborne, Henderson, MacPherson, Szu, Wild, & Yao, 2016) do not include (a) evidence evaluation based on criteria; (b) re-evaluating evidence in light of new evidence; and (c) combining pieces of

evidence to make a new body of evidence. Improvements to normative accounts of good reasoning in science classes could be made with the inclusion of these practices.

3.6 Acknowledgments

I would like to thank the many teachers, administrators, and research assistants who have had a hand in shaping, refining and contributing to the course of the learning environment design I have presented here. This material is based upon work supported by the National Science Foundation under Grant No. 1008634. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

3.7 References

- Anmarkrud, Ø., Bråten, I., & Strømsø, H. I. (2014). Multiple-documents literacy: Strategic processing, source awareness, and argumentation when reading multiple conflicting documents. *Learning and Individual Differences*, 30, 64-76.
- Ault, C. R. (1998). Criteria of excellence for geological inquiry: The necessity of ambiguity. *Journal of Research in Science Teaching*, 35(2), 189-212.
- Berland, L. K., & McNeill, K. L. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. *Science Education*, 94(5), 765-793.
- Braasch, J. L., Rouet, J. F., Vibert, N., & Britt, M. A. (2012). Readers' use of source information in text comprehension. *Memory & Cognition*, 40(3), 450-465.
- Bråten, I., Britt, M. A., Strømsø, H. I., & Rouet, J. F. (2011). The role of epistemic beliefs in the comprehension of multiple expository texts: Toward an integrated model. *Educational Psychologist*, 46(1), 48-70.
- Britt, M. A., & Aglinskias, C. (2002). Improving students' ability to identify and use source information. *Cognition and Instruction*, 20(4), 485-522.
- Britt, M. A., Richter, T., & Rouet, J. F. (2014). Scientific literacy: The role of goal-directed reading and evaluation in understanding scientific information. *Educational Psychologist*, 49(2), 104-122.
- Bromme, R., & Goldman, S. R. (2014). The public's bounded understanding of science. *Educational Psychologist*, 49(2), 59-69.
- Bromme, R., Kienhues, D., & Porsch, T. (2010). Who knows what and who can we believe? Epistemological beliefs are beliefs about knowledge (mostly) to be attained from others. In L. D. Bendixen & F. Feucht (Eds.), *Personal epistemology in the classroom: Theory, research, and implications for practice*. Cambridge University Press.
- Bromme, R., Thomm, E., & Wolf, V. (2015). From understanding to deference: laypersons' and medical students' views on conflicts within medicine. *International Journal of Science Education, Part B*, 5(1), 68-91.
- Cerdán, R., & Vidal-Abarca, E. (2008). The effects of tasks on integrating information from multiple documents. *Journal of Educational Psychology*, 100(1), 209.
- Chinn, C. A., & Brewer, W. F. (2001). Models of data: A theory of how people evaluate data. *Cognition and Instruction*, 19(3), 323-393.
- Chinn, C. A., & Buckland, L. A. (2011). Differences in epistemic practices among scientists, young earth creationists, intelligent design creationists, and the scientist-creationists of Darwin's era. In R. S. Taylor & M. Ferrari (Eds.), *Epistemology and science education: understanding the evolution vs. intelligent design controversy*, (pp. 38-76). Routledge.

- Chinn, C. A., Buckland, L. A., & Samarapungavan, A. L. A. (2011). Expanding the dimensions of epistemic cognition: Arguments from philosophy and psychology. *Educational Psychologist*, 46(3), 141-167. doi: 10.1080/00461520.2011.587722
- Chinn, C. A., & Rinehart, R. W. (2016). Commentary: Advances in research on sourcing—source credibility and reliable processes for producing knowledge claims. *Reading and Writing*, 29(8), 1701-1717.
- Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education*, 84(3), 287-312.
- Duschl, R. A. (1990). *Restructuring science education: The importance of theories and their development*. Teachers College Press.
- Duschl, R. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of Research in Education*, 32(1), 268-291.
- Elby, A., & Hammer, D. (2001). On the substance of a sophisticated epistemology. *Science Education*, 85(5), 554-567.
- Feinstein, N. (2011). Salvaging science literacy. *Science Education*, 95(1), 168-185.
- Feinstein, N. W., Allen, S., & Jenkins, E. (2013). Outside the pipeline: Reimagining science education for nonscientists. *Science*, 340(6130), 314-317.
- Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., ... & Strijbos, J. W. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research*, 2(3), 28-45.
- Garcia-Mila, M., Gilabert, S., Erduran, S., & Felton, M. (2013). The effect of argumentative task goal on the quality of argumentative discourse. *Science Education*, 97(4), 497-523.
- Goldman, S. R., & Bisanz, G. L. (2002). Toward a functional analysis of scientific genres: Implications for understanding and learning processes. In J. Otero, J. León, & A. C. Graesser (Eds.), *The psychology of science text comprehension*, (pp. 19-50). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- Hapgood, S., Magnusson, S. J., & Palincsar, A. S. (2004). Teacher, text, and experience: A case of young children's scientific inquiry. *The Journal of the Learning Sciences*, 13(4), 455-505.
- Hardwig, J. (1985). Epistemic dependence. *The Journal of Philosophy*, 82(7), 335-349.
- Hardwig, J. (1991). The role of trust in knowledge. *The Journal of Philosophy*, 88(12), 693-708.
- Hofer, B. K. (2000). Dimensionality and disciplinary differences in personal epistemology. *Contemporary Educational Psychology*, 25(4), 378-405.
- Kelly, G. J., & Takao, A. (2002). Epistemic levels in argument: An analysis of university oceanography students' use of evidence in writing. *Science Education*, 86(3), 314-342.
- Keselman, A., Hundal, S., Chentsova-Dutton, Y., Bibi, R., & Edelman, J. A. (2015). The relationship between biology classes and biological reasoning and common health

- misconceptions. *The American Biology Teacher*, 77(3), 170-175.
- Kuhn, Deanna. "Thinking as argument." *Harvard Educational Review* 62.2 (1992): 155-179.
- Lackey, J. (2008). *Learning from words: Testimony as a source of knowledge*. Oxford University Press on Demand.
- Lackey, J. (2011). Testimonial knowledge. In S. Bernecker & D. Pritchard (Eds.), *The Routledge companion to epistemology* (pp. 316-325). Routledge.
- Linn, M. C., Clark, D., & Slotta, J. D. (2003). WISE design for knowledge integration. *Science Education*, 87(4), 517-538.
- McNeill, K. L., & Krajcik, J. (2007). Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations. In M. C. Lovett & P. Shah (Eds.), *Thinking with data*, (pp. 233-265). Mahwah, NJ: Lawrence Erlbaum Associates.
- Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction—what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching*, 47(4), 474-496.
- Muis, K. R. (2007). The role of epistemic beliefs in self-regulated learning. *Educational Psychologist*, 42(3), 173-190.
- Origgi, G. (2004). Is trust an epistemological notion?. *Episteme*, 1(01), 61-72.
- Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching*, 41(10), 994-1020.
- Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S. Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching*, 53(6), 821-846.
- Palincsar, A. S., & Magnusson, S. J. (2001). The interplay of first-hand and second-hand investigations to model and support the development of scientific knowledge and reasoning. In S. M. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 151-193). Mahwah, NJ: Lawrence Erlbaum Associates.
- Phillips, L. M., & Norris, S. P. (1999). Interpreting popular reports of science: what happens when the reader's world meets the world on paper?. *International Journal of Science Education*, 21(3), 317-327.
- Phillips, L. M., & Norris, S. P. (2009). Bridging the gap between the language of science and the language of school science through the use of adapted primary literature. *Research in Science Education*, 39(3), 313-319.
- Rinehart, R. W., Duncan, R. G., & Chinn, C. A. (2014). A scaffolding suite to support evidence-based modeling and argumentation. *Science Scope*, 38(4), 70-77.
- Rinehart, R. W., Duncan, R. G., Chinn, C. A., Atkins, T. A., & DiBenedetti, J. (2016). Critical design decisions for successful model-based inquiry in science classrooms. *International Journal of Designs for Learning*, 7(2).

- Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education*, 92(3), 447-472.
- Sandoval, W.A., & Çam, A. (2010). Elementary children's judgments of the epistemic status of sources of justification. *Science Education*, 95(3), 383-408.
- Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and instruction*, 23(1), 23-55.
- Schwarz, B. B., Neuman, Y., Gil, J., & Ilya, M. (2003). Construction of collective and individual knowledge in argumentative activity. *The Journal of the Learning Sciences*, 12(2), 219-256.
- Sober, E. (2008). *Evidence and evolution: The logic behind the science*. Cambridge University Press.
- Suthers, D. D., & Hundhausen, C. D. (2003). An experimental study of the effects of representational guidance on collaborative learning processes. *The Journal of the Learning Sciences*, 12(2), 183-218.
- Tenopir, C., King, D. W., & Bush, A. (2004). Medical faculty's use of print and electronic journals: changes over time and in comparison with scientists. *Journal of the Medical Library Association*, 92(2), 233.
- Tenopir, C., King, D. W., Boyce, P., Grayson, M., & Paulson, K. L. (2005). Relying on electronic journals: Reading patterns of astronomers. *Journal of the Association for Information Science and Technology*, 56(8), 786-802.
- Tenopir, C., King, D. W., Edwards, S., & Wu, L. (2009, January). Electronic journals and changes in scholarly article seeking and reading patterns. In *Aslib proceedings 61(1)*, 5-32. Emerald Group Publishing Limited.
- Thagard, P. (2000). *How scientists explain disease*. Princeton University Press.
- Thomm, E., & Bromme, R. (2012). "It should at least seem scientific!" Textual features of "scientificness" and their impact on lay assessments of online information. *Science Education*, 96(2), 187-211.
- Toth, E. E., Suthers, D. D., & Lesgold, A. M. (2002). "Mapping to know": The effects of representational guidance and reflective assessment on scientific inquiry. *Science Education*, 86(2), 264-286.
- Toulmin, S. (1958). *The uses of argument*. Cambridge University Press.
- Weber, K., Inglis, M., & Mejia-Ramos, J. P. (2014). How mathematicians obtain conviction: Implications for mathematics instruction and research on epistemic cognition. *Educational Psychologist*, 49(1), 36-58.
- Wiley, J., Goldman, S. R., Graesser, A. C., Sanchez, C. A., Ash, I. K., & Hemmerich, J. A. (2009). Source evaluation, comprehension, and learning in Internet science inquiry tasks. *American Educational Research Journal*, 46(4), 1060-1106.
- Windschitl, M. (2008). What is inquiry? A framework for thinking about authentic scientific practice in the classroom. In J. Luft & R. L. Bell (Eds.), *Science as inquiry in the secondary setting* (pp. 1-20). NSTA Press.

- Windschitl, M., Thompson, J., & Braaten, M. (2008a). Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. *Science Education*, 92(5), 941-967.
- Wineburg, S. S. (1991). Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology*, 83(1), 73.
- Yarden, A. (2009). Reading scientific texts: Adapting primary literature for promoting scientific literacy. *Research in Science Education*, 39(3), 307-311.
- Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review*, 20(1), 99-149.

Chapter 4: Using Evidence to Develop and Refine Models of Inheritance

Abstract

The Next Generation Science Standards emphasize learning science through engagement with authentic scientific practices such as modeling and reasoning with evidence. This paper presents the results of a three-day modeling activity in which seventh-grade life science students developed their own models of inheritance in response to multiple evidence sets. Research in genetics education has shown that developing robust conceptual understandings of inheritance patterns and their underlying mechanism is challenging for students due to the abstract and invisible nature of genes. I found that students were capable of developing models that: (a) were consistent with evidence; (b) were internally consistent; (c) increased in their use of causal mechanisms; and (d) increased in their consistency with normative explanations of inheritance. Students' abilities to correctly make predictions about novel inheritance problems significantly increased over time. This modeling activity facilitated the development of more sophisticated student thinking about genes, traits, and patterns of inheritance.

4.1 Introduction

Developing explanatory models is a core scientific practice aimed at helping scientists represent and make sense of the world (Giere, 2004; Godfrey-Smith, 2006). These models often include unobservable (theoretical) entities and processes that are part of the normative explanations of phenomena, and thus can be used to test hypotheses and make predictions about the natural world (Machamer, Darden, & Craver, 2000). When a model fails to explain some aspect of existing data, it can help pinpoint gaps in our understanding of the phenomenon; thus, modeling can support the development of new knowledge (Driver, Leach, Millar, & Scott, 1996; Lesh & Lehrer, 2003; Zohar & Nemet, 2002).

Modeling, the development and refinement of explanatory models, is a challenging practice for biology students because many models describe causal relations that include unobservable (theoretical) entities that exist at different levels of biological organization (sub-cellular, cellular, tissue etc.) and processes that exist across multiple temporal and spatial scales (Duncan, 2007; Horwitz, 1996). Reasoning about models and evidence related to genetics and inheritance has proven to be particularly difficult for students because they include macro level phenomena (e.g., traits), unfamiliar micro level phenomena (e.g., genes, alleles), multiple forms of representation (e.g., pedigrees, Punnett Squares), unfamiliar terminology (e.g., heterozygous, homozygous) and span multiple temporal scales stretching from days to multiple generations (Bahar, Johnstone, & Hansell, 1999; Cartier, 2000). Developing understanding and facility with foundational knowledge in genetics and inheritance is a significant cognitive achievement for students.

The research presented here involved seventh-grade students who were just beginning to learn genetics and inheritance. The aim of this lesson was to engage students

with developing explanatory models of the rules of inheritance for simple Mendelian traits. Students reasoned about observable traits and attempted to develop rules for the “unseen” level of mechanisms that drive genetic phenomena. They did this over several days as they attempted to develop and revise models to explain evidence of increasing complexity, including evidence that presented new anomalies that needed to be resolved if the students were to be able to successfully explain the phenomenon and make novel predictions based on their models.

4.1.1 Designing for Learning from Anomalous Data

The conceptual underpinning for the design of this lesson was that students might be able to successfully revise their models of the rules of inheritance by repeatedly encountering carefully designed, anomalous evidence, and then engaging in model revision until their models could account for the anomalies. Recognizing and resolving anomalies is a core practice for scientists (Chinn & Brewer, 1993; Darden, 1992) including biologists and geneticists (Machamer, Darden, & Craver, 2000). The introduction of subtle anomalies, like those described later in this paper, for students to resolve while engaging in model revision creates a significant potential for students to fail on four different fronts. First, students may not recognize that an anomaly exists (Darden, 1992). Second, even if students recognize the anomaly, they may engage with a variety of stances that allow them to avoid resolving it, by ignoring, rejecting, or reinterpreting the anomaly to maintain their present set of beliefs, thereby obviating any need for revising the models they are developing (Chinn & Brewer, 1993). Third, even if they recognize that such an anomaly exists and make a good faith effort to resolve it, they may simply be unable to do so because they lack some element of prior knowledge that is

necessary for successful resolution. Finally, they may be unable to conceive of the right class of explanatory model that would allow them to resolve the anomaly.

It was conceivable that seventh-grade students might be unable to construct a scientifically normative theory of inheritance that could account for counterintuitive observations like traits skipping generations. Such explanations need to leverage conceptions of alleles, independent assortment, and the probabilistic nature of inheritance that students without prior genetics instruction, as was the case with these students, do not have. However, I anticipated that even if students were not initially successful in developing fully normative models of Mendelian inheritance, their initial failures could possibly become productive (Kapur, 2008; Kapur & Bielaczyc, 2012) as they continued with the modeling activity, because their initial failures would highlight the gaps in their understanding and the conceptual issues that a successful model must solve.

The design of this lesson was intended to facilitate the development of more sophisticated student thinking about genes, traits, and patterns of inheritance by progressively constraining the range of theoretical options for explaining inheritance with carefully designed evidence that systematically introduced anomalies. The general approach followed several steps: (a) students generated a model based on simple evidence; (b) students were then confronted with new anomalous evidence that their existing models could not explain; then (c) students revised their models in light of this new evidence. Steps (b) and (c) can be repeated until students have reached a desired end state with their models; for this particular study students engaged with two cycles of (b) and (c).

Much of the success or failure of this kind of modeling hinges on the design and presentation order of the evidence. The evidence presented during each repetition of steps (b) and (c) needs to rule out, or constrain, certain unproductive elements of students' models. Ordering of the evidence is crucial so that just enough of the students' models still remain that they can rebuild upon them to form newly revised models. Also, well designed evidence, and tasks for using it, provide refinement pressure in two directions. There is external pressure to align the model with the evidence, and there is internal pressure to make sure that the elements of the model are coherent. Repeating the process can lead to a refined model. Next, I will describe the design rationale and principles for an instructional approach that is commensurate with a productive success approach (Kapur, 2016) to model-based inquiry.

4.1.2 Designing for Productive Success

Designing for productive success using model-based inquiry techniques requires a learning environment design that strikes a balance between success and failure, scaffolding and challenge. This approach exists within the “vast design space” between direct instruction at one extreme and purely unguided discovery at the other (Kapur, 2016, p. 289). Kapur identified four categories of design possibilities including: productive failure, unproductive failure, productive success, and unproductive success. Productive success is intended to promote student learning in both the short and long term (Kapur, 2016). A prototypical example is Problem Based Learning (PBL) environments where students' prior knowledge might be low, but appropriate scaffolds are provided so that with time, students develop problem solving success and learn about targeted concepts (Kapur, 2016). The lesson presented here is commensurate with

productive success in that it is a form of guided inquiry intended to promote successful problem solving that leads to learning.

Prior research in genetics education has suggested that empirical and conceptual assessment of models is valuable to teach to students (Cartier, 2000), and I argue it may also serve well for evaluating the quality of student modeling. Empirical assessment involves explaining data and making predictions with a model, while conceptual assessment involves coordinating a model with other models and established knowledge in the domain (Cartier, 2000). With respect to empirical assessment for this lesson, I was interested in the consistency between students' models of inheritance and the evidence sets. Prior research has suggested that distinguishing between evidence and models is challenging for students (Kuhn, 1989). Therefore, it was an open question as to whether students would do well with inquiry in this particular learning environment. For conceptual assessment, I was interested in the internal consistency (i.e., the consistency of one model element in relation to others) of students' models. The lesson design for this research was based on the assumption that if students could experience success with developing their models, success being that the models were consistent with the available evidence and that elements within the model were coherent (i.e., non-contradictory), then students might be able to develop the disciplinary knowledge (e.g., bi-parental contribution of genetic material, connections between genotypes and phenotypes, and alleles) needed to fully explain all of the evidence and make predictions about new inheritance problems. The evidence was developed so that only a normative model of simple Mendelian inheritance could explain all of the features of the evidence. Particulars of the design of the evidence are described in greater detail in the methods. Next I review

several criteria that good models for inheritance should adhere to; these criteria were used to analyze the quality of students' models, including (a) evidence to model connections; (b) internal coherence; and (c) linking phenotypes with genotypes.

4.1.3 Evidence to Model Connections

In their review of criteria that scientists use for judging models, Pluta, Chinn & Duncan (2011) asserted, among other criteria, that “good models are consistent with empirical evidence” (Pluta et al., 2011, p. 486). However, the simultaneous coordination of evidence and explanatory model construction is a cognitively demanding task. Some researchers have argued that inquiry-based instruction, like the lesson presented here, is not productive for reasons relating to the cognitive load placed on the learner (Kirschner, Sweller, & Clark, 2006). In this research, student performance was taken to be sophisticated when students' models were consistent with the available evidence. Successfully using a model to account for all of the evidence would be a sign that such instruction is productive for learning.

The evidence used by students in this study was in the form of pedigrees. Pedigrees can be used to find out if a trait is autosomal dominant, autosomal recessive, sex-linked dominant, and so on across different modes of inheritance. In this case students were tasked with reasoning about multiple pedigrees that reflected an autosomal recessive pattern of inheritance. A trait that is autosomal recessive is a trait that can be passed from two parents who do not have the trait, only a single gene for the trait, to an offspring who inherits a recessive gene from each parent and then expresses the trait. This pattern of inheritance was selected because students would need to explain the “skip a generation” phenomenon that can occur with simple Mendelian recessive traits.

Successfully explaining this phenomenon requires thinking about theoretical entities such as alleles, and generating inferences based on those entities such as the link between a genotype and phenotype based on a particular combination of alleles. Students who exhibit such thinking have made progress in developing knowledge in the domain of genetics and inheritance.

4.1.4 Developing Internally Consistent Models

Internal consistency is an important criterion for scientific models (Cartier, 2000). When a model is internally consistent “...all the elements or assumptions of the model fit with one another without contradiction...” (Cartier, 2000, p. 4). In their review of model criteria used by scientists, Pluta and colleagues wrote that scientists believe that “Good models have high levels of conceptual coherence and clarity” (Pluta et al., 2011, p. 486). When describing the work of scientists, Windschitl and colleagues argued that “...they are all engaged in the same knowledge-building pursuit—the development of coherent and comprehensive explanations through the testing of models” (Windschitl, Thompson, & Braaten, 2008, p. 945). Successfully developing a coherent model would be evidence that the form of inquiry-oriented instruction used in this lesson is productive for learning, contrary to claims that this might be an ineffective form of instruction (Kirschner et al., 2006).

In her review of strategies used to resolve scientific anomalies, Darden argued that when models are revised, some elements are maintained, whereas others are modified or removed. This balancing of retained, revised, and rejected elements provides additional constraints on the problem solver (Darden, 1992). Productively working with these constraints in model revision activities in genetics has proven challenging for some

high school students (Cartier, 2000; Johnson & Stewart, 2002). Given that the students in this study were middle schoolers there were reasons to believe that their success might be more limited.

Pluta et al. (2011) found that middle-school students value coherent models; a value that is commensurate with those of practicing scientists. Valuing this property of models, however, does not entail the skill of producing an internally consistent model. Prior research has shown that students struggle with developing models that are internally consistent while engaged in modeling about problems in genetics (Cartier, 2000). Cartier found that while high school students (Grade 10) studying genetics displayed skill at making evidence to model connections, they were less apt to evaluate models in terms of their internal consistency. It was plausible that students in this study would generate models of inheritance (i.e., collections of multiple rules for inheritance) that were internally inconsistent. It was an open question as to whether or not seventh-grade students could develop internally consistent models in a learning environment that was designed to present them with a sequence of anomalous data that needed to be explained by their models.

4.1.5 Linking the “Seen” and the “Unseen”

Recognizing the role of unseen mechanisms, and generating models that can explain patterns in data based on hypothetical entities, are important steps in developing sophisticated conceptions in the domain of genetics, and more broadly speaking in science (Stewart, Cartier, & Passmore, 2005; Windschitl et al., 2008). Students often struggle to develop robust understandings of foundational concepts in genetics such as independent assortment, the roles of alleles, and linking genotypes to phenotypes

(Venville, Gribble, & Donovan, 2005; Venville & Treagust, 1998). There are many reasons for these difficulties. First, it is the case that many of the entities involved are invisible and unfamiliar (e.g., genes, alleles, chromosomes). Second, it is also the case that reasoning well about problems in genetics involves processes that span multiple time and space scales (Bahar, Johnstone, & Hansell, 1999; Horwitz, 1996; Tsui & Treagust, 2003). Third, reasoning well in genetics requires the coordination of concepts across two ontologically distinct levels, an information level and a physical level (Duncan & Reiser, 2007). Both the information and physical levels can contain unseen elements. Genes exist at the informational level, while “proteins, cells, tissues, etc.” operate at the level of physical organization in organisms (Duncan & Reiser, 2007, p. 939).

Given the challenging nature of teaching and learning in this domain (Kindfield, 1994; Lewis & Wood-Robinson, 2000; Stewart & Van Kirk, 1990), most of the interventions designed to support learning of inheritance through a model-based inquiry approach have targeted the high-school level and have often involved specialized pieces of software like Biologica or the Genetics Construction Kit (Buckley et al., 2004; Cartier & Stewart, 2000). The approach taken here, however, does not use specialized software and instead focuses on carefully crafted pedigrees to support middle-school students’ sense making about genes, alleles, and inheritance based on the pedigree evidence.

Venville and colleagues have investigated how high school students’ conceptions about genes change over time with instruction (Venville, Gribble, & Donovan, 2005; Venville & Treagust, 1998). Based on interviews with Grade 10 students taking a ten week genetics course, they found that students’ conceptions about the role of genes in producing particular phenotypes existed at five levels: (a) no definite conception of the

link between genes and phenotype; (b) genes as a passive particle; (c) genes as an active particle involved in producing a particular phenotype; (d) genes as a sequence of instructions; and (e) genes as a productive sequence of instructions for producing proteins (Venville et al., 2005; Venville & Treagust, 1998). Most students either had no definite conception of the link between genotype and phenotype or had a view of genes as passive particles.

This is not to say that students' prior conceptions are not useful, but rather that they may serve either to facilitate or impede the development of a more sophisticated view of the inheritance process (Mbajjorgu et al., 2006). For example, a student may have prior knowledge that will facilitate learning if they know that both parents contribute an equal amount of genetic material to their offspring. Conversely, a student may have heard through popular media that genes can "cause" certain traits. Such a view does not embrace the probabilistic nature of genetics and thus may impede learning. Consequently, helping students develop robust models of genetic inheritance that involve genotypic thinking and the probabilistic nature of inheritance is not a trivial instructional goal.

Being able to develop genotypic thinking is a step toward developing an understanding of how genes function as information that impacts phenotypes. A lack of genotypic thinking becomes apparent when students "...do not understand that a single phenotype could map to two genotypes" (Slack & Stewart, 1990, p. 64). Johnson and Stewart (2002) found that among high school students solving inheritance problems, the least successful groups tended to operate at a very shallow phenotypic level. This

underscores the need to help students develop genotypic thinking if the aim is to help them see the deeper mechanisms that underlie genetic phenomena.

In this research I looked for evidence that students could connect genotypes with phenotypes, and perhaps more importantly, that they could mobilize the concept of an allele to explain novel inheritance patterns. While the intent was for students to experience some initial success with the modeling elements of the lesson, there was also a significant potential for students to fail in this task. First, prior research has shown that students often operate at only the phenotypic level (Johnson & Stewart, 2002), and it would have been possible for students to reason about information in a pedigree at the level of the phenomenon (i.e., the mother has a particular trait and gives it to her daughter). Second, students had not previously learned about alleles, but rather were given the opportunity to develop the concept on their own through attempting to explain data that would be anomalous to someone armed with a non-allelic mode of inheritance. Being able to describe patterns in pedigrees as a function of the inheritance of different alleles allows students to expand their models to encompass “unseen” mechanisms that have greater explanatory power than a merely phenotypic model. Overall, the lesson was designed with the aim of helping students move from models that were largely phenotypic to more mechanistic genotypic models.

4.1.6 The Present Study

This study was part of the larger Promoting Reasoning and Conceptual Change in Science (PRACCIS) project, which engaged multiple middle schools, their teachers, and students, to promote inquiry learning. The project included the development of numerous instructional units on a variety of biology topics such as evolution, cell organelles, and

genetics. Units included a suite of instructional scaffolds to promote students' engagement with the practices of model-based inquiry (Rinehart, Duncan, & Chinn, 2014).

This study investigated how seventh-grade students developed new understandings of the biological mechanisms that govern gene-trait inheritance patterns by modeling the “rules” of inheritance, using three sets of scientific evidence in the form of family trees (pedigrees). Further, in this study I examined how middle-school students engaged in the practices of modeling, such as model development, evidence to model coordination, and model revision, while developing their knowledge of genetics and inheritance.

The research questions for this study included:

1. Are students' rules consistent with the available evidence and does consistency change as the complexity of the evidence increases over time?
2. Are students' models internally coherent (i.e., do they contain rules that contradict one another?) and does the degree of coherence change as the complexity of the evidence increases over time?
3. To what extent do student models change over time with respect to their ability to develop causal accounts of inheritance?
4. Can students make useful predictions about novel inheritance problems based on their rules?

4.2 Methods

This instructional intervention was carried out with 242 students in the classes of four 7th grade science teachers in a middle-class suburban middle school. Students

eligible for free and reduced lunch made up 14% of the total population.

Demographically the school's students were 61% Caucasian, 28% Asian, 6% Hispanic, and 5% African-American. The lessons were conducted in the classrooms of four teachers, who had a wide range of teaching experience. The least experienced teacher was in her 4th year of teaching at the school, the second teacher was in her 7th year, and the most experienced teacher had more than fifteen years of experience. Two teachers completed all three days of Lesson 3 in all of their class periods. One teacher completed the lesson in three of her five classes, but because of other scheduling considerations could not complete the final day in her other two classes. The fourth teacher enacted the lesson but failed to retain students' written data. From the remaining pool of students, 141 were fully consented and completed all three days of the lesson. These students were included for analysis.

4.2.1 Timeline and Noteworthy Activities

The important activities that made up the first five days of the genetics unit are chronicled in detail in Table 4.1. Students began Lesson 1, which occurred on Day 1, with a brief activity to stimulate their thinking about which traits are genetic and which are not, as well as how inheritance occurs. Students developed their own initial models of inheritance and wrote a brief explanation. For Lesson 2, which occurred on Day 2, students' models were collected and some were categorized to fit common alternative concepts about inheritance, for example that girls only get their traits from their mothers. A handout of four common alternative conceptions, as drawn and described by the students, was provided to students. Students used an evidence packet that was specially designed to rule out the alternative conceptions. One model was left that was similar to a

simple Mendelian model of inheritance, although it was highly incomplete. It focused on the notion that sometimes there is a strong gene and that people get a trait when they get the strong gene. Lesson 3, which occurred over Days 3, 4, and 5, is the critical lesson and will be the focus of the analysis for this paper.

During the first day of Lesson 3, students learned how to read and interpret pedigrees; then they received evidence in the form of three pedigrees of three different families with Cystic Fibrosis (CF), an autosomal recessive trait (i.e., an inherited disorder) that affects the lungs. A total of nine pedigrees were used by students over the course of the lesson to develop their rules of inheritance. Each day the pedigrees became more complex and included more generations and family members. On Day 3 students received Evidence Set 1, as shown in Figure 4.1. This evidence set has three pedigrees that only include children and parents. The pedigrees follow the standard form of representation which includes the following:

1. Females are shown as circles.
2. Males are shown as squares.
3. Affected individuals (i.e., people with cystic fibrosis) are shown in gray.
4. Unaffected individuals (i.e., people without cystic fibrosis) are shown in white.
5. Generations are indicated with branched vertical lines.
6. Parent sets are indicated by connecting horizontal lines.

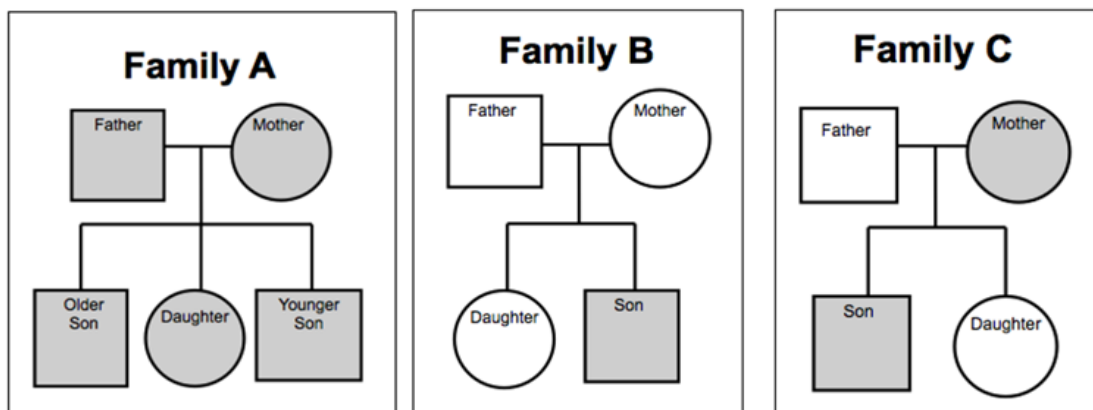


Figure 4.1. Evidence Set 1, which students received on Day 3

Atop the pedigree page were the following instructions:

“Each pedigree shows a family of a mother, father and three children. As scientists your job is to figure out how the disorder is being passed on from one generation to the next. You need to figure out what gene or genes each person has and which gene or genes cause the disorder. Below are three different pedigrees. For each person in each family write down what gene or genes you think they have in their box or circle.”

One aim of these instructions was to provide students with a subtle hint to try to separate out the phenotype that is shown in the color of the object (circle or square) from the genes that are driving the expression of that phenotype.

After seeing the pedigrees, students responded to the prompt “Write down your rules for how the strong gene model works. These rules should describe how genes and traits are inherited in families. Remember, the rules have to work for all individuals in all three families!” The aim of these instructions was to provide students with an opportunity to generate an initial model (a model built out of rules) with principles that could later be revised. Further, students were directed to solve pedigree problems about family members that were not included in their initial pedigrees, as seen in Figure 4.2. This type of prediction questions was designed to find out if students had reached a point where

they could correctly solve novel pedigrees. As students had opportunities to revise their models of the rules of inheritance, it was hypothesized that their ability to correctly solve pedigrees might increase.

Question 3: For Family B, based on your model, what would you predict about the father's parents? Will they have the disorder? What genes will they have? Be sure to give reasons for your answers!

Father's Father: _____

Father's Mother: _____

Family B

```

graph TD
    FF[Father's Father ?] --- FM[Father's Mother ?]
    FM --- F[Father]
    F --- M[Mother]
    F --- M --- D[Daughter]
    F --- M --- S[Son]
    style FF stroke-dasharray: 5 5
    style FM stroke-dasharray: 5 5
    style S fill:#ccc
  
```

Figure 4.2. A prediction question from the first day of Lesson 3

On Day 4 students received Evidence Set 2, which had the three pedigrees shown in Figure 4.3. These pedigrees followed the same families (A, B, C) from the previous day but also included the grandparent generation. The inclusion of the grandparent generation was critical, because it highlighted what can be called a *skip-a-generation* phenomenon. In this phenomenon, a phenotypic appearance of a trait seems to disappear between generations. So, for example, cystic fibrosis may be present in the grandparent generation and the grandchild generation, but it might completely skip the parent generation in the middle of the pedigree. Family B exemplifies this and was designed to provide students with the opportunity to reason about how genes might transfer information between generations even when the trait does not seem to be present. Family C was also carefully designed so that there was a partial skip-a-generation phenomenon

where the paternal line, when considered in isolation from the maternal line, also contained a generational skip. The grandmother had the disease but the grandfather did not, the father did not but one of the grandchildren did. Being able to explain the skip-a-generation phenomenon was important because a successful explanation, one that could explain features of the other pedigrees as well, required students to leverage the deeper mechanisms of inheritance, namely the idea that alleles can be “hidden” and that certain combinations of alleles result in the expression of different phenotypes.

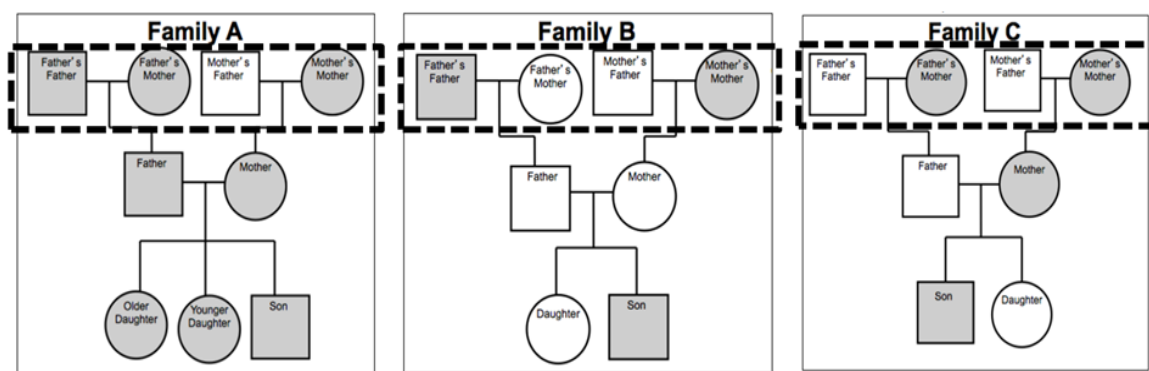


Figure 4.3. Evidence Set 2, which includes the family members from Day 1 plus the older grandparent generation

Evidence Set 3, shown in Figure 4.4, was composed of three pedigrees further elaborated to include aunts, uncles, and cousins. All the pedigrees reflected a recessive inheritance pattern (students were not told this) and had at least one member of each family affected with CF.

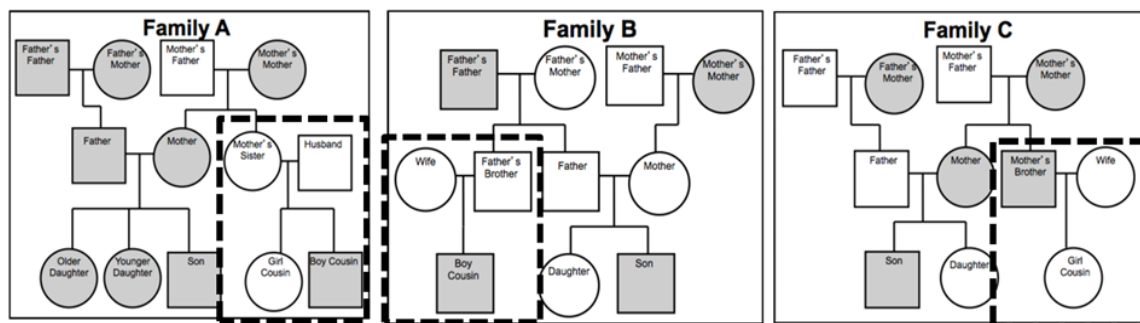


Figure 4.4. Evidence Set 3, which included the family members from Evidence Sets 1 and 2, plus aunts, uncles and cousins

Each day during the course of the lesson the students were asked to generate or revise a set of rules of inheritance (their conceptual models) that could be used to explain the accumulating evidence; thus students revised their models three times. Students generated their rules individually but discussed them in pairs or groups of four. The type of work, individual, pair, group, or class, is indicated in Table 4.1 for each of the activities across Lesson 3.

Table 4.1

A brief summary of the first five days of the genetics lesson

<u>Activity</u>	<u>Sequence</u>	<u>Summary</u>	<u>Noteworthy Details</u>
Lesson 1: Initial models of inheritance	Day 1	Students examined a list of traits like hair color, eye color, having kidneys, speaking French, and riding a bike, and were asked which traits are genetic and which are not. They discussed their choices in pairs. Then students individually responded to the prompt shown in Activity 2 (to the right), drew a model of inheritance, and explained it in writing.	<p>Activity 1: What is a genetic trait?</p> <p>Activity 2: Develop a model that explains how a genetic trait, like dimples, gets passed on from parents to their children. Write or draw your model.</p>
Lesson 2: Ruling out some initial models of inheritance	Day 2	Students were presented with four student generated models from the previous day. Students were given an evidence packet with 3 pieces of evidence in the form of pedigrees and asked to rule out any models that could not explain the evidence.	<p>Activity 3: Rule out alternative models.</p> <p>One of the models was a “sex model” in which girls get traits only from their mothers. The evidence packet was constructed to rule out this possibility as well as others. The “strong gene” model (a simple Mendelian model) was the only model not ruled out. This model was drawn and explained by a student, so typical genetics terminology (recessive, homozygous etc.) was absent.</p>

<u>Activity</u>	<u>Sequence</u>	<u>Summary</u>	<u>Noteworthy Details</u>
Lesson 3: Developing the rules of inheritance	Day 3	<p>The teacher presented a 16 slide interactive PowerPoint that contained materials designed to help students learn to read and reason about pedigrees. Students began with simple instructions on the representational elements (i.e., circle is a female, a shaded square means that someone has Cystic Fibrosis) and built toward relational elements (i.e., vertical lines indicate generations and parentage).</p> <p>A reminder slide at the beginning of the PowerPoint set the stage for what had been learned up to that point including:</p> <ol style="list-style-type: none"> 1. There is a connection between genes and some traits. 2. We know that some traits are determined by our genes, some by our environment and some traits are a mix of both. 3. We get our genes from our parents. 4. The strong gene model is the best model <u>so far</u> to explain how children inherit traits from their parents. <p>Students completed Activities 4 and 5 described on the right.</p>	<p>*Activity 4: Rules of inheritance, version 1</p> <p>Students were introduced to pedigrees of three families who have an inherited disease expressed in some family members and not in others (cystic fibrosis).</p> <p>Students responded individually to the following prompt: “Write down your rules for how the strong gene model works. These rules should describe how genes and traits are inherited in families. <u>Remember, the rules have to work for all individuals in all three families!</u>”</p> <p>Students were given five lines to write the rules. A few students wrote more than five rules and some wrote less.</p> <p>**Activity 5: Solve pedigree problems, version 1</p> <p>After writing their rules, students were asked to solve three pedigrees. The pedigrees asked students to make predictions about the grandparent generation (whose disease status was unknown to the students at this time). Students solved them individually first and then shared their answers with their table partner.</p> <p>The teacher did not give feedback to students about the correctness of their pedigree problems.</p>

<u>Activity</u>	<u>Sequence</u>	<u>Summary</u>	<u>Noteworthy Details</u>
Lesson 3: Developing the rules of inheritance	Day 4	<p>The class period began with a “do now” problem (projected at the front of the room) in which students tried to solve a “skip a generation” problem, where grandparents and grandchildren have a trait but the parent generation in the middle does not.</p> <p>The teacher used a 3 slide PowerPoint to build on students’ knowledge of pedigrees to include additional family relations like aunts, uncles and cousins.</p> <p>Students completed Activities 6-10 described on the right.</p>	<p>Activity 6: Do Now #1 “The grandfather and the son do not have dimples, but the father does. What is your best explanation for how this happens? Write your answer on the lines provided.” Students answered in their packets. The teacher did not provide feedback for the do now.</p> <p>Activity 7: Explain an inheritance pattern After completing the do now, students received new pedigrees for the three families which included the disease status of the grandparents, which was unknown the day before. Students individually responded to the following prompt for Families A and B: “Since you now know whether the parents of the father have the disorder, what gene or genes do you think they have? Be sure to give reasons for your answers!”</p> <p>*Activity 8: Rules of inheritance, version 2 Next students were asked to individually write their rules, the same as in Activity 4: “Write down your revised rules for how the strong gene model works. These rules should describe how genes and traits are inherited in families. <u>Remember, the rules have to work for all individuals in all three families!</u>”</p> <p>**Activity 9: Solve pedigree problems, version 3 Students worked individually on a pedigree with extended family members (aunts, uncles, cousins) and shared their answer after completing the problem.</p> <p>Activity 10: Rules of inheritance, version 3 Students formed groups of four and wrote out a revised set of inheritance rules using the same prompt as before.</p>

<u>Activity</u>	<u>Sequence</u>	<u>Summary</u>	<u>Noteworthy Details</u>
Lesson 3: Developing the rules of inheritance	Day 5	<p>The class period began with a “do now” problem (projected at the front of the room) in which students tried to solve a pedigree problem for a new family.</p> <p>Students received new pedigrees for the three families that now included aunts, uncles, and cousins. They had made predictions about the disease status and genes some of these family members on the previous day.</p> <p>Students completed Activities 11-15 described on the right.</p>	<p>Activity 11 Do Now #2 Students began class with a new pedigree that included grandparent, parent, and child generations (the 4th family to be introduced).</p> <p>**Activity 12: Solve pedigree problems, version 4 Students made predictions about the possible offspring of two parents from one of the families. They were also asked to explain how when one parent has the disorder and another does not, sometimes the children have the disorder and sometimes they do not.</p> <p>*Activity 13: Rules of inheritance, version 4 Students discussed their rules in groups or with a partner, then wrote their rules individually.</p> <p>Activity 14: Rules of inheritance, version 5 Students shared the rules of inheritance they had developed with the teacher and contributed to a class list on the board.</p> <p>Activity 15: Solve pedigree problems, version 5 Students were asked to reason about a final pedigree for a 5th family that was new to the students. They were asked to answer questions about the genes and disease status of some of the children in the family.</p>
<p>* Indicates items that were coded and analyzed</p> <p>** One of the pedigrees was analyzed</p>			

4.2.2 Data Collection and Analysis

I collected multiple sources of data from participating teachers' classrooms, including students' written artifacts and classroom video. I analyzed written artifacts from 141 students enrolled in three different teachers' classrooms. Two kinds of written data were analyzed. These are indicated in Table 4.1 and will be described here. The first kind of written data were students' rules of inheritance. These were completed five times over three days as indicated in Table 4.1. Students responded to the prompt "Write down your rules for how the strong gene model works. These rules should describe how genes and traits are inherited in families. Remember, the rules have to work for all individuals in all three families!" on the first day and were asked on subsequent days to revise their initial models. I analyzed three sets of rules generated by students in response to Evidence 1, 2, and 3. Rule set one was developed on Day 3, rule set two was developed on Day 4 and rule set three was developed on Day 5, the final day of the lesson. The analyzed rule sets are marked with a single asterisk in Table 4.1.

The second kind of written data were the prediction questions that students responded to each day. Three of the prediction questions were analyzed, one from each day of Lesson 3, and all three are included in Appendix A. The prediction questions that were analyzed are indicated in Table 4.1 with a double asterisk.

4.2.3 Coding

Each individual rule was coded for: (a) consistency with the available evidence (shown in Table 4.2 below); (b) internal consistency within the set of rules (i.e., did the rule contradict any of the student's other rules in that set, shown in Table 4.3 below); (c) connections between the proposed genotype (genes) of individuals and their traits (shown

in Table 4.4 below); and (d) references to a notion of 2-allele combinations (e.g., a pair of genes) for each trait (shown in Table 4.5 below). Two independent raters coded about 80% of the data and about 20% of the data were coded by a third rater. Intercoder reliability for each rule was between 86-97% for each of the four categories of codes. Any disagreements were discussed and resolved and code assignments revised accordingly.

4.2.3.1 Coding for consistency with evidence.

Students' individual rules within their rule sets were coded for how well their models matched with the pedigree evidence for that particular day. Some of the rules within a given model were simple but accurate, like "Both parents contribute genes to the child." Other rules were inconsistent with the evidence, like "Children get the disorder from the parent of the opposite gender." Rules were given a "Yes" code for accurate statements and a "No" code for inaccurate statements.

Table 4.2		
<i>Is the student's rule consistent with the available evidence?</i>		
<u>Code</u>	<u>Definition</u>	<u>Examples</u>
Yes	The student's rule is consistent with the evidence they have seen up to that point.	"The kids get the hidden gene for the disease from the parents." "There is a hidden part of the disease."
No	The student's rule is inconsistent with the evidence they have seen up to that point.	"A child is going to have it." "Boys get the disorder from the mother."

4.2.3.2 Coding for internal coherence.

Students' individual rules within a rule set were coded for internal coherence.

Rules were considered to be coherent when they did not contain statements that were mutually exclusive with other rules. A rule received a "Yes" code when the rule contradicted other rules in the set and a "No" code when the rule did not contradict any other rules. A rule set was fully coherent if it received all "No" codes. Examples of contradictory and non-contradictory rules are in Table 4.3.

Table 4.3		
<i>Does the student's rule contradict other rules within the rule set?</i>		
<u>Code</u>	<u>Definition</u>	<u>Examples</u>
Yes	Two or more of the student's rules contradict each other.	<p>Rule 1: "The disease can skip generations but not always."</p> <p>Rule 2: "At least one person in each generation has it."</p>
No	The student's rule does not contradict other rules within the rule set.	<p>Rule 1: "The disease can skip generations."</p> <p>Rule 2: "It is possible to get the disease if the parents have it."</p>

4.2.3.3 Coding for genotype to phenotype connections.

Students' individual rules within a rule set were coded for connections between genotype and phenotype. A rule received a "Yes" code when it clearly connected the concept of a gene with a phenotype (trait). It received a "No" when the rule did not connect a genotype with a phenotype. Examples of rules that made genotype to

phenotype connections, and rules that did not make this connection, are included in Table 4.4.

Table 4.4		
<i>Does the rule make a connection between the proposed genotype (genes) of individuals and their traits?</i>		
<u>Code</u>	<u>Definition</u>	<u>Examples</u>
Yes	The student's rule clearly maps between genotype and phenotype.	<p>"At least one parent carries the hidden gene if they have the disorder."</p> <p>"People with the disorder have two weaker genes."</p>
No	The student's rule does not map between genotype and phenotype.	<p>"Parents don't have to have it for the child to have it."</p> <p>"Someone had to pass down the gene to the next generation."</p>

4.2.3.4 Coding for pairs of genes (alleles).

Students' individual rules within a rule set were coded for the number of genes described in the rule. A rule received a "Yes" code when it clearly described two genes for a trait. It received a "Mixed" code if the rule only mentioned a single gene. It received a "No" code when the rule did not mention any genes. Examples of rules that received these codes are included in Table 4.5.

Table 4.5		
<i>Does the student's rule reflect the notion of 2-allele combinations (e.g., a pair of genes) for each trait?</i>		
<u>Code</u>	<u>Definition</u>	<u>Examples</u>
Yes	The student's rule uses a "genes-in-pairs" mode of inheritance.	"To get the disease you must get both of the weak genes." "Both parents give a gene to the child."
Mixed	The student's rule is a single gene model of inheritance.	"They pass on the disease gene to the kids." "Even if they don't have it they could still pass the gene."
No	The student's rule does not have a clear connection to genes as mediators of inheritance. This is a no-gene model of inheritance.	"Someone in the family must have the disease." "Both parents have the disease then all of the kids have it."

4.2.3.5 Prediction Question Coding: Correctness.

Students' prediction questions were coded for correctness. A student response received a "Correct" code when it made a correct prediction for that problem. It received a "Partially correct" code if the student's prediction was less than fully correct. It received a "No" code when prediction was clearly wrong. Examples are included in Table 4.6. All three pedigree problems are included in Appendix A.

Table 4.6		
<i>Does the student make a correct prediction?</i>		
<u>Code</u>	<u>Definition</u>	<u>Examples</u>
Correct	The student correctly answers the prediction question. If more than one answer choice was possible a single correct answer was given full credit.	“Yes, because both parents have the disease they can both only pass down a Y gene. This way the child should have two Y genes therefore he or she will have the disease.”
Partially Correct	The student only partially correctly answers the prediction question.	“One parent can only have the strong gene so one kid can only get the disorder because it’s only dominant in one unless both parents have the disorder.”
Incorrect	The student does not make a correct prediction.	“the disease was not strong enough to be passed down”

4.3 Results

Based on the written data from students’ worksheets, field notes, observations, and teachers’ reports, students were able to independently generate rules even for the first set of evidence. Analysis of the data indicates that students could readily come up with initial sets of rules with little teacher intervention. Students’ models typically included between 3-5 rules. An example of one student’s initial rule is, “If both parents have it, all the children can have it” where “it” refers to the disease cystic fibrosis. Later rules for some students tended to be more sophisticated and included concepts about the relationship between genes and visible traits as well as the probability of the genes being passed on. These will be described in more detail later.

I analyzed students’ rules at two different levels. First, I analyzed the content of students’ rules at the level of an individual rule (i.e., a single statement). A rule level analysis provides a fine grained account of the shifts in the frequency and predominance

of certain kinds of rules. Second, I analyzed students' rules at the level of a model (i.e., a set of rules taken together to be a rule-based model of genetics). Analysis at the student level allows for investigations into how the overall model changed (or did not change) over time.

4.3.1 Changes in Students' Rules: Consistency with Evidence

The pedigrees students encountered in the first evidence set were less complex, in terms of the number of generations and family members featured, than the pedigrees they received as part of the second and third evidence sets. Students' rules were consistent with the evidence 78% of the time for the first evidence set, as shown in Table 4.7. For the third evidence set, 82% of their rules were consistent with the evidence. This was a counterintuitive result because the evidence became significantly more complex on subsequent days, and one might predict that performance on this measure would decrease as there was more evidence to contend with. I checked for consistency both at a descriptive and predictive level. Descriptive consistency concerns the fit between the student's rule and the pedigree evidence. As an example, a statement like "The disease skipped the parent generation" is descriptive. Predictive statements include a speculation about something not depicted in the pedigree evidence. To be clear, these are not statements about the prediction problems that students completed later in their packets, rather they are statements made in reference to the pedigrees they used as evidence to generate their rules and construct their model of inheritance. Examples of predictive statements include "If one parent has the disease and one does not, it is possible for the offspring to have the disease" and "If neither parent has the disease it is still possible for a child to have the disease."

Table 4.7			
<i>Are the rules consistent with the evidence?</i>			
<u>Code</u>	<u>Evidence Set 1</u>	<u>Evidence Set 2</u>	<u>Evidence Set 3</u>
	(n = 521)	(n = 607)	(n = 565)
Yes	78%	79%	82%
No	22%	21%	18%
<i>Note.</i> Students were encouraged to develop 3-5 rules each but could develop as many as they wanted. Therefore the sample size for each evidence set is different.			

An analysis of the individual students' rule sets at three time points is presented in Table 4.8. A rule set was considered to be fully consistent when all of the rules in the rule set were consistent with the evidence. A rule set was mostly consistent when more than half of the rules were consistent with the evidence and only partly consistent when half or fewer of the rules were consistent with the evidence. A fully inconsistent rule set was one where the student had no rules that were consistent with the evidence. Results showed that 80% or more of students' rule sets were fully or mostly consistent with the evidence across all three days.

Table 4.8			
<i>Are students' models consistent with the evidence?</i>			
<u>Code</u>	<u>Evidence Set 1</u>	<u>Evidence Set 2</u>	<u>Evidence Set 3</u>
	(n = 141)	(n = 141)	(n = 141)
Fully consistent	41%	40%	50%
Mostly consistent	45%	45%	30%
Partly consistent	11%	11%	14%
Fully inconsistent	3%	3%	4%
<i>Note.</i> Percentages may not equal 100% due to rounding.			

4.3.2 Changes in Students' Rules: Internal Coherence

Given the difficulty of this task, it was plausible that students would generate sets of rules that were internally inconsistent. For example, if a student's rule stated "Every generation must have someone with the disease" and they also, within the same rule set, stated that "The disease sometimes skips a generation," then they would both be coded as internally contradictory. However, the students in this study were able to generate sets of rules that were internally consistent, and they showed some improvement in this ability with extended practice with evidence. For Evidence Set 1 students' rules were consistent 93% of the time, and for Evidence Set 3 their rule sets were internally consistent 97% of the time, as shown in Table 4.9, despite a substantial increase in the evidence's complexity.

Table 4.9			
<i>Does the rule contradict other rules?</i>			
<u>Code</u>	<u>Evidence Set 1</u>	<u>Evidence Set 2</u>	<u>Evidence Set 3</u>
	(n = 521)	(n = 607)	(n = 565)
Yes	7%	4%	3%
No	93%	96%	97%
<i>Note.</i> Students were encouraged to develop 3-5 rules each but could develop as many as they wanted. Therefore the sample size for each evidence set is different.			

Students' models were analyzed for internal coherence as shown in Table 4.10. If a model (i.e., set of rules) contained any internal contradictions it was coded as "contains contradictions." Most students' models contained no contradictions, with 85% of models internally consistent after the first evidence set, rising to 94% by the third evidence set. Prior research has identified internal consistency of a model as an important consideration (Cartier et al., 2005). This shows that even though the complexity of the evidence increased, many students were capable of producing models that were internally consistent.

Table 4.10			
<i>Does the student's model have internal contradictions?</i>			
<u>Code</u>	<u>Evidence Set 1</u>	<u>Evidence Set 2</u>	<u>Evidence Set 3</u>
	(n = 141)	(n = 141)	(n = 141)
Contains contradictions	15%	9%	6%
Contains no contradictions	85%	91%	94%

4.3.3 Changes in Students' Rules: Genotype to Phenotype Connections

Initially students generated rule sets that were largely phenotypic; only after reasoning about further evidence did some move toward more mechanistic accounts. For example, Figure 4.5 shows a set of rules derived from Evidence Set 1 that was purely phenomenological; the rules only described the potential disease status of an individual (or the generation, i.e., parents) without specifying anything about the role of invisible mechanisms and theoretical entities like the inheritance of different combinations of alleles. That is, they described what you can see, and not why it happens. The second rule set in Figure 4.5 shows a shift from phenomenological rules to rules that attempted to account for the underlying mechanism, that of invisible genes that determine traits. The first rule for Evidence Set 2 shows that the student considered genes as occurring in pairs (i.e., two alleles), whereas in the initial rule set from Evidence Set 1 there was no evidence of this form of thinking.

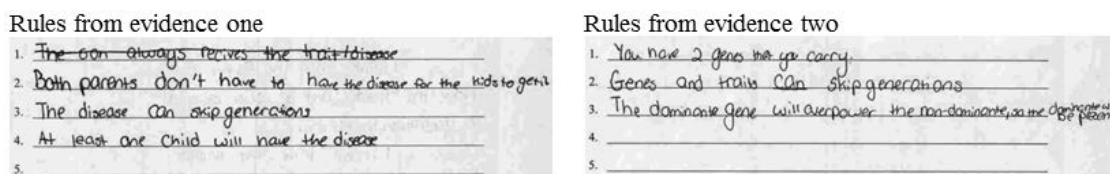


Figure 4.5. A single student's initial and intermediate rule sets

Other shifts in student thinking are also evident in Figure 4.5. Based on Evidence Set 1, the student wrote that “The disease can skip generations.” In the next iteration, for Evidence Set 2, the student upgraded this rule to say “Genes and traits can skip generations.” Three important shifts have happened here. First, the student acknowledged the role of “genes” as the unobservable mechanistic entity driving trait determination. Second, the student changed from using the term “disease” to the term “traits”, potentially suggesting the development of a more generalized notion of inheritance. Finally, the student noted that the “genes” and the “traits” can both “skip” a generation, whereas they initially said that the “disease” can skip a generation.

There are at least two conceptual shifts that must occur for students to see this connection: (a) that a theoretical entity (i.e., genotype) is possible; and (b) that it has a causal connection to the trait expressed (i.e., phenotype). Moving from a phenotype-only concept (i.e., only recognizing the feature, which in this case is a disease) to a genotype-to-phenotype view is a major conceptual change for middle-school students and represents a significant cognitive achievement. As shown in Table 4.11, there was a steady increase in the number of students' rules that connected phenotype to genotype. It is important to note that not all rules needed to connect genotype to phenotype for a rule to be productive. For example, a student could have a rule (and in fact many had this rule) that “A trait can skip a generation.” This rule did not connect genotype to phenotype

and yet it was still productive. In fact, a professional genetic counselor could make this statement. In this instance students' thinking was quite similar to expert thinking.

Table 4.11			
<i>Does the rule connect the genotype to the phenotype?</i>			
<u>Code</u>	<u>Evidence Set 1</u>	<u>Evidence Set 2</u>	<u>Evidence Set 3</u>
	(n = 521)	(n = 607)	(n = 565)
Yes	15%	25%	43%
No	85%	75%	57%
<i>Note.</i> Students were encouraged to develop 3-5 rules each but could develop as many as they wanted. Therefore the sample size for each evidence set is different.			

When analyzed at the level of students' rule sets, students' work progressed in a way that was similar to the individual rule level. Their rules increasingly showed a more genotypic approach, with fewer rule sets focused only on phenotype. The percentage of students who had at least one rule that connected genotype to phenotype increased from 33% to 67%, as shown in Table 4.12. Being able to engage in genotypic thinking is a hallmark of increasing sophistication of students learning genetics (Slack & Stewart, 1990) and the data shows that while not all students made this shift, approximately one-third of them did make the shift.

Table 4.12			
<i>Does the student's model connect genotype to phenotype?</i>			
<u>Code</u>	<u>Evidence Set 1</u>	<u>Evidence Set 2</u>	<u>Evidence Set 3</u>
	(n = 141)	(n = 141)	(n = 141)
At least one rule in the model connects genotype to phenotype	33%	56%	67%
No rules in the model connect genotype to phenotype	67%	44%	33%

4.3.4 Changes in Students' Rules: Genes Occurring in Pairs

Very few student rules initially made reference to genes occurring in pairs. Rather, many students' first rule sets could be described as being no-gene models of inheritance. In these sets of rules it was the trait that was inherited, with no conception of genes being the mediators of inheritance. For example, students had rules like "The trait is passed down" or "Sons get the disease from their mothers." Between the no-gene kind of rule and a fully normative 2-allele rule was an intermediate ground with at least two different kinds of rules, which were: (a) single gene rules; and (b) rules with an unspecified number of genes. For a single gene rule, the traits were described as being attached to a single discrete gene; a trait was inherited from one parent or the other, but not both. An example rule would be "The sons get the diseased gene from their mother." The second type of intermediate rule had an unspecified number of genes as mediators for the disease. For example, "The diseased genes are passed on to the children" or "They passed

on the disease genes.” These types of intermediate rules were not particularly productive because they made it difficult to account for how traits skip a generation. The single gene model could not explain how, if neither parent had the gene for a particular trait, it would be possible for their children to have the trait. The unspecified number of genes model was less explicitly problematic in explaining the skipped generation phenomenon, but its lack of precision rendered it less productive as well.

As students continued working with evidence, they showed increasing levels of sophistication by abandoning less productive no-gene or intermediate rules and developing, adopting or adapting a greater number of rules that treated genes as pairs. The total number of two-allele rules used by students increased from approximately 5% for Evidence Set 1 to about 31% for Evidence Set 3, as shown in Table 4.13. Correspondingly, the number of no-gene rules dropped from 71% for Evidence Set 1 to 60% for Evidence Set 3.

Table 4.13			
<i>How many genes are described by the rule?</i>			
<u>Code</u>	<u>Evidence Set 1</u>	<u>Evidence Set 2</u>	<u>Evidence Set 3</u>
	(n = 521)	(n = 607)	(n = 565)
Two	5%	22%	31%
Intermediate	24%	18%	9%
Zero	71%	59%	60%
<i>Note.</i> Students were encouraged to develop 3-5 rules each, but could develop as many as they wanted. Therefore the sample size for each evidence set is different.			

When analyzed at the level of an individual student, as shown in Table 4.14, the number of students who included in their model at least one rule that reflected an allelic conception of inheritance increased from 11% to 67%. There was a corresponding decrease in the number of students who had a zero gene concept of inheritance from an initial level of 51% down to 18% for evidence set 3. The pattern of results shows that students were making a shift from thinking of traits as being disconnected from genes (zero gene rules) to traits as being determined by particular combinations of alleles.

Table 4.14			
<i>Does the student's model make use of the concept of alleles?</i>			
<u>Number of Genes</u>	<u>Evidence Set 1</u>	<u>Evidence Set 2</u>	<u>Evidence Set 3</u>
	(n = 141)	(n = 141)	(n = 141)
Two	11%	56%	67%
Intermediate	38%	30%	15%
Zero	51%	14%	18%

4.3.5 Prediction Correctness

After reviewing the findings above one might ask, “If students have productive, or at least semi-productive, rules, can they make useful predictions based on those rules?” On each of the three days when students were asked to generate rules, as described above, they were also asked to make predictions. These predictions involved students’ reasoning about family members they had not seen yet. On each day students were asked

to complete one or more prediction questions based on a pedigree. Generally, each prediction question asked students about one or more people that had not previously been included in the three family pedigrees.

Students' answers to the prediction questions were scored based on a three part scale. A correct answer was worth two points, a partially correct answer was worth one point, and an incorrect answer was scored zero points. For example, if a student were to say that the mother's brother (i.e., an uncle) received a gene that would not lead to CF from his father and a gene that would lead to CF from his mother, then the student's response would be scored correct (two points) because it contained one possible solution to the problem. There were two prediction questions scored for each day of the intervention; therefore on any given day a student could score between 0 and 4 points for the correctness of their predictions. Students' responses ($N = 141$) were scored each day, for a total possible 564 points each day. The points earned and the percentage of the total possible earned are shown in Table 4.15. Overall, 126 of the 141 students showed score improvements on Day 5 compared to their first score on Day 3.

Table 4.15		
<i>Students' points earned on pedigree predictions</i>		
<u>Day</u>	<u>Points Earned</u>	<u>Percent of Total Possible Points Earned</u>
Day 3	67	11.9%
Day 4	168	29.8%
Day 5	435	77.1%

A one-way within subjects repeated measures ANOVA was conducted to compare students' scores on the prediction questions across all three days. Students' scores were

square root transformed to improve the normality of the data. Mauchley's test of sphericity did not show a significant result (Mauchley's $W = 0.967$, $df = 2$, $p = 0.097$) indicating that the assumption of sphericity was met with respect to the ANOVA test. There was a significant change in students' scores across all three days ($F = 235.546$, $df = 2$, $p < .001$) with a large effect size ($\eta^2 = 0.627$). Pairwise comparisons showed that there was a significant increase in students' scores from Day 3 to Day 4 ($p < .001$) and from Day 4 to Day 5 ($p < 0.001$). Results from the test show that students significantly improved in their ability to make correct predictions about novel pedigree problems over time.

4.4 Discussion and Implications

Empirical fit between model and evidence is an important criterion for evaluating the quality of a model (Cartier, 2000). Students in this study initially developed models that were largely consistent with the evidence, and their performance did not diminish as the evidence increased in complexity. Students were able to generate rules—elements within their models—that successfully accounted for the available evidence. Many students had initial success with coordinating evidence and models, a finding that is not always supported in the literature (Kuhn, 1989).

Beyond making sure that their models fit with the evidence, a large percentage of students in this study developed models that were highly internally coherent. Internal coherence is an important property of scientific models (Pluta et. al, 2011; Windschitl et al., 2008). When contrasting the relative occurrence of empirical versus conceptual assessment (i.e., conceptual assessment includes internal consistency of a model), Cartier (2000) reported that well-conducted empirical assessments of a model were more

common among high school students than conceptual assessments, even after explicit instruction on conceptual assessments. The middle-school students in this study did not receive explicit instruction on conceptual assessment, but they did receive prompts pushing them to align their models with the evidence. It was therefore somewhat surprising that so many students developed models that were conceptually coherent.

Many students had success with the empirical and conceptual elements of modeling, but did success with modeling entail later gains with developing conceptual knowledge in genetics? Prior research on the differences between novices and experts has shown that novices tend to focus on surface features when explaining a phenomenon, whereas experts look for deeper structural elements of a phenomenon (Chi, Feltovich, & Glaser, 1981). It was possible for students in this study to end where they began, thinking only about phenotype (i.e., a surface feature) in the absence of genotype (i.e., a deeper feature). Genotypic thinking, seeing the connections between a genotype and a phenotype, is an important step in developing knowledge about genetics and inheritance (Slack & Stewart, 1990). Students' genotypic thinking was evident in their models when they described particular combinations of alleles leading to one outcome or another. An implication of this research is that genotypic thinking can follow from, rather than be required for, successful model-based inquiry in genetics and inheritance. Phrased differently, genotypic thinking can be a form of disciplinary thinking developed by students during model-based inquiry designed with productive success in mind.

Prior research has shown that developing genotypic thinking in high school students is difficult (Slack & Stewart, 1990). By the end of this three day intervention, the majority of middle-school students developed rules that connected genotype to phenotype

and considered the allelic nature of genes. This is not to say that students' genotypic thinking was robust across a wide range of genetic phenomena. A limitation of this study is that it involved only the first few days of genetics instruction and students had not encountered other inheritance patterns (e.g. sex linked, codominant and so on). However, the results do indicate that some students started to develop genotypic thinking, at least for the limited range of phenomena they had encountered.

An alternative to thinking about increasing sophistication in students' understanding of genetics is to focus on their ontological commitments about the role of genes in genetics and inheritance rather than genotypic thinking. In Venville & Treagust's (1998) study of the ontological shifts in students' theoretical commitments with respect to the roles of genes in inheritance, they found five distinct levels including (a) no definite conception; (b) genes as a passive particle; (c) genes as an active particle; (d) genes as a set of instructions; and (e) genes as productive instructions. At the end of a naturalistic survey of a 10 week genetics course for grade 10 students they found that most, 20 out of 29 students, had a view of genes as active particles. Very few students (only 7) were found to be at the two higher levels, and two students still viewed genes as a passive particle (i.e. one of the lowest levels). In short, many of the students were in the transitional intermediate levels of the framework. Many of the seventh-grade students in this study, 94 in total, attained the view of genes as an active particle. Instruction within the scope of this lesson, which was only three days long, was not intended to promote genes as instructions or teach the molecular side of genes as a productive set of instructions for making proteins with certain functions. In all it was encouraging to see that seventh-grade students could rapidly progress to a view of genes as active particles.

As students' models increased in sophistication and explanatory scope, their ability to solve novel pedigree problems increased. At the end of the first day of instruction only 11% of the prediction problems were solved correctly, however, there was a steady increase in the number of correct predictions made across all three days of instruction, culminating in 77% of students making correct predictions about novel problems at the end of the final day. This provides some support for the idea that modeling provided students with opportunities to develop normative knowledge of inheritance that could help them solve new problems. A limitation of this study is that no longer term measures of maintenance were administered to see how students would perform on a delayed post-test.

The instructional approach in this study was intended to promote learning in a way that is commensurate with conceptualizations of productive success, the goal of which is "...to learn through a successful problem solving activity itself" (Kapur, 2016, p. 294). The design of this study drove student learning by first providing students with the opportunity to engage in modeling and then providing students with a carefully designed sequence of anomalous evidence that would rule out some of the least productive elements of their models. Then, by attending to the sequence of evidence, students were able to fill gaps in their models while still aligning their models with the evidence. Part of the power of this approach lies not just in the construction of a model and its alignment with evidence, but also in the push to rule out elements of the model that are no longer viable. An implication of this research is that it takes time and careful lesson design to help students rule out unworkable ideas and refine their remaining workable ones.

A limitation of this study is that students were only offered a single resolution strategy when faced with anomalous evidence: to change their theory. Darden (1992) posited that anomaly resolution is a central feature of science and that scientists employ a variety of techniques for anomaly resolution, some of which rely on experimental procedures like verifying that an anomaly actually exists by attempting to reproduce it. Students in this study were using non-experimental data and as a result could not reproduce anomalies. It might be the case that a full exploration of the range of students' anomaly resolution strategies is better addressed in a real experimental setting or at the very least a simulated experimental setting, like those found in the Genetics Construction Kit Software (Slack & Stewart, 1990). Further research would be needed to address the potential interaction between model modifications and various anomaly resolution strategies.

Overall, I was encouraged to find that students readily engaged with the task and were able to develop rules, and these rules became better aligned with the canonical scientific account of inheritance patterns as students continued to revise them. In sum, this paper has presented evidence that engagement in scientific modeling can shift students' thinking from less productive phenomenological descriptions toward more sophisticated and causal accounts of inheritance. Given the strong emphasis on the melding of content and practice in the Next Generation Science Standards (NGSS Lead States, 2013), providing accounts of how students can learn novel content by engaging in relatively authentic modeling tasks is useful. This study illustrates one successful account of such learning.

4.5 Acknowledgments

I would like to thank the many teachers, administrators, and research assistants who have had a hand in shaping, refining and contributing to the course of the learning environment design I have presented here. This material is based upon work supported by the National Science Foundation under Grant No. 1008634. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

4.6 References

- Bahar, M., Johnstone, A. H., & Hansell, M. H. (1999). Revisiting learning difficulties in biology. *Journal of Biological Education*, 33(2), 84-86.
- Buckley, B. C., Gobert, J. D., Kindfield, A. C. H., Horwitz, P., Tinker, R. F., Gerlits, B., Wilensky, U., Dede, C., & Willett, J. (2004). Model-based teaching and learning with BioLogica: What do they learn? How do they learn? How do we know? *Journal of Science Education and Technology*, 13(1), 23-41.
- Cartier, J. (2000, April). Assessment of explanatory models in genetics: Insights into students' conceptions of scientific models (Res. Rep. 98-1). Madison, WI: National Center for Improving Student Learning and Achievement in Mathematics and Science.
- Cartier, J. L., & Stewart, J. (2000). Teaching the nature of inquiry: Further developments in a high school genetics curriculum. *Science & Education*, 9(3), 247-267.
- Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of educational research*, 63(1), 1-49.
- Darden, L. (1992). Strategies for anomaly resolution: Diagnosis and redesign. In R. Giere (Ed.), *Cognitive Models of Science* (pp. 251-273). Minnesota Studies in the Philosophy of Science, Minneapolis, MN: University of Minnesota Press.
- Driver, R., Leach, J., Millar, R., & Scott, P. (1996). *Young people's images of science*. Bristol, PA: Open 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.
- Duncan, R. G., Rogat, A. D., & Yarden, A. (2009). A learning progression for deepening students' understandings of modern genetics across the 5th-10th grades. *Journal of Research in Science Teaching*, 46(6), 655-674.
- Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science*, 71, 742-752.
- Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology and Philosophy*, 21, 725-740.
- Hafner, R. & Stewart, J. (1995). Revising explanatory models to accomdate anomalous genetic phenomena: Problem solving in the "context of discovery". *Science Education*, 79(2) 111-146.
- Horwitz, P. (1996). Teaching science at multiple space time scales. *Communications of the ACM*, 39(8), 100-102.
- Kapur, M. (2008). Productive failure. *Cognition and Instruction*, 26(3), 379-424.
- Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *The Journal of the Learning Sciences*, 21,(1), 45-83.
- Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist*, 51(2), 289-299.
- Kindfield, A. C. (1994). Biology diagrams: Tools to think with. *The Journal of the Learning Sciences*, 3(1), 1-36.

- Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75-86.
- Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review, 96*(4), 674-689.
- Lesh, P., & Lehrer, R. (2003). Models and modeling perspectives on the development of students and teachers. *Mathematical Thinking and Learning, 5*, 109-129.
- Lewis, J., & Wood-Robinson, C. (2000). Genes, chromosomes, cell division and inheritance —Do students see any relationship? *International Journal of Science Education, 22*(2), 177-195.
- Machamer, P. K., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1-25.
- Mbajiorgu, N., Ezechi, N., & Idoko, C. (2006). Addressing nonscientific presuppositions in genetics using a conceptual change strategy. *Science Education, 9*(3), 419-438.
- NGSS Lead States. 2013. *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.
- Pluta, W. J., Chinn, C. A., & Duncan, R. G. (2011). Learners' epistemic criteria for good scientific models. *Journal of Research in Science Teaching, 48*(5), 486-511.
- Rinehart, R. W., Duncan, R. G., & Chinn, C. A. (2014). A scaffolding suite to support evidence-based modeling and argumentation. *Science Scope, 38*(4), 70-77.
- Slack, S. J., & Stewart, J. (1990). High school students' problem-solving performance on realistic genetics problems. *Journal of Research in Science Teaching, 27*(1), 55-67.
- Stewart, J., Cartier, J. L., & Passmore, C. M. (2005). Developing understanding through model-based inquiry. In M. S. Donovan and J. D. Bransford (Eds), *How students learn: Science in the classroom* (pp. 515-565). Washington, DC: The National Academies Press.
- Stewart, J., & Kirk, J. V. (1990). Understanding and problem-solving in classical genetics. *International Journal of Science Education, 12*(5), 575-588.
- Tsui, C., & Treagust, D. F. (2003). Genetics reasoning with multiple external representations. *Research in Science Education, 33*(1), 111-135.
- 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.
- Venville, G., Gribble, S., & Donovan, J. (2004). An exploration of young children's understandings of genetics concepts from ontological and epistemological perspectives. *Science Education, 89*(4), 614-633.
- von Aufschnaiter, C., Eruduran, S., Osborne, J., & Simon, S. (2008) Arguing to learn and learning to argue: Case studies of how students' argumentation relates to their scientific knowledge. *Journal of Research in Science Teaching, 45*, 101-131.
- Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching, 39*(1), 35-62.

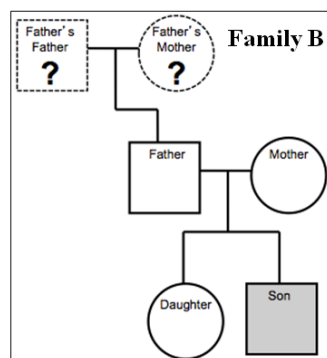
4.7 Appendix A

Day 3 Prediction Questions

Question 3: For Family B, based on your model, what would you predict about the father's parents? Will they have the disorder? What genes will they have? Be sure to give reasons for your answers!

Father's Father: _____

Father's Mother: _____

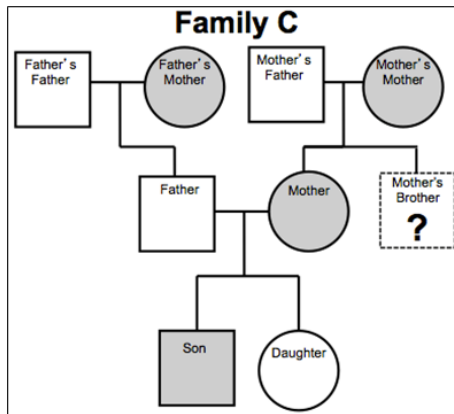


Day 4 Prediction Questions

Question 5: In Family C, the mother has a brother who is shown on the right by the box with the “?”

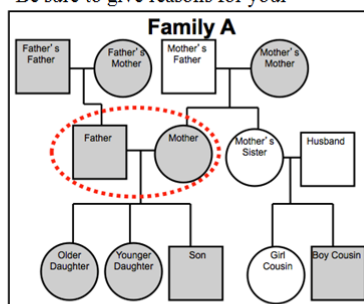
Based on your model, what would you predict about the genes of the mother's brother? Be sure to give reasons for your answers!

Will the mother's brother have the disorder? Be sure to give reasons for your answer!

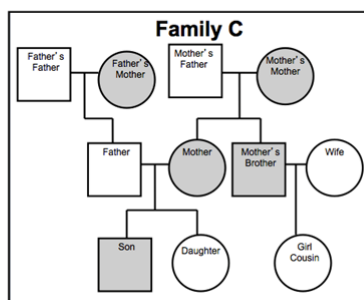


Day 5 Prediction Questions

Question 1: The two parents in family A (shown in the circle) decided to have a fourth child. Would you predict that this child would have the disorder? Be sure to give reasons for your answer!



Question 2: In family C when one parent has the disorder but the other does not, sometimes the children have the disorder and sometimes they do not. Explain why this occurs.



Chapter 5: Conclusion

5.1 Introduction

A confluence of, until recently, separate research traditions has contributed to a need for research in the domain of how students reason about and reason with evidence while learning science (Britt, Richter, & Rouet, 2014; Bromme & Goldman, 2014; Feinstein, Allen, & Jenkins, 2013; Phillips & Norris, 2009). Work on the science pipeline (Feinstein, 2011; Feinstein et al., 2013) underscores the need for additional research, design principles, lesson plans and learning materials that embrace a broader range of science preparation for those inside and outside the STEM pipeline. Work on the division of cognitive labor has shown that much of what we know comes from experts and that many of our everyday decisions derive not from firsthand experience but from secondhand evidence (Bromme, Kienhues, & Porsch, 2010). The secondhand evidence that laypeople encounter is often highly variable in content, quality, and manner of presentation. Making sense of this complicated patchwork of evidence requires skills at coordinating multiple documents (Britt et al., 2014). As has been recently recognized by multiple documents researchers, domain general factors (e.g., sourcing and corroboration) and epistemic cognition play strong roles in how people process information from multiple documents (Braten, Britt, Stromso, & Rouet, 2011; Braten, Stromso, & Samuelstuen, 2008). This picture of domain general multiple documents processes, strategies, and factors pertaining to epistemic cognition is still emerging, particularly as theoretical constructs around epistemic cognition continue to be refined.

Moreover, researchers are just beginning to examine the domain specific and situationally relevant factors of epistemic cognition that influence students' strategic processing of multiple documents in classrooms that strongly feature the epistemic practices of science. For example, recent research has called for greater attention to the

intersection of three factors, specifically the role of (a) multiple pieces of evidence of variable quality; (b) domain and topic specific criteria; and (c) argumentation (Britt et al., 2014). I would add that understanding students' epistemic cognition is an important and under-researched consideration. This dissertation has been aimed at exploring the intersection of these four factors.

5.2 Findings and Implications

In this dissertation, particularly in Chapter 2, I have articulated a set of challenges faced by learning environment designers and I have recommended a number of guidelines and principles for resolving these design challenges. The guidelines and principles described in Chapter 2 were derived from the experience of designing lessons for the classroom based research that has characterized the PRACCIS project. During the PRACCIS project I encountered numerous challenges that needed to be overcome in the design of learning environments that embrace a reformed vision of science focusing on the epistemic practices of science itself and bringing them into the classroom. There were four broad categories of challenges including: (a) identifying phenomena worthy of investigation; (b) developing models for inquiry; (c) developing evidence for students to use; and (d) enhancing engagement during inquiry. The empirical research presented in Chapter 3 and Chapter 4 has leveraged a number of these design principles in different ways.

Findings from Chapter 3 (the HIV lesson) demonstrate that students evaluated and re-evaluated evidence and engaged in written evidence-based argumentation in ways that are poorly accounted for by prior research. Beyond very domain general criteria (i.e., Does the evidence come from a trustworthy source?) students used a large variety of their

own implicit criteria to reason about evidence. Further, students don't engage in reasoning about evidence in isolation but rather often think about evidence in light of other evidence.

A distinction can be made between students reasoning about evidence versus students reasoning with evidence. As students engage with the evaluative task of reasoning about evidence they might use one or more evaluation strategies. They could compare one piece of evidence against another as a means of ranking the quality of the evidence they have at hand, or evaluate the quality of evidence against particular criteria. Some criteria might be domain general, like whether a controlled experiment was used, but some criteria might be domain specific, like knowing which variables are the relevant ones in need of control. The results from Chapter 3 show that some of these criteria were very domain specific, like comparing the utility and appropriateness of one kind of experimental organism against another. Furthermore, results indicated that evidence re-evaluation could positively impact students' ability to critically evaluate evidence.

As students reason with evidence they are looking for connections between the evidence at hand and the claims that are being advanced. Reasoning with evidence engages students with the opportunity to develop their own personal epistemic stances on a topic. Students in these studies used evidence to advance arguments in favor of one claim or model over another and they used evidence as part of a rebuttal against counterclaims. Reasoning with evidence engages students with thinking about the interconnections of models and evidence, including the relevance of the evidence to the models and the strength of support or contradiction to the models.

I also found that students, in some cases, made meaningful conceptual links between different pieces of evidence. Once connected, these individual pieces of evidence become a body of evidence that students used to differentiate between competing claims. Further, results of an analysis of the four major argument types, (a) no overarching structure; (b) counterargument; (c) body of evidence; and (d) body of evidence with a counterargument, showed that the number of argument components (i.e., the amount of evidence cited and the number of reasons provided) significantly increased when students used a body of evidence.

The results of this research have implications for multiple documents research, epistemic cognition research, and research on model-based inquiry. First, evidence re-evaluation plays an important role in how students reason about competing claims. In other words, seeing new evidence, and revising their beliefs about old evidence, causes shifts in students' thinking about which claims are true and which are false. Second, evidence re-evaluation seems to provide students with the opportunity to calibrate their sense of what counts as good and bad evidence. It is clear that students shifted their critical evaluation of the quality of the texts over time. Being able to flexibly adapt to new evidence could be considered to be an important part of scientific literacy. Further, existing frameworks for multiple documents cognition could be enhanced by a more nuanced consideration of the conceptual links between evidence made by students. The role of evidence-to-evidence coordination, the development of bodies of evidence, and the role of implicit (or explicit) evidence criteria could productively be included in a revised account of learning progressions for argumentation in science education. Students

showed latent capacities for generating and using bodies of evidence; a practice that likely could be promoted through appropriate scaffolding in science classrooms.

Several implications for science educators can be drawn from this research. First, students engaged in productive reasoning about evidence that was not collected first hand in the classroom. It may be the case that for some, if not many, topics in science, second-hand evidence is the best evidence available to students. In my research it would not have been possible to run controlled trials of simians exposed to SIV or lions exposed to FIV in the classroom, nor would it have been possible to engage in the molecular and cellular biology experimental techniques needed to verify that a person is in fact resistant to HIV. However, understanding how scientists use these processes can be an important part of science education; one that might have implications for how people reason outside the classroom. Second, scientists read a great deal as part of their professional practice. Reading and evaluating, and re-reading and re-evaluating, evidence is significant part of what scientists do (Tenopir et al., 2005; Tenopir et al., 2009). Elevating the role of reading evidence in the context of model-based inquiry could play a productive and more prominent role in science classrooms. Third, providing students with scaffolds that encourage them to engage in reasoning with evidence and reasoning about evidence in light of models provides them with opportunities to engage with science in ways that are commensurate with a reformed view of education; one where students are expected to evaluate evidence and engage in model-based inquiry.

In Chapter 4 (the inheritance lesson) I explored how students used evidence over time with some significant differences from Chapter 3. In the HIV lesson presented in Chapter 3 students used evidence to select (i.e., indicate a belief in a model) or reject a

model. In Chapter 4 students used evidence to generate and revise models of inheritance over time. Instead of evaluating the quality of the evidence, students evaluated the quality of their models in light of evidence. Results show that despite the increasing complexity of the evidence, students' rules of inheritance were consistent with the evidence and their inheritance rule sets were largely internally coherent. Further, students' rules increasingly shifted from phenomenological accounts of inheritance toward a more causal account involving two alleles. This research stands as a refutation to claims (Kirschner, Sweller and Clark, 2006) that such inquiry tasks would be unproductive in such an environment.

Several implications can be drawn with respect to the teaching of science and the design of science learning environments. First, students can work productively with models, and model revision tasks, even when their domain knowledge is low. In short, they can generate a theory themselves based on the data, rather than the typical cookbook lab style of explanation in which knowing the correct theory precedes any data interpretation. This is important because it provides students with opportunities to grapple with science like scientists do; constructing theories to test and revise rather than devising experimental procedures to verify already known theories. Second, this method of instruction is a form of productive success (Kapur, 2016), and in this case initial success with modeling activities led to later gains in students' disciplinary knowledge. This provides at least some evidence that modeling itself can be an avenue for learning domain knowledge in science, rather than needing to have command of the domain knowledge first before engaging in any modeling.

5.3 Future Research

Findings from across this dissertation suggest several productive avenues for future research. First, a limitation of Chapter 3 (the HIV lesson) is that it deals with just a single problem in the domain of life science. If it is the case that much of the work of evaluating evidence is through domain specific criteria rather than domain general criteria, then more work that makes use of firsthand and secondhand evidence is needed to uncover students' implicit criteria. A second limitation of Chapter 3 is that it relies on implicit criteria. A learning environment that makes use of explicit criteria for evidence, similar to Pluta, Chinn, and Duncan's (2011) study of model criteria, might uncover what criteria students would find most useful. Moreover, it might show how reasoning can be scaffolded by the explicitness of the criteria (e.g., class list, distributed to everyone, on a piece of paper, a digital wiki, a poster). Given the domain specific nature of reasoning in science, a greater range of topics and domains might uncover novel insights into the evidence evaluation and evidence integration practices of students. Finally, it was the case that the role of experimental organisms (i.e., domestic and wild cats, monkeys) and their taxonomic relationships to humans played a significant role in the reasoning of students in the study. Additionally the similarity (or dissimilarity) of FIV, SIV, and HIV drove students' reasoning about evidence quality and the conceptual links between pieces of evidence. Further research on the role of experimental organisms and the agents and objects (e.g., pathogens, drugs, new biotechnologies) they interact with might further reveal how laypeople reason about biology and health topics.

Learning about inheritance is challenging. A limitation of the Chapter 4 study (the genetics lesson) is that it focuses on Mendelian inheritance. Most traits are non-Mendelian in nature. Learning about other modes of inheritance through model-based

inquiry could present significant conceptual challenges that have not been explored in this dissertation. Research on how students reason about inheritance has often been supported with computer programs like the Genetics Construction Kit which enable students to generate data that they then evaluate. Using sophisticated software that can mimic populations might provide some additional insights into how students engage in anomaly resolution beyond simply changing their theory.

5.4 References

- Bråten, I., Britt, M. A., Strømsø, H. I., & Rouet, J. F. (2011). The role of epistemic beliefs in the comprehension of multiple expository texts: Toward an integrated model. *Educational Psychologist, 46*(1), 48-70.
- Bråten, I., Strømsø, H. I., & Samuelstuen, M. S. (2008). Are sophisticated students always better? The role of topic-specific personal epistemology in the understanding of multiple expository texts. *Contemporary Educational Psychology, 33*(4), 814-840.
- Britt, M. A., Richter, T., & Rouet, J. F. (2014). Scientific literacy: The role of goal-directed reading and evaluation in understanding scientific information. *Educational Psychologist, 49*(2), 104-122.
- Bromme, R., & Goldman, S. R. (2014). The public's bounded understanding of science. *Educational Psychologist, 49*(2), 59-69.
- Bromme, R., Kienhues, D., & Porsch, T. (2010). Who knows what and who can we believe? Epistemological beliefs are beliefs about knowledge (mostly) to be attained from others. In L. D. Bendixen & F. C. Feucht (Eds.), *Personal epistemology in the classroom: Theory, research, and implications for practice* (pp. 163-193). Cambridge: Cambridge University Press.
- Feinstein, N. (2011). Salvaging science literacy. *Science Education, 95*(1), 168-185.
- Feinstein, N. W., Allen, S., & Jenkins, E. (2013). Outside the pipeline: Reimagining science education for nonscientists. *Science, 340*(6130), 314-317.
- Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75-86.
- Phillips, L. M., & Norris, S. P. (2009). Bridging the gap between the language of science and the language of school science through the use of adapted primary literature. *Research in Science Education, 39*(3), 313-319.
- Pluta, W. J., Chinn, C. A., & Duncan, R. G. (2011). Learners' epistemic criteria for good scientific models. *Journal of Research in Science Teaching, 48*(5), 486-511.
- Tenopir, C., King, D. W., Boyce, P., Grayson, M., & Paulson, K. L. (2005). Relying on electronic journals: Reading patterns of astronomers. *Journal of the Association for Information Science and Technology, 56*(8), 786-802.
- Tenopir, C., King, D. W., Edwards, S., & Wu, L. (2009, January). Electronic journals and changes in scholarly article seeking and reading patterns. In *Aslib proceedings 61*(1), 5-32. Emerald Group Publishing Limited.