International Journal *of* Designs for Learning
2016 | Volume 7, Issue 2 | Pages 17-40

CRITICAL DESIGN DECISIONS FOR SUCCESSFUL MODEL-BASED INQUIRY IN SCIENCE CLASSROOMS

Ronald W. Rinehart[1], Ravit Golan Duncan[1], Clark A. Chinn[1], Trudy A. Atkins[2], & Jessica DiBenedetti[2]
[1]*Rutgers, The State University of New Jersey;* [2]*East Brunswick Public Schools, New Jersey*

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.

Ronald W. Rinehart is a Ph.D. candidate at the Rutgers University Graduate School of Education and former middle school science teacher. His research interests include designing and investigating learning environments that promote productive reasoning and epistemic practices. He will be joining the Educational Psychology and Foundations faculty at the University of Northern Iowa in the fall of 2016.

Ravit Golan Duncan is an Associate Professor of science education with a joint appointment in the Graduate School of Education and the School of Environmental and Biological Sciences at Rutgers University. She currently has two main research strands: designing and studying inquiry-based learning environments in life sciences that engage students with modeling and argumentation, and studying learning progressions in genetics.

Clark A. Chinn is Professor and Chair of Educational Psychology at Rutgers, the State University of New Jersey. His research focuses on reasoning and argumentation, epistemic practices and epistemic cognition, conceptual change, and collaborative learning. He was Editor of the journal Educational Psychologist from 2011 to 2015.

Trudy A. Atkins is the K-12 Supervisor of Science for the East Brunswick Public Schools. She obtained her BSEd and M.S. in Molecular Biology from West Chester University.She frequently collaborates with Rutgers Graduate School of Education on matters of teaching and learning in the Life Sciences and has promoted the use of inquiry-oriented learning throughout evidence-based practices (EBPs).

Jessica DiBenedetti has been a science teacher for the East Brunswick Public Schools for six years; she has taught 6th and 7th grade science and currently teaches 9th grade Biology. She has a B.S. in Biology from Felician University and a M.A. in Education from Georgian Court University. She is a collaborator on the Promoting Reasoning and Conceptual Change in Science (PRACCIS) project and incorporates her work on the project into her science classroom.

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 1. Each PRACCIS element presents the designer with particular challenges. Each element and associated design challenge is presented in greater detail later.

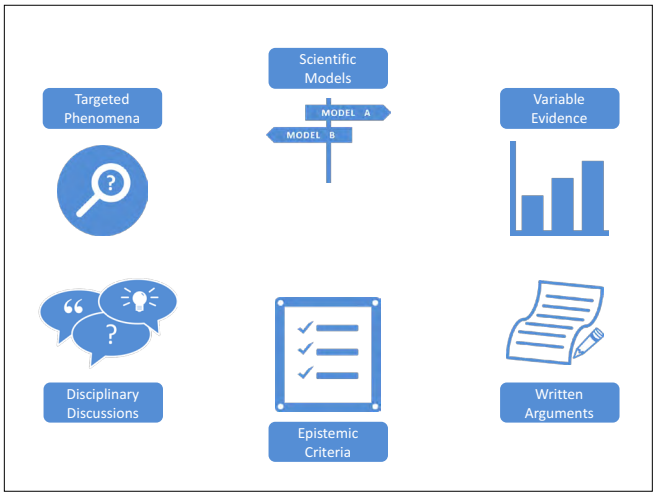


FIGURE 1. Most lessons and units developed for the PRACCIS project include these six major elements. Each element presents a suite of challenges that we had to consider during our design process.

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.

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 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.

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.

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

DESIGN CHALLENGE: CHOOSING PHENOMENA	PRINCIPLES
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.	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). 1b. To the extent possible, the designers should choose phenomena that are meaningful and relevant to students. 1c. It can be advantageous for a designer to choose mysterious, counterintuitive, and non-obvious phenomena, which can enhance engagement (Hidi & Baird, 1986). 1d. Mechanisms relevant to the phenomenon are more accessible if they have real world analogues that students are familiar with, especially macro-scale analogues.
GUIDELINE 2: Modeling should promote mechanistic understandings of phenomena.	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). 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.
GUIDELINE 3: There should be a significant base of evidence that supports the existence of the phenomenon and underlying mechanisms.	3a. Models of candidate phenomena should be grounded in a significant amount of evidence. 3b. Designers should carefully develop evidence so that it is accessible to students.

TABLE 1. Guidelines for choosing phenomena for scientific modeling activities.

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.

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.

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.

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.

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

DESIGN CHALLENGE: DEVELOPING MODELS	PRINCIPLES
GUIDELINE 1: We recommend that models generated by a designer are, at least initially, comprehensible, plausible, compelling, and of comparable quality.	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. 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. 1c. When possible designers should choose incorrect models that reflect misconceptions that have been identified in the research literature (Pfundt & Duit, 1988).
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.	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: 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 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.

TABLE 2. Guidelines for developing models for scientific modeling activities.

what can be left out?). Table 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.

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 and 3 show both models of the second HIV lesson.

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.

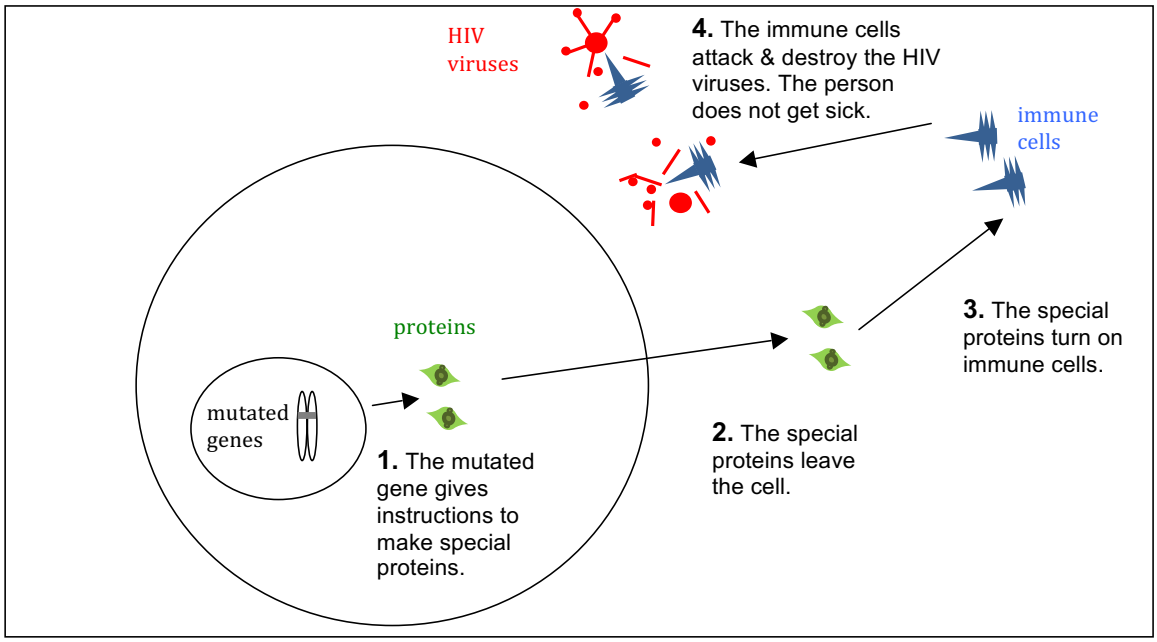
Finally, it is advisable that when possible, lesson designers use modeling as an opportunity to address common student misconceptions. A common misconception about mutations is that they typically add a new function to the body (Nehm & Ha, 2011). Using the HIV models from above, some students think that 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.

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

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.

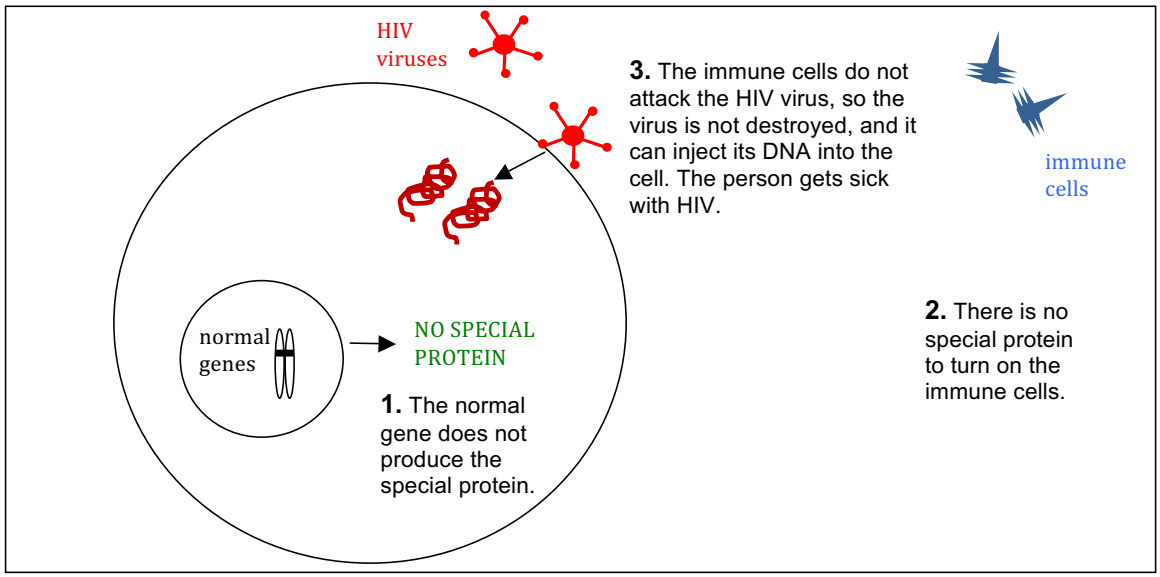
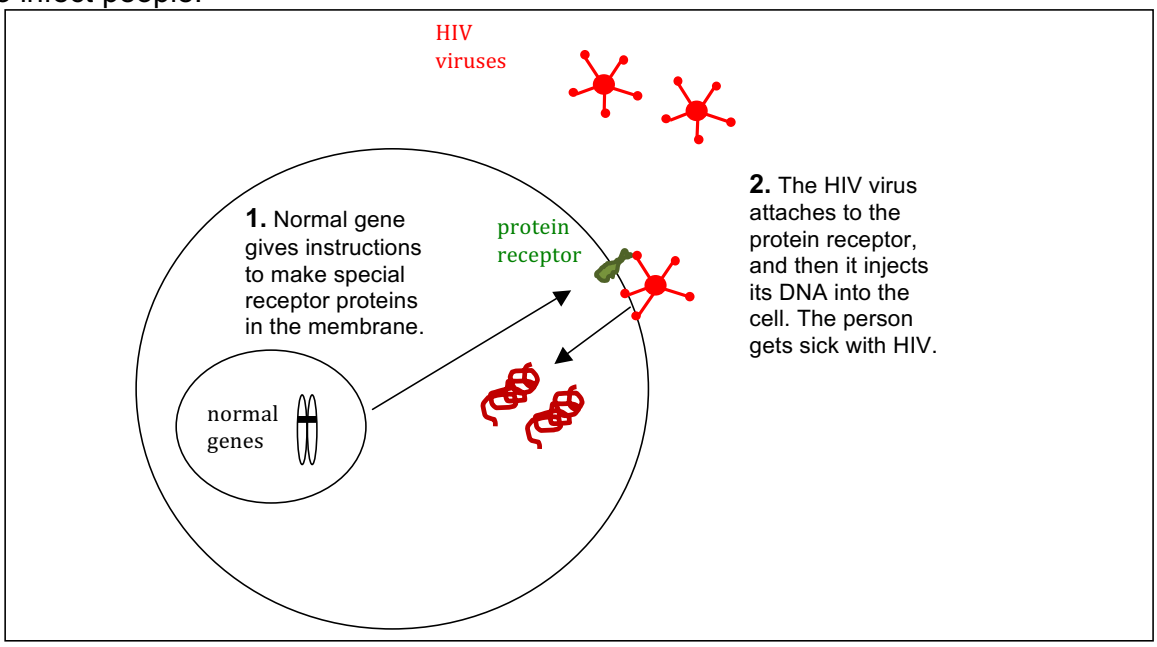


FIGURE 2. The "Attack-and-destroy" model shows that people resist HIV because of a mutated gene that makes a special protein that activates the immune system to fight the HIV.

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.

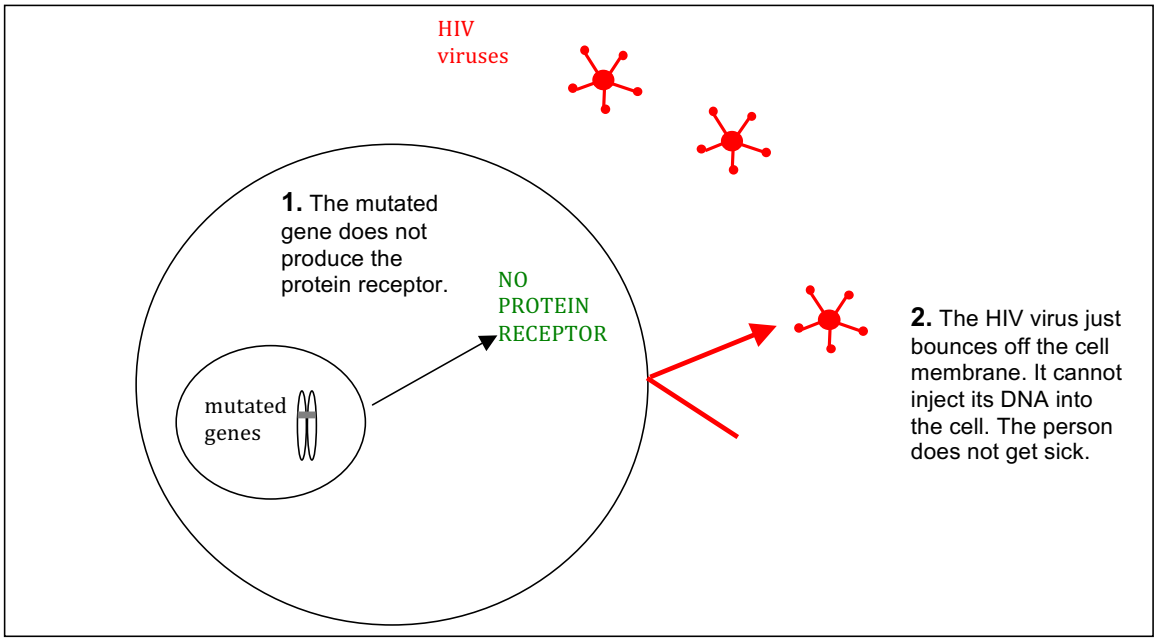


FIGURE 3. The "Keep-it-out" model shows that people resist HIV because of a mutated gene that fails to make a protein receptor that HIV needs to enter the cell.

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).

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

DESIGN CHALLENGE: DEVELOPING EVIDENCE	PRINCIPLES
GUIDELINE 1: Designers should take into account the variety of evidence features that can be varied along two continua: (a) complexity and (b) quality.	1a. Evidence exists along two continua: (a) simple to complex evidence and (b) high quality to low quality. 1b. Designers can foster students' evidence evaluation skills by designing evidence that exists along the full range of both the complexity and quality continua.
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.	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. 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.
GUIDELINE 3: Evidence should often, but not always, contain data.	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). 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. 3c. Data can include qualitative evaluations by experts and non-experts.

TABLE 3. Guidelines for developing evidence for scientific modeling activities.

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 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.

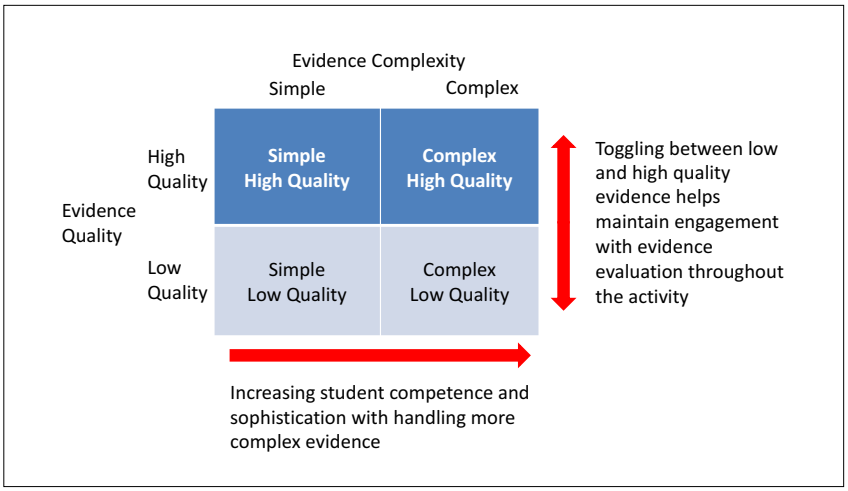


FIGURE 4. This is a 2 x 2 heuristic for the combinations of evidence quality and evidence complexity.

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 4. We offer Figure 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.

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.

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.

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 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 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.

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 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.

Some students tend to think of this as higher quality evidence because it involves medical professionals. However, once they encounter

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

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.

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 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.

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.

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 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.

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

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.

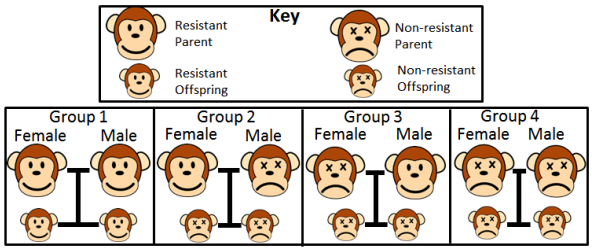
Method and Results: Scientists did four breeding experiments with eight parent monkeys. The groups were completely separated so that they did not have contact with monkeys outside of their group. All monkeys were tested for SIV resistance using high-quality blood tests.

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.

_____

_____

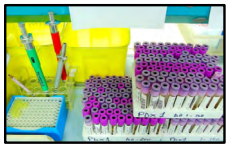
FIGURE 7. Evidence 3, the results of an experiment using SIV in monkeys, is an example of complex low quality evidence.

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 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).

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.

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.

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.

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 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).

DESIGN CHALLENGE: GENERATING PRODUCTIVE DISCIPLINARY ENGAGEMENT	PRINCIPLES
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).	1a. Learning environment designers can promote autonomy by putting students in the role of decision makers and problem solvers. 1b. We recommend that designers guide students toward developing discussion stems that foster disciplinary talk (Michaels, O'Connor, & Resnick, 2008). 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).
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.	2a. Designers are encouraged to incorporate scaffolds that promote systematic examination of the relationship between evidence and models (Lombardi, Sibley, & Carroll, 2013; Rinehart, Duncan, & Chinn, 2014; Suthers & Hundhausen, 2003; Toth, Suthers, & Lesgold, 2002). 2b. Designers can incorporate scaffolds that promote model and evidence quality evaluation (Rinehart et al., 2014).
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.	3a. Evidence exists along two continua: (a) low relevance to high relevance and (b) low diagnosticity to high diagnosticity. 3a. 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.
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.	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. 4b. We encourage designers to develop assessments and activities that effectively capture students' facility with the scientific practices and content of the modeling activities.

TABLE 4. Guidelines for generating productive disciplinary engagement with scientific modeling activities.

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 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.

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

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? **Giving reasons and developing arguments** I think _____ because _____ . _____ because _____ . Why do you __(agree/disagree/think)__? I agree with _____ because _____. **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 _____.	**Purpose** Why did they _____ ? What was the purpose of _____ ? **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 _____ . **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 9. Above is a sample of the discussion stems used in an earlier iteration of the project.

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).

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

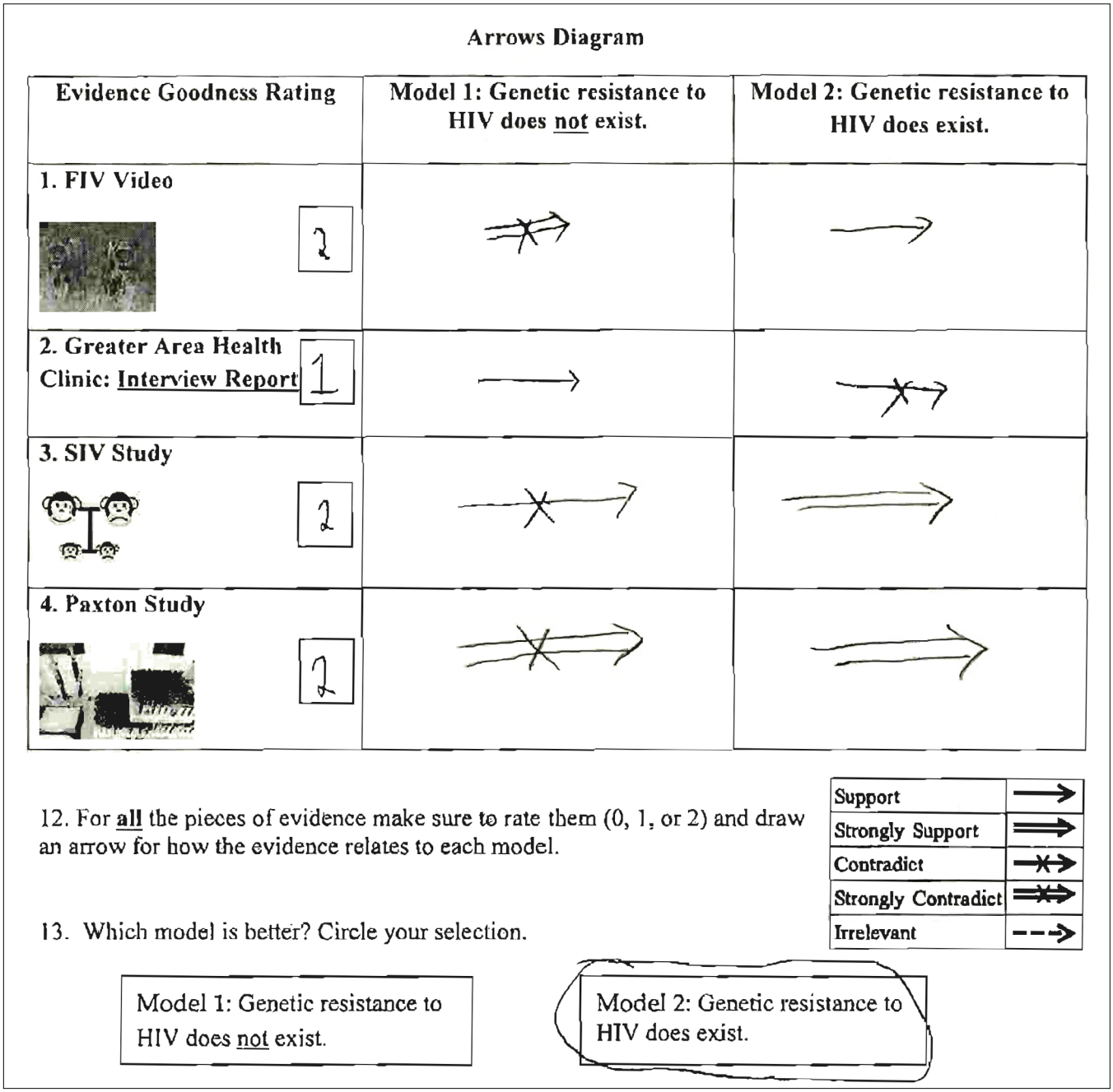


FIGURE 10. The MEL Matrix for HIV Lesson 1 including the arrows diagram, evidence quality boxes, and student model selection boxes.

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 10.

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 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 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).

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.

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.

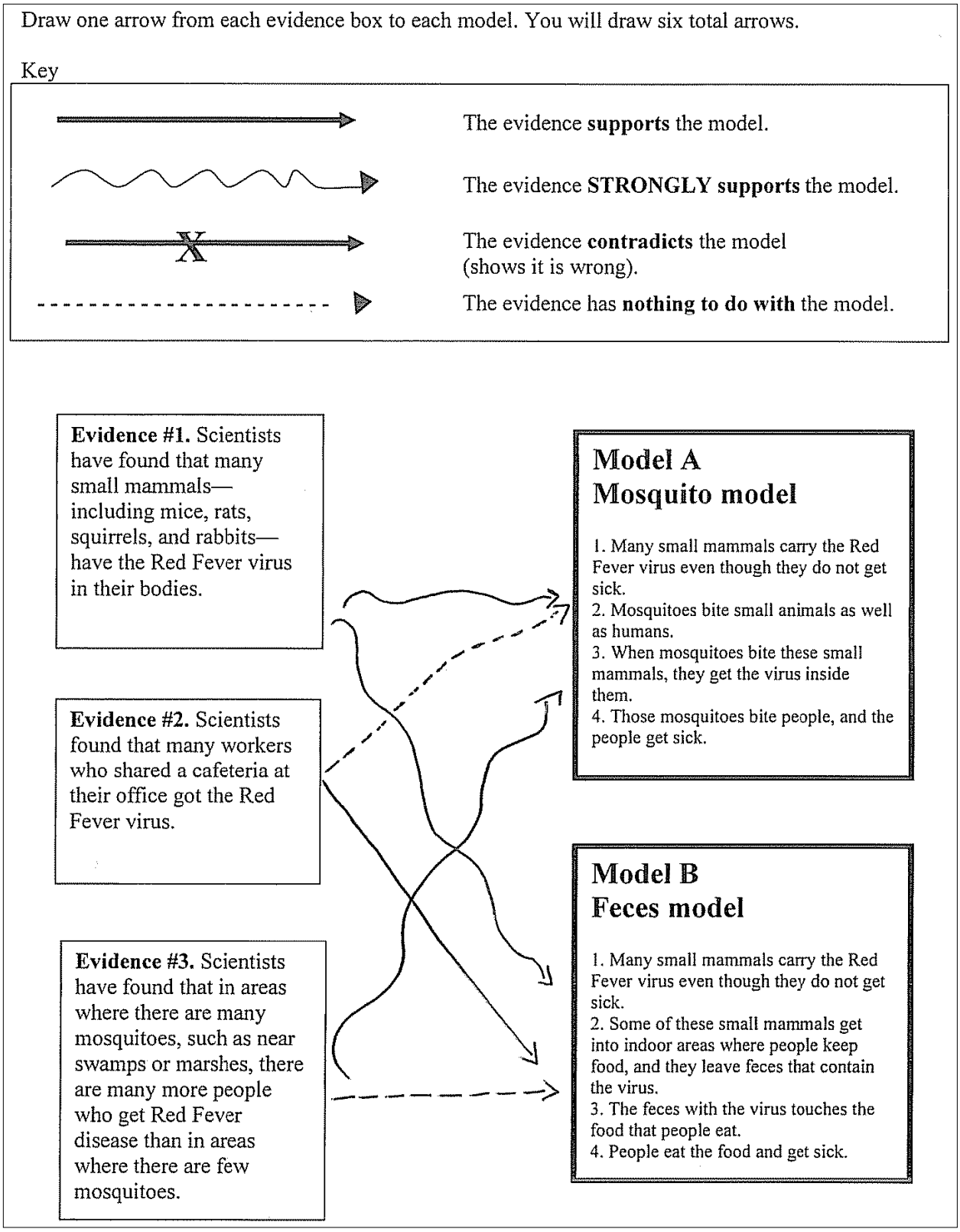


FIGURE 11. This is a typical example of student work using the MEL 1.0 for a modeling activity about the cause of a disease. The HIV lesson described throughout this design case did not make use of the MEL 1.0 so we had to use a representation from another activity. For our purposes here the key features 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.

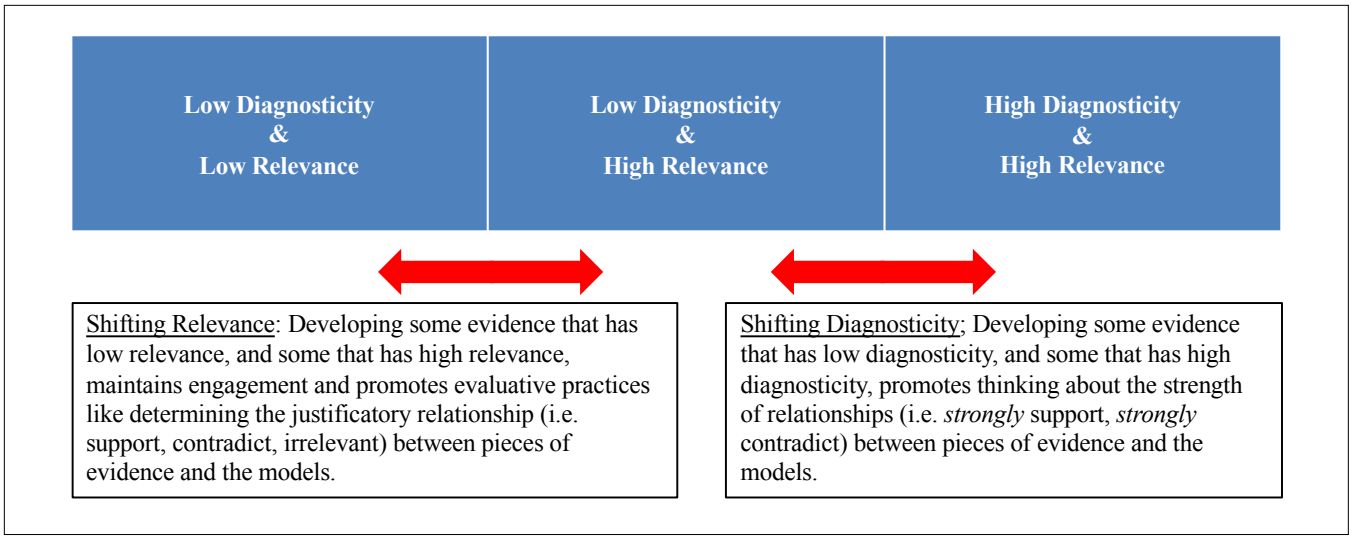


FIGURE 12. This represents the three basic combinations of relevance and diagnosticity, and shows the two major design decisions: (a) *Shifting Relevance* and (b) *Shifting Diagnosticity.*

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 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 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 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.

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.

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.

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.

REFERENCES

Allers, K., Hütter, G., Hofmann, J., Loddenkemper, C., Rieger, K., Thiel, E., & Schneider, T. (2011). Evidence for the cure of HIV infection by CCR5Δ32/Δ32 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). T*he 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.