Developing Assessments That Measure Core Ideas and Scientific Practices: Challenges and Insights

Veronica L. Cavera, Ravit Golan Duncan, Clark A. Chinn, and Moraima Castro-Faix
Veronica.Cavera@gse.rutgers.edu, Ravit.Duncan@gse.rutgers.edu, Clark.Chinn@gse.rutgers.edu,
Moraima.Castro@gse.rutgers.edu
Rutgers University

Abstract: Current science reforms in the US and elsewhere advocate that students should learn core disciplinary ideas through engagement in scientific practices. Engaging students in practices allows for a deeper learning of content and a more productive understanding of how scientific knowledge is developed. In order to evaluate the effectiveness of student learning in practice, assessments have to simultaneously address students' understanding of content and practice. In order to investigate the kinds of heuristics that may assist in developing such assessments, we present data from parallel phenomenon-based modeling assessments given to 11th grade biology students. While students constructed structurally similar arguments, the amount and how content knowledge was used differed. Students were able to bring in content knowledge more readily when presented decontextualized phenomenon-based models and used this knowledge to make productive connections between evidence and provided models. We discuss heuristics and implications for design.

Background

A critical aspect of current science reform is content learning through the practices of science. Doing so gives students the opportunity to not only demonstrate what they know (the content) and how they know it (the practice) but to also do so in an authentic environment (Pellegrino, 2012). One such document is the *Framework for K-12 Science Education* (NRC, 2012) accompanied by the *Next Generation Science Standards* (NGSS Lead States, 2013). Competence within the framework comprises three major elements: (1) Disciplinary Core Ideas, the big ideas within the domain, (2) Scientific and Engineering Practices, the practices that scientists engage in such as argumentation, modeling, investigation, etc., and (3) Cross Cutting Concepts, ideas that define the nature of what it means to do and know science that crosses disciplinary boundaries (NRC, 2012). Each of these individually is insufficient to define competence, but the combination can more accurately express it through performance expectations. Performance expectations give a more specific definition of what is expected of students at specific grades and grade bands as expertise develops. These expectations provide more specificity about the practices of science used and demonstrate how these integrate with core content. Ideally, these performance expectations should lead to coherence among curriculum, instruction, and assessment design (Mislevy, et al., 2017). Assessments are designed with the purpose of assessing whether students are reaching these desired outcomes. They can provide a clear and powerful tool in improving science education. However, if assessments are not well-aligned to the content or the practices under evaluation, they can have unintended consequences (e.g. Pellegrino et al., 2001; Ruiz-Primo et al., 2012). The concept of a performance expectation does provide some guidance in what to assess particularly as science assessments are often viewed as measuring whether students know a grade-level topic (Pellegrino, 2013).

As a community, we recognize the challenges around designing assessments and have begun to develop ways that help rigorously evaluate student cognition (e.g. Mislevy, et al., 2017; NRC, 2014; Pellegrino, 2001; 2012). While there has been a push for making the process of design more rigorous, there is less work around what these designs need to look like, and how such designs should be framed. More explicitly, we need to make the heuristics for designing science assessments more salient so that as a community we can communicate what features and how these features should be made explicit to students. In our own work, we tried to assess two dimensions in the context of genetics using an extended task in which students wrote an argument wherein they evaluated two models of a phenomenon. We found that students engaged with our task in different ways depending on how the mechanism within the models were explained and provide heuristics when students engage in a modeling task.

Theoretical framework

Current science reforms task designers with creating assessments that assess students' ability to integrate content knowledge with practice. These tasks are often given under the larger presentation of some complex phenomenon in order to make the activity authentic (NGSS Lead States, 2013; NRC, 2012). In order to contend with this, guidelines are provided with detailed bounds of what students should know and how they should demonstrate that knowledge. In the case of the NGSS, these are provided in the form of performance expectations (Pes). PEs provide details such as the boundaries of what content students should know at each grade band, clarification about the underlying scientific practice under evaluation and how both should be evaluated in the context of the overarching disciplinary idea. Each PE is helpful but leaves several questions about the claims that can be made about what students should know and how to assess those claims. The design of valid and reliable assessments requires not just understanding the boundaries of the required practice and content under focus but going beyond that level of content for how students will engage with the materials. Pellegrino (2013) notes that beyond content, designing science assessments requires attention to aspects like how students will conceptually engage with models and evidence, and how students will engage at different grade levels and across various medium (e.g. computer-based or paper and pencil). Collectively, it is critical to consider the kinds of evidence obtained from assessments and the conclusions about student proficiency.

In order to more rigorously link the content and practice with the task, designers have turned to more systematic approaches like Evidence Centered Design (ECD) (Mislevy & Haertel, 2009). The process can be summarized into three larger components; the claim made about the knowledge expected from students and how they are expected to demonstrate it, evidence that students have demonstrated knowledge, and the tasks performed by students that show their knowledge (NRC, 2014; Pellegrino, 2013). PEs serve as claims about student proficiency while student performance on items provides evidence of student competency. This relationship needs to be unambiguous so that valid interpretations can be made about student performance.

Previous work into this kind of rigorous assessment design has been largely focused on providing a roadmap of how to go about ECD (e.g. NRC, 2014; Pellegrino, 2012, 2013; Mislevy, 2017). This work is important for exploring how to better integrate and connect these dimensions. This framework and others like it provide a critical process for design but do not help designers address the specific heuristics that underpin how students should and can understand specific aspects of the task. Increasing the rigor between each aspect of the design of assessments is critical, and cannot be understated. However, how students interpret the specific feature of science assessments, and how these features can be designed to contend with student interpretation is less well studied. As it stands, we need to provide insight into the kinds of guiding principles that help with the interplay between designing assessments that elicit evidence of content knowledge and practice by making clear guidelines about the features and aspects of tasks that students attend to during such tasks. Doing so should help not only identify how students interact with assessments, but with improving the sensitivity of assessments to engage students around a variety of topics and contexts. Providing heuristics that help bridge these sorts of gaps will help better connect what we as designers need to know for design so that we can create more efficacious assessments that better assess students understanding of science.

In our own work, we assessed students about a genetics context (Duncan, Choi, & Castro-Faix, 2016). We developed assessments that aim to evaluate students' competencies and reasoning with models, and the construction of evidence-based arguments about core genetics topics in molecular and classical genetics. We designed counter-balanced, structurally similar assessments that had students arguing about two models of a genetics phenomenon. One assessment focused on molecular genetics ideas, and the other focused on classical genetics ideas. We designed these phenomenon-based assessments to provide evidence of student proficiency with both evidence-based argumentation and understanding of genetic concepts. Our research questions are: How did the assessments vary in terms of the affordances and constraints of using aspects of modeling practices and content knowledge? What are some design heuristics that can inform the design of NGSS-aligned assessments?

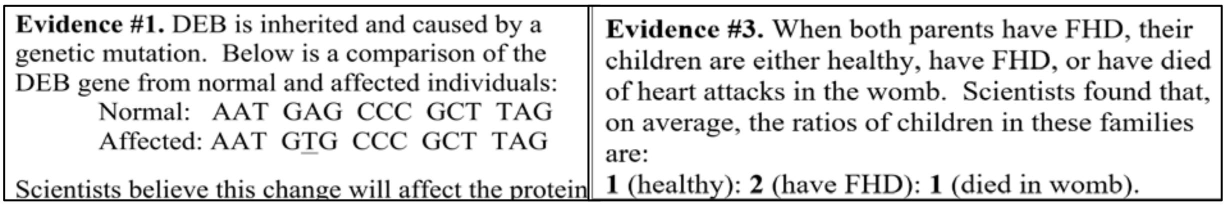
Methodology

Study context

The study took place in a North Eastern suburban high school with five biology teachers and their 11th grade students (n = 271). The school was relatively diverse and 34% of the students were eligible for free or reduced lunch. Instruction lasted 10 weeks and addressed topics in genetics. Of these, five weeks were dedicated to molecular-centered lessons and included topics about the central dogma of DNA and the fundamental role of proteins. The other five weeks of instruction included key concepts about classical genetics and covered topics related to patterns of inheritance. Students were given multiple opportunities to develop extended evidence-based arguments in support of a chosen model (from 2-3 alternative models). During instruction, students were provided scaffolding which afforded opportunities to build a class consensus list of what qualifies as a good model (Pluta, Chinn, & Duncan, 2011). A feature of the instruction tasked students to make connections between the evidences and models (e.g. Chinn, Reinhart, & Buckland, 2014). Students used this list in order to engage in epistemic reasoning of what qualities are important in scientific models. These public criteria for good models was displayed for the duration of instruction. All curriculum engaged students in the development, evaluation, and revision of models of genetic phenomena.

Instrument design and data collection

The assessment included two parts and lasted for 60 minutes. In the first part of the assessment, students were provided three pieces of evidence and tasked with generating a model of a phenomenon. Students were then given three more pieces of evidence and two models, and tasked with building an argument in support of the model they thought was best. We will be discussing data from the second part of the assessment, the argument. For brevity, we will refer to these assessments as the Molecular Modeling Assessment (*MMA*) and the Classical Modeling Assessment (*CMA*). The evidence provided to students in each assessment were either of high quality (e.g. scientific studies), provided information on the phenomenon (e.g. a description of the symptoms) or was an anecdotal account. Figure 1 provides a sample evidence from each assessment. Evidence 1 (left) from the *MMA* describes that a mutated gene causes DEB (the disease under study) and this mutation may change the protein. Evidence 3 (right) describes the ratio of healthy, to individuals with the disease, to those who died from FHD (the disease in the *CMA)*. These evidences are designed to relate (support or contradict) to one or both of the models.



Evidence #1. DEB is inherited and caused by a genetic mutation. Below is a comparison of the DEB gene from normal and affected individuals:

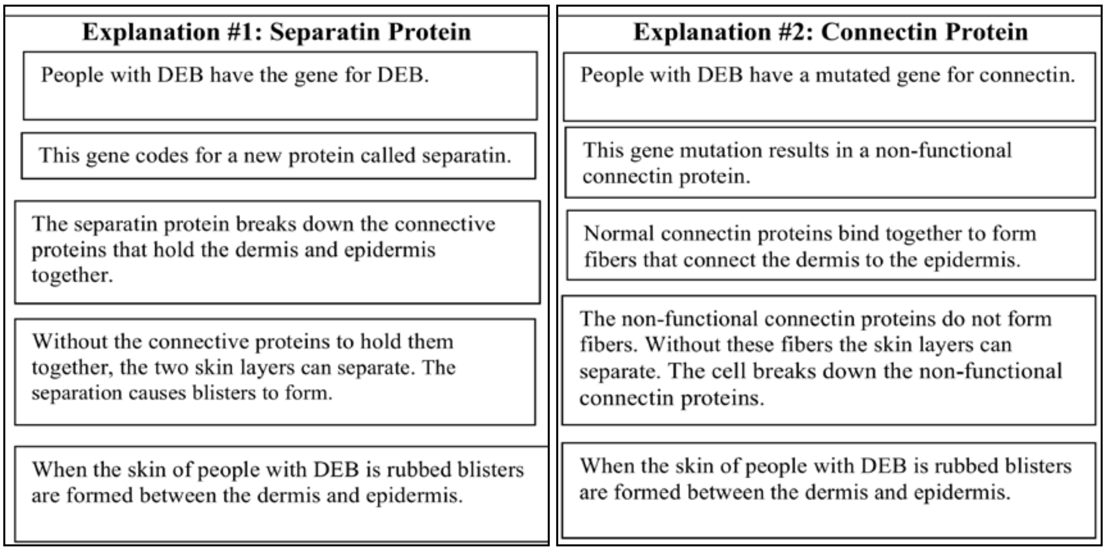
Normal: AAT GAG CCC GCT TAG
Affected: AAT GTG CCC GCT TAG

Scientists believe this change will affect the protein

Evidence #3. When both parents have FHD, their children are either healthy, have FHD, or have died of heart attacks in the womb. Scientists found that, on average, the ratios of children in these families are:
1 (healthy): **2** (have FHD): **1** (died in womb).

Figure 1. Evidence provided to students. Evidence #1 is from the Molecular Modeling Assessment, Evidence #3 is from the Classical Modeling Assessment.

In the *MMA*, students developed a model for the hypothetical skin disorder DEB while in the *CMA* students developed a model for FHD, a disorder that causes an irregular heartbeat. In the *MMA,* the first model postulated that a gene codes for a new protein, *separatin*, that breaks down the skin layers (Figure 2). This gain of function of an entirely new protein is not biologically plausible. The second and correct model postulated that a mutated gene lead to a non-functional protein, *connectin*, that can no longer hold the skin layers together. This loss of function is correct; the mutated gene will lead to a protein with a different structure and the disorder.



Explanation #1: Separatin Protein

People with DEB have the gene for DEB.

This gene codes for a new protein called separatin.

The separatin protein breaks down the connective proteins that hold the dermis and epidermis together.

Without the connective proteins to hold them together, the two skin layers can separate. The separation causes blisters to form.

When the skin of people with DEB is rubbed blisters are formed between the dermis and epidermis.

Explanation #2: Connectin Protein

People with DEB have a mutated gene for connectin.

This gene mutation results in a non-functional connectin protein.

Normal connectin proteins bind together to form fibers that connect the dermis to the epidermis.

The non-functional connectin proteins do not form fibers. Without these fibers the skin layers can separate. The cell breaks down the non-functional connectin proteins.

When the skin of people with DEB is rubbed blisters are formed between the dermis and epidermis.

Figure 2. Explanations presented in the *MMA.*

In the *CMA,* students were presented with two competing explanations (Figure 3) about the mode of the disorder. One suggested a recessive mode in which two parents could be healthy carriers of the disorder resulting in a 0.25 likelihood that their children will have the disorder (which is not the correct explanation based on the provided evidence). While the other, incomplete dominance, explained that individuals can be healthy, mildly or severely sick, leading to three different versions of the trait (phenotypes). The key distinction is the presence of three phenotypes, which can only be explained by the *incomplete dominance* explanation. This is supported by evidence which described that FHD has three phenotypes and therefore must be caused by *incomplete dominance.* In both cases, once students were presented with all six pieces of evidence and the models, they then responded to the prompt: "Which do you think is the better explanation of DEB, the *separatin explanation* or the *connectin explanation*? Write *at least four (4)* detailed reasons for your answer. Write to someone who may disagree with you, but who has not seen the evidence."

Explanation #1: Incomplete Dominance	**Explanation #2: Recessive**
1. Children receive one allele from each parent. The combination of alleles that they end up with determines whether they will be sick.	1. Children receive one allele from each parent. The combination of alleles that they end up with determines whether they will be sick.
2. Individuals with [AA] alleles are healthy and do not show any symptoms.	2. Individuals with [AA] alleles are healthy and do not show any symptoms.
3. Individuals with [Aa] alleles are moderately sick. They show symptoms later in life.	3. Individuals with [Aa] alleles are carriers of the disorder, but are not sick themselves.
4. Individuals with [aa] are severely sick and become sick at a much younger age.	4. Individuals with [aa] are sick and show symptoms of the disorder at a young age.

Figure 3. Explanations presented in the *CMA.*

We reported data from students who completed the argument for both the *MMA* and *CMA* assessment (n = 271). Each assessment was given to students following five weeks of inquiry-based genetics instruction. Therefore, the *MMA* was given following five weeks of molecular genetics instruction, and the *CMA* was given following five weeks of classical instruction. Students were given both sets of instruction, though half of the students (n = 144) were first given the classical form of instruction before switching to the molecular units, while the other half (n = 127) were given the molecular instruction first before switching to the classical units. All students completed both curricular units. Any students that did not complete both assessments were removed from data analysis.

Analysis

In order to assess how students constructed their arguments we coded along two dimensions; (1) structural and epistemic dimensions of evidence and models, and (2) the type and quantity of content knowledge. In terms of structural components, we looked for a claim, the amount of evidentiary backing for a provided claim, and the relationship between the evidence and claim (reasoning) (Krajcik, McNeill, & Reiser, 2008; McNeill & Krajcik, 2011). We defined using evidence as describing a connection between a specific piece of evidence to one of the models. A student response could be: *scientists injected DEB patients with a working connectin protein and 80% of them got better and their skin didn't produce blisters when rubbed anymore. Which proves that a healthy connectin was missing all along.* This student described a piece of evidence (wherein scientists provided patients with a connectin-containing medication that should repair damaged skin), and connects this back to one of the models. Reasons are defined here as a justification between the model and evidence that leverages scientific knowledge (McNeill & Krajcik, 2011). Understanding how students used evidence within their arguments helps to position how students construct their argument and what was attended to during argument construction. Students can further explain the evidence, describe the relationship between the evidence and the model, and explore how the evidence either supports or refutes a model. In terms of epistemic dimensions, we analyzed whether students critiqued the evidence and/or the models and the relationship between these features. These included such responses as: *incomplete dominance is very confusing, its [SIC] not give a straight forward answer. The recessive explanation clearly explains whether it's a dominant or recessive disorder.*

Content knowledge was coded as any articulation of domain knowledge that went beyond what was stated in the evidence or models. We coded prepositions and categorized how the content related to aspects of the assessment (Ruppert, Duncan, & Chinn, 2017). For example, a student during the *MMA* who stated *the codons show a mutation which must code for a mutated gene*, discussed the concept of codons. This student is describing a piece of evidence in the *MMA*. Three independent coders coded 36% of the *CMA* data and 32% of the *MMA* data. Interrater reliability was 91% and 87%, *CMA*, *MMA*, respectively. All disagreements were discussed.

Results

In order to understand how the assessments differed in eliciting evidence about students' ability to write evidence-based arguments use of content-knowledge, we first analyzed dimensions of evidence use, epistemic considerations, and whether or not students used core ideas from instruction in their arguments. We discuss the features of student arguments and identify the kinds of content knowledge that students provided. We then categorized the role between content knowledge, evidence, and models in order to highlight differences across the two assessments. We identify these differences in order to discuss heuristics for design of modeling assessments.

Argument features

In looking at differences in argument structure and epistemic reasoning, we found no significant differences in selection of the canonically correct model (74%, 69%; *CMA*, *MMA*), average number of evidences used (3.5 pieces), amount of reasons (36%, 32%; *CMA*, *MMA*) or epistemic reasoning (19%, 11%; *CMA, MMA*). However, we found a small, but significant difference by students who used 5-6 pieces of evidence with students in the *MMA* more likely to use most of the provided evidence. Students who used less evidence were more likely to provide more epistemic reasoning, and students who used more evidence provided less epistemic reasoning. However, in general, our students did not generate structurally different arguments across these assessments.

Role of content knowledge

We next examined how students incorporated content knowledge. In general, students brought in 32% and 16% content knowledge into their arguments, *CMA, MMA*. When we investigated how students leveraged their knowledge we found content knowledge was brought in specific ways within each assessment. For each assessment, two of the types of knowledge were generally not helpful in leading students towards the canonically correct model with the third being overwhelmingly helpful.

Analysis of the MMA provided examples of explanatory prepositions. In this case, one of the evidences showed a mutated DNA sequence that may impact a protein. Some students added information stating that the DNA sequence was a sequence of codons, or added information on RNA (e.g., [this]…*leads to a bad RNA and that can cause the wrong protein to be made*). Students also discussed how the protein causing a problem was no longer at the cell or tissue level (e.g. *I believe the connectin protein can't connect the skin layers because it isn't there*). Like the *CMA*, these prepositions helped students explain a concept not explored in detail in the evidence or models. These did not help students choose the correct model, and in the case of students who brought in ideas about the protein moving from or out of the tissue layers, it overwhelmingly negatively impacted their model selection. The third type of preposition that students brought into the *MMA* was that the production of a new protein seemed unlikely (e.g. *Usually, when someone has a disorder they are missing a gene that a normal person does. Like in albinism. I haven't heard of someone with a disorder because of something extra*). This piece of knowledge allowed students to identify the key difference between the models in terms of the disciplinary distinction at play, which is the presence of a missing non-functional protein. This is only made possible by the connectin protein story.

During the *CMA*, students explained carriers (e.g. *in class we talked about how sometimes you can have a gene but not show it*) or identified that the diseased allele must have skipped a generation (e.g. *Evidence #6 said that FHD does not skip a generation. So that means even if you have the correct copy of genes, the symptoms still show up just not as severe*). These two uses of outside content serve as warrants for students' arguments. Meaning, students took a concept, explained it, and made a connection between the evidence and model. We noted that while some students explained these concepts correctly, the conclusion drawn from the interpretation either led them to select the incorrect model or did not have an impact on their model selection. These two pieces of content did not help inform the underlying mechanism critical to distinguishing between the two models. The third content preposition students used was the presence of multiple phenotypes (e.g. *...the way genes work normally is that you either have this or that, but with incomplete dominance you can either have one thing, the other, or something in between, like it says in evidence 4. Which leaves the recessive invalid. In terms of flowers, if you cross-breed a red and white flower- you either get red, white, or pink*). This bringing in of ideas, like the aforementioned carriers or skip generation prepositions, had students making a connection between the presented evidence and the models, but was overwhelmingly helpful in distinguishing between the two models. Students could identify that a critical idea in genetics (here the presence of multiple phenotypes) could not be explained by simple dominance (which has two phenotypes).

While students could bring in prepositions of content knowledge in one of three ways in each assessment, in each case only one was productive in allowing students to make a connection between what they know about genetics and identify the mechanistic difference between models. Beyond that, the way in which students leveraged this knowledge was different. In a simplified Toulmin argument (Toulmin, 1958), an individual makes a claim, and must back that claim with grounds. Students in these assessments made a claim when they selected a model and provided grounds in the form of evidentiary support. In the *CMA,* students described how the phenomenon had three phenotypes and connected the evidence that demonstrated this (one evidence piece showed a ratio of very sick, healthy, moderately sick individuals) with the model. This is leveraging knowledge as a warrant—making a connection between evidence and model (Figure 4, right). In the *MMA*, students who brought in content knowledge did so as backing—providing substantiation for either the model or evidence (Figure 4, left). This difference indicated that students leveraged knowledge in specific ways in each assessment. In order to identify the affordances and constraints of each assessment we next investigated the kinds of approaches students took when structurally arranging their arguments.

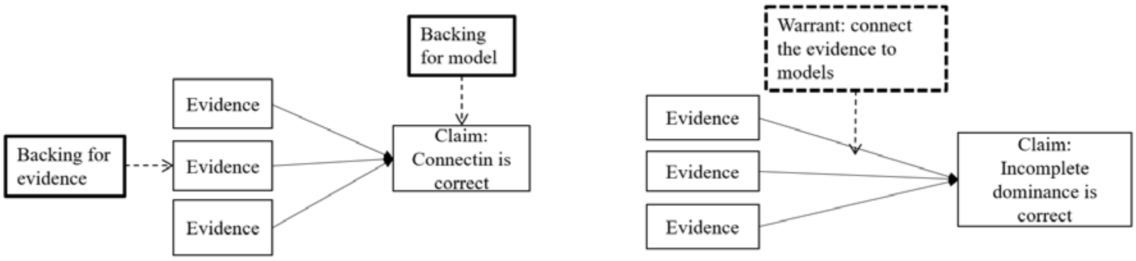

Backing for model
Evidence
Backing for evidence
Evidence
Evidence
Claim: Connectin is correct
Warrant: connect the evidence to models
Evidence
Evidence
Evidence
Claim: Incomplete dominance is correct

Figure 4. Student uses for content knowledge in arguments.

Approaches to argument structure

In both assessments we noted that in general students took one of two approaches—either their arguments were structurally sound and used evidence but not content knowledge or they used content knowledge but tended to use less evidence. A smaller proportion of students combined these approaches and used both. We first detail students who used one approach.

Students were expected to make connections between the models and evidence on the basis of content as well as evaluations of how the models and the evidence related in terms of amount of supporting evidence or the quality of the evidences that provided support. These two approaches should have happened in tandem and not at the expense of one another as the purpose of our assessment was for students to demonstrate their understanding of content through the scientific practice of argumentation. However, we noted that students were more likely to make connections between the evidence and models (argue) instead of using their domain knowledge, which misses the disciplinary point of writing arguments in science. Students should feel compelled to take both of these approaches in order to have a well-founded argument. We provide examples of students who took each of these approaches in Figures 5 and 6. Student A produced a coherent argument that contains the structural and epistemic features of arguments but does not contain any outside content knowledge. This student refutes the competing model based on the quality of evidence that supports it and then shows how two pieces of evidence relate to their chosen model.

> Student A: I believe the better explanation is the connectin explanation. I believe this is true partly because the only piece of evidence to the contrary would be Evidence #4 wear a boy was given medication to break down the separatin protein by his mother. While this piece of evidence is valid this is one boy who was given something found on the internet, it could have been anything. Another reason I believe this explanation is correct is because evidence 6 is a scientific study in which scientists injected 25 patients with connectin and their skin improved. Thirdly evidence 5 is another scientific study that shows that affected people are missing

Figure 5: Student argument that applies structural and epistemic features.

Conversely, Student B (Figure 6) relied entirely on content knowledge and critiqued the validity of one of the models without the use of evidence. This is apt as the student supplies content knowledge but this student ignores the relationship between the models and evidence. This student does note that a gene coding for a new protein is something unusual but does not identify the distinction between the two models (that is—the presence of a missing protein). If students cannot see the disciplinary issue at hand they will not attend to it, as seen with Student A. Or, if in the case of Student B, they focus instead on the content underneath the assessment, they will not attend to the models. One of the key aspects of this assessment was tasking students with need to attend to both the content and the argumentation practice.

> Student B: People who have DEB have a mutated gene called connectin. The gene results in a non-functional connectin protein. The non-functional connectin protein doesn't form fibers that connect the dermis to the epidermis. The separatin explanation says that there is a new protein for DEB. Since when does a gene code for a new protein?

Figure 6. Student argument that applies content knowledge.

Student C (Figure 7) took both approaches in order to refute the competing model at two major critical faults (what this student refers to as *red flags*). The student addressed that the only supporting piece of evidence that provides backing for the competing model is not good quality. This falls in line with an expected connection between models and evidence. The student then follows this up by addressing that the presence of multiple phenotypes is only possible in one model (the incomplete dominance model) and not in the competing model (recessive model, which this student refers to as *Dominant and Recessive genes*). The leveraging of content knowledge in such a way—using it to show how only one model is capable of explaining the phenomenon in conjunction with the provided evidence—is a powerful argument. While all three students provide sound arguments, Student C's argument includes both content knowledge and an evidence-based argumentation in a coherent and integrated fashion. This combination is critical because NGSS-aligned assessments should have students include structural and content dimensions in order to demonstrate their competency.

> Student C: I believe explanation #1 is the better explanation. I believe this mostly because of evidence #4. In evidence #4 scientist is a reliable source took a direct survey on 115 patients with the disease. The result was out of all 115 only 12 patients between the ages of 0-3 were severely sick. While 103 were moderately sick. Now red flags start flying with this piece of evidence. The biggest one however, is in explanation #2. It talks about Dominant and Recisive [SIC] genes. You can only have two phenotypes with this kind of inheritance yet in Evidence 4 there's people who are sick, and there are people who are moderately sick. Also of course there are the people who are healthy. Three distinct phenotypes. The second biggest flag is that Explanation #2 there is no talk about someone with a moderate case of the disease. Yet in evidence #4 the majority of the people have the moderate case. Making Explanation #2 invalid.

Figure 7. Student argument that use both approaches.

Discussion

We presented data from arguments generated during modeling assessments. We noted similarities in terms of students who selected the canonically correct model, average number of evidences, reasons, and epistemic reasoning. Students in the *MMA* were more likely to use all of the evidences, however, these same students brought in half the amount of content knowledge to support their arguments as compared to students during the *CMA*. This difference represents a critical distinction in how students responded to the assessments. The *CMA* made the relevant disciplinary distinction between the competing models more salient to students. The *CMA* afforded students the opportunity to bring in more content knowledge by providing nonspecific mechanistic models. While these models presented a mechanism for a disorder, the steps included in the models were not based around the disorder itself and instead focused on describing the mechanism of each inheritance pattern. Students in the *CMA* not only brought in more content knowledge but leveraged it as warrants. Conversely, the *MMA* provided mechanisms that were contextualized making the underlying difference less distinct. Students who brought in content knowledge focused on the *context* of what the mechanism meant and provided further backing for evidence *or* the models but *not* make a connection.

Heuristics

In thinking about our findings in more general terms, we have tried to extract implications in the form of heuristics that can guide the design of assessments that can elicit evidence about student understanding in regards to both content knowledge and scientific practices. This first heuristic is to *make disciplinary distinctions between models explicit*. As has been stated previously, a goal in the NGSS and elsewhere is for students to demonstrate their competency in the context of some practice. This is possible if students can identify the content under investigation and can apply that in order to identify differences between models. The disciplinary goal has to be clear enough to identify and argue about. If students cannot make a clear distinction between the models in terms of the content they will apply what they understand about the relationship between models and evidences in terms of the quality of those pieces. And while these arguments may be coherent they do not fulfill the goal of having students demonstrate their knowledge simultaneously with practice. However, the distinction cannot be too overt, as this runs the risk of students too easily identifying what the context is, and focusing on the content. In short, students should be able to see what the issue is at stake but this issue should not be made so overly accessible that students are easily provided the solution. We recognize that this is no easy challenge, and only intend to imply that the more explicit the distinction the more likely students will attend to disciplinary issue at the heart of the assessment.

A way to help make the distinction explicit without losing nuance is in the presentation of models itself. In short, the second heuristic is to *decontextualize mechanistic models*. The mechanisms that underlie each model must be explicit but should reinforce the disciplinary content under question by making the content more prevalent than the story used to convey that content. Looking at the two sets of models presented here, students were overwhelmed by the details to the point where it obscured the overall context in the *MMA*. The models were explanatory but far too contextualized for students to find what to attend to. The models presented in the *CMA* presented decontextualized mechanisms that allowed students to identify the disciplinary distinction. It perhaps would be better if these models were even more decontextualized so students may bring in a higher percentage of content knowledge. However, getting students too involved in the details so that they miss the overarching situation is equally problematic. Moving towards the production of mechanistic, decontextualized models that make their disciplinary content distinct presents an elegant solution for developing science assessments that could elicit student responses that integrate science content and practice. These should make the disciplinary idea under study more salient and more engaging.

References

Chinn, C. A., Rinehart, R. W., & Buckland, L. A. (2014). Epistemic cognition and evaluating information: Applying the AIR model of epistemic cognition. *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences*, 425-453.

Duncan, R. G., Castro-Faix, M., & Choi, J. (2016). Informing a Learning Progression in Genetics: which Should BE Taught First, Mendelian Inheritance or the Central Dogma of Molecular Biology?. *International Journal of Science and Mathematics Education*, *14*(3), 445-472.

Krajcik, J., McNeill, K. L., & Reiser, B. J. (2008). Learning‐goals‐driven design model: Developing curriculum materials that align with national standards and incorporate project‐based pedagogy. *Science Education*, *92*(1), 1-32.

McNeill, K. L., & Krajcik, J. S. (2011). Supporting Grade 5-8 Students in Constructing Explanations in Science: The Claim, Evidence, and Reasoning Framework for Talk and Writing. *Pearson*.

Mislevy, R. J., Haertel, G., Riconscente, M., Rutstein, D. W., & Ziker, C. (2017). Evidence-centered assessment design. In *Assessing Model-Based Reasoning using Evidence-Centered Design* (pp. 19-24). Springer, Cham.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

National Research Council. (2014). *Developing assessments for the next generation science standards*. National Academies Press.

NGSS Lead States. (2013). *Next generation science standards. for states, by states: The standards: Arranged by disciplinary core ideas and by topic*. Washington, D.C.: National Academies Press.

Pellegrino, J. W. (2012). Assessment of Science Learning: Living in Interesting Times. *Journal of Research in Science Teaching, 49* (6), 831-841.

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.

Ruiz‐Primo, M. A., Li, M., Wills, K., Giamellaro, M., Lan, M. C., Mason, H., & Sands, D. (2012). Developing and evaluating instructionally sensitive assessments in science. *Journal of research in science teaching*, *49*(6), 691-712.

Ruppert, J., Duncan, R. G., & Chinn, C. A. (2017). Disentangling the Role of Domain-Specific Knowledge in Student Modeling. *Research in Science Education*, 1-28.

Toulmin, S. E. (1958). *The philosophy of science* (Vol. 14). Genesis Publishing Pvt Ltd.

Acknowledgements

This research is supported by a National Science Foundation Award #1053953.