Developing NGSS-Aligned Assessments to Measure Crosscutting Concepts in Student Reasoning of Earth Structures and Systems

By

Gary Weiser

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY
2019

Abstract

Developing NGSS-Aligned Assessments to Measure Crosscutting Concepts in Student Reasoning of Earth Structures and Systems

Gary Weiser

The past two decades of research on how students develop their science understandings as they make sense of phenomena that occur in the natural world has culminated in a movement to redefine science educational standards. The so-called *Next Generation Science Standards (or NGSS)* codify this new definition into a set of distinct *performance expectations*, which outline how students might reveal to what extent they have sufficient understanding of disciplinary core ideas (DCIs), science practices (SEPs), and crosscutting concepts (CCCs). The latter of these three dimensions is unique both in being the most recent to the field and in being the least supported by prior science education research. More crucially, as a policy document, the *NGSS* alone does not provide the supports teachers need to bring reforms to their classrooms, particularly not summative assessments. This dissertation addresses both of these gaps using a combination of quantitative and qualitative techniques. First, I analyze differential categorization of problems that require respondents to engage with their CCC understandings via confirmatory factor analysis inference. Second, I use a set of Rasch models to measure preliminary learning progressions for CCCs evident in student activity within a computer-assisted assessment experience. Third, I analyze student artifacts, think-aloud interviews, and post-task reflective interviews via activity theory to adapt the progression into a task model in which students explain and predict aspects of Earth systems. The culmination of these three endeavors not only sets forth a methodology for researching CCCs in a way that is more integrative to the other dimensions of the NGSS, but

also provides a framework for developing assessments that are aligned to the goals of these new standards.

TABLE OF CONTENTS

Table of Figures, Graphs, and Tables

Acknowledgements

My exceeding thanks go out to Prof. Emdin and Prof. Anderson for their continuing expertise and guidance along my journey to the dissertation. Additional thanks are due to Prof. Rivet for the initial push that set my work in motion and to Prof. Mensah for seeing it at the end. This work would not have been possible without the mentorship and friendship of Lei Liu and the rest of the DAT-CROSS team. I extend to them sincere appreciation. Lastly, I want to thank my family and friends whose continued dedication made this work possible.

*The project described henceforth was partially supported by the **U.S. Department of Education. Institute of Education Sciences, National Center for Education Research**, through grant number R305A170456. The content is solely the responsibility of the author and does not necessarily represent the official views of the Institute of Education Sciences.*

Chapter 1

INTRODUCTION

The Next Generation Science Standards (NGSS) take a novel approach to the ways the science education community thinks about what counts as acceptable expressions of students' science understandings (Krajcik & Merritt, 2012). Previous reform documents such as the American Association for the Advancement of Science's *Benchmarks for Science Literacy* (henceforth AAAS; AAAS, 1993) described expectations in terms of what students should *know* rather than what students *can do* (Duschl, Schweingruber, & Shouse, 2007). The NGSS reforms, however, describe expectations for student understanding in terms of a performance, something students *need to do* as a way of expressing competencies in use of relevant disciplinary core ideas, science and engineering practices, and crosscutting concepts (National Research Council, 2012). Though these standards are not, themselves, assessments, they play an essential role in informing assessment design (DeBarger, Penuel, Harris, & Kennedy, 2015) by helping to constrain possible perspectives on the assessment triangle: Cognition, Observation, and Interpretation (Pellegrino, Chudowsky, & Glaser, 2001). The three-dimensional view of science understanding (Duschl et al., 2007) which fuse disciplinary core ideas (DCIs), crosscutting concepts (CCCs), and science-engineering practices (SEPs) detailed in the National Research Council's (henceforth referred to as the NRC) *A Framework for K-12 Understanding* (NRC, 2012) (henceforth referred to as *Framework*) sets expectations for how the community should think about what students know (the cognition corner of the assessment triangle model; Pellegrino et al., 2001). Similarly, the expectations, themselves, make salient what student understanding of science should look like (the observation corner of the triangle model; Pellegrino et al., 2001).

However, where assessment observations should be found, and how to interpret observations made (Pellegrino et al., 2001), remain underspecified in reform literature.

Uniquely underspecified in existing literature are supports for eliciting and interpreting observations of the crosscutting concepts (CCCs), which has left some questioning why the dimension appears at all (Osborne, Rafanelli, & Kind, 2018). While the subtlety of CCCs is undeniable, they have immense power in shaping how students use language to explain phenomena (Weiser, Lyu, & Rojas-Perilla, 2017) and, by extension, how assessors value the responses students provide. Nonetheless, if my philosophical position is that student reasoning of the crosscutting concepts adds value to their science understandings in a way that the other dimensions do not, then failing to capture that reasoning is failing to capture the full richness of their performances. It is, therefore, a continuing charge placed on the NGSS-supporting science education community to design assessments that align to all parts of the new standards. In the following sections, I describe a framework that the evaluation community might use to design assessments for the NGSS, filling in the gaps on the assessment triangle along the way.

This research study to follows a paradigm known as design-based research, which emerged from the learning sciences during the late 1990s (Barab & Squire, 2004). Unlike clinical intervention research, design-based research makes a tradeoff; i.e., sacrificing typical treatment controls in favor of data that can only be collected in the bustle of the real world, but no less in a systematic and evidence-based way (Hoadley, 2004). DiSessa et al. (2004) describe the goals of these endeavors as 'ontological innovation', in which the research adds something new to an important theory that could only have been added with data collected from the nonclinical setting. Part of this goal of innovation (as DiSessa and his colleagues describe it) is a new form of validity dubbed consequential validity by Hoadley (2004) that defines the quality both of theory and research in terms of their ability to solve a problem that actually exists in the real world. Simply put, a good educational theory ought to successfully support good design principles that in turn should successfully support a good educational product or practice. The framers of the Next Generation Science Standards certainly have an abundance of good educational theory (set forth exceptionally by Duschl, 2008) and evidence-centered design is a growing set of design principles that are often described in terms of their alignment to these new standards (DeBarger et al., 2016). The time now comes to engage with these principles in order to produce needed assessment and, in so doing, innovate on (as Pellegrino, Chudnowsky, and Glaser put it) how best to know that students know (Pellegrino et al., 2001).

The validity argument that foregrounds the evidence-centered design framework hinges on the ability of the assessment to successfully support a particular claim regarding what students know (Mislevy et al., 2017). However, before evidence backing a claim can be elicited, designers first need to define what claims are useful for relevant stakeholders; such as teachers, students, administrators, and policymakers (Debarger et al., 2016). While the broad strokes presented in the Next Generation Science Standards (NGSS; NGSS Lead States, 2013) may be sufficient for administrators looking to make overarching claims about the science proficiency of students in their local purview (Pellegrino, 2013), teachers and students require more detailed information about their present understandings that are capable of being mapped to a learning trajectory that ends at some proficiency goal (Harris et al., 2016). In ideal cases, prior literature provides empirically-backed models for what claims ought to be made at certain grade bands (Duncan & Rivet, 2013). For cross-cutting concepts (CCCs), however, the lack of research into what students know as they build overall science competency makes the determination of appropriate claims difficult. For this reason, the primary research question of my dissertation is: *What does student demonstrations of the crosscutting conceptual reasoning aspect of their three-dimensional science understandings look like at different levels of overall science understanding?*

To address this broad question, a project titled: Developing Assessments and Tools to Support the Teaching and Learning of Science Crosscutting Concepts (DAT-CROSS) was created in collaboration between the Educational Testing Service (ETS) and Indiana University. The project was supported by a grant (*R305A170456*) from the National Council for Education Research (a division of the U.S. Department of Education's Institute of Education Sciences) to develop an empirically-validated progression for CCC understanding (particularly regarding the CCC's *Systems and System Models* and *Structure and Function*; NGSS Lead States, 2013c). Building on my previous working relationships with several ETS researchers, as well as one of my prior developed expertise on Crosscutting concepts (Rivet et al., 2016; Weiser et al., 2017), I was invited to design an assessment suite capable of meeting the learning progression goals of the project. As with many IES-funded grant projects, the DAT-CROSS endeavor is designed to continue across several years, culminating in materials that can be readily provided to teachers to formatively assess students' CCC understandings. The data for my research question derived from the first two years of the project, in which the team investigated the 'usability' (Zaharias & Poylymenakou, 2009) of the assessment in order to validate the chosen tasks (Johnstone, Bottsford-Miller, & Thompson, 2006) and to minimize the role of non-focal skills (such as ELA proficiency or computer-use competency) in mediating students' ability to engage with the tasks (Hoadley, 2004).

In the process of working on the DAT-CROSS project, it became clear that part of the challenge in the design of assessments for crosscutting concepts was the lack of clarity over how the concepts were distinct from one another. As recently as 2018, researchers in science education have expressed doubts over the utility of even attempting to measure crosscutting concepts in-use (Osborne, Rafanelli, & Kind, 2018). Answering my initial research question entailed, first, answering a brand-new question: *Is there evidence that the crosscutting concepts of the NGSS are distinct constructs that can be measured as students use them?* Simultaneously, the goal of design research is not merely to confirm that the designed product works as intended. The goal of design research, particularly when student assessment is involved, is to identify all the impediments to making a locally functional solution while keeping to existing design principles (Joseph, 2004). This goal manifested a third research question, described below, that focuses on the challenges faced by my subjects as they interacted with assessment task elements.

Research Questions

There are three research questions for this study:

1. *Is there evidence that the crosscutting concepts of the NGSS are distinct constructs that can be measured as students use them?* Answers to this question undergird the possibility of a measurement model.
2. *What does student demonstrations of the crosscutting conceptual reasoning aspect of their three-dimensional science understandings look like at different levels of overall*

science understanding? Answers to this question will inform the development the evidence model.

3. *What challenges do middle-school-aged subjects face in presenting their CCC understandings while engaging with an interactive suite designed to target Systems-Models and Structure-Function dimensions of their three-dimensional science understandings?* While CCCs represent an important constituent of students' science reasoning, the dimension is often the subtlest constituent of a given performance. As students face challenges in the task, determining that their barriers stem from underdeveloped CCC understanding is critical to determining the usefulness of the assessment.

Chapter 2

LITERATURE REVIEW

Dimensions of the NGSS

The first step in understanding the rationale for this thesis study comes from unpacking the major dimensions of the Next Generation Science Standards (NGSS).

Performance Expectations: The structure of the NGSS*.*

The Next Generation Science Standards are composed of performance-based tasks which combine all three dimensions of successful science education: a) disciplinary core ideas that make up the relevant science content (e.g., the relationship between energy and forces), b) scientific practices that tie instruction to activity of the scientific community (e.g., developing and using models), and c) crosscutting concepts that reflect the commonalities of the task and main ideas to all science disciplines (e.g., cause and effect) (Krajcik & Merritt, 2012; Krajcik et al., 2014; NRC, 2012). The synthesis of these three dimensions at a relevant grade-band results in a performance expectation (or PE) stated like this: "HS-PS3-5. Students who demonstrate understanding can: develop and use a model of two objects interacting through electric or magnetic fields to illustrate the forces between objects and the changes in energy of the objects due to the interaction." (NGSS Lead States, 2013a). This reflects a radically altered view of the nature of science, and its relation to science education, compared to that of the *National Science Education Standards* (National Research Council, 1996) and the *Benchmarks for Scientific Literacy* (AAAS, 1993). Both focused on developing inquiry practices and on replacing students' alternative conceptions with those AAAS considered more accurate (Kesidou & Roseman, 2002). Where earlier national standards suggested that students have sufficient knowledge when they “know” particular science content, the NGSS states that students reveal their knowledge by “doing” a task that requires use of that knowledge (Krajcik et al., 2014).

Disciplinary Core Ideas

Disciplinary core ideas (DCIs) are the domain-specific content of science knowledge writ-large, best exemplified by the information contained within textbooks or flashcards. Ideas like Newton’s laws or the process of natural selection are examples of this dimension that represent key building blocks to modern scientific understandings (Krajcik, 2015). These building blocks have been featured in science standards for about as long as the concept of educational standards have existed (DeBoer, 1991). Unsurprisingly, the historic salience of the content, now categorized as DCI, too often dominates in the classroom at the expense of the other dimensions of the NGSS (Stroupe, 2015). Successfully planning instruction aimed towards achieving the goals of the NGSS entails a careful balancing act between helping students develop sophistication around these core ideas and helping them understand the ways these core ideas are actually used by scientists (Krajcik et al., 2014). Without additional supports for teachers, particularly in assessing student understandings of the other two dimensions of the NGSS, it is suspected that instructors will default to a focus on DCIs (Pellegrino et al., 2014).

Science and Engineering Practices

Much of the recent literature on science learning that emerged from the learning science field has centered on how knowledge gets used (Brown, Collins, & Duguid, 1989; Cognition and Technology Group at Vanderbilt University, 1992; Edelson, 2001; Krajcik & Merritt, 2012; Nersessian, 2002; Stroupe, 2015). Though described in earlier standards in terms of scientific inquiry (DeBoer, 1991; Edelson, 2001), the various means by which scientists (and by extension science students) engage with their knowledge and build new understanding are featured as the second dimension of the NGSS: Science and Engineering Practices (SEPs) (Krajcik & Merritt, 2012). There are eight such practices: a) asking questions and defining problems; b) developing and using models; c) planning and carrying out investigations; d) analyzing and interpreting data; e) using mathematical and computational thinking; f) constructing explanations and designing solutions; g) engaging in argument from evidence; and h) obtaining, evaluating and communicating information (Osborne, 2014). While each of these practices have appeared in a litany of past educational standards (Osbourne, 2014), prior research has been uneven on what those practices look like both at the expert and novice level (Stroupe, 2015). Overwhelmingly, research has centered on the development of science models (Pluta, Chinn, & Duncan, 2011; Schwarz et al., 2009; Schwarz & White, 2005; Svoboda & Passmore, 2011) and the construction of scientific arguments (Berland & Reiser, 2009; Osborne et al., 2016; Sampson & Clark, 2009). While this focus is a reasonable function of the nature of the scientific enterprise (McComas, Clough, & Almazroa, 2002), it means that there is a dearth of empirically-validated instructional approaches to helping students develop practical skills along many of the other SEPs.

Crosscutting Concepts.

Crosscutting concepts are a comparatively new addition to the core components of science standards (first appearing as part of the “unifying concepts and processes” in the National Science Education Standards produced by the NRC in 1996 [p.104]) and represent themes that underlay the scientific enterprise (Rivet et al., 2016) in all domains and disciplines. These seven themes are: a) patterns; b) cause and effect; c) scale, proportion and quantity; d) systems and systems models; e) energy and matter; f) structure and function; and g) stability and change (NGSS Lead States, 2013b). The NGSS prides itself on its empirical backing, and its framers routinely claim that the scope and sequence encouraged by the NGSS and its supporting documents was derived from evidence around how students learn science concepts (Duschl, 2008; Duschl et al., 2007; NRC, 2012). While this claim is mostly true for disciplinary core ideas, and partly true of science and engineering practices, there is a well-established lack of evidence for crosscutting concepts (Rivet et al., 2016). For this reason, some critics have claimed that CCCs make a poor addition to the new standards (Osborne et al., 2018), but I disagree with the assessment that the current absence of evidence for CCCs in student work is evidence that the dimension is not equally important in student learning (Weiser et al., 2017).

As in the case of SEPs, what prior research does exist regarding student understanding of the concepts at play for the CCC dimension of the NGSS is limited to just a few of the enumerated themes. While much has been written about matter and energy (Jin & Anderson, 2012; Neumann, Viering, Boone, & Fischer, 2013; Stevens, Delgado, & Krajcik, 2009), this research generally treats such concepts as bounded by the uses that are particular to a science domain (akin to DCIs). One of the few instances in which research into CCCs seems to consider the researched construct as a concept that might apply across contexts is with regards to students' use of systems and systems models. Examining this literature across biological, physical, and earth sciences (Breslyn et al*.,* 2016; Gunckel et al., 2012; Jin & Anderson, 2012: Mohan et al., 2009; Songer et al*.,* 2009), I found four general pathways by which sophistication in systems thinking builds over time as follows.

1. System Phenomena – As students build sophistication in their thinking around systems, they are better able to describe phenomena as a system of many simultaneous interactions.
2. System Components – As students build sophistication in their thinking around systems, they are better able to break down the components of a system into their constituent, dynamic parts.
3. System Relationships – As students build sophistication in their thinking around systems, they are better able to describe the relationships between previously identified components of the system.
4. System Boundaries – As students build sophistication in their thinking around systems, they are better able to define the boundaries of the system and track the flows of inputs and outputs across those boundaries.

These pathways will later serve as key indicators of progress used in the development of the tasks that make up my research endeavor (see Appendices A and B).

Given the philosophic and epistemic nature of the CCCs, I believe it is more likely that appropriate instruments for finding CCCs still need to be developed (Rivet et al., 2016). In previous research, my team found a set of four roles that students expressed in their understanding of CCCs when constructing scientific explanations (Weiser et al*.,* 2017) as listed below.

1. CCCs as Lenses – The role of the CCC is to highlight salient features that may not be immediately obvious due to scale, scope, or size.
2. CCCs as Bridges – The role of the CCC is to draw connections between two entities, facilitating transfer of understanding.

3. CCCs as Levers – The role of the CCC is to combine several, related ideas, entities, or representations in order to make understanding take on a new form towards a particular goal.
4. CCCs as Rules – The role of the CCC is to validate a representation's utility in explaining or predicting the natural world.

Whichever role students employ, it radically alters the relationship between the evidence used in the explanation and the subject-phenomena relationship. By thinking about CCCs in terms of these roles, I believe I can construct assessments that are both better able to elicit evidence of CCCs, and better contextualize that evidence in terms of students' three-dimensional science understandings.

Learning Progressions

One of the advertised features of the NGSS is their basis in *empirically derived* (Duschl et al., 2007) research into how students learn. Exemplified in the work of Gotwals and Songer (2013), learning progressions (LPs) are the dominant framework for identifying what that empirical evidence looks like (Corcoran, Mosher, & Rogat, 2009). While discussions that appear in Berland and McNeill's (2010) work paint a very rosy picture of the positive implications of the LP framework, there is also much discussion focusing on the notion of a "messy middle" (Gotwals & Songer, 2013, p. 599) where the pathways between well-established stepping-stone ideas are less clear. The dominant perspective on learning progressions and learning science research (Lehrer et al., 2001; Rivet & Krajcik, 2008) suggest that progressions are anchored by the contexts of a particular phenomenon, making it hard to generalize one learning progression (for example, on moon phases [Plummer & Krajcik, 2010]) to another about a different topic, such as Newton's laws (Alonzo & Steedle, 2009). Some existing progressions attempt to be applicable across disciplinary contexts by focusing on science practices, as in the case of Berland and McNeill (2010); or on concepts that appears in all science domains, as in the case of Neumann et al. (2013). However, such progressions are few and far between (Duschl et al., 2011). LPs continue to be a useful paradigm for research and theory, but incorporation into instructional practice requires greater elucidation.

Contemporary Critiques

No reform movement is without its constructive critics, and while there has been longstanding agreement that past standards have failed to live up to their lofty goals (Eisenhart, Finkel, & Marion, 1996; Lee, 1997), the best step forward always remains a topic of debate. For the NGSS, an early critique of its emphases was the diminished focus on the nature of science, which was initially absent from the standards until its inclusion as an addendum document (Appendix H of the NGSS [NGSS Lead States, 2013]). For science educators who argued that nature of science is a critical content element that needs to be explicitly discussed in classrooms (Abd-El-Khalick, Bell, & Lederman, 1998; Lederman & Zeidler, 1987), relegating nature of science content to an appendix that only appears on some performance expectations would lead teachers to avoid such topics when weighed against their other instructional responsibilities (Lederman & Lederman, 2014). Coupled to this discussion were critiques over the presence of new dimensions and the uneven degree to which existing learning progressions could describe how these dimensions grew in sophistication as students mature (Duschl, Maeng, Sezen, 2011). Returning to the notion of consequential validity, the true test of the theories undergirding the NGSS will be the degree to which instructional-material developers (including assessment designers) can produce the tools capable of helping students reach the supposed goals of the new standards (Furtak, 2017).

Concerns of equity in science learning have also been a locus of critical review of educational standards reform. As researchers like Lee (1997) and Brown (2005) investigated during the rollout and, later, enactment of *The Benchmarks for Science Literacy* (AAAS, 1993), the definitions for terms like "science literacy" (Lee, 2005) become definitions for who has access to effect resources. As states begin the process of implementing the NGSS, a new generation of researchers like Mutegi (2011) and Elam-Respass (2018) are wondering how new standards will affect historically under-represented teachers and students. Equity in the NGSS, as a set of standards driven by student doing over student knowing, is uniquely tied to assessment equity (Rodriguez, 2015). Now more than ever, alignment to both the text and the goals of the NGSS requires the design of assessments responsive to students' linguistic skills and socially-mediated cognitions (Mislevy, 2016).

Anchoring Learning and Assessment Contexts

The concept of learning as cognitive apprenticeship, developed by Brown, Collins, and Duguid (1989), tells us that learning represents a master-novice relationship between the teacher and the student in which the student becomes enculturated in authentic practice through the doing of relevant activity, much as the apprentice of old practiced his craft under the tutelage of a master craftsman. To this end, they argue that all learning is situated in the context in which it was taught, with this context being key to the learner's ability to transfer concepts from the learning environment to new scenarios (J. Brown et al., 1989). Critical to developing mastery of transfer is "authenticity." Authenticity reflects the degree to which a cognitive tool that develops during learning in the classroom context is, later, usable to students as expert practitioners might wield it. To use an example developed by Brown et al. (1989), if someone uses a hammer as a paperweight, it is a valid use of the tool; but, it is not an authentic use. In contrast, developing the use of a hammer in context of building a birdhouse represents authentic practice of using the tool as a carpenter or craftsman would use it. The important difference is that our earlier example would not be extendable to many other instances in which a hammer might be useful, while our latter example provides useful knowledge about the kinds of scenarios in which authentic use of a hammer may be called for. In the most fundamental sense, every instance in which a tool may be used authentically has some commonality with every other possible instance of authentic use. Contextualizing instruction in terms of this authentic use helps the novice learner recognize commonalities, which assists in both integration (Rivet & Krajcik, 2008) and transfer (Yilmaz, Eryilmaz, & Geban, 2006) of new learning.

However, cognitive apprenticeship is more a description of what it means to learn than an explicit means of evaluating the quality of a given learning environment. There is some sense in which all learning environments are instances of cognitive apprenticeship (Brown et al., 1989) and, moreover, such instances are not inherently sufficient for developing a successful learning environment. Extending their theory on cognitive apprenticeship, Lave and Wenger (1991) presented a more complex analysis of learning environments; i.e., legitimate peripheral participation (LPP), which examines learning not only as apprenticeship, but as a means of moving peripheral members of a community of practice into full participants. Legitimate peripheral participation represents an analytical perspective for evaluating cognitive apprenticeship environments for their ability to produce desirable changes not just in novices' abilities, but in their role within a community of practice and in their self-identity in relation to that community. This is no small task, even a well-established apprenticeship can fail if the novices do not perceive themselves as moving from peripheral participation to central participation (Lave & Wenger, 1991).

In the ensuing decades of research following that of Lave and Wenger, key ideas from the frames of cognitive apprenticeship and legitimate peripheral participation have continued to prove useful. Examining the role of science identity (as influenced by peer and by teacher evaluation of language use), Brown et al. (2005) reiterate the importance of identity shift (Lave & Wenger, 1991). Understanding use of the established technology of a community of practice is needed in all learning environments, and LPP presents such understanding through the concept of *transparency*. Transparency represents a conjunction of the ability to recognize the use of a particular tool and the scenarios in which it has utility (visibility) and the ability to use that tool in order to achieve a desired end (invisibility). Lave and Wenger (1991) present the example of a window, which one uses to look outside from the inside of a confined space (it is invisible); but simultaneously, we must recognize it as something we can look out of in order to do so (it has visibility). A tool must be transparent to the learner for them to effectively use it in both peripheral and central tasks. The medical school student must be able to use a stethoscope both in their novice, peripheral stages working with dummies and in their central role on actual humans once they become residents. Failures to see a cognitive tool both in its present, novice use, and how experts similarly use it can often have disastrous results for students' views on the utility of their science knowledge (Sandoval, 2005). This idea remerged in critique of then-present science standards by Eisenhart, Finkel, and Marion (1996) who highlighted the importance of couching what is taught in the light of activities students find relevant to them and, by extension, are better able to see the desired end(s) of a piece of science knowledge. Authority also matters in terms of classroom discourse. Legitimate learning sets itself apart from traditional instruction in which the master will *talk about* a practice, opting instead for encouraging peripheral members to t*alk within* the established rules of the community (Lave & Wenger, 1991). Ford (2008) reiterates this idea in suggesting that a teacher's authority should extend from their ability to help students engage with the scientific discipline (and its rules) rather than from their position within the school hierarchy.

Rich Performance Tasks

For assessment development, all this research has served to show that the things students have learned are much more contextually dependent than might have been suspected by prior reform movements (which focused more on what scientists know than what student know; Eisenhart et al., 1996). Consequently, assessment tasks that measure learning ought to mirror the complexity of the learning process by accepting a wide array of potential performances that all successfully meet the criteria of a task (Mislevy, 2017). Such tasks are known as 'rich performance tasks' and are well aligned to the philosophical underpinnings of the NGSS (Gorin & Mislevy, 2013). Unfortunately, they are also more resource intensive (in capital, cognitive load, and time) than more traditional tasks. Determining and designing learning to achieve the appropriate balance between richness and domain range remains a topic of continued research (Weiser & Liu, 2018).

Socio-scientific issues, cases where there is a salient intersection between the understandings generated by scientists and ongoing human behavior (Kitcher, 2010), are productive contexts for creating rich tasks (Morin, Simonneaux, & Tytler, 2017).

Unfortunately, students struggle to readily engage with their science knowledge once a political lens has been applied to a problem context (Morin, Simonneaux, & Tytler, 2017) and teachers are hesitant to wade into a context that may prove rife with political pitfalls (Kilinc, Demiral, & Kartal, 2017). Designing towards a socio-scientific context for an assessment may prove fruitful for eliciting robust evidence (which may be needed when the target construct is so subtle as are CCCs [Weiser et al., 2017]), but care should be taken to choose those contexts where the instructor already has strong, positive affect (Kilinc et al., 2017).

Connection to the Dimensions of the NGSS.

The Next Generation Science Standards represent a major shift away from Bruner's inquiry (DeBoer, 1991) and towards more transformative, social-construction based activity (Blumenfeld et al., 2000). This shift, in addition to the redevelopment of benchmarks as performance expectations (Krajcik et al., 2014) represents clear attempts to establish technologic transparency (Lave & Wenger, 1991) within science content and to induce knowledge cycles among peers and near-peers through discourse, cooperation, and critique (Lave & Wenger, 1991; NRC, 2012). Along these dimensions, the NGSS represent a step forward from Project 2061 in establishing legitimate peripheral participation within the classroom. However, the choice to present these standards as being wholly differentiated from curriculum materials (Duschl et al., 2007), while conducive to teacher-student centered instructional design, also places a sizable burden on instructors to redevelop their lessons without as much instructional support as AAAS provided (AAAS, 2000).

Elements of Rigor in the Development of Assessments

Validity and Generalizability

Assessments exist in terms of their validity (Pellegrino et al., 2001): do they measure what they claim to measure? In order for my assessments to have content validity, they need to provide the supports that will allow assessors to measure students' CCC understandings. However, the three-dimensional framework of student science learning that is the basis for the NGSS (NRC, 2012) is clear in its assertion that evaluators cannot measure any one dimension in isolation, because the other two dimensions always affect the context of the evaluation. Thus, assessments designed to align to the NGSS need a very particular type of structural validity in which the relationships between the three dimensions are considered. This is not a simple task and the central concept of my dissertation research is that all currently existing assessments lack this form of validity, because they do not consider the role of CCCs. Generalizability is another challenge for my research project. All design projects are built upon developing solutions for local problems; this is especially true of frameworks, like activity theory, which take ethnographic approaches. Barab and Squire (2004) suggest a reconceptualization of generalizability that takes the form of what they call "consequential validity" (p. 8). For consequential validity, the designed solution is not the end goal of research. Rather, it is a tool for evaluating a theory of design (Cobb, Zhao, & Dean, 2009). Therefore, my research seeks generalizability not by suggesting that the particular assessment I have designed should be used in novel contexts, but rather by showing a successful method for design (Clarke & Dede, 2009) of science assessments that makes conclusions about students' three-dimensional understandings in a more structurally valid way.

The Assessment Triangle

The assessment triangle (Pellegrino, Chudowsky, & Glaser, 2001) sits at the center of all claims regarding the consequential validity of any psychometric instrument. The goal for any assessment is to make a conclusion about a latent cognitive process from a collected set of observable evidence. While evidence from observation derives from cognition processes, they are not the cognition itself. An interpretation scheme must be used to convert from the observations to the desired claims. These three corners of the assessment triangle (cognition, observation, and interpretation) need to agree if the assessment is to be valid – observables should be targeted to best elicit relevant evidence, the interpretation scheme should be able to account for all meaningful observations, and the claims about cognition should be limited to those where observables could be reasonably generated.

Evidence-Centered Design

From the evidence-centered design perspective, assessment serves as an opportunity for students to provide evidence of what they know, and as an opportunity for assessors to collect, evaluate, and draw conclusions from that evidence (DeBarger et al., 2015; Mislevy & Haertel, 2007). Mislevy and Haertel (2007) suggest that an early step in developing assessment is the construction of a conceptual assessment framework which dictates what you are interested in about a subject (the student model), how you will create opportunities for subjects to provide assessors with evidence (the task model), and what collected evidence needed for the analysis looks like (the evidence model). The evidence model can be further subdivided into an observation model that dictates what evidence will count as applicable, and a measurement model that dictates how observations will be converted into a form suitable for analysis (Mislevy & Haertel, 2007). The evidence model (and its subdivisions), which seems to correlate with the underspecified interpretation corner of the assessment triangle (Pellegrino et al., 2001), is the focus of my research.

Models of the evidence-centered design (ECD) process (Figure 2.1) frequently take a form similar to that of Toulmin arguments (Mislevy, 2017) in order to highlight how the validity of an assessment is not some metric like precision to be quantified, but rather is an ongoing argument that the instrument is successfully supporting useful claims about the subjects (Baxter & Mislevy, 2004).

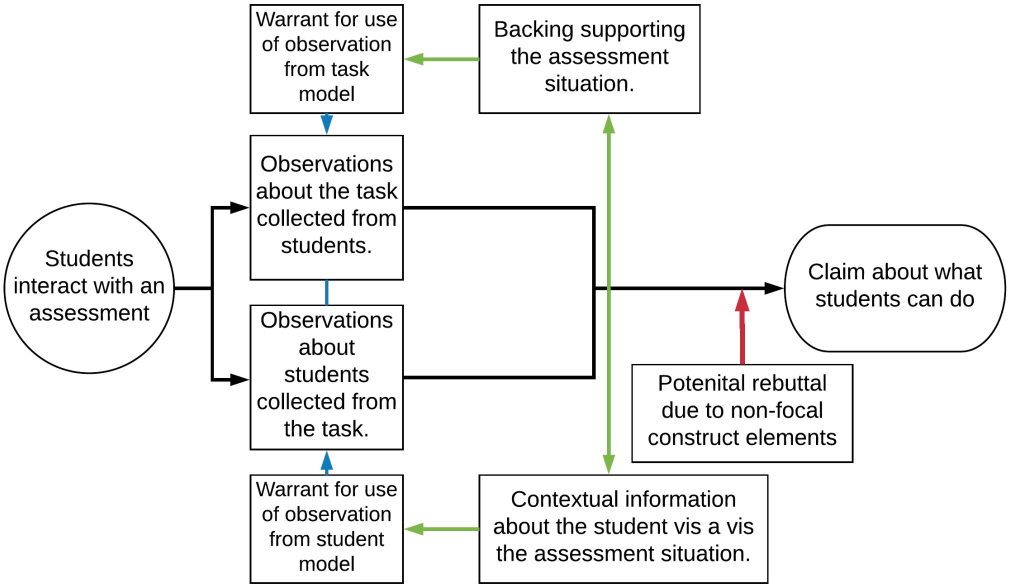

Warrant for use of observation from task model
Backing supporting the assessment situation.
Observations about the task collected from students.
Students interact with an assessment
Claim about what students can do
Observations about students collected from the task.
Potenital rebuttal due to non-focal construct elements
Warrant for use of observation from student model
Contextual information about the student vis a vis the assessment situation.

Figure 2.1 *Fundamentals of the ECD Validity Argument. Adapted from Mislevy et al. (2017)* "Assessing model-based reasoning using evidence-centered design: A suite of research-based design patterns." Cham, Switzerland: Springer International Publishing.

ECD and NGSS Performance Expectations

The goal of assessment in education is to measure current levels of proficiency in a particular domain (in this case, science) in order to support future growth. The NGSS have redefined what it means for students to be proficient in science (Pellegrino, 2013) by listing a particular set of claims regarding what proficient students can do (elucidated in the Evidence Statements which accompany each performance expectation [NGSS Lead States, 2013]). This makes the PEs of NGSS (or a bundle of them) well suited to the evidence-centered design approach, which frames assessments as the instruments that elicit evidence in support of proficiency claims (DeBarger et al., 2016). Once we bundle a set of PEs, so also, we have selected a set of claims to make about students; i.e., known as the student model (Mislevy & Haertel, 2006). From there the PEs can be unpacked into component dimensions as part of a domain analysis (Harris et al., 2016) that seeks to outline the many ways different parts of the dimensional constructs integrate as students demonstrate three-dimensional understanding. The next step is to construct an evidence model (Mislevy & Haertel, 2006) capable of delineating valuable evidence of student proficiency from evidence attributable to the non-focal knowledge, skills, or abilities that a given task might require. Our evidence model takes the form of a design pattern (Mislevy & Haertel, 2006), which highlights recurrent themes in the assessment of any one PE from the bundle linked by a viable sequence of science/engineering practices. Finally, we construct a task model (Mislevy & Haertel, 2006) that defines the way the item/task set will elicit evidence in the light of a provided scenario. In later subsections, I will further elaborate on the use of ECD in the creation of tasks targeting CCC constructs from a bundle of performance expectations.

Ethics of Assessment

Part of the Toulmin-esque validity argument made via the evidence-centered design approach to assessment development is accounting for alternative explanations by establishing that the tasks used were appropriate for both the use case (Gorin & Mislevy,

2013) and the population (Cohen & Lotan, 1995). Students with minimal experience with the terminology that is presented within a task may struggle to engage with task items (Noble et al., 2012). This struggle, therefore, may be a result of a lack of language rather than a lack of conceptual understanding (Brown et al., 2005). Equity in education considers what is socially relevant to students (Freire & Macedo, 1995) so that they have the best possible opportunities to learn. Likewise, equity in testing considers what is socially relevant to students (Wiliam, 2010) so that they have the best possible opportunity to present what they know to assessors. To assuage some, though not all, concerns all DAT-CROSS tasks were passed through an equity and fairness review process that considered context relevance and ELA difficulty in determining the appropriateness of the tasks. Additional

Designing the DAT-CROSS Assessment Suite

Crosscutting concepts are broad themes that consistently appear in similar forms across disciplinary boundaries and, despite differences in conceptual content, seem to perform similar roles within student understanding across contexts (Rivet et al., 2016). These roles are subtle and frequently lose salience to assessors when compared to the students' more obvious disciplinary core idea competency. Assessing CCCs, therefore, requires rich performance tasks (Gorin & Mislevy, 2013) that allow students to bring a wider swath of personal experiences and understandings to bear during the activity. Mislevy (2017) provides some insight into how rich tasks can still be implemented despite their high cognitive load. Although the performance of any one student on any one rich performance task may be too messy to successfully make individual-level construct-relevant conclusions, it is possible to use repeated measures of the construct across many contexts and many individuals that combine (or overlap) to produce a clearer picture of students' CCC understandings.

DAT-CROSS was designed to use three such rich tasks, each with a storyline around a familiar system: infestation of a farm by a dangerous insect (ecological systems), the shrinking water access of a farm town (earth-hydrology systems), and the consequences of starch-rich diets on human health (human body systems). Each storyline is grounded in real problems that scientists seek to solve (Brown et al., 1989) that students can contextualize (Noble et al., 2012) in light of their own experiences with insects, waterways, and food consumption. The following paragraphs exemplify the design process that I used to develop the ecological systems storyline.

Once I selected the ecological system of the farm as being a viable context for anchoring the assessment, the next step is to find appropriate performance expectations (PEs) that target the relevant CCC constructs, science practices, and core ideas applicable both to the scenario and to the objective of the research project (the CCCs 'system and system models' and 'structure and function'; NGSS Lead States, 2013c). I selected a bundle of MS-LS2-3, MS-LS2-4, and MS-LS2-5 (NGSS Lead States, 2013a) which jointly describe a task of modeling energy and matter flows within an ecosystem, how those flows can be affected by changes to the physical components of the ecosystem, and how humans might engineer a solution that seeks to maintain a pre-existing ecology. It should be noted that the PEs of this bundle do not fully align to the desired CCC constructs (a common problem due to uneven distribution of CCCs across the PEs of the NGSS). However, addendum documents to the NGSS focusing on CCCs (NGSS Lead States, 2013c) note that CCCs rarely exist in isolation and can frequently be grouped thematically (i.e. 'energy and matter' can be a useful lens in the understanding of systems or 'stability and change' can be critical to the continued functioning of a particular structure). Appendix A highlights how the PEs were adapted from their original form to place greater emphasis on the new focal CCCs.

In process of making sense of the phenomenon (Krajcik & Merritt, 2012) regarding the consequences of an insect infestation to a pre-existing ecology, and the steps a local community might take to protect their farmland, students will naturally present their understandings of systems (Breslyn, et al., 2016; Gunckel, et al., 2012; Jin & Anderson, 2012: Mohan, et al., 2009; Songer et al., 2009; Yoon et al., In Press), food webs (Griffiths & Grant, 1985; Hogan, 2000; Hokayem, Ma, & Jin, 2015), argumentation (Berland & McNeill, 2010), and modeling (Schwarz et al., 2009). The prior literature on practices, DCI learning progression, and expected use of the CCC *Systems and Systems Models* served to set the claims regarding what it means for students to be scientifically literate in terms of the real-world scenario. This acted as a form of domain analysis (Mislevy & Haertel, 2006) that facilitated creation of a task model describing a broad set of tasks that could potentially be constructed for this research study; only a subset of those possible were produced. Figure 2.4 outlines the process by which the three selected PEs were unpacked and translated into a task model (described in Appendix A) appropriate towards the CCC-assessing goals of the DAT-CROSS project.

One of the biggest challenges associated with the design process was how to maintain the salience of CCCs even after integrating the more well-supported dimensions of the NGSS. Unlike the integration process used by Harris et al. (2016), which combined all three dimensions at the same time following their domain analysis, I staggered the integration of the DCI. This permitted greater in the design process toward the interaction of SEPs and CCCs, before the salience of the DCI overwhelmed the design process.

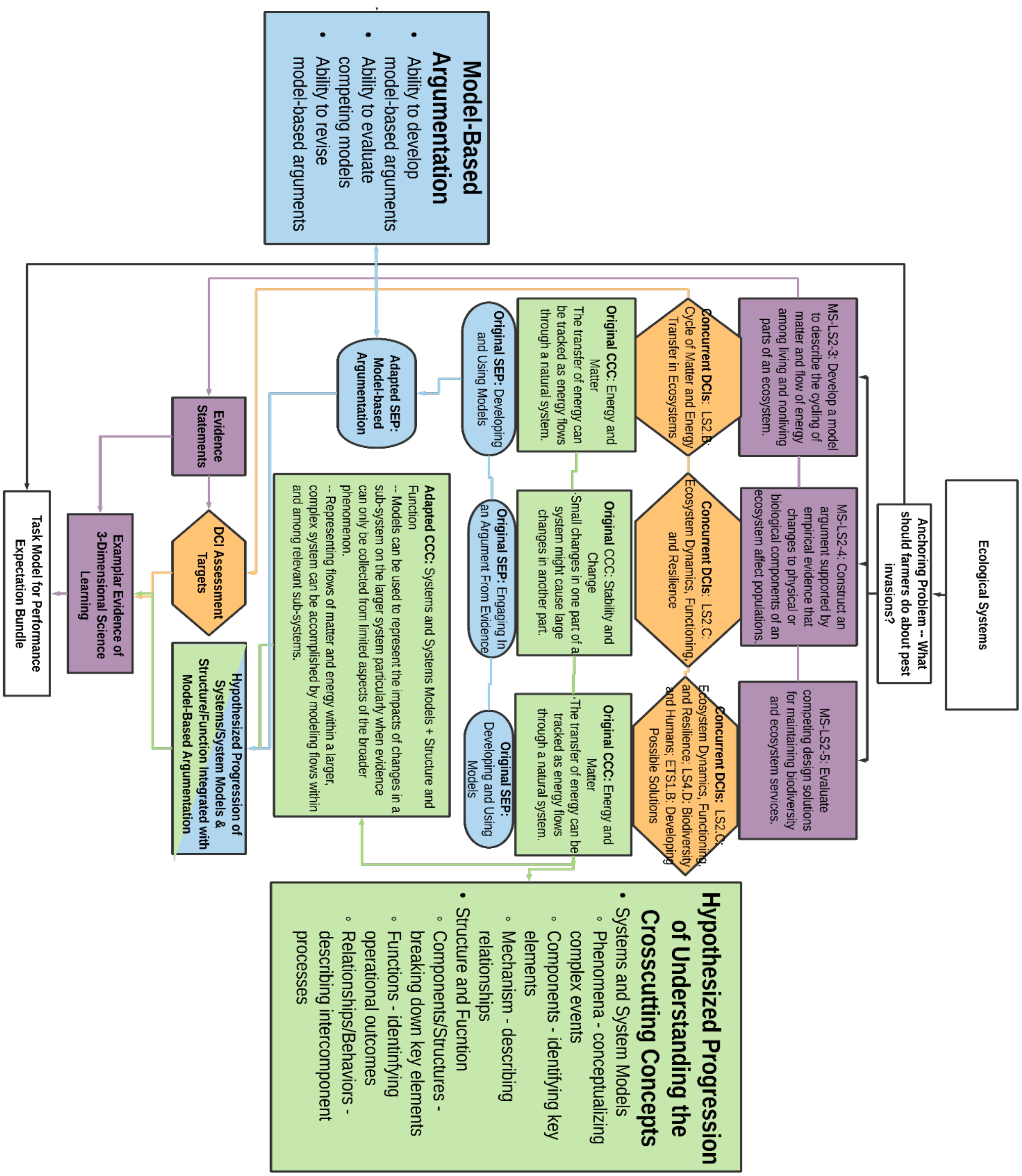
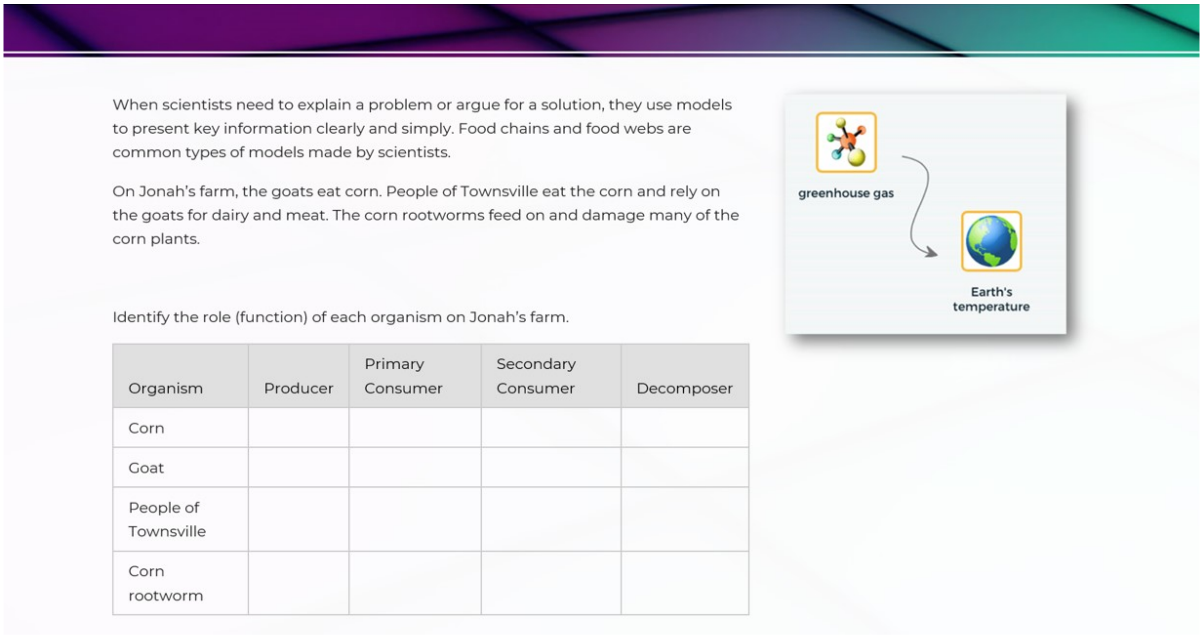

Ecological Systems
Anchoring Problem -- What should farmers do about pest invasions?
MS-LS2-3: Develop a model to describe the cycling of matter and flow of energy among living and nonliving parts of an ecosystem.
MS-LS2-4: Construct an argument supported by empirical evidence that changes to physical or biological components of an ecosystem affect populations.
MS-LS2-5: Evaluate competing design solutions for maintaining biodiversity and ecosystem services.
Concurrent DCIs: LS2.B: Cycle of Matter and Energy Transfer in Ecosystems
Concurrent DCIs: LS2.C: Ecosystem Dynamics, Functioning, and Resilience
Concurrent DCIs: LS2.C: Ecosystem Dynamics, Functioning, and Resilience; LS4.D: Biodiversity and Humans; ETS1.B: Developing Possible Solutions
Original CCC: Energy and Matter
The transfer of energy can be tracked as energy flows through a natural system.
Original CCC: Stability and Change
-Small changes in one part of a system might cause large changes in another part.
Original CCC: Energy and Matter
-The transfer of energy can be tracked as energy flows through a natural system.
Original SEP: Developing and Using Models
Original SEP: Engaging In an Argument From Evidence
Original SEP: Developing and Using Models
Model-Based Argumentation
Ability to develop model-based arguments
Ability to evaluate competing models
Ability to revise model-based arguments
Adapted SEP: Model-based Argumentation
Adapted CCC: Systems and Systems Models + Structure and Function
-- Models can be used to represent the impacts of changes in a sub-system on the larger system particularly when evidence can only be collected from limited aspects of the broader phenomenon.
-- Representing flows of matter and energy within a larger, complex system can be accomplished by modeling flows within and among relevant sub-systems.
Hypothesized Progression of Understanding the Crosscutting Concepts
Systems and System Models
Phenomena - conceptualizing complex events
Components - identifying key elements
Mechanism - describing relationships
Structure and Function
Components/Structures - breaking down key elements
Functions - identinfying operational outcomes
Relationships/Behaviors - describing intercomponent processes
Evidence Statements
DCI Assessment Targets
Hypothesized Progression of Systems/System Models & Structure/Function Integrated with Model-Based Argumentation
Exemplar Evidence of 3-Dimensional Science Learning
Task Model for Performance Expectation Bundle

Figure 2.2 Diagram of Ecological Systems Task Model Development

Developing Cognitive Lab Protocols

In assessment development, the role of cognitive labs – in which students of target age and/or content familiarity work through the task while "thinking aloud" (Johnstone et al., 2006) – is to validate test elements that will not be readily accessible once the task moves into production stages. It is of critical importance that test items have construct validity (Peterson et al., 2017). However, in multiple choice questions especially, it may not be possible to know if the reasons why students select a particular response is actually reflective of reasoning within the relevant construct (Howell et al., 2013). Simultaneously, the target audience of the final product needs to be viewed as usable and useful (O'Brien et al., 2008). Accordingly, the final research protocol was manifested in two halves: a content validation half and a reflective use half. These two halves, described in subsequent subsections, are presented in Appendices C and D.

A protocol for content validation. Howell et al. (2013) describe two research goals addressed by the cognitive interview method. The first goal is confirming items are appropriately keyed such that the stated best answer is the only one that students could successfully justify within the context of the question. The second goal is to review responses to distractor options to ensure they reflect a reasoned guess rather than the result of some non-focal barrier. This requires the support of "design rationales" (Howell et al., 2013), which describe both the intended reasoning that each task of an assessment is targeting, and the plausible reasonings that would reflect a well-reasoned inaccuracy rather than a use issue. The answer to these questions support what Snow and Katz (2009) call the "interpretive argument," a stage of the evidence-centered validity argument in which the designer can argue that the interpretation of evidence collected during the task is actually useful towards claims about what students know or can do. Peterson et al. (2017) notes a cog-lab research cycle in which good understanding of the goals of each item supports identification of both successes and barriers to student performance. Below, I exemplify an item design rationale for question 2 of the DAT-CROSS assessment:



When scientists need to explain a problem or argue for a solution, they use models to present key information clearly and simply. Food chains and food webs are common types of models made by scientists.

On Jonah's farm, the goats eat corn. People of Townsville eat the corn and rely on the goats for dairy and meat. The corn rootworms feed on and damage many of the corn plants.

Identify the role (function) of each organism on Jonah's farm.

Organism	Producer	Primary Consumer	Secondary Consumer	Decomposer
Corn				
Goat				
People of Townsville				
Corn rootworm				

Figure 2.3 Question 2 of DAT-CROSS with Item Design Rationale, below

Key: Corn – Producer, Goat – Primary Consumer,
Corn rootworm – Primary Consumer
Human – Primary Consumer *and* Secondary Consumer

Item Design Rationale

The major decisions to be made in the design of items is addressed below.

What is the assessment item asking? This item presents a variety of trophic roles organisms can take on within an ecosystem and asks respondents to correctly associate story elements (corn, goats, people, and corn rootworms [CRWs]) with those roles.

What information is important? It is important to pay attention to respondents' ability to differentiate primary and secondary consumption, identify which kind of organisms are considered producers, and to see CRWs as invasive consumers rather than key decomposers.

What is the rationale associated with each possible answer? Note that rationales also exist for some select distractor responses. Further rationales may need to be investigated with interview probes.

1. **The corn are producers –** Corn produces edible biomass by converting matter (carbon in the air and water from air/soil) by using energy from the sun.
2. **The goats are primary consumers. –** The goat directly consumes the corn.
3. **The people of Townsville are both primary consumers *and* secondary consumers. –** Humans consume corn both by directly eating it and by eating organisms (the goats) that eat the corn.
4. **The people of Townsville are one of either primary *or* secondary consumers -** Humans consume corn *either* by directly eating it *or* by eating organisms (the goats) that eat the corn. (**Note:** check respondent's reading of the story and whether they noted that both consumptions occur).
5. **The corn rootworms are primary consumers. –** The CRWs directly consume the corn.
6. **The corn rootworms are decomposers. –** The CRWs cause the corn to die (decompose).

In addition to outlining the tasks undertaken by students, Appendix C includes my documentation of reasoning and design rationale that undergirds each response option.

A protocol to inspect usability barriers. Nonetheless, it is not enough that questions are construct relevant. They must also be perceived as authentic by students (O'Brien et al., 2008) and align to what they should be able do at their stage of development (Johnstone et al., 2006). Inauthentic tools will feel frustrating and foreign, potentially limiting ability or willingness to engage with a task; while tools that are beyond the scope of students' skills will not provide construct-relevant information. Benson et al. (2002) provide a series of heuristics for identifying usability, particularly in computer-assisted tasks (like DAT-CROSS), which served to inform a second, semi-structured protocol. During these interviews, I asked students to reflect on task elements and identify challenges in task clarity, ability to error check, and their comprehension of the content in the task storyline with questions such as: "Was the video useful to help you understand the relationships in the system?”; “Based on the video, what was the purpose of the simulation?”; and “What did you learn from the video?” A more thorough detailing of the usability reflection protocol is provided in Appendix D.

Framework for Analysis

Confirmatory Factor Analysis

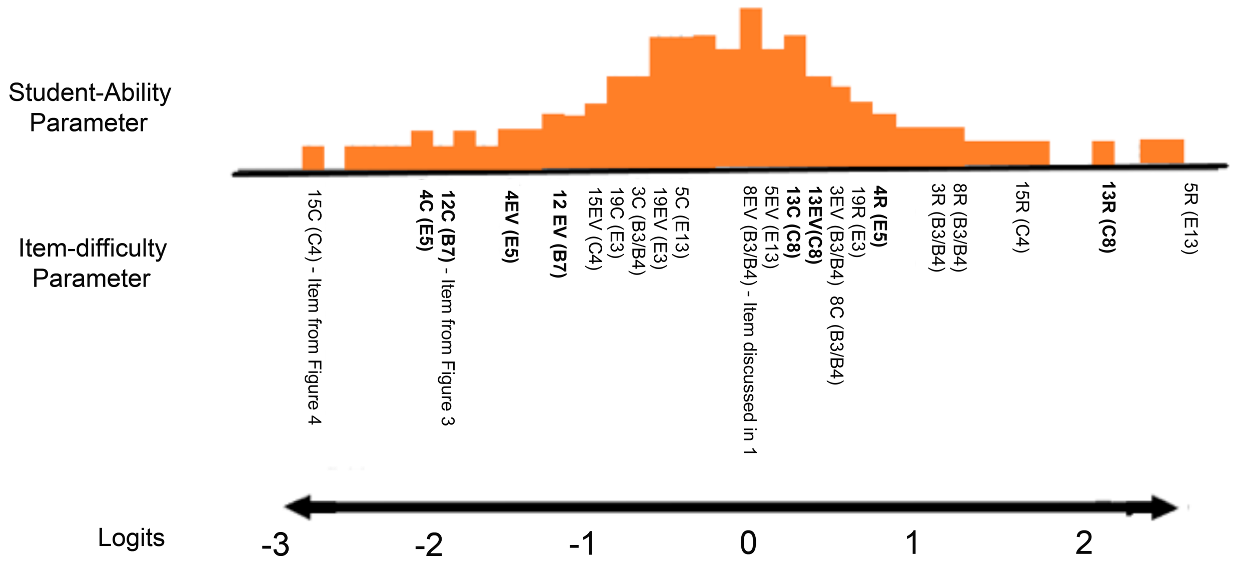
Factor analysis is a critical tool in this process of supporting a validity argument (Kline, 1994). Factor analysis uses linear algebra to break down the scores assigned to each item into a lesser number of latent variables which should (if the instrument has construct validity) map to the target constructs. Confirmatory factor analysis (as opposed to exploratory factor analysis) is used to ‘validate’ a set of constructs; whereas, exploratory factor analysis is used to analyze data to potentially discover underlying constructs or factors.

Rasch Models and Wright Maps

While many analysis approaches fit within the learning progression paradigm (Rivet & Duncan, 2013), a current dominant means of interpreting student responses in order to empirically support a hypothesized progression makes use of the Rasch model (Briggs & Alonzo, 2012). The Rasch model is a special class of item-response-theory models which supposes that the only meaningful factors in determining the likelihood that student i gets task item n correct is the relative magnitudes of the student’s latent ability (δ_i) and the difficulty of the item (β_n). As in Equation 1, below, these two factors combine to dictate a logit model.

$$\Pr\{X_{ni} = 1\} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}} \qquad (Equation\ 2.1)$$

The Rasch model can be amended to consider partial credit (as in Gotwals & Songer, 2013) and some examples exist in which a greater number of item-parameters (such as discriminability) are included (as in Wilson & Draney, 2005), but the primary end-goal of all Rasch models is the Wright Map. Since the two factors (ability and difficulty) are modeled onto a shared interval scale, we can graph them along a shared axis (e.g., Figure 2.2), whereby each item (the right side of the map) sits at the same height at which a student of corresponding ability should have a 50% chance of getting the item correct. By coding questions to particular construct levels, the Wright map can provide some visual evidence of patterns in the difficulty of items relative to the construct levels (Rivet & Kastens, 2012). In the figure below (adapted from Gotwals and Songer [2013]), questions of code E5 consistently sit below questions of code C8, suggesting that construct E5 (regarding ability to detect scientific evidence) is easier than construct C8 (regarding ability to detect scientific claims). This visual evidence can be further supported by ANOVA of difficulty scores to determine if observed difficulty differences are statistically significant (Lee & Liu, 2010).



Legend: The orange histogram bars indicate the number of accurate responses for each evaluation item in a series of increasing difficulty, some of which are indicated by labels listed at the base of the histogram.
Figure 2.4 Example of a Wright Map. Adapted from Gotwals & Songer (2013). Validity evidence for learning progression-based assessment items that fuse core disciplinary ideas and science practices. *Journal of Research in Science Teaching*, 50(5), 597–626.

Activity Theory

Developing an observation model requires deep thinking regarding what kind of evidence assessors can observe and what kind of evidence is useful to observe. A careful consideration of contextual factors can help assessors make these needed distinctions between the many observations. From the situated action perspective (Nardi, 1996), context plays the most important role in transforming a potential assessment space into an actualized assessment space, often transforming these spaces in ways the assessor cannot possibly predict. In contrast, the distributed cognition perspective (Nardi, 1996) suggests that assessment spaces exist as instances of shared cognition between assessors (who put thinking into making a problem solvable) and subjects (who put thinking into producing the solution). Activity theory sits as a sort of middle ground between the two (Nardi, 1996) by implying that while context ultimately directs the subject's cognition, and that cognition can be constrained by the assessor towards a particular goal. Engestrom (2000), as the founder of activity theory, describes the framework as a method for analyzing the interaction between subjects, the object of their performances, the instruments or tools used to operate on the objects, the rules for acceptable performance, the norms of the community, and the division of labor among relevant parties. These six factors, which all contribute to the achievement of a desired outcome, are frequently depicted as vertices of interconnected triangles. Student work, too, is a form of activity – depicted in hypothetical triangle form as Figure 2.3.

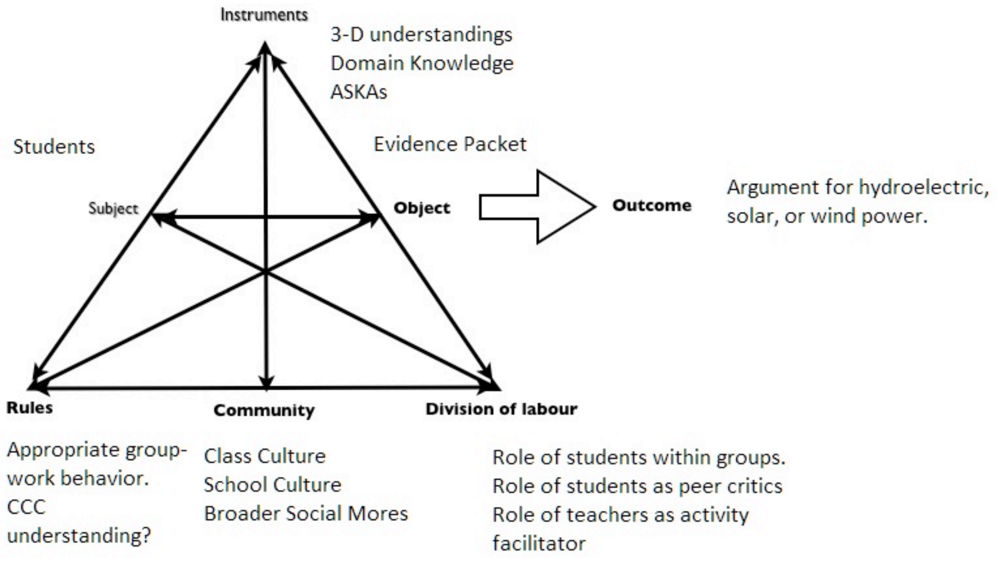


Figure 2.5 Example of Engestrom's (2000) Activity Theory Model in Student Work

Jonassen and Rohrer-Murphy (1999) explains how activity theory can serve as a framework for evaluating and developing learning environments. While they focus on activity theory as a mechanism for evaluating design interventions, the ideas can carry-over to evaluation of student performances (as in the activity theory-based design process by Sparks and Deane, 2015). Successful student performances involve effectively using a model as a tool for transforming the object to align with the performance goal, activity theory can be used to understand when these performances are disrupted. Cohen (1998) suggests that these disturbances are a function of mismatches between expectations of subjects and objects, usually due to the differing contexts by which the two live. Anchoring context (Cognition and Technology Group at Vanderbilt, 1992) may, therefore, prove effective in constraining context while making explicit to students the kinds of performances deemed acceptable by the community.

I believe the findings that emerged by employing the activity theory framework were invaluable in directing much of the analysis into how the DAT-CROSS assessment might be improved. Indeed, much has been written on the affordances of activity theory (Nardi, 1996; Roth & Tobin, 2002). Since the early 2000s, however, increased attention has been placed on adapting activity theory to account for the roles culture and history play (Nussbaumer. 2012). The cultural-historical activity theory model (CHAT) considers that what scripts for activity exist for the same tools develop under different contexts and manifest in different forms under the influence of both the lived experiences of the subject but also the collective experiences of the culture the subject inhabits. More nuance is required, therefore, in the claims I make regarding the kinds of students who might be willing to perform usability tests for a company that specializes in assessment design.

Accordingly, I cannot claim that participants are representative of the general population of future users but I can suppose that these participants are at least best-case use-participants for their grade band. That is, any usability challenges faced by this best-case sample of users are likely to be faced by most other students, and possibly in a more demanding way. Moreover, for the very first implementation of a novel research enterprise, it is often useful to begin with more tractable participants to better establish a foundation for broader future research. It is left to future research endeavors to examine how students use CCCs not only across disciplines (as set forth by the NRC [2014]), but also across cultural contexts.

Chapter 3

METHODOLOGY

In this chapter, I describe the procedures entailed in both collecting and analyzing the data for my dissertation research endeavor. I used different forms of data and, by extension, employed different analytic procedures for each of my three questions interwoven amongst each other in what Creswell would call an "analytic spiral" (Creswell, 2009, p.183). Below, I restate my three research questions.

1. *Is there evidence that the crosscutting concepts of the NGSS are distinct constructs that can be measured as students use them?*
2. *What does student demonstrations of the crosscutting conceptual reasoning aspect of their three-dimensional science understandings look like at different levels of overall science understanding?*
3. *What challenges do middle-school-aged subjects face in presenting their CCC understandings while engaging with an interactive suite designed to target Systems-Models and Structure-Function dimensions of their three-dimensional science understandings?*

A more succinct overview of these methods can be found in the data collection matrix (Appendix F).

Addressing these research questions required two forms of data: a) the multiple-choice responses directly produced by respondents when they engaged with the test suite, and b) selected interview responses obtained from the respondents' narrative as they participated in a think-aloud task that involved a review of all the components of the activity (Hoadley, 2004). The first of these data sets is quantitative in nature while the second is qualitative; thus, all-together requiring a mixed-methods approach commonly employed in the design-based research paradigm (Brown, 1992). However, unlike some narrative forms of qualitative research that emphasize the semantic meanings of the narrative as a record of the individual's lived experiences (as in the case of ethnography, case study, or grounded theory (Creswell, 2009); the goal of these interviews is to confirm that the prompts and stimuli used in the assessment are successfully inducing students to provide evidence of relevant cognitive processes. Ideally, the assessment situation establishes a positive climate for activity, where the student is not faced with significant barriers from the non-focal processes involved in the doing of tasks (i.e., the subjects' knowing the correct buttons to click on screen to log their desired multiple-choice selection, etc.). Prior lived experiences play a role – as they do for all activities of life – in how students both engage with assessments-writ-large and how they do so in the specific, somewhat rare context of usability interviews. However, the analysis of such impacts remains a topic of future study.

Data Collection

Participants and Experimental Setting.

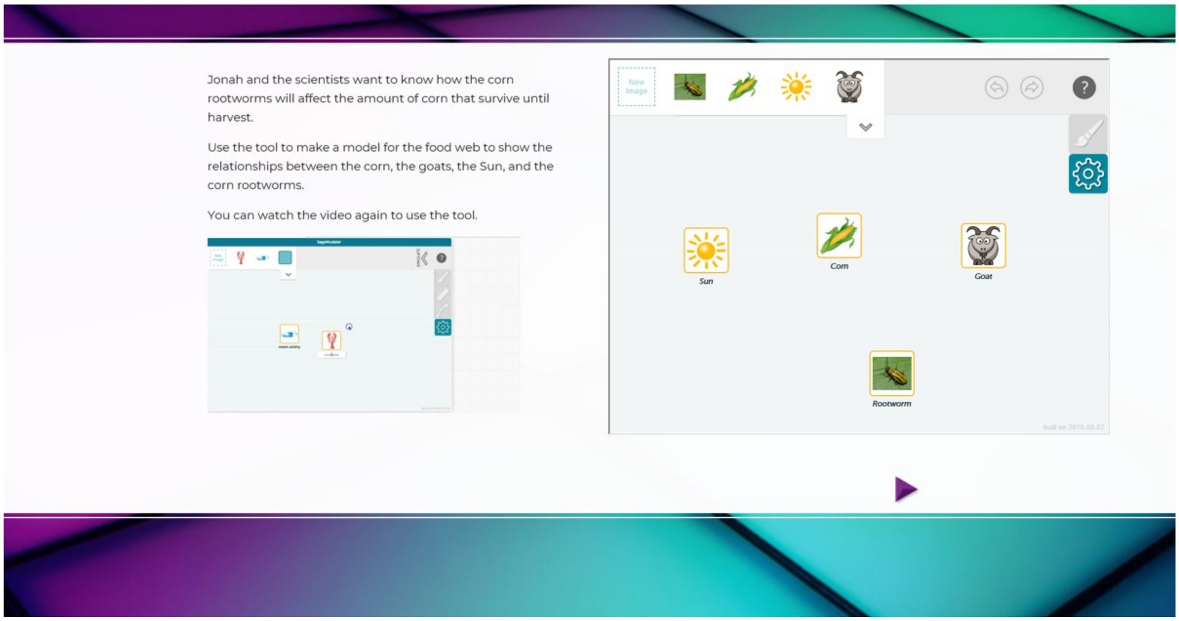
The cognitive labs that make up the data collection for my dissertation research took place at ETS headquarters in Princeton, NJ during the autumn of 2018. The DAT-CROSS team recruited 45, middle-school-age children, primarily by reaching out to ETS staff who have appropriately aged offspring. Middle-schoolers are the target participants, because they represent the target age of both the DAT-CROSS assessment suite and the grade band of the NGSS performance expectations used during the development process. Ultimately, about 15 participants from each of grades 6, 8 and 10 were recruited, generating a spread in student performances. Participants were remunerated for their time with a Visa gift card at an ETS standard rate of $30.

Regarding my participants, 47% percent were Female, 44% were white, 38% were Asian or Asian-Indian, 11% were African-American or Black, while the remaining 7% were Hispanic or of other racial/ethnic background. A more

Process/Procedures

The format of the DAT-CROSS suite is detailed in Appendix C, including questions used during the assessment (which were a mixture of multiple choice, multiple select, drag-and-drop, and subject-constructed style items). Each question has a dedicated stimulus designed to map successful completion of the task to the key progress indicators, while still being capable of mapping potential distractors to lower levels of the learning progression. During the assessment, subjects engaged with a storyline that required them to interact with simulations and models in order to draw the evidence needed to support socio-scientific arguments (like those of Eggert and Bögeholz, 2010). These arguments take the general form of how best to improve the functioning of an engineered solution to a local problem (like an infestation of crop-destroying insects to a farm). Figure 3.1 exemplifies one such question (Question 5) from the activity.

During this phase, the assenting participants (consent also provided from their legal guardian) will be asked to complete the tasks while verbalizing their thinking about each of their choices. Hammer (1994) has noted that even advanced students struggle to spontaneously engage in these types of interviews, so a DAT-CROSS team member (though not the author of this proposal) was present to guide the subjects through the activities in order to help participants recognize instances in which subject thinking may have occurred.



Jonah and the scientists want to know how the corn rootworms will affect the amount of corn that survive until harvest.

Use the tool to make a model for the food web to show the relationships between the corn, the goats, the Sun, and the corn rootworms.

You can watch the video again to use the tool.

Figure 3.1 Question 5 of DAT-CROSS Ecosystem Storyline

> Question text: Jonah and the scientists want to know how the corn rootworms will affect the amount of corn that survive until harvest. Use the tool to make a model for the food web to show the relationship between the corn, the goats, the Sun, and the corn rootworms.
> Key: Sun -> Corn; Corn -> Goat; Corn -> Rootworm.

Each of the subjects responded in a 'think aloud' format of the assessment suite. Additional questions were posed by the researcher targeting places where subjects struggled (Hoadley, 2004). This was crucial to establishing evidence of the usability of the assessment suite, so that the future iterations of the design could more effectively devise assessments with minimal interruptions by a proctor to explain aspects of the task. The think-aloud process was more time consuming than a non-verbal participation approach – closer to two hours of participation, rather than one.

I scored (with the assistance of fellow researchers) all non-constructed responses according to their key – listed below each item in figures 2.5, 3.1, and the remainder in Appendix C. I scored each constructed response (or CR) question (Questions 10, 12, and 13) according to the rubric in Appendix G, and then I assigned various subsets of the responses to five assistant raters. For each of the three CRs, 35 of the 45 responses were rated by at least three people. Since constructed response scores are ordinal, and measured by nonunique sets of multiple raters, a weighted kappa (Maclure & Willett, 1987) seemed appropriate. I also considered some other measures of interrater reliability suggested by Hallgren (2012), including percent agreement and intraclass coefficient. Table 3.1 describes these ratings for each constructed response question.

Table 3.1 Inter-rater Reliability Ratings for Constructed Responses

Question	Percent Agreement	Krippendorf's Alpha	Fleiss' Kappa (weighted)	Intra-Class Coefficient
Question 10	90.5%	0.506	0.501	0.603
Question 12	90.3%	0.444	0.439	0.533
Question 13	91.6%	0.608	0.603	0.671

I should also note that my role as a consultant precluded the proctoring of these usability/cognitive labs. My analysis in this research study was derived from pseudonymized data of these labs, simultaneously diminishing possible sources of bias that I might otherwise bring into the think-aloud process, but also limiting my agency over what questions were asked by the lab proctors.

Data Analysis Methodology

The overall goal of my research questions is to develop an effective evidence model for students' understandings of crosscutting concepts, particularly regarding two of the CCCs

Systems and System Models and *Structure and Function* (NGSS Lead States, 2013c). As mentioned in the introductory chapter, however, developing an evidence model requires the ability isolate relevant observations of CCC reasoning from observations of the other knowledges, skills, or abilities that are evoked by the task (addressed by Research Question 3) while also differentiating evidence reflecting progressing levels of CCC understanding (addressed by Research Questions 1 and 2). Addressing all three research questions requires a mix of both quantitative and qualitative analytical methods.

Quantitative Analysis – Research Questions 1 and 2

Data gathered to address Research Questions 1 and 2 form the quantitative aspect of the research to be answered via factor analysis (Kline, 1994) and item response theory (Gotwals & Songer, 2013). Factor analysis serves as the first step in construct validation, in which we examine if variation in performance across a broad set of questions can be accounted for in terms of a lesser number of latent dimensions (Kline, 1994). Ideally, this analysis reveals that a variable that cuts across contexts (much as Crosscutting Concepts do) is a significant explainer of overall performance. Using an item-response Rasch model, I then place both students and CCC performance along the same dimension, reflecting an increase in item difficulty associated with complexity of CCC understanding elicited by the task (the equation for this model can be found in Chapter 2 of this thesis).

Qualitative Analysis – Research Question 3

Research Question 3 fits squarely within a phenomenological paradigm (as defined by Creswell, 2009). Like the approach taken by Duncan and Tseng (2011) and by Hogan and Fisherkeller (1996), phenomenological study regarding the experience of interacting with curricular or assessment materials can serve as productive frames for evaluating the usability of educational materials. Both groups of researchers made use of etic coding (where thematic codes are generated prior to analysis; Bricker & Bell, 2008) to check alignment of student work to the designed goals of the instructional materials and emic codes (in which themes are allowed to emerge in the process of analysis) in order to probe instances in which misalignment was found. I derived etic codes for my analysis from prior published research on the nature of CCCs (Rivet et al., 2016; Weiser et al., 2017) as well as known avenues of progressions in system thinking (Breslyn et al., 2016; Gunckel et al., 2012; Jin & Anderson, 2012: Mohan et al., 2009; Songer et al., 2009). These codes can be found in Appendix G.

Hermeneutics. Translating the words of students into conclusions about their thinking involves hermeneutic analysis (Eger, 1992) in which the language choice of students in their crafting of arguments and explanations is evidence for their level of understanding. Eger (1992) exemplifies this by noting that the claim that an object 'has' a force may sound like an object 'exerts' or 'is acted upon by' a force, thus entailing a very different understandings of Newton's laws. As in the work of Rivet et al. (2016), such analyses draw their array of conclusions from subtle differences in wording (allowing for a host of alternative interpretations of the data). In order to further support my analysis, the findings were analyzed quantitatively in terms of their alignment to the hypothesized evidence model and member checked with the other facilitators of the usability study.

Activity theory as an analytical framework. Within the previous sections, I have described the broad problem of designing assessment appropriate to the NGSS (DeBarger et al., 2015) and have outlined activity theory as a framework for thinking deeply about potential solutions to that problem. Figure 2.3 from the literature review (reprinted as

Figure 3.2, below) exemplifies one possible activity theory triangle that demonstrates the interaction between the responses/activities of the students and the evidence:

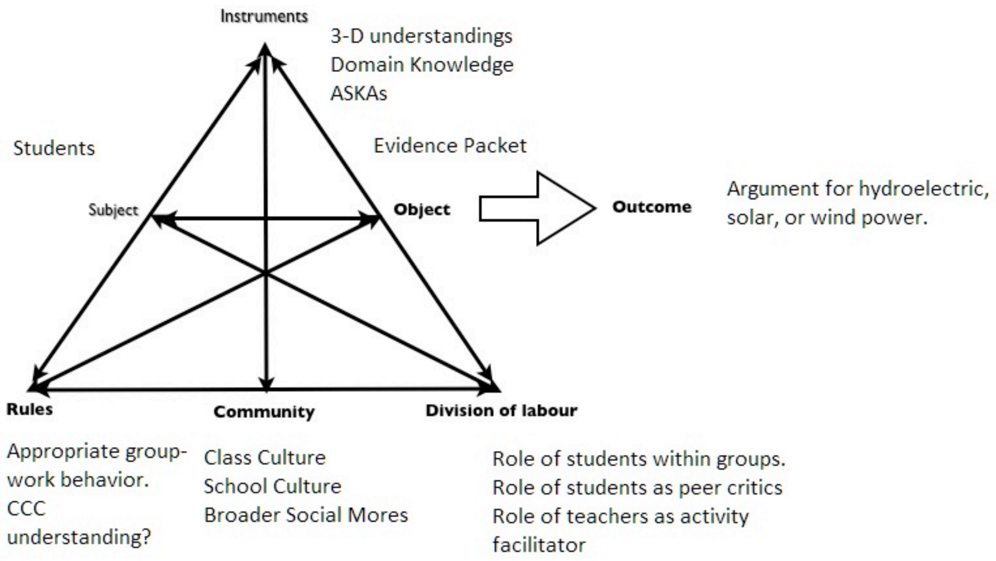


Figure 3.2 Example of Activity Theory Model at Play in Classroom Work

These triangles serve to inform the evidence model of the conceptual assessment framework (Mislevy & Haertel, 2007) for this research, and highlight a clear problem space for the design of an assessment. My design intervention required: (a) assessment which makes the evidence dictated by the activity theory model salient (Mislevy & Haertel, 2007) and (b) tools for the teacher that helps them measure that evidence (Joe, Tocci, Holtzman, & Williams, 2013; van Es & Sherin, 2002).

Roth and Tobin (1992), in their use of activity theory to improve teacher preparedness programs, describe several different forms of contradictions that impede successful engagement in a particular activity. Of these, two are of particular relation to this research project: (a) differences between the language of the classroom and that of the student, and (b) differences in motive between assessment activity and the activity of expert scientists. In the former contradiction, the greatest challenge to eliciting evidence of CCCs is getting students to understand and recognize CCC use in their own work. CCCs like patterns and systems have meanings in the sciences that are distinct from their everyday, English-language meanings (NGSS Lead States, 2013b). Because assessment materials are linguistically constrained, students receive some prompt in English and, then, have to interpret what that means relative to their science knowledge (Lee, Quinn, & Valdes, 2013). In pilot studies, looking for evidence of student's CCC understandings – as in Weiser et al. (2017) – students routinely asked for clarification of the meaning of patterns and systems. This put research into a dilemma: clarify meaning at the cost of possibly imposing our own understandings of the CCCs onto students, or refuse to clarify meaning at the cost of students' ability to provide evidence of their reasoning. By combining field notes of the confederate think aloud facilitators with the post-task interview questions, I used activity theory as a lens to examine if the language of the assessment was successful in eliciting students' display of their CCC understandings. In the second-referenced contradiction (differences in motive between assessment activity and the activity of expert scientists), students generate epistemic cleavages between their own way of knowing and the ways of knowing appropriate to scientists (further expounded in Sandoval, 2005). Activity theory serves as a useful framework for ensuring that the finalized, three-dimensional assessment is an effective stepping-stone towards the practice of expert scientists (Stroupe, 2015). For both contradictions of interest, activity theory is not the framework for how the assessor interprets evidence, but rather how it solicits evidence (the task model of DeBarger et al., 2015). Returning to the notions of validity, I have discussed in a previous section, activity theory serves as a framework for analyzing field notes, student responses, and teacher probes to ensure that the assessments have construct validity (whereby the construct is the activity done by practicing scientists) and consequential validity (whereby the assessment activities do not have glowing contradictions which inhibit their use in the classroom).

Ethics and Reflexivity

The approach used to answer my research questions has several features that pose both general and unique ethical challenges falling into three broad categories: (a) the concerns associated with any study carried out on human subjects, (b) the concerns associated with remunerating subjects for their participation, and (c) the concerns associated with the disconnect between my role as data analyzer/research coordinator and the role of other DAT-CROSS team members who were the PIs of the usability study. Of the foremost, this study was approved by the appropriate Institutional Review Boards as fulfilling their requirements and standards.

Chapter 4

RESULTS

Summary Statistics

Table 4.1 presents performance averages for each assessment item (listed as sub-items where relevant) of the DAT-CROSS ecosystem task (Appendix C) categorized by both grade and gender. The first item, Q1, was unscored as it served as an activator that could not be keyed to some best-possible-response. Differences between grades reflect some evidence of a learning progression, while the absence of gender differences is surface evidence that items were not gender biased in content or structure.

Table 4.1 Average Performance on Assessment Items by Subgroup

	OVERALL AVG	G6 AVG	G8 AVG	G10 AVG	MALE AVG	FEMALE AVG
Q2 CORN	.91 (.28)	.88 (.33)	.94 (.25)	1.00 (0.00)	1.00 (0.00)	.81 (.40)
Q2 GOAT	.62 (.48)	.71 (.47)	.44 (.51)	.82 (.4)	.71 (.46)	.52 (.51)
Q2 PEOPLE	.04 (.21)	0.00 (0.00)	0.00 (0.00)	**.18** **(.4)**	.04 (.2)	.05 (.22)
Q2ROOTWORM	.04 (.21)	.06 (.24)	0.00 (0.00)	.09 (.30)	.08 (.28)	0.00 (0.00)
Q2SUM	1.62 (.77)	1.65 (.70)	1.38 (.62)	2.09 (.83)	1.83 (.70)	1.38 (.80)
Q3	.93 (.25)	.88 (.33)	.94 (.25)	1.00 (0.00)	1.00 (0.00)	.86 (.36)
Q4	.33 (.47)	.24 (.44)	.38 (.50)	.36 (.50)	.25 (.44)	.43 (.51)
Q5	.39 (.49)	.35 (.49)	.20 (.40)	**.73** **(.47)**	.39 (.49)	.38 (.5)
Q6PT1	.73 (.44)	.71 (.47)	.75 (.45)	.82 (.4)	.71 (.46)	.76 (.44)
Q6PT2	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
Q6SUM	1.73 (.44)	1.71 (.47)	1.75 (.45)	1.82 (.4)	1.71 (.46)	1.76 (.44)

Continued next page

Table 4.1 Average Performance on Assessment Items by Subgroup, continued

	OVERALL AVG	**G6**	**G8**	**G10**	**MALE**	**FEMALE**
		AVG	**AVG**	**AVG**	**AVG**	**AVG**
Q6SUM	1.73 (.44)	1.71 (.47)	1.75 (.45)	1.82 (.4)	1.71 (.46)	1.76 (.44)
Q7CORN	.96 (.21)	.88 (.33)	1.00 (0.00)	1.00 (0.00)	.96 (.2)	.95 (.22)
Q7 GOATS	.87 (.34)	.76 (.44)	.88 (.34)	1.00 (0.00)	.83 (.38)	.90 (.30)
Q7 PEOPLE	.33 (.47)	.41 (.51)	.25 (.45)	.27 (.47)	.25 (.44)	.43 (.51)
Q7SUM	2.16 (.67)	2.06 (.90)	2.13 (.50)	2.27 (.47)	2.04 (.69)	2.29 (.64)
Q8	.6 (.49)	.65 (.49)	.5 (.52)	.73 (.47)	.67 (.48)	.52 (.51)
Q9	.18 (.38)	0.00 (0.00)	**.38** **(.50)**	.18 (.40)	.17 (.38)	.19 (.40)
Q10	1.38 (.85)	1.06 (.75)	**1.63** **(1.02)**	1.55 (.69)	1.25 (.85)	1.52 (.87)
Q11	.02 (.15)	0.00 (0.00)	0.00 (0.00)	.09 (.30)	.04 (.20)	0.00 (0.00)
Q12	1.11 (.77)	.94 (.90)	1.13 (.62)	1.36 (.81)	1.08 (.78)	1.14 (.79)
Q13	1.16 (.89)	.94 (.83)	1.25 (.86)	1.45 (1.04)	1.21 (.93)	1.1 (.89)
TOTAL	11.6 (2.95)	10.47 (3.26)	11.63 (2.45)	**13.64** **(2.25)**	11.63 (2.92)	11.57 (3.12)

Note 1: Standard deviations in parenthesis. Bold text signifies statistically significant difference (α=0.05) from grade 6 scores

Research Question 1

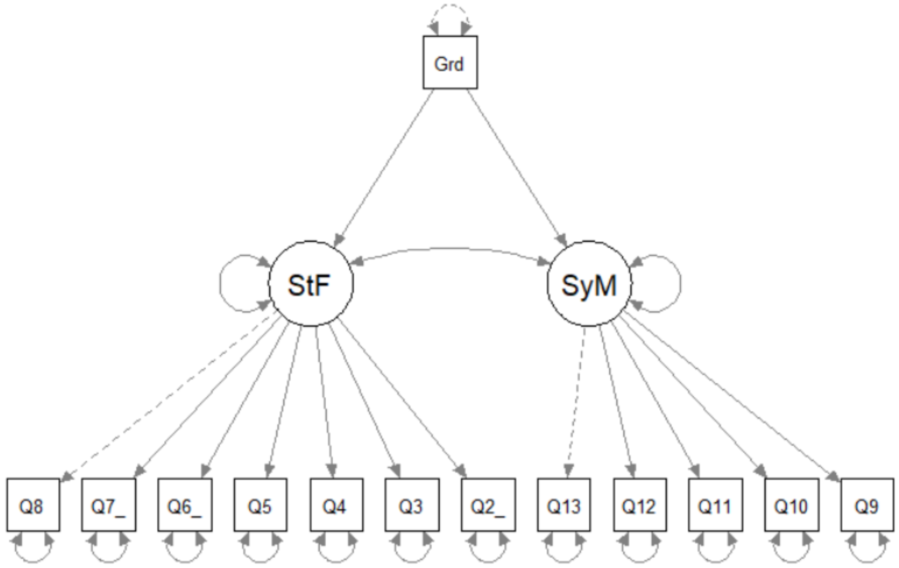
Is there evidence that the crosscutting concepts of the NGSS are distinct constructs that can be measured as students use them?

Research question 1 addressed the issue if crosscutting concepts like *Structure and Function* (StF in Figures 4.1 and 4.2) or *Systems ans System Models* (SyM in Figures 4.1 and 4.2) are different and measureable constructs that can be assessed. For DAT-CROSS, questions 2 through 8 were designed to target the StF CCC. These questions entailed the following general aspects: a) constructing a model to understand the trophic struture of the farm, b) identifying the role of the corn as a trophic producer, and c) predicting how the introduction of a new consumer (the Western corn rootworm, *D. virgifera*) might threaten that trophic structure.

Following those questions, students interacted with a simulation that demonstrated the outcome of implementing a two alternative strategies for controlling the infestation of the farm. Throught that interaction, students responded to Questions 9 through 13, which were designed to target the SyM CCC. These five questions entailed the following aspects: a) understanding the consequencs of not implementing any infestation control strategy, b) comparing and explaining the efficacy of the demonstrated control strategies, and c) making a recommendation regarding how the control strategy ought to be used in the future.

Including grade as a possible factor influencing performance, I hypothesize from the design of the DAT-CROSS assessment factor structure illustrated in Figure 4.1.

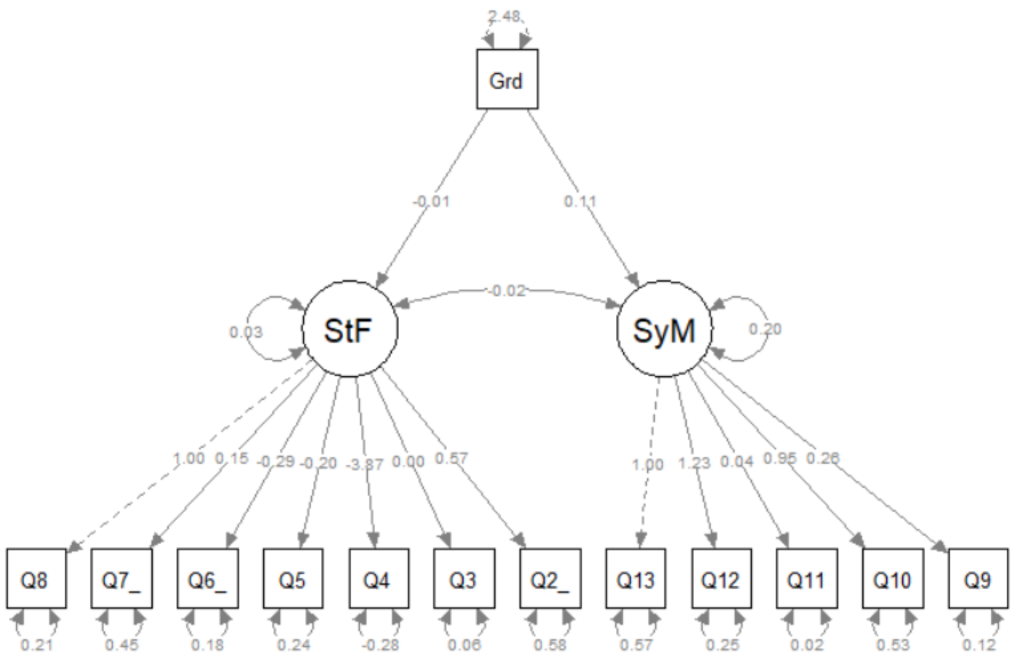
Figure 4.2 describes the resulting standardized parameters from the confirmatory factor analysis. The number within each arrow represents the standardized correlation between variables.



Legend:
Grd=Student's Grade-year; StF=Structure and Function; SyM=Systems and System Models; Q#=the score students got on an item number #.
Square boxes signify observed variables while circles convey latent variables

Figure 4.1 Hypothesized Factor Structure of DAT-CROSS Ecosystem Assessment.

The highlight of this analysis is the lack of correlation between StF and SyM with confidence interval [-0.070, 0.033]. Following the convention outlined by Garver and Mentzer (1999), a standardized inter-construct correlation that excludes 1 within its confidence interval is evidence for discriminability of constructs.



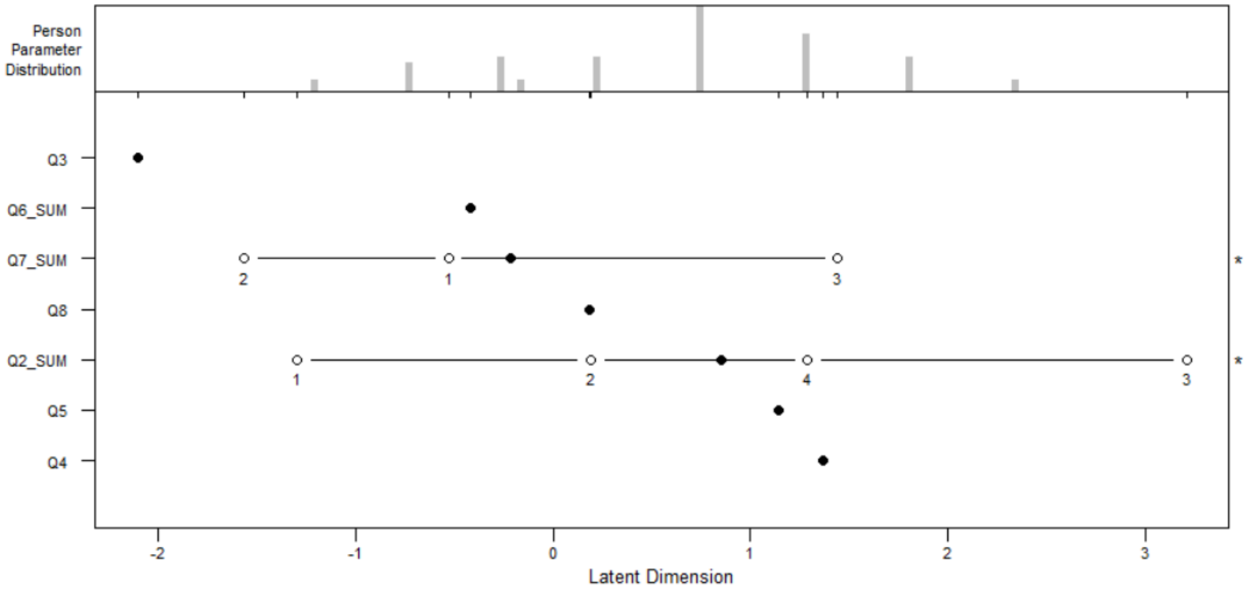
Legend:
Grd=Student's Grade-year; StF=Structure and Function; SyM=Systems and System Models; Q#=the score students got on an item number #.
Square boxes signify observed variables while circles convey latent variables

Figure 4.2 Resulting Correlations from CFA Analysis

Research Question 2

What does student demonstrations of the crosscutting conceptual reasoning aspect of their three-dimensional science understandings look like at different levels of overall science understanding?

Preliminary Learning Progression. Following the results of the CFA analysis, I used separate Rasch partial credit models to estimate student performance on both the *Structure-Function*-targeting and *Systems-and-System-Models*-targeting items. These resulted in Wright Maps (Figures 4.3 and 4.4, respectively) and fit statistics (Tables 4.2 and 4.3, respectively). In general, using 45 participants is on the low end of acceptable for validating a Rasch model (de Ayala, 2010), so these results should be viewed as preliminary and as a means of finding avenues of improvement rather than proof unto itself.



Legend:
Top: The ordinate dimension is a histogram of number of students in relation to Logits of difficulty/ability on the abscissa (the Latent Dimension shared by items and people).
Bottom: The ordinate dimension lists *Structure-Function*-targeting questions (with partial scores as applicable) in relation to Logits of difficulty/ability on the abscissa (the Latent Dimension shared by items and people).

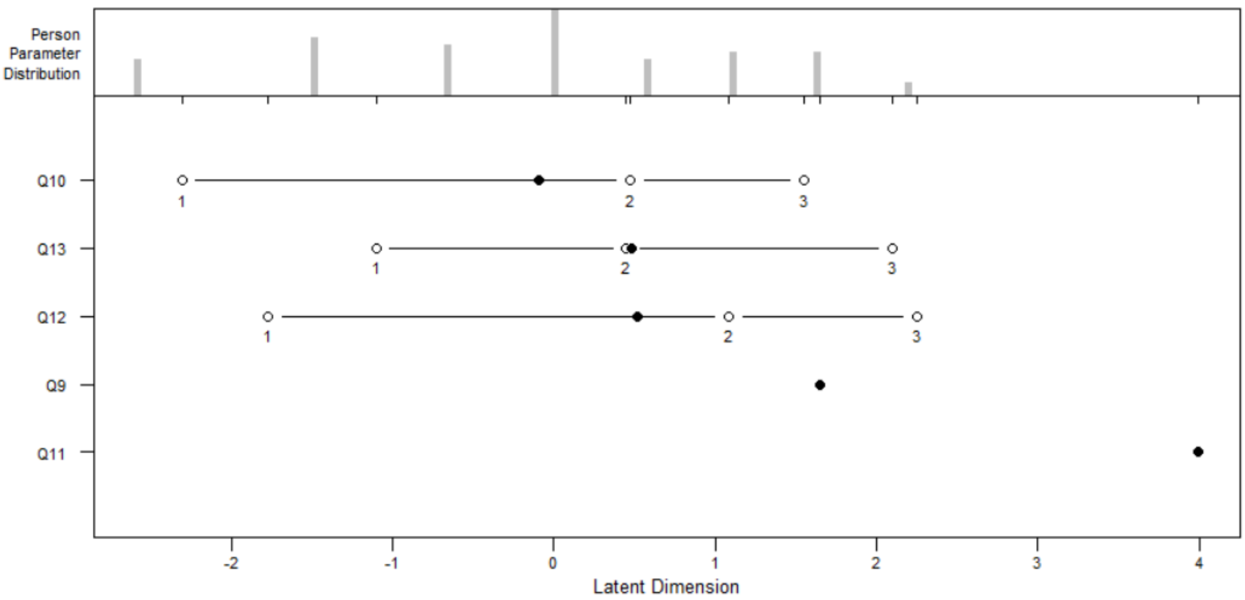
Figure 4.3 Person-Item (Wright) Map of Structure Function Items

From Figure 4.3, the first stand out are Questions 2 (in which students identify the trophic level of various organisms on the farm) and 7 (in which students make a prediction about the outcome of the oncoming rootworm infestation). The Rasch partial credit model (adapted from Equation 2.1 from the literature review) estimates that it was easier to get four responses out of four correct in Question 2, than it was to get three responses correct and, similarly, estimates that it was easier to get two responses out of three correct in Question 7, than it was to get just one. These strange results bear out in the fit statistics as Question 2 had an outfit T beyond acceptable levels (de Ayala, 2010). Questions 2 and 7 were also somewhat unique in that they prompted multiple select responses (i.e., students were asked to select all that apply).

Table 4.2 Fit Statistics of Structure and Function Items

	CHISQ	DF	P-VALUE	OUTFIT MSQ	INFIT MSQ	OUTFIT T	INFIT T
Q2_SUM	22.588	44	0.997	0.502	0.469	-2.59	-2.86
Q3	21.636	44	0.998	0.481	0.802	-0.78	-0.34
Q4	54.053	44	0.142	1.201	1.147	1.07	1.16
Q5	30.923	43	0.915	0.703	0.741	-2.02	-2.49
Q6_SUM	34.631	44	0.843	0.77	0.838	-1.01	-0.99
Q7_SUM	31.003	44	0.93	0.689	0.725	-1.43	-1.26
Q8	54.678	44	0.13	1.215	1.174	1.38	1.39

Overall, the *Systems and System Models* items seem to behave well with some decent evidence. For example, the responses to DAT-CROSS Question 9 (in which students were asked to describe patterns in the data produced from the simulation as it simulated the efficacy of the first strategy for controlling the rootworm infestation – adding new predators) and the constructed response questions (questions 10, 12, and 13 in Appendix C) aligns to the hypothesized learning progression (detailed in Appendix B). That is, Question 9 was designed to align to level 4 on the Systems learning progression (in which students are able to make predictions about the future state of a system), about equal to the target alignment of the constructed responses that earned a three-point score based on the scoring rubrics (further detailed in Appendix H).



Legend:
Top: The ordinate dimension is a histogram of number of students in relation to Logits of difficulty/ability on the abscissa (the Latent Dimension shared by items and people).
Bottom: The ordinate dimension lists *Systems-and-System-Models*-targeting questions (with partial scores as applicable) in relation to Logits of difficulty/ability on the abscissa (the Latent Dimension shared by items and people).

Figure 4.4 Person-Item (Wright) Map of Systems and System Model Items.

In Figure 4.4, we find exactly that the alignment behavior bears out in the modeled item difficulty (the latent dimension axis). For example, Question 9 has a near equivalent item difficulty measure as three-point scores in Question 10 (the first constructed response question in which students are asked to explain why the predator-introduction strategy that was also the subject of Question 9 was not as effective as predicted). Examining the fit statistics (Table 4.3), further examination is needed to understand why Question 12 (in which students explain why a second control strategy – planting a trap crop – was effective at reducing the impacts of the infestation) has poor outfit T, and why students found Question 11 so difficult (which was similar in form to Question 9 but regarding the efficacy of the second, trap-crop strategy rather than the first).

Table 4.3 Fit Statistics of Systems and System Model Items

	CHISQ	DF	P-VALUE	OUTFIT MSQ	INFIT MSQ	OUTFIT T	INFIT T
Q9	29.343	42	0.93	0.682	0.973	-0.51	-0.05
Q10	37.621	42	0.663	0.875	0.846	-0.57	-0.74
Q11	19.178	42	0.999	0.446	0.827	0.2	0.04
Q12	23.946	42	0.989	0.557	0.562	-2.45	-2.42
Q13	35.926	42	0.734	0.835	0.83	-0.75	-0.82

These Rasch models allowed me to isolate six subjects (three each from StF and SyM) who performed at the best, worst, and median of participants. Their responses both during the think aloud and in the reflective interview are examined in the discussion chapter.

More on the Structure and Function and the Systems and Systems Models Constructs. To gain further evidence that the *Structure and Function* construct is unique compared to the *Systems and System Models* construct, I compared the modeled student ability scores (delta from Equation 2.1, along the latent dimension axis from Figures 4.3 and 4.4) from their performance on the questions from each construct. If the constructs were equivalent, we would expect that the distribution among the population of each was equivalent. Using a Wilcoxon signed-rank test (Kerby, 2014) to account for potentially nonparametric values, the distributions of students' ability in the two constructs rejected the null hypothesis of equivalent distributions (W = 830, p << .01).

Research Question 3

What challenges do middle-school-aged subjects face in presenting their CCC understandings while engaging with an interactive suite designed to target Systems-Models and Structure-Function dimensions of their three-dimensional science understandings?

Understanding Disruptions in Student Performance

In light of the findings from Research Question 2, in which certain items were found to be poorly functioning, understanding the non-construct relevant challenges students faced becomes all the more critical. To answer Research Question 3, I return to Engestrom's (2000) activity theory model as a means for analysis. Within his framework, the student who were my subjects have developed a script for being assessed in their science classroom that has been crafted, much like the constructivist schema, in response to their prior experiences. As they worked within the novel context (understanding ecosystems) of the assessment and in a novel setting (on a computer outside of their normal classroom), students may find that their prior scripts are unproductive or counterproductive – leading to a disruption in their engagement with the task. In the following subsections, I outline three patterns of disruption that commonly occurred as students worked with the assessment – each diagrammed using activity theory triangles with red linkages.

Disruption 1: Paring the Food Web. The first set of activities in the farm ecosystem assessment was designed to introduce students to modeling ecosystems as a set of relationships between organisms that might grow in complexity as new organisms were added into the model. While students had minimal difficulty constructing the initial model (Q3 of the task as described in Appendix C) with 44 of 45 correctly connecting model elements, they expressed greater challenge in understanding the role of the model when asked to incorporate the rootworm into their prior creation. Usability Participant 15 expresses their confusion...

> **Facilitator:** So, the first is about the first model that you created. So, was it easy to interact with the components?
> **Usability 15:** Yes. It was easy, but I think you should like explain more about what you want us to do next time. Like, I got the part that you wanted us to show like the relationship between them, but like what kind of relationship? Like the food chain

> relationship or like, is it like the sun feeds the corn or is it the goat eats the corn? The corn takes from the sun or is it the sun feeds the corn? The corn feeds the goat? Is it backwards or is it forwards?

Student's challenges in these early models could propogate into later items despite leveling (in which students were shown the correct answer as part of interaction with the task). Usability Participant 30 initially explains why the harvest men strategy might not have been effective…

> **Usability 30:** Adding the harvestmen did not help increase the percent corn yield because the number of rootworm eggs were still increasing also.

Only on later reflection, during the post-task interview portion, did they describe the phenomenon using ideas from the model…

> **Facilitator:** So, we have the second part that it was related to the harvestmen. So, what did you notice in terms from the data from the simulation video and from the chart and the graphs? Why does it remain? Think the way you have to explain to someone.
> **Usability 30:** The corn harvested was going down but it wasn't going down as graphically as before. And the number of rootworm eggs wherein as high as before.
> **Facilitator:** Do you think that this strategy was effective?
> **Usability 30:** Not really, because they're still losing corn yield and the rootworm eggs are still, like, they are still increasing and multiply as much.
> **Facilitator:** Why do you think that the number of harvestmen introduced by the scientist was the same every year from year three to five?
> **Usability 30:** Because they want the harvestmen becoming -- they want the harvestmen having the predators also and taking over the -- disturbing the food chain.
> **Facilitator:** Okay. What do the harvestmen do in this ecosystem?
> **Usability 30:** They gather up the rootworms.
> **Facilitator:** And how do they get rid of the rootworms?
> **Usability 30:** I think they would eat them.

I call this pattern (as diagrammed in Figure 4.5) "paring the foodweb", in which students paid selective attention to ideas from the modeling task which did not transfer into making sense of the simulation.

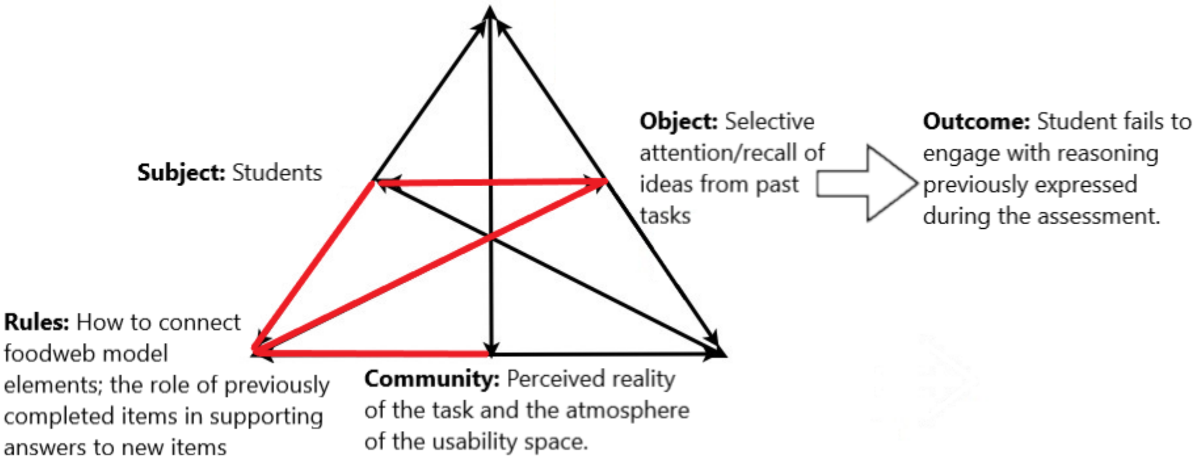

Subject: Students
Object: Selective attention/recall of ideas from past tasks
Outcome: Student fails to engage with reasoning previously expressed during the assessment.
Rules: How to connect foodweb model elements; the role of previously completed items in supporting answers to new items
Community: Perceived reality of the task and the atmosphere of the usability space.

Figure 4.5 Diagram of Activity Disruption When Student Interacts with Modeling Tool

Disruption 2: Failure to Interact with Evidence Available in the Simulation. As students move into the second phase of the task, they are asked to interact with simulated implementations of the control strategies in order to collect evidence about each strategy's efficacy. Drawing a successful conclusion entailed merging evidence available from several sources – information provided to students before the simulation regarding how each strategy might work, a visualization of organisms in the farm ecosystem interacting according to the control strategy, a data table indicating the population numbers of each organism by year, and a set of graphs (see Appendix C). However, in practice, students often restricted themselves to only one set of evidence, usually a single graph or the data table. Below, Usability Participant 27 explains how he used the evidence in reasoning about the harvestmen strategy.

> **Facilitator:** Yeah. So, the term harvestmen, are you okay with that term? Harvestmen?
> **Usability 27:** Well, I didn't really know it before.
> **Facilitator:** Okay, but when you see this, it's kind of bug
> **Usability 27:** It just seemed like, (00:42:50) is just a bug.
> **Facilitator:** Yeah, okay. So, for the second video, do you recall what it was trying to tell you? It's about the harvestmen.

Usability 27: How they grew when added when like you added the harvestmen with the rootworms.
Facilitator: Yeah.
Usability 27: I feel like it didn't really make a change that much.
Facilitator: So, here is the data table here.
Usability 27: I still didn't look at the data table. I looked at the graphs.
Facilitator: You still looked at the graphs?
Usability 27: Yeah.
Facilitator: So, you are looking at whether the -- I think you look at this graph a lot, right?
Usability 27: Yeah.
Facilitator: You are looking whether the rootworm eggs are growing or decreasing.
Usability 27: Because in this one, it increased (00:43:37) like went a little bit and went back up.
Facilitator: Still increased?
Usability 27: Yeah.
Facilitator: Okay. And for the -- did you -- can you explain why do you think the harvestmen methods didn't work?
Usability 27: Because I don't think the harvestmen worked because here, if you look at -- I looked at this one a lot because it started up this high, but then it looks like it's slowly starting to decrease.

These types of behaviors (diagrammed in Figure 4.6) resulted in constructed responses which were based on changes in the population of rootworms, or on changes in the population of corn, but not both (corresponding to a level 1 score in the rubric).

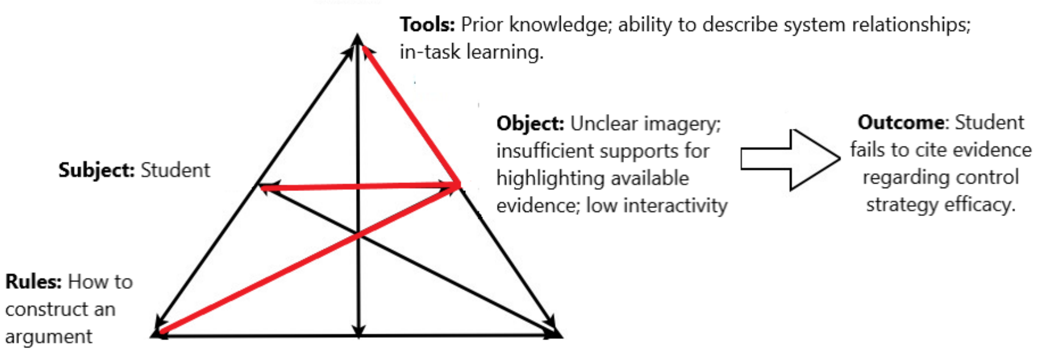

Tools: Prior knowledge; ability to describe system relationships; in-task learning.
Object: Unclear imagery; insufficient supports for highlighting available evidence; low interactivity
Outcome: Student fails to cite evidence regarding control strategy efficacy.
Subject: Student
Rules: How to construct an argument

Figure 4.6 Diagram of Activity Disruption During Student Interaction with Simulation

Disruption 3: Salience of Systemic Relationships. The final disruption pattern (depicted in Figure 4.7) came when students lacked relevant prior knowledge to engage with key task components in their everyday life experiences. Much as occurred in disruption 1, students were unable to describe ecosystem functioning in terms of system relationships (like predation or energy usage); instead they resorted to accounts, best described as layperson terms. For example, Usability Participant 26 describes how the harvestmen scare away the rootworms…

> **Facilitator:** Okay, perfect. We're moving next to the two control methods, the harvestmen and the alfalfa. This video and this video was about the harvestmen. Based on the data, what is the main takeaway from this situation?
> **Usability 26:** That the harvestmen, they helped but not that much because I feel like if there (00:47:30) had more harvestmen and we had more harvestmen eggs like that then that would kind of help because they were only like ten and the number of eggs were increasing. It was hard for the harvestmen to go to each and help the corn so I feel that if we had more harvestmen, that would have been easier and we have had more corn harvested.
> **Facilitator:** How do the harvestmen help the corn?
> **Usability 26:** I think they helped because I think they kind of scare away the corn rootworms because they're bigger or like, yeah.

While such lay accounts could sometimes allude to complex interaction dynamics between the corn, the rootworm, and the control strategy, more often the result was explanations that misread available information. Usability Participant 1 inaccurately describes the trap crop method as ineffective, because "The alfalfa increased the rootworms more just by providing more shelter for them to plant their eggs."

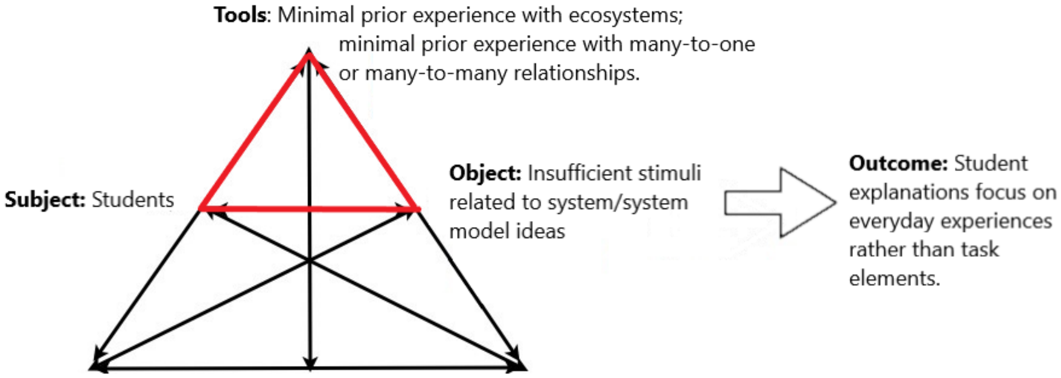

Tools: Minimal prior experience with ecosystems; minimal prior experience with many-to-one or many-to-many relationships.
Subject: Students
Object: Insufficient stimuli related to system/system model ideas
Outcome: Student explanations focus on everyday experiences rather than task elements.

Figure 4.7 Diagram of Activity Disruption During Student Constructed Responses

Chapter 5

FINDINGS and DISCUSSION

The design-based nature of my research questions led me to adopt a methodology akin to other design-based research (Brown, 1992) by blending quantitative and qualitative elements. Mirroring the "analytic spiral" described by Creswell (2009, p.183), my qualitative data served as a means of phenomenologically validating the activities that produced quantitative findings, while the quantitative data served as a tool to identify and triangulate new avenues of qualitative analysis. More specifically, the qualitative data from the think-alouds and reflective interviews helped to support the claim that students were using the relevant constructs in their reasoning while the quantitative data from student performance on the assessment helped identify target assessment items in need of revision; and target students for whom a deeper analysis of their statements would be most fruitful. This spiraling analysis helped identify places where the assessment could be revised as well as provided some guidance for the development of brand-new items. In the following sections, I discuss the results of my study by research question in relation to this overarching goal of implementing CCC-conscious assessment.

Research Question 1

Is there evidence that the crosscutting concepts of the NGSS are distinct constructs that can be measured as students use them?

The highlight finding from the CFA used for Research Question 1 was the indistinguishable-from-zero correlation between the *Structure and Function* (StF in Figures 4.1 and 4.2) and the *Systems and System Models* (SyM in Figures 4.1 and 4.2). The remaining results from the CFA are only moderate, with some questions having much better correlation to the relevant CCC construct than others. While the p-value of the chi-square goodness of fit was much greater than 0.05, the root-mean-squared-error of approximation (Mueller & Hancock, 2015) was just barely within the acceptable range with a 90% confidence interval of [0, 0.11]. Overall, the items were better correlated and more reliable in the *System and System Models* CCC than in *Structure and Function*. Overall, these findings point to two implications for the DAT-CROSS assessment: a) more items are needed in order to improve intra-construct reliability, and b) understanding of how to design systems-reasoning-oriented assessment exceeds understanding of how to do so for assessments oriented towards *Structure-and-Function*-based reasoning.

Implication a) is being addressed by developing two new sets of items under a parallel design process as was used for the usability study's set. These two new sets are designed to use the relevant CCCs in similar ways but in new contexts (hydrologic systems and human-body systems). Further discussion on the development of the new items appears later in this chapter.

Implication b) reflects, in my estimation, the current state of the science education field on *Structure and Function* compared to *Systems and Systems Models*. It is common, in existing learning progression literature, to see phenomena (and the learning thereof) portrayed as a set of relationships that organize into a system at sufficiently sophisticated levels of student reasoning (Breslyn et al., 2016; Gunckel et al., 2012; Jin & Anderson, 2012: Mislevy, 2016; Mohan et al., 2009; Songer et al., 2009). These phenomena can span many relevant disciplines, from geoscience (Breslyn et al., 2016) to ecology (Hokayem, Ma, & Jin, 2015; Jin & Anderson, 2012) to hydrology (Gunckel et al., 2012) and beyond (Mislvey 2016), facilitating transfer into new contexts or content areas. In contrast, *Structure and Function* has not been the focus of as much science-education literature. The outcome of the CFA reveals that while the design framework can design items that are *Systems and Systems Models* oriented and can design items that are *not Systems and Systems Models* oriented, more revision is necessary to transform the *not Systems and Systems Models* construct into one that can more directly elicits observations pertaining to *Structure and Function.*

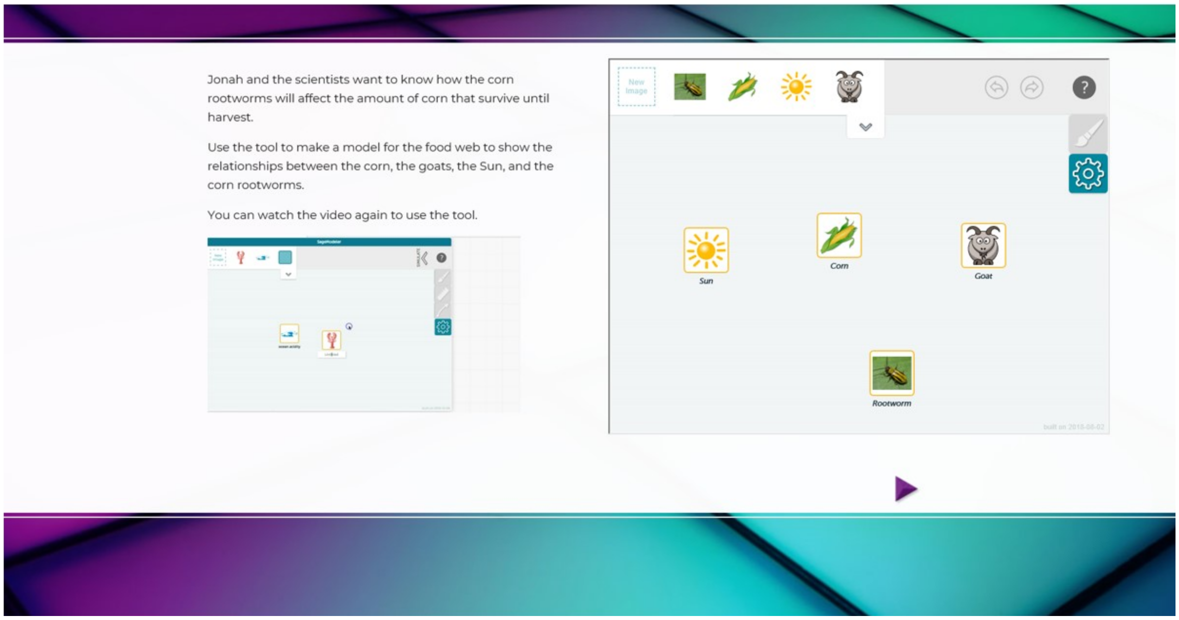
Research Question 2

What does student demonstrations of the crosscutting conceptual reasoning aspect of their three-dimensional science understandings look like at different levels of overall science understanding?

One of the main affordances of the Rasch model used to analyze student performance is the ability to place students along a dimension of ability that is of the same scale as the dimension of item difficulty (Briggs & Alonzo, 2012). That allows me to estimate student's position along the hypothesized learning progression (described in Appendix B). In the following subsections I discuss student think-aloud responses to a selected *Structure and Function* item and a selected *Systems and System Models* item for students who have the best, the worst, and the median ability score in the corresponding construct. Returning to the four roles of CCCs (Weiser et al., 2017) from the literature review, I analyze how the roles selected by each student differed as they responded to the same item.

Student Responses at Different *Structure and Function* Progression Levels.

Question 5 from the DAT-CROSS assessment (Figure 5.1) was the final click and drag item, in which students amend a prior model of the farm to account for the introduction of rootworms to the ecosystem. It was also modeled to have a difficulty score of about 1.14 logits, making it one of the harder items in the *Structure and Function* set. The interactive elements of the task make it a productive avenue for examining think-aloud responses.



Jonah and the scientists want to know how the corn rootworms will affect the amount of corn that survive until harvest.

Use the tool to make a model for the food web to show the relationships between the corn, the goats, the Sun, and the corn rootworms.

You can watch the video again to use the tool.

Figure 5.1 DAT-CROSS Question 5
Note: Question Key: Sun -> Corn; Corn -> Goat; Corn -> Rootworm

High *Structure and Function* ability (Usability 32). Usability 32 was a 10th grader who had the highest modeled ability score at about 2.34 logits (corresponding to a 75% chance of getting Question 5 correct via the Rasch equation [Equation 2.1]). Here is them thinking aloud with artifact illustrated in Figure 5.2.

> **Usability 32:** So, now the corn rootworms will affect the amount of corn survive until harvest. To make one food web of corn (00:07:26) corn goes to the sun and the corn rootworms. The sun power the corn, so the sun is really far away, so it's got to be up (00:07:37). And this goat and this rootworm, they're both next. So the corn is going into the goat, but this corn is also going into the rootworm, so, that makes the problem.

In this response, Usability 32 can correctly describe the teleological function each entity played (either beneficially or detrimentally) and use their understanding of a discrete set of relationships to build up to the whole structure. From the perspective of the four roles of CCC reasoning (Weiser et al., 2017), this is an example of 'CCCs as Levers' employed successfully.

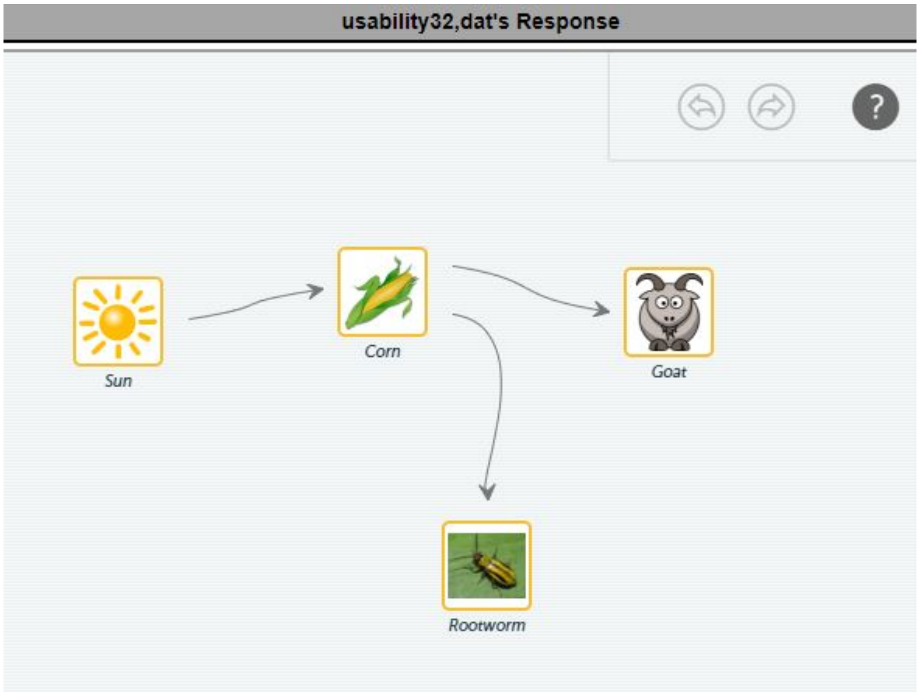


Figure 5.2 Usability 32's response to DAT-CROSS Question 5

Interestingly, while Usability 32 had the best performance on the *Structure and Function* items (Questions 2 through 8), their performance on the later, Systems and Systems Models items (Questions 9 through 13) were middling at best – with all three of their constructed responses scoring a 1 based on the scoring rubric.

Median *Structure and Function* ability (Usability 41). Usability 41 was an 8th grader who had the median modeled ability score at about 0.74 logits (corresponding to a 40% chance of getting question 5 correct via the Rasch equation [Equation 2.1]). Here is them reflecting on their response with their artifact illustrated in Figure 5.3:

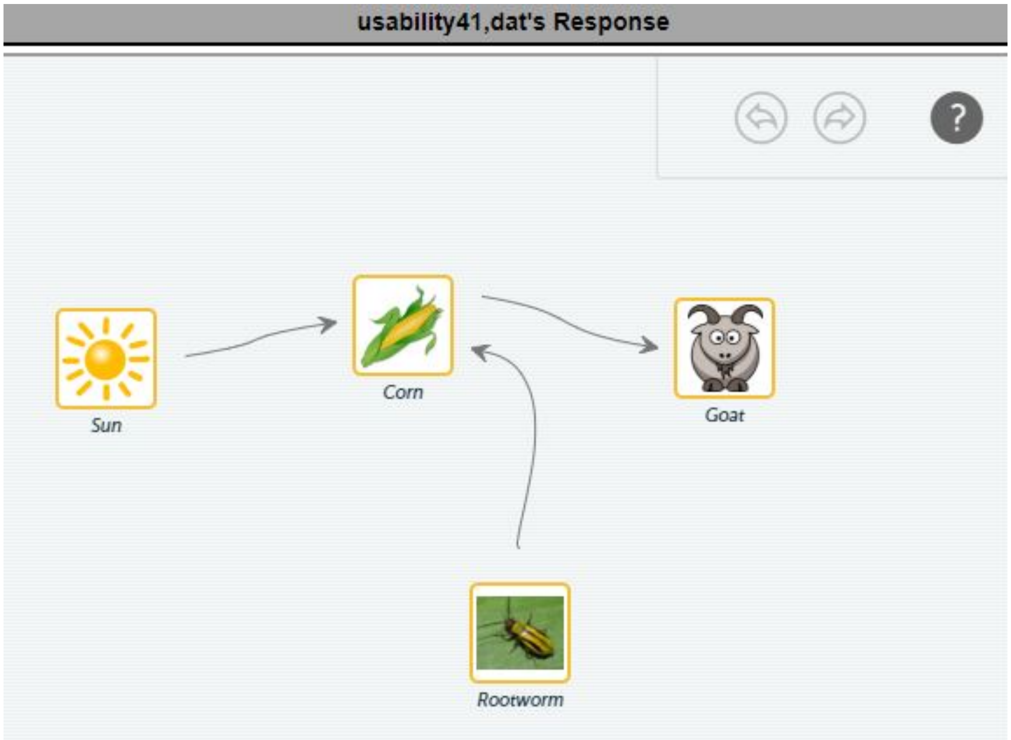


Figure 5.3: Usability 41's response to DAT-CROSS Question 5

Usability 41: I knew rootworm would -- like that more people were using the corn. It was like taking corn away from the other people who were using it.
Facilitator: Okay. So, you think rootworm is taking corn away?
Usability 41: Yeah. It was like decomposing it.
Facilitator: Decomposing that. Okay. So, is this the model that you drew? Or the arrows are different?
Usability 41: I think the arrows are a little different.
Facilitator: So, what's your arrow? Yeah, your direction is from rootworm to the corn?
Usability 41: Yeah.
Facilitator: What does that mean?
Usability 41: That the rootworms were like eating the corn.
Facilitator: Eating the corn. Okay. However, here, the corn to goat, is also goat eating the corn, why the arrow is different?
Usability 41: I think I did that because of the previous one. I'd meant a different thing.

Rather than construct a coherent structure in which the arrows always mean the same thing, Usability 41 uses his 'CCC as a Bridge' (Weiser et al., 2017) – with discrete connections between relevant entities that are sensible (given their content knowledge) only as a set of pieces. Identification of discrete relationships is a common theme at lower stages of many LPs (Mislevy, 2016) including my hypothesized one (Appendix B).

Low *Structure and Function* ability (Usability 7). Usability 7 was a 6th grader who had the lowest modeled ability score at about -1.21 logits (corresponding to a 9% chance of getting question 5 correct via the Rasch equation [Equation 2.1]). Here is them thinking aloud with artifact illustrated in Figure 5.4.

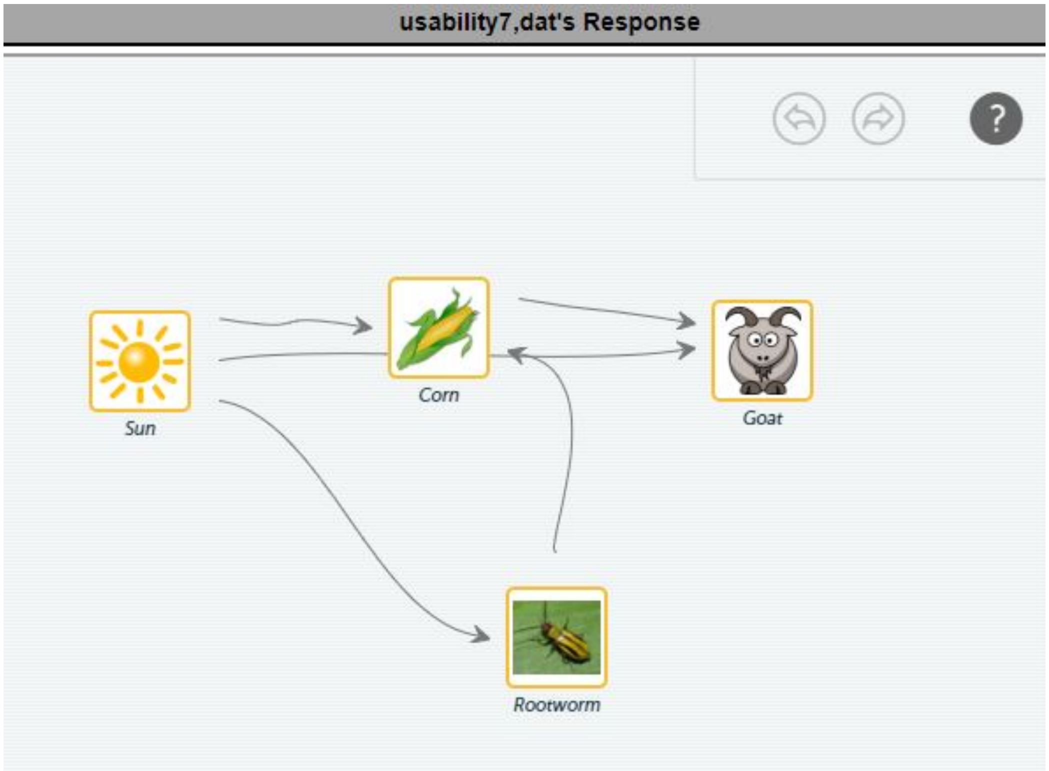


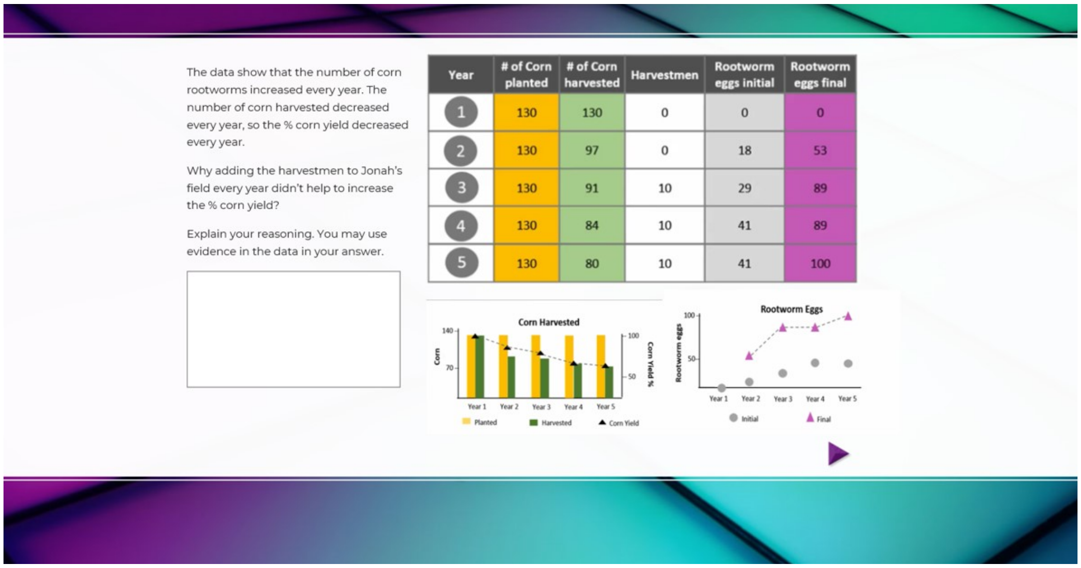
Figure 5.4 Usability 7's response to DAT-CROSS Question 5

> **Usability 7:** I think that it reflects your relationships between – the feeding relationships because if you don't have the sun, then you can't have corn and if you can't corn, then there's nothing to feed the goats so they would die. So, if one is if you don't have one, the other one would die or not do well.

In their model, Usability 7 seems to develop an organizing principle, that the sun needs to exist to support the other organisms, but then struggles to coordinate that principle into a coherent model. Interestingly, under the four-role framework (Weiser et al., 2017), this participant's behavior is associated with using the 'CCC as a Rule,' which (in the task model) was placed at the higher end of the learning progression. Still, in the context of the question, it was inappropriate and led the student to create a model in which the arrows meant different things across different entities.

Student Responses at Different *Systems and System Models* Progression Levels.

Question 10 from the DAT-CROSS assessment (Figure 5.5, below) was the first constructed response item, in which students were asked to explain why the control strategy in which predators were introduced to prey on the rootworms was not effective.



The data show that the number of corn rootworms increased every year. The number of corn harvested decreased every year, so the % corn yield decreased every year.

Why adding the harvestmen to Jonah's field every year didn't help to increase the % corn yield?

Explain your reasoning. You may use evidence in the data in your answer.

Year	# of Corn planted	# of Corn harvested	Harvestmen	Rootworm eggs initial	Rootworm eggs final
1	130	130	0	0	0
2	130	97	0	18	53
3	130	91	10	29	89
4	130	84	10	41	89
5	130	80	10	41	100

Figure 5.5 DAT-CROSS Question 10

As a constructed response, Question 10 was scored by a rubric (Appendix H) meaning that each potential score (from 0 to 3 points) is associated with a distinct difficulty measure. Specifically a 1-point score had a difficulty of -2.31 logits, a 2-point score had a difficulty of 0.47 logits, and the top score (3 points) had a difficulty of 1.55 logits. The constructed nature of students' responses entails 'intentionality' (Jonassen & Rohrer-Murphy, 1999), which make them a productive avenue for thinking about student activity.

High *Systems and System Models* ability (Usability 20). Usability 20 was a 10th grader who had the highest modeled ability score at about 2.19 logits (corresponding to a 99%, 85%, and 65% chance of scoring 1, 2, or 3-points, respectively, via the Rasch equation [Equation 2.1]). Here is their constructed response:

> **Usability 20:** There is the same number of harvestmen through years 3 to 5 but the rootworm eggs still increased in number overall. If the harvestmen could only keep 91 cornstalks harvestable in year 3, they are not going to keep more cornstalks harvestable if they have more rootworm eggs to deal with in the next years.

Usability 20's ability to detect a trend in the system and, from that trend, predict a future state is indicative of the 'CCC as a Rule" role (Rivet et al., 2016) used successfully.

Median *Systems and System Models* ability (Usability 45). Usability 45 was a 10th grader who had the median modeled ability score at about 0.01 logits (corresponding to a 91%, 39%, and 18% chance of scoring 1, 2, or 3-points, respectively, via the Rasch equation [Equation 2.1]). Here is their constructed response:

> **Usability 45:** Adding the Harvestmen in the feild every year didn't change the corn yeild becasue they number of harvest men stayed the same and the rootworms only addapted to their new envirnment. The harvestmen number stayed the same while the rootworms had eggs each year and only continued this cycle. So because the harvestmen were added instead of increasing corn yeild, the corn yeild only decresed less rapidly {sic}.

In contrast to Usability 20's ability to construct organizing principles that dictated the state of the farm ecosystem, Usability 45 makes the CCC act as a lens by identifying relevant parts that could help support an explanation. While that role was successful in identifying evidence relevant to their claim, we see that their response has no underlying reasoning and described no mechanism by which the harvestmen and the rootworms are related to each other.

Low *Systems and System Models* ability (Usability 36). Usability 36 was a 6th grader who had the lowest modeled ability score at about -3.75 logits (corresponding to a 19%, 1%, and 0.5% chance of scoring 1, 2, or 3-points, respectively, via the Rasch equation [Equation 2.1]). Here is their constructed response:

> **Usability 36:** Because there were too many Rootworms for the HarvestMen {sic} to fight off.

These sorts of partial responses are commonplace both in our assessment and in past assessments involving student explanations (Beck, Bookbinder, Lee, & Rivet, 2017) in which some relevant claim is made without citing any meaningful evidence or providing sufficient reasoning.

Research Question 3

What challenges do middle-school-aged subjects face in presenting their CCC understandings while engaging with an interactive suite designed to target Systems-Models and Structure-Function dimensions of their three-dimensional science understandings?

Role Selection and Disruptions in Student Activity.

From the discussion of the findings for Research Question 2, it seems that while the Roles (Rivet et al., 2016) do not always align to learning progression levels (as supposed by the task model [Appendix A]), students' selection of Role does influence their ability to construct explanations for phenomena. This is consistent with my findings from prior research (Weiser et al., 2017). It stands to reason, given the influence of 'Role Selection' (Weiser et al. 2017), to conceptualize the disruptions from the results chapter in the light of the 4 Roles and which ones had been selected when disruptions occurred.

Disruption 1. In disruption 1, I found that students, despite interacting with the food-web model during questions 3-6, rarely carried ideas from that set of tasks into their explanations for the efficacy of the rootworm-infestation-mitigating strategies. Instead, they would focus only on the current material available on screen, identify what mattered from that set, and attempt to construct an explanation from there (often missing out on the relationships, mechanisms, or reasoning required). Here, students used their CCCs as lenses to identify components they found relevant to the task, but then struggled in bridging those components together. This behavior is consistent with prior research regarding student views on behaviors and functions (Hmelo-Silver, Liu, & Jordan, 2009) in which students can make sense of components and the individual behaviors of components before they can describe how those behaviors impact other entities.

***Disruption 2*.** In disruption 2, students failed not only to carry ideas from past screen into later items, they also failed to cite evidence that was present right in front of them. The best explanation for this disruption was that they did not have sufficient prior knowledge about relevant science practice, and by extension could not determine how that content was meant to serve as evidence for their explanations. From a role selection perspective, this disruption underscores the three-dimensionality of science understanding. Even when students are able to select an appropriate role for the CCCs (e.g., to use their understanding of the system to construct hierarchy in the relationships of system components), without sufficient understanding of the relevant disciplinary core ideas or science and engineering practice, it is not possible to perform to the expectation of the standards.

Disruption 3. Disruption 3 was a counterpoint to disruption 2. Where disruption 2 occurred when students failed to cite evidence on screen, disruption 3 occurred when evidence was cited but the explanation constructed followed a lay account of the phenomenon. This kind of behavior is common when students have developed sophistication in their CCCs and SEPs before they have done so in their DCIs (Beck et al., 2017). In Usability 26's response to Question 10 (as described in Chapter 4), they are able to effectively use the right role, CCC as Lever, to construct a sophisticated, albeit lay, account of the predation strategy. The linguistic success of Usability 26, contrasted with the content failing, reveals the how important language is to the CCCs. As noted in the literature review, the CCCs, while important in unique ways to the scientific domain, also play out in the everyday sense of pattern, cause and effect, systems, etc. That is, students develop these concepts both in coordination with their learning experiences in the science classroom and in their everyday experiences when they might not be attending to other science content.

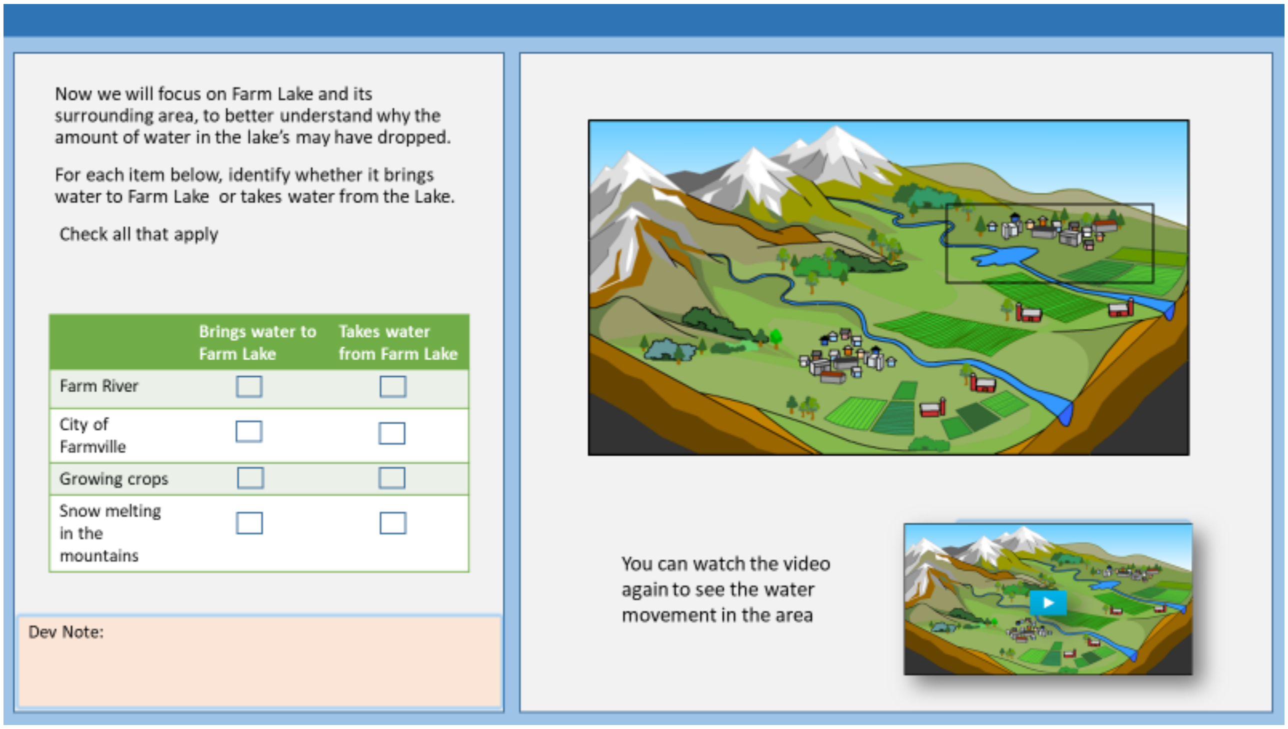
Making Amendments to the Ecosystem Task Storyboard

Following my analysis on the usability of the existing DAT-CROSS items, revisions were put forth to address usage concerns from Research Question 3 and make construct elements (as described in the hypothesized learning progression) more salient, particularly in the items meant to target *Structure and Function*. Additional helper text was included in Questions 2 and 7 to make clear that students should click on more than one option as appropriate. Language, regarding the idea that organisms on the farm serve a function in the overall functioning of the farmyard ecosystem, to text on screen between questions 2 and 3 and between questions 3 and 4. That text provides the answers to previous questions so that all students have the necessary content knowledge even if they responded to previous questions incorrectly). Later in the assessment, more emphasis was placed on making sure students understood why the farmer was introducing the different control strategies and what he expected each strategy to do.

Implications for the Design of Tasks in New System Contexts

In addition to informing revisions to the existing items of DAT-CROSS, my research sought to inform larger design ideas that could affect future CCC-targeting assessment development work. In the light of the IES grant and the findings from Research Question 1, new items were designed to improve intra-construct reliability while also accounting for the 'cutting across disciplines' (NRC, 2014, p. 37) nature of the crosscutting concepts. From the findings of Research Question 2, it was clear that students needed greater in-assessment support for thinking about the functions of structural components and how they support the overall goal of a structure. These supports not only include more screens between questions to introduce content, but also questions that are targeted to the function and behavior progression pathways from the learning progression (Appendix B). Figure 5.6 exemplifies one of these new questions.

To minimize usage disruptions across the whole set of tasks, improved support text and instructional videos were developed to help explain how various interactive elements worked.



Now we will focus on Farm Lake and its surrounding area, to better understand why the amount of water in the lake's may have dropped.

For each item below, identify whether it brings water to Farm Lake or takes water from the Lake.

Check all that apply

	Brings water to Farm Lake	Takes water from Farm Lake
Farm River	□	□
City of Farmville	□	□
Growing crops	□	□
Snow melting in the mountains	□	□

Dev Note:

You can watch the video again to see the water movement in the area

Figure 5.6 Example Question from Water Use System Storyline

Limitations to My Findings

My research for this dissertation came from only a small slice of a broader ongoing project, as usability testing is often the first step in the process of assessment validation (Zaharias & Poylymenakou, 2009). Unlike later stages, which may involve many more students (and accordingly, less rich data), usability tests involve fewer participants (my study had 45 middle-school-aged students) that may be selected on a more intentional, nonrandom criteria. The choice to use what I have previously called a "best case" sample, particularly in light of the framework of activity theory, impacts some distinct limitations on the generalizability of my findings regarding the exact usage problems all students will face when interacting with the DAT-CROSS Suite. For example, one of the highlight findings from the use of activity theory to investigate research question 3 was preliminary evidence that many students have naïve, proto-conceptions about crosscutting concepts that have built up in an evolutionary reasoning style (as described by Osborne *et al.*, 2018). These primitive notions regarding structures and their functions seem present in the words that students use to describe phenomena where such concepts are useful towards making meaning of the world, perhaps serving as a useful lever for teachers. However, it is very likely that the "best case" students who participated were also most likely to attempt to vocalize even their underdeveloped ideas without fear of judgement by their peers (who were not present in the testing setting) or the facilitating adult. Indeed, such willingness was found in other investigations of CCC reasoning where likely-affluent populations served as subjects (Beck *et al*., 2016). In contrast, students of color often struggle to put forth their primitive ideas for fear of the teacher-student and student-student interactions that can manifest (Brown, 2004). It may be the case in future uses with more vulnerable populations of students, that these vocalizations do not occur, hindering educators' ability to capitalize on them.

Another limitation is students' baseline content familiarity. Given the recentness of the NGSS (NGSS Lead States, 2013), few students have had any explicit and meaningful instruction regarding the crosscutting concepts. Thus, the preliminary evidence for a learning progression seen from the result of research question 2 can be best thought of as a progression of informal learning in that all the learning experiences related to the content and phenomena entailed in the progression have likely only occurred in informal, lived experience settings. Students of similarly little formal learning opportunities regarding ideas like systems or structure and function but of different lived experiences may progress along a very different trajectory. It is, therefore, crucial to understand my findings only as a proof-of-concept for assessing CCCs.

Chapter 6

CONCLUSION

Within the NGSS, the salience of DCIs and SEPs often overwhelms the value of the CCCs in displaying expected performance. Even among current research in developing assessment for the NGSS (DeBarger et al., 2015), it is rare to see rubrics that value the presence of CCCs in student reasoning beyond a binary condition. However, previous research findings suggest that students' explanations were influenced not only by which phenomena they were asked to explain, but also by the CCCs they were asked to use and the wording on the questions they were asked (Weiser et al., 2017). Simultaneously, the language of the NGSS seems to emphasize students *doing* over students *knowing* (NGSS Lead States, 2013a; NRC, 2012). The science education community needs new assessments that are both *doing*-oriented and capable of drawing evidence of student understanding across all three dimensions of the science-learning framework (Duschl, 2008). To this end, activity theory, when coupled with quantitative metrics, can be a powerful analytic tool in understanding and improving the kind of work dynamics (Engestrom, 2000) seemingly desired by the NGSS (Duschl et al., 2007; NRC, 2012). In setting a novel observation model for eliciting evidence from students, the AT framework provides teacher and student alike to recognize opportunities for expression of three-dimensional understandings that may not be clearly available to them for practical or socio-cultural reasons (van Es & Sherin, 2002; Farenga & Joyce, 1997; Jonassen & Rohrer-Murphy, 1999). As I described in previous sections, activity theory fits well into the design of assessments, especially ones designed to measure summative knowledge interactively (Pea, 1993) or to provide formative learning opportunities to students (Black & Wiliam, 2009). However, the task of aligning system-wide assessments to the NGSS remains. I believe that activity theory, as a method for developing evidence models for assessments (Mislevy & Haertel, 2007) has merit, but work still needs to be done to develop such models at appropriate scale for this purpose.

References

Abd-El-Khalick, F., Bell, R. L., & Lederman, N. G. (1998). The nature of science and instructional practice: Making the unnatural natural. *Science Education*, *82*(4), 417–436. doi:10.1002/(SICI)1098-237X(199807)82:4<417::AID-SCE1>3.0.CO;2-E

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education*, *93*(3), 389–421. doi:10.1002/sce.20303

American Association for the Advancement of Science (AAAS). (1993). *Benchmarks for science literacy*. *Advancement Of Science*. New York: Oxford University Press. Retrieved from http://www.project2061.org/publications/bsl/online/index.php

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences*, *13*(1), 1–14. doi:10.1207/s15327809jls1301_1

Baxter, G. P., & Mislevy, R. (2004). *The Case for an Integrated Design Framework for Assessing Science Inquiry*. *Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA*.

Beck, A., Bookbinder, A., Lee, M.J., & Rivet, A.E. (2017, April). *Identifying Early Productive Stepping Stone Conceptions of Three-dimensional Earth Science Understanding by High School Students.* Paper present at NARST, San Antonio.

Benson, L., Elliott, D., Grant, M., Holschuh, D., Kim, B., Kim, H., … Reeves, T. C. (2002). Usability and instructional design heuristics for e-learning evaluation. In P. Barker & S. Rebelsky (Eds.), *ED-MEDIA 2002--World Conference on Educational Multimedia, Hypermedia & Telecommunications* (pp. 1615–1621). Denver, CO: Association for the Advancement of Computing in Education (AACE). Retrieved from https://eric.ed.gov/?id=ED477005

Berland, L. K., & McNeill, K. L. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. *Science Education*, *94*(5), 765–793. doi:10.1002/sce.20402

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education*, *93*(1), 26–55. doi:10.1002/sce.20286

Blumenfeld, P., Fishman, B. J., Krajcik, J., Marx, R. W., & Soloway, E. (2000). Creating Usable Innovations in Systemic Reform : Scaling Up Technology-Embedded Project-Based Science in Urban Schools. *Educational Psychologist*, *35*(3), 149–164. doi:10.1207/S15326985EP3503

Breslyn, W., McGinnis, J. R., McDonald, R. C., & Hestness, E. (2016). Developing a learning progression for sea level rise, a major impact of climate change. *Journal of Research in Science Teaching*, *53*(10), 1471–1499. doi:10.1002/tea.21333

Bricker, L. A., & Bell, P. (2008). Conceptualizations of argumentation from science studies and the learning sciences and their implications for the practices of science education. *Science Education*, *92*(3), 473–498. doi:10.1002/sce.20278

Briggs, D. C., & Alonzo, A. C. (2012). The psychometric modeling of ordered multiple-choice item responses for diagnostic assessment with a learning progression. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning Progressions in Science* (pp. 293–316). Rotterdam: SensePublishers. doi:10.1007/978-94-6091-824-7_13

Brown, A. L. (1992). Design Experiments: Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings. *Journal of the Learning Sciences*, *2*(2), 141–178. doi:10.1207/s15327809jls0202_2

Brown, B. A., Reveles, J. M., & Kelly, G. J. (2005). Scientific literacy and discursive identity: A theoretical framework for understanding science learning. *Science Education*, *89*(5), 779–802. doi:10.1002/sce.20069

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, *18*(1), 32–42. doi:10.3102/0013189X018001032

Clarke, J., & Dede, C. (2009). Design for Scalability: A case study of the river city curriculum. *Journal of Science Education and Technology*, *18*(4), 353–365. doi:10.1007/s10956-009-9156-4

Cobb, P., Zhao, Q., & Dean, C. (2009). Conducting Design Experiments to Support Teachers' Learning: A Reflection From the Field. *Journal of the Learning Sciences*, *18*(2), 165–199. doi:10.1080/10508400902797933

Cognition and Technology Group at Vanderbilt University. (1992). Anchored instruction in science and mathematics: Theoretical bases, developmental projects, and initial research findings. In R. A. Duschl & R. J. Hamilton (Eds.), *Philosophy of science, Cognitive psychology, and Educational theory and practice* (pp. 244–273). New York: SUNY Press.

Cohen, E. (1998). Making cooperative learning equitable. *Educational Leadership*, *56*(1), 18–21. Retrieved from http://web.b.ebscohost.com.ezproxy.elib10.ub.unimaas.nl/ehost/pdfviewer/pdfviewer?sid=d7594b1a-abc6-4dfd-9d76-a2cf6bc85224@sessionmgr115&vid=1&hid=101

Cohen, E. G., & Lotan, R. A. (1995). Producing Equal-Status Interaction in the Heterogeneous Classroom. *American Educational Research Journal*, *32*(1), 99–120. doi:10.3102/00028312032001099

Corcoran, T. B., Mosher, F. A., & Rogat, A. (2009). Learning Progressions in Science: An Evidence- Based Approach to Reform Learning Progressions in Science: An Evidence-Based Approach to Reform. *CPRE Research Reports*. Retrieved from http://repository.upenn.edu/cpre_researchreports

Creswell, J. W. (2009). *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*. *Book* (3rd ed., Vol. 2nd ed). Los Angeles: SAGE Publications. doi:10.1177/1524839915580941

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.

de Ayala, R. (2010). Item response theory. In G. R. Hancock, R. O. Mueller, & L. M. Stapleton (Eds.), *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (p. 155). New York, NY: Routledge.

DeBarger, A. H., Penuel, W. R., Harris, C. J., & Kennedy, C. A. (2016). Building an assessment argument to design and use Next Generation Science assessments in efficacy studies of curriculum interventions. *American Journal of Evaluation*, *37*(2), 174–192. doi:10.1177/1098214015581707

DeBoer, G. E. (1991). *A history of ideas in science education: Implications for practice*. New York: Teachers College Press.

DiSessa, A. A., Cobb, P., & Andrea, A. (2004). Ontological innovation and the role of theory in design experiments. *The Journal of the Learning Sciences*, *13*(1), 77–103. doi:DOI 10.1207/s15327809jls1301_4

Duncan, R. G., & Rivet, A. E. (2013). Science Learning Progressions. *Science*, *339*(6118), 396–397. doi:10.1126/science.1228692

Duncan, R. G., & Tseng, K. A. (2011). Designing project-based instruction to foster generative and mechanistic understandings in genetics. *Science Education*, *95*(1), 21–56. doi:10.1002/sce.20407

Duschl, R. (2008). Science Education in Three-Part Harmony: Balancing Conceptual, Epistemic, and Social Learning Goals. *Review of Research in Education*, *32*(1), 268–291. doi:10.3102/0091732X07309371

Duschl, R. (2012). The second dimension — Crosscutting concepts. *Science and Children*, *79*(2), 10–14.

Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: a review and analysis. *Studies in Science Education*, *47*(2), 123–182. doi:10.1080/03057267.2011.604476

Duschl, R., Schweingruber, H. A., & Shouse, A. W. (2007). *Taking science to school: Learning and teaching Science in grades K-8*. *Taking Science to School*. Washington, D.C.: National Academies Press. doi:10.17226/11625

Edelson, D. C. (2001). Learning-for-Use: A framework for the design of technology-supported inquiry activities. *Journal of Research in Science Teaching*, *38*(3), 355–385. doi:10.1002/1098-2736(200103)38:3<355::AID-TEA1010>3.0.CO;2-M

Eger, M. (1992). Hermeneutics and science education: An introduction. *Science and Education*, *1*(4), 337–348. doi:10.1007/BF00430961

Eggert, S., & Bögeholz, S. (2010). Students' use of decision-making strategies with regard to socioscientific Issues: An application of the rasch partial credit model. *Science Education*, *94*(2), 230–258. doi:10.1002/sce.20358

Eisenhart, M., Finkel, E., & Marion, S. F. (1996). Creating the conditions for scientific literacy: A re-examination. *American Educational Research Journal*, *33*(2), 261–295. doi:10.3102/00028312033002261

Elam-Respass, T. (2018). *Effective Instructional Practices that Engage the African American Male in Middle School Science*. University of Maryland, College Park.

Engestrom, Y. (2000). Activity theory as a framework for analyzing and redesigning work. *Ergonomics*, *43*(7), 960–974. doi:10.1080/001401300409143

Farenga, S. J., & Joyce, B. A. (1997). What Children Bring to the Classroom: Learning Science From Experience. *School Science and Mathematics*, *97*(5), 248–252. doi:10.1111/j.1949-8594.1997.tb17270.x

Ford, M. (2008). Disciplinary authority and accountability in scientific practice and learning. *Science Education*, *92*(3), 404–423. doi:10.1002/sce.20263

Freire, P., & Macedo, D. P. (1995). A dialogue: Culture, langague and race. *Harvard Educational Review*, *65*(3), 377–403. doi:10.17763/haer.65.3.12g1923330p1xhj8

Furtak, E. M. (2017). Confronting dilemmas posed by three-dimensional classroom assessment: Introduction to a virtual issue of Science Education. *Science Education*. doi:10.1002/sce.21283

Garver, M. and Mentzer, J. (1999). Logistics research method: Employing structural equation modeling to test for construct validity. *Journal of Business Logistics*, *20*(1), 33. https://doi.org/10.1002/(ISSN)2158-1592

Gorin, J. S., & Mislevy, R. J. (2013). Inherent measurement challenges in the Next Generation Science Standards for both formative and summative assessment. In *Invitational Assessment Symposium* (pp. 2–39). Washington, D.C.

Gotwals, A. W., & Songer, N. B. (2013). Validity evidence for learning progression-based assessment items that fuse core disciplinary ideas and science practices. *Journal of Research in Science Teaching*, *50*(5), 597–626. doi:10.1002/tea.21083

Griffiths, A. K., & Grant, B. A. C. (1985). High school students' understanding of food webs: Identification of a learning hierarchy and related misconceptions. *Journal of Research in Science Teaching*, *22*(5), 421–436. doi:10.1002/tea.3660220505

Gunckel, K. L., Covitt, B. A., Salinas, I., & Anderson, C. W. (2012). A learning progression for water in socio-ecological systems. *Journal of Research in Science Teaching*. doi:10.1002/tea.21024

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. https://doi.org/10.1016/j.biotechadv.2011.08.021.Secreted

Hammer, D. (1994). Epistemological beliefs in introductory physics. *Cognition and Instruction*, *12*(2), 151–183. doi:10.1207/s1532690xci1202_4

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & McElhaney, K. W. (2016). *Constructing assessment tasks that blend disciplinary core ideas, crosscutting concepts, and science practices for classroom formative applications*. Menlo Park, CA: SRI International.

Hewson, P. W., & Hewson, M. G. A. B. (1984). The role of conceptual conflict in conceptual change and the design of science instruction. *Instructional Science*, *13*(1), 1–13. doi:10.1007/BF00051837

Hmelo-Silver, C. E., Liu, L., & Jordan, R. (2009). Visual representation of a multidimensional coding scheme for understanding technology-mediated learning about complex natural systems. *Research and Practice in Technology Enhanced Learning*, *04*(03), 253–280. https://doi.org/10.1142/S1793206809000714

Hoadley, C. M. (2004). Methodological Alignment in Design-Based Research. *Educational Psychologist*, *39*(4), 203–212. doi:10.1207/s15326985ep3904

Hobson, A. (1993). Ozone and Interdisciplinary Science Teaching - Learning to Address the Things that Count Most. *Journal of College Science Teaching*, *23*(1), 33–37.

Hogan, K. (2000). Assessing students' systems reasoning in ecology. *Journal of Biological Education*, *35*(1), 22–28. doi:10.1080/00219266.2000.9655731

Hogan, K., & Fisherkeller, J. (1996). Representing students' thinking about nutrient cycling in ecosystems: Bidimensional coding of a complex topic. *Journal of Research in Science Teaching*, *33*(9), 941–970. doi:10.1002/(SICI)1098-2736(199611)33:9<941::AID-TEA1>3.0.CO;2-V

Hokayem, H., Ma, J., & Jin, H. (2015). A learning progression for feedback loop reasoning at lower elementary level. *Journal of Biological Education*, *49*(3), 246–260. doi:10.1080/00219266.2014.943789

Howell, H., Phelps, G., Croft, A. J., Kirui, D., & Gitomer, D. (2013). Cognitive interviews as a tool for investigating the validity of content knowledge for teaching assessments. *ETS Research Report Series*, *2013*(1), i-97. https://doi.org/10.1002/j.2333-8504.2013.tb02326.x

Jin, H., & Anderson, C. W. (2012). A learning progression for energy in socio-ecological systems. *Journal of Research in Science Teaching*. doi:10.1002/tea.21051

Joe, J., Tocci, C., Holtzman, S., & Williams, J. (2013). *Foundations of observation: Considerations for developing a classroom observation systlem that helps districts achieve consistent and accurate scores*. *MET Project Policy and Practice Brief: Bill and Melinda Gates Foundation*. Princeton, NJ. Retrieved from http://www.metproject.org/downloads/MET-ETS_Foundations_of_Observation.pdf

Johnstone, C. J., Bottsford-Miller, N. a., & Thompson, S. J. (2006). Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and english language learners. *Technical Report 44 University of Minnesota, National Center on Educational Outcomes*. Retrieved from http://www.cehd.umn.edu/nceo/OnlinePubs/Tech44/%5Cnhttp://education.umn.edu/NCEO/OnlinePubs/Tech44/

Jonassen, D. H., Rohrer-Murphy, L., & Jonassen H., D. (1999). Activity theory as a framework for designing constructivist learning environments. *Educational Technology Research and Development*, *47*(1), 61–79. doi:10.1007/BF02299477

Joseph, D. (2004). The practice of design-based research: uncovering the interplay between design, research, and the real-world context. *Educational Psychologist*, *39*(4), 235–242. https://doi.org/10.1207/s15326985ep3904_5

Kanter, D. E. (2010). Doing the project and learning the content: Designing project-based science curricula for meaningful understanding. *Science Education*, *94*(3), 525–551. doi:10.1002/sce.20381

Kelly, G. J., McDonald, S., & Wickman, P.-O. (2012). Science learning and epistemology. In B. J. Fraser, K. Tobin, & C. J. McRobbie (Eds.), *Second International Handbook of*

Science Education (pp. 281–291). Dordrecht: Springer Netherlands. doi:10.1007/978-1-4020-9041-7

Kerby, D. S. (2014). The simple difference formula: an approach to teaching nonparametric correlation. *Comprehensive Psychology*, *3*(1), 11.IT.3.1. https://doi.org/10.2466/11.IT.3.1

Kesidou, S., & Roseman, J. E. (2002). How well do middle school science programs measure up? Findings from Project 2061's curriculum review. *Journal of Research in Science Teaching*, *39*(6), 522–549. doi:10.1002/tea.10035

Kilinc, A., Demiral, U., & Kartal, T. (2017). Resistance to dialogic discourse in SSI teaching: The effects of an argumentation-based workshop, teaching practicum, and induction on a preservice science teacher. *Journal of Research in Science Teaching*, *54*(6), 764–789. doi:10.1002/tea.21385

Kline, P. (1994). Factor Analysis in Test Construction. In *An Easy Guide to Factor Analysis* (pp. 125–139). New York: Routledge.

Krajcik, J. (2015). Three-dimensional instruction: Using a new type of teaching in the science classroom. *The Science Teacher*, *82*(8).

Krajcik, J., & Merritt, J. (2012). Engaging Students in Scientific Practices: What does constructing and revising models look like in the science classroom? *Science and Children*, *49*(7), 10.

Krajcik, J., Codere, S., Dahsah, C., Bayer, R., & Mun, K. (2014). Planning instruction to meet the intent of the Next Generation Science Standards. *Journal of Science Teacher Education*, *25*(2), 157–175. doi:10.1007/s10972-014-9383-2

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. doi:10.1002/sce.20240

Lakoff, G., & Johnson, M. (1997). *Metaphors We Live By*. *The production of reality: essays and reading on social interaction* (2nd ed.). Chicago: University of Chicago Press.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. New York: Cambridge University Press. Retrieved from https://books.google.co.uk/books?id=ZVogAwAAQBAJ&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

Lederman, N. G., & Lederman, J. S. (2014). Is nature of science going, going, going, gone? *Journal of Science Teacher Education*, *25*(3), 235–238. doi:10.1007/s10972-014-9386-z

Lederman, N. G., & Zeidler, D. L. (1987). Science Teachers' Conceptions of the Nature of Science: Do They Really Influence Teaching Behavior? *Science Education*, *71*(5), 721–734. doi:10.1002/sce.3730710509

Lee, H. S., & Liu, O. L. (2010). Assessing learning progression of energy concepts across middle school grades: The knowledge integration perspective. *Science Education*, *94*(4), 665–688. doi:10.1002/sce.20382

Lee, O. (1997). Guest editorial: Scientific literacy for all: What is it, and how can we achieve it? *Journal of Research in Science Teaching*, *34*(3), 219–222. doi:10.1002/(SICI)1098-2736(199703)34:3<219::AID-TEA1>3.0.CO;2-V

Lee, O., Quinn, H., & Valdes, G. (2013). Science and language for English language learners in relation to Next Generation Science Standards and with implications for Common Core state standards for English language arts and mathematics. *Educational Researcher*, *42*(4), 223–233. doi:10.3102/0013189X13480524

Lehrer, R., & Schauble, L. (2006). Cultivating model Based reasoning in science education. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 371–387). Cambridge, UK: Cambridge University Press.

Lin, C., & Hu, R. (2003). Students' understanding of energy flow and matter cycling in the context of the food chain, photosynthesis, and respiration. *International Journal of Science Education*, *25*(1loving2), 1529–1544. doi:10.1080/0950069032000052045

Liu, L., Rogat, A., & Bertling, M. (2013). A CBAL™ Science Model of Cognition: Developing a Competency Model and Learning Progressions to Support Assessment Development.

Maclure, M., & Willett, W. C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology*, *126*(2), 161–169. https://doi.org/10.1093/aje/126.2.161

McComas, W. F., Clough, M. P., & Almazroa, H. (2002). The role and character of the nature of science in science education. In W. F. McComas (Ed.), *The nature of science in science education: Rationale and strategies* (pp. 3–39). Dordrecht: Kluwer Academic Publishers. doi:10.1007/0-306-47215-5_1

Mislevy, R. J. (2016). How developments in psychology and technology challenge validity argumentation. *Journal of Educational Measurement*, *53*(3), 265–292. https://doi.org/10.1111/jedm.12117

Mislevy, R. J. (2017). Resolving the paradox of rich performance tasks. In *Test Fairness in the New Generation of Large-Scale Assessment* (pp. 1–46). Charlotte, N.C.: Information Age Publishers.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20. doi:10.1111/j.1745-3992.2006.00075.x

Mislevy, R. J., Haertel, G., Riconscente, M., Wise Rutstein, D., & Ziker, C. (2017). *Assessing model-based reasoning using evidence-centered design: A suite of research-based design patterns*. Cham, Switzerland: Springer International Publishing. doi:10.1007/978-3-319-52246-3

Mislevy, R. J., Riconscente, M. M., & Rutstein, D. W. (2009). Design patterns for assessing model-based reasoning (Large-Scale Assessment Technical Report 6). *Menlo Park, CA: SRI International*.

Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching*, *46*(6), 675–698. doi:10.1002/tea.20314

Morin, O., Simonneaux, L., & Tytler, R. (2017). Engaging with socially acute questions: Development and validation of an interactional reasoning framework. *Journal of Research in Science Teaching*, *54*(7), 825–851. doi:10.1002/tea.21386

Mueller, R. O., & Hancock, G. R. (2015). Factor analysis and latent structure analysis: confirmatory factor analysis. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (Second Edi, pp. 686–690). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.25009-5

Mutegi, J. (2011). The inadequacies of science-for-all and the necessity and nature of a socially transformative curriculum approach for African American science education. *Journal of Research in Science Teaching*, *248*(3), 301–316.

Nardi, B. A. (1996). Studying context: A comparison of Activity Theory, Situated Action Models , and Distributed Cognition. In B. A. Nardi (Ed.), *Context and Conciousness: Activity Theory and Human Computer Interaction* (pp. 69–102). Cambridge, MA: The MIT Press. doi:citeulike-article-id:1292510

National Research Council. (1996). *National Science Education Standards*. Washington, D.C.: National Academies Press. doi:10.17226/4962

National Research Council. (2011). *Successful K-12 STEM Education*. Washington, D.C.: National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, D.C.: National Academies Press. Retrieved from http://www.nap.edu/catalog/13165/a-framework-for-k12-science-education-practices-crosscutting-concepts-and

National Research Council. (2013). *Monitoring progress toward successful K-12 STEM education: A nation advancing? National Academies Press*. Washington, D.C.: National Academies Press. doi:10.17226/13509

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards*. Washington, D.C.: National Academies Press. doi:10.17226/18409

Nersessian, N. J. (2002). The cognitive basis of model-based reasoning in science. In P. Carruthers, S. Stich, & M. Siegal (Eds.), *The Cognitive Basis of Science* (pp. 133–153). New York: Cambridge University Press. doi:10.1017/cbo9780511613517.008

Neumann, K., Viering, T., Boone, W. J., & Fischer, H. E. (2013). Towards a learning progression of energy. *Journal of Research in Science Teaching*, *50*(2), 162–188. doi:10.1002/tea.21061

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

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

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

Noble, T., Suarez, C., Rosebery, A., O'Connor, M. C., Warren, B., & Hudicourt-Barnes, J. (2012). "I never thought of it as freezing": How students answer questions on large-scale science tests and what they know about science. *Journal of Research in Science Teaching*, *49*(6), 778–803. doi:10.1002/tea.21026

Nussbaumer, D. (2012). An overview of cultural historical activity theory (chat) use in classroom research 2000 to 2009. *Educational Review*, *64*(1), 37–55. https://doi.org/10.1080/00131911.2011.553947

O'Brien, H. L., & Toms, E. G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, *59*(6), 938–955. https://doi.org/10.1002/asi.20801.1

Osborne, J. (2014). Teaching scientific practices: Meeting the challenge of change. *Journal of Science Teacher Education*, *25*(2), 177–196. doi:10.1007/s10972-014-9384-1

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S. Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching*, *53*(6), 821–846. doi:10.1002/tea.21316

Osborne, J., Rafanelli, S., & Kind, P. (2018). Toward a more coherent model for science education than the crosscutting concepts of the next generation science standards: the affordances of styles of reasoning. *Journal of Research in Science Teaching*, *55*(7), 962–981. https://doi.org/10.1002/tea.21460

Pellegrino, J. W. (2013). Proficiency in Science: Assessment Challenges and Opportunities. *Science*, *340*(6130), 320–323. doi:10.1126/science.1232065

Pellegrino, W., Chudowsky, N., & Glaser, R. (2001). The nature of assessment and reasoning from evidence. In *Knowing What Students Know: The Science and Design of Educational Assessment* (pp. 37–54). Washington, D.C.: National Academies Press. Retrieved from http://books.google.com/books?hl=en&lr=&id=t5mcAgAAQBAJ&pgis=1

Peterson, C. H., Peterson, N. A., & Powell, K. G. (2017). Cognitive interviewing for item development: validity evidence based on content and response processes. *Measurement and Evaluation in Counseling and Development*, *50*(4), 217–223. https://doi.org/10.1080/07481756.2017.1339564

Plummer, J. D., & Krajcik, J. (2010). Building a learning progression for celestial motion: Elementary levels from an earth-based perspective. *Journal of Research in Science Teaching*, *47*(7), 768–787. doi:10.1002/tea.20355

Plummer, J. D., Palma, C., Flarend, A., Rubin, K., Ong, Y. S., Botzer, B., … Furman, T. (2015). Development of a learning progression for the formation of the solar system. *International Journal of Science Education*, *37*(9), 1381–1401. doi:10.1080/09500693.2015.1036386

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. doi:10.1002/tea.20415

Quellmalz, E. S., & Pellegrino, J. W. (2009). Technology and Testing. *Science*, *323*(5910), 75–79. doi:10.1126/science.1168046

Ríordáin, M. N., Johnston, J., & Walshe, G. (2015). Making mathematics and science integration happen: key aspects of practice. *International Journal of Mathematical Education in Science and Technology*, *47*(September), 1–23. doi:10.1080/0020739X.2015.1078001

Rivet, A. E., & Kastens, K. A. (2012). Developing a construct-based assessment to examine students' analogical reasoning around physical models in Earth Science. *Journal of Research in Science Teaching*, *49*(6), 713–743. doi:10.1002/tea.21029

Rivet, A. E., & Krajcik, J. S. (2004). Achieving standards in Urban systemic reform: An example of a sixth grade project-based science curriculum. *Journal of Research in Science Teaching*, *41*(7), 669–692. doi:10.1002/tea.20021

Rivet, A. E., & Krajcik, J. S. (2008). Contextualizing instruction: Leveraging students' prior knowledge and experiences to foster understanding of middle school science. *Journal of Research in Science Teaching*, *45*(1), 79–100. doi:10.1002/tea.20203

Rivet, A. E., Weiser, G., Lyu, X., Li, Y., & Rojas-Perilla, D. (2016). What are crosscutting concepts in science? Four metaphorical perspectives. In C. K. Looi, J. L. Polman, U. Cress, & P. Reimann (Eds.), *Proceedings of International Conference of the Learning Sciences, ICLS* (Vol. 2). Signapore: International Society of the Learning Sciences. doi:10.22318/icls2016.149

Rodriguez, A. J. (2015). What about a dimension of engagement, equity, and diversity practices? A critique of the next generation science standards. *Journal of Research in Science Teaching*, *52*(7), 1031–1051. https://doi.org/10.1002/tea.21232

Roseman, J., Kesidou, S., & Stern, L. (1996). Identifying curriculum materials for science literacy: A Project 2061 evaluation tool. Retrieved November 27, 2014, from http://www.project2061.org/publications/articles/roseman/roseman2.htm

Roth, W.-M., & Tobin, K. (2002). Redesigning an "Urban" Teacher Education Program: An Activity Theory Perspective. *Mind, Culture, and Activity*, 9(2), 108–131. doi:10.1207/S15327884MCA0902

Rutherford, F. J., & Ahlgren, A. (1990). *Science for All Americans*. New York: Oxford University Press.

Sandoval, W. A. (2005). Understanding students' practical epistemologies and their influence on learning through inquiry. *Science Education*, *89*(4), 634–656. doi:10.1002/sce.20065

Schwarz, C. (2009). Developing preservice elementary teachers' knowledge and practices through modeling-centered scientific inquiry. *Science Education*, *93*(4), 720–744. doi:10.1002/sce.20324

Schwarz, C. V., & White, B. Y. (2005). Metamodeling Knowledge: Developing Students' Understanding of Scientific Modeling. *Cognition and Instruction*, *23*(2), 165–205. doi:10.1207/s1532690xci2302_1

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D., ... & Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching*, *46*(6), 632-654.

Shavelson, R. J., Phillips, D. C., Towne, L., & Feuer, M. J. (2003). On the Science of Education Design Studies. *Educational Researcher*, *32*(1), 25–28. doi:10.3102/0013189X032001025

Snow, E., & Katz, I. R. (2009). Using cognitive interviews to validate an interpretive argument for the ETS iSkills™ assessment. *Communications in Information Literacy*, *3*(2), 99–127.

Songer, N. B., Kelcey, B., & Gotwals, A. W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. *Journal of Research in Science Teaching*, *46*(6), 610–631. doi:10.1002/tea.20313

Songer, N. B., Lee, H.-S., & McDonald, S. (2003). Research towards an expanded understanding of inquiry science beyond one idealized standard. *Science Education*, *87*(4), 490–516. doi:10.1002/sce.10085

Sparks, J. R., & Deane, P. (2015). Cognitively Based Assessment of Research and Inquiry Skills: Defining a Key Practice in the English Language Arts. *ETS Research Report Series*, *2015*(2), 1–55. doi:10.1002/ets2.12082

Stevens, S. Y., Delgado, C., & Krajcik, J. S. (2009). Developing a hypothetical multi-dimensional learning progression for the nature of matter. *Journal of Research in Science Teaching*, *47*(6), 687–715. doi:10.1002/tea.20324

Stroupe, D. (2015). Describing “Science Practice” in learning settings. *Science Education*, *99*(6), 1033–1040. doi:10.1002/sce.21191

Svoboda, J., & Passmore, C. (2011). The Strategies of Modeling in Biology Education. *Science & Education*, *22*(1), 119–142. doi:10.1007/s11191-011-9425-5

van Es, E. A., & Sherin, M. G. (2002). Learning to Notice: Scaffolding New Teachers’ Interpretations of Classroom Interactions. *Journal of Technology and Teacher Education*, *10*(4), 571–596.

Weiser, G., & Liu, L. (2018, March). *A design framework for the development of scenario-based assessments for summative assessment settings.* Poster paper present at NARST, Atlanta.

Weiser, G., Lyu, X., & Rojas-Perilla, D.F. (2017, April). *What Are Crosscutting Concepts in Science? Four Metaphorical Perspectives.* Paper present at NARST, San Antonio.

Wiliam, D. (2010). What counts as evidence of educational achievement? The role of constructs in the pursuit of equity in assessment,. *Review of Research in Education*, *34*(1), 254–284. doi:10.3102/0091732X09351544

Wilson, M., & Draney, K. (2005). Some links between large-scale and classroom assessments: The case of the BEAR assessment system. *Yearbook of the National Society for the Study of Education*, *103*(2), 132–154. doi:10.1111/j.1744-7984.2004.tb00051.x

Yilmaz, S., Eryilmaz, A., & Geban, O. (2006). Assessing the impact of bridging analogies in mechanics. *School Science and Mathematics*, *106*(6), 220–230. doi:10.1111/j.1949-8594.2006.tb17911.x

Yoon, S. A. (2008). An evolutionary approach to harnessing complex systems thinking in the science and technology classroom. *International Journal of Science Education*, *30*(1), 1–32. doi:10.1080/09500690601101672

Yoon, S. A., Goh, S.-E., & Park, M. (2018). Teaching and learning about complex systems in K–12 science education: A review of empirical studies 1995–2015. *Review of Educational Research*, *88*(2), 285–325. doi:10.3102/0034654317746090

Yore, L., Bisanz, G. L., & Hand, B. M. (2003). Examining the literacy component of science literacy: 25 years of language arts and science research. *International Journal of Science Education*, *25*(6), 689–725. doi:10.1080/09500690305018

Zaharias, P., & Poylymenakou, A. (2009). Developing a usability evaluation method for e-learning applications: Beyond functional usability. *International Journal of Human-Computer Interaction*, *25*(1), 75–98. doi:10.1080/10447310802546716

APPENDICES

Appendix A: Task Model for Assessment of Ecological Systems, Structures, and Functions – Adapted from MS-LS2-3, MS-LS2-4, & MS-LS2-5

Performance Expectations and Their Dimensions

Students who demonstrate understanding can:	
MS-LS2-3 (Original)	Develop a model to describe the cycling of matter and flow of energy among living and nonliving parts of an ecosystem. [Clarification Statement: Emphasis is on describing the conservation of matter and flow of energy into and out of various ecosystems, and on defining the boundaries of the system.] [Assessment Boundary: Assessment does not include the use of chemical reactions to describe the processes.]
MS-LS2-3 (Adapted)	Develop a model to describe relationships among living and nonliving parts of an ecosystem that is capable of tracking the cycling of matter and flow of energy among ecosystem components.
MS-LS2-4 (Original)	Construct an argument supported by empirical evidence that changes to physical or biological components of an ecosystem affect populations. [Clarification Statement: Emphasis is on recognizing patterns in data and making warranted inferences about changes in populations, and on evaluating empirical evidence supporting arguments about changes to ecosystems.]
MS-LS-4 (Adapted)	Construct an argument supported by empirical evidence that changes the functioning of an ecosystem (including its ability to support population growth) can be driven by changes to the physical or biological components of the ecosystem.
MS-LS2-5 (Original)	Evaluate competing design solutions for maintaining biodiversity and ecosystem services. [Clarification Statement: Examples of ecosystem services could include water purification, nutrient recycling, and prevention of soil erosion. Examples of design solution constraints could include scientific, economic, and social considerations.]
MS-LS2-5 (Adapted)	Evaluate competing design solutions for achieving some human want based on criteria that consider maintaining biodiversity and ecosystem services.

Science and Engineering Practices	**Disciplinary Core Ideas**	**Crosscutting Concepts**

Developing and Using Models Modeling in 6–8 builds on K–5 experiences and progresses to developing, using, and revising models to describe, test, and predict more abstract phenomena and design systems. • Develop a model to describe phenomena. (MS-LS2-3) **Engaging in Argument from Evidence** Engaging in argument from evidence in 6–8 builds on K–5 experiences and progresses to constructing a convincing argument that supports or refutes claims for either explanations or solutions about the natural and designed world(s). • Construct an oral and written argument supported by empirical evidence and scientific reasoning to support or refute an explanation or a model for a phenomenon or a solution to a problem. (MS-LS2-4) • Evaluate competing design solutions based on jointly developed and agreed upon design criteria. (MS-LS2-5)	**LS2.B: Cycle of Matter and Energy Transfer in Ecosystems** • Food webs are models that demonstrate how matter and energy is transferred between producers, consumers, and decomposers as the three groups interact within an ecosystem. Transfers of matter into and out of the physical environment occur at every level. Decomposers recycle nutrients from dead plant or animal matter back to the soil in terrestrial environments or to the water in aquatic environments. The atoms that make up the organisms in an ecosystem are cycled repeatedly between the living and nonliving parts of the ecosystem. **LS2.C: Ecosystem Dynamics, Functioning, and Resilience** • Ecosystems are dynamic in nature; their characteristics can vary over time. Disruptions to any physical or biological component of an ecosystem can lead to shifts in all its populations.	**Systems and Systems Models (Adaptation):** • Representing flows of matter and energy within a larger, complex system can be accomplished by modeling flows within and among relevant sub-systems. (MS-LS2-3) • Models can be used to represent the impacts of changes in a sub-system on the larger system particularly when evidence can only be collected from limited aspects of the broader phenomenon. (Might be adaptable into Structure/Function if we add ESS2.A or ESS2.C DCIs) (MS-LS2-4) **Structure and Function (Adaptation):** • The structure of engineering solutions are designed to serve particular functions through the physical properties of the materials used and the environment the solution inhabits (MS-LS2-5) **Energy and Matter (Original)** • The transfer of energy can be tracked as energy flows through a natural system.

	- Biodiversity describes the variety of species found in Earth's terrestrial and oceanic ecosystems. The completeness or integrity of an ecosystem's biodiversity is often used as a measure of its health. **LS4.D: Biodiversity and Humans** - Changes in biodiversity can influence humans' resources, such as food, energy, and medicines, as well as ecosystem services that humans rely on—for example, water purification and recycling. (secondary) **ETS1.B: Developing Possible Solutions** - There are systematic processes for evaluating solutions with respect to how well they meet the criteria and constraints of a problem. (secondary)	**Stability and Change (Original)** - Small changes in one part of a system might cause large changes in another part.

Practice-oriented Focal KSAs (Note: "ability" here means capability to reason as described about given DCIs and CCCs. No claim is made for "abilities" as decoupled from DCIs and CCCs.)	<u>Ability to develop model-based arguments:</u> - Ability to develop models that represent natural events systems, aspects of a theory and evidence, or design solutions. - Ability to determine the components as well as connections and relationships among multiple components of the event, system, or design solution to include in the model and those to omit. - Ability to determine scope, scale, and grain-size of the model, as appropriate to its intended use.

This section adapted from Model-based Argumentation Design Pattern Document.	• Ability to represent mechanisms, relationships, and connections to explain the event, system, or design solution with multiple types of models. Ability to evaluate competing models: • Ability to evaluate the explanatory or predictive power of competing models taking into account evidence and empirical data associated with a phenomenon under investigation. • Ability to evaluate how competing models describe mechanisms and processes related to the target event or system. • Ability to collect evidence to reason qualitatively or quantitatively about concepts and relationships represented in models. • Ability to use evidence or empirical data to generate explanations and predictions about the behavior of a scientific phenomenon. • Ability to apply science concepts or principles to reason why the data support or refute an argument. • Ability to select appropriate model representations that are most useful for supporting or refuting an argument. Ability to revise model-based arguments: • Ability to revise models in light of empirical evidence to improve their explanatory and predictive power. • Ability to apply alternative science concepts/principles to support an argument. • Ability to refine the data collection approach to improve the appropriateness, accuracy, or sufficiency of the empirical data.
Viable Indicators of Cross-Dimensional Progress (Note: "cross-dimensional" here means reasoning that manifests in the process of engaging with a practice in a particular way about a particular phenomenon. No claim is made for "abilities" as decoupled from DCIs.) **This section adapted from Design**	**Progress in Systems and System Models** SM.1. **Lowest performing students – Ability to identify relevant features but not provide linkages connecting several features nor provide explicit reasoning why features are or are not relevant.** SM.1.a. Fragmented identification of system components and relationships within and across scales or scopes that do not map to each other. SM.1.b. System components may be represented as "boxed in" by system boundary and immutable to change. SM.1.c. Limited ability to represent the boundaries of the system and system may be inappropriately categorized as open or closed to external inputs. SM.1.d. Identification of salient features may be based on familiarity (or similarity to familiar features/phenomena) rather than on what best suits the goal.

Patterns for CCC x SEP	SM.1.e. Goal of model development is visuospatial illustration phenomenon with overemphasis on representing "the look" of the system rather than relationships among entities. SM.2. **Lower Intermediate performing students – Able to create single entity-entity relationship connections. Able to find several instances in which the same connection appears.** SM.2.a. In model development, may identify multiple system entities and relationships at multiple system levels, but not represent interdependence among those system levels. SM.2.b. In model use, may be able to describe that system levels are interdependent without being able to provide driving mechanism. SM.2.c. Able to identify movement of energy or matter across system boundaries (though inciting act will rely on anthropomorphic agents). SM.2.d. May be inconsistently able to create a chain of single cause-single effect relationships within the system linked one after the next. i. Represented chains will always reinforce the effect (positive feedback loops only). SM.2.e. Reasoning for represented relationships may be just-so or overly emphasize anthropomorphic agents. Goal of model development is to describe "the way things are" rather than to have some broader argumentative/predictive goal SM.3. **Higher Intermediate performing students – Able to identify how micro-level causal relationships can produce higher-level system behavior. Able to contextualize a model in terms of some explanatory or argumentative goal.** SM.3.a. In models and explanations, indicate relationship mapping in which many causes combine to produce any one effect. i. At the higher end of this benchmark, mapping may begin to become many-to-many SM.3.b. In models and explanations, interdependent relationships across system/subsystem levels are represented in limited form. ii. Student can identify/represent evidence at the subsystem level for relationships at the system level.

	iii. Student can identify evidence at the system level for relationships that primarily exist at the subsystem level. SM.3.c. Can describe/represent both positive and negative feedback loops in matter/energy (or generalized input/output) flows. SM.3.d. Continues to explain emergent phenomena in terms of central control or adherence to a coherent, rigid framework. SM.3.e. Can explain how features of model support a broader explanatory goal and make choices regarding model feature selection pursuant to that goal. SM.4. **Highest performing students – Able to recognize constraints and limitations of a model** SM.4.a. Able to identify feature mediators (e.g. temperature, topography, available inputs) that may alter existing system relationships. SM.4.b. Able to identify subsystem relationships mediate (or create dynamism) among system relationships SM.4.c. –May already be at the highest level by SM.3.c SM.4.d. In models and explanations, students represent how emergent properties at higher system levels manifest from the rules governing lower-level system interactions without requiring centralized agents/actors. SM.4.e. Goal of the model is to support critique of provided arguments, suggest new sources of evidence, and make complex predictions about how relationships fit into a broader system context **Progress in Structure and Function** SF.1. **Lowest performing students – Ability to identify relevant features but not provide linkages connecting several features nor provide explicit reasoning why features are or are not relevant.** SF.3.a. Fragmented identification of structural features and entity functions within and across scales or scopes that do not map to each other. SF.3.b. In models and model use, students primarily identify macro-scale structures or functions and only some micro-scale functions (micro-scale structure may not be expressed). SF.3.c. Identification of salient features may be based on familiarity (or similarity to familiar

	features/phenomena) rather than on what best suits the goal. SF.3.d. Goal of model development is visuospatial illustration phenomenon SF.2. **Lower Intermediate performing students (Bridges?) – Able to create single structure-function connections (or chain of that connection). Able to find several instances in which the same connection appears.** SF.2.a. In models and explanations, indicate structures and functions are mapped 1:1 SF.2.b. In models and explanations may identify multiple structures and functions at multiple system levels SF.2.c. In models and explanations, fail to identify how structures and interrelationships among structures afford function. SF.2.d. Reasoning for structure may be just-so or overly emphasize anthropomorphic agents. SF.3. **Higher Intermediate performing students – Able to identify how macro-level functional properties interact with micro-level structures and how model can be used to support goal** SF.3.a. In models and explanations, indicate structure-function mapping (may be many to one) SF.3.b. In models and explanations, identify that structures and functions are inter- related across system levels SF.3.c. Can use the model to make an argument in support of a driving mechanism that makes structures afford function. SF.3.d. Can explain how features of model support a broader argumentative goal and make choices regarding model feature selection pursuant to that goal. SF.4. **Highest performing students – Able to recognize constraints and limitations of a model** SF.4.a. In models and explanations, identify structure-function relationships are dynamic SF.4.b. Able to identify mediators (e.g. temperature, topography, available inputs) that may alter existing structure-function relationships. SF.4.c. Able to use model of structure and function to make predictions about how relationships fit into a broader system context. SF.4.d. Able to critique a model or ask a question of a model that highlights failure to reach an explanatory/predictive/argumentative goal.

DCI Assessment Targets	Students can: LS2.B.a. Differentiate between organisms based on whether they are producers, consumers, or decomposers. LS2.B.b. Describe the role nonliving parts of an ecosystem (like water, air, and minerals) play in mediating the interactions between producers, consumers, and decomposers within an ecosystem. LS2.B.c. Tracks and quantify the cycling of matter and energy through various reservoirs (both living and nonliving) within an ecosystem. LS2.B.d. Describe relationships between the consumption of matter and energy by both consumer and producer organisms in terms of matter/energy conservation principles. LS2.C.a. Identify evidence that changes are occurring among the physical or biological components of an ecosystem and describe the rate at which such changes are occurring. LS2.C.b. Identify evidence that changes are occurring to the functioning of an ecosystem (particularly its ability to support population growth) and describe the rate at which such changes are occurring. LS2.C.c. Identify keystone ecosystem components/relationships that, if changed, have disproportionately large effects on the continued functioning of ecosystem. LS2.C.d. Describe mechanisms by which various ecosystem components act to support population growth, quantify the strength of correlations between those mechanisms and population sizes, and the describe the resilience of those mechanisms to change. LS2.C.e. Use biodiversity as a measure of the health of an ecosystem and the ability of an ecosystem to support population growth. LS4.D.a. Identify various 'Ecosystem services' in which humans act to support the stability/resilience of an ecosystem. LS4.D.b. Describe the tradeoffs associated with various ecosystem services in light of a local context (both local desires and local constraints). LS4.D.c. Identify potential unintended consequences of a given ecosystem service endeavor and describe ways that can mitigate the risk of unintended consequences occurring. ETS1.B.a. Identify relevant constraints and criteria that apply to designing an engineering solution to a problem faced by a local community. ETS1.B.b. Describe tradeoffs between solving only a limited number of the total set of problem-pieces.

	ETS1.B.c. Evaluate the effectiveness of multiple (three or more) potential design solutions to an engineering problem based on provided data. ETS1.B.d. Identify appropriate evidence sources that could be used to evaluate potential design solutions on their ability to meet the criteria of the problem definition while staying within relevant constraints.
Possible Phenomena or Contexts*	Potential criteria/constrains that may govern the viability of design solutions include: • Relevant physical principles • Cost of development • Features of the local geography • Differential features of local climate • Trends in population growth or per-capita resource consumption • Local and global trends in human behavior
Examples of Integration of Assessment Targets and Evidence	1. **Systems and System Models –** Task provides a model representing food web relationships among organisms within an ecosystem. Students are asked to identify consumers, producers, and decomposers as well as identify keystone organisms (whereby changes to their populations have dramatic effects on the other populations in the ecosystem). a. **Successful Performance:** Students are able to identify and categorize organisms appropriately in light of a provided context **(SM.1.a)**. b. **Unsuccessful Performance**: Students suggest that irrelevant system components are most salient (i.e. fish are keystone organisms because humans will feel sad if they die out) **(SM.1.d)**. 2. **Systems and System Models –** Task expands on the food web model to include the tracking of transfer of energy and matter via the production and consumption of organisms' biomasses (this may be a simulated task). Students are asked to quantify matter/energy stored in different reservoirs. **a.** **Successful Performance:** Students are able to track the transfer of energy among ecosystem components via a chain of solar source -> producer via photosynthesis -> primary consumer -> secondary consumer -> etc **(SM.2.c)**. They may exhibit confusion about the role of decomposers as part of a nutrient transport cycle **(SM.2.b & SM.2.d)**. b. **Unsuccessful Performance:** Student fails to identify energy/matter transport pathways or fails to amend model to represent transport **(SM.1.c)**.

3. **Systems and System Models –** Task expands on the context grounding the food web model to provide details of a change that may be affecting the physical or biological subsystem components of the ecosystem. Students are asked to reason about the higher-level effects these changes might have on existing energy/matter transport pathways. (Note that, at this level, students will still focus on centralized control mechanisms).
    a. **Successful Performance:** Students are able to express bi-directional, interdependent relationships between subsystem and system components **(SM.3.b)**. They are able to use these relationships to explain whether or not the ecosystem is resilient to the change provided by the task **(SM.3.c)**.
    b. **Unsuccessful Performance:** Student fails to describe ways in which changes to system components can result in changes to existing food-web relationships and/or fails to describe ways in which changes to one or more food-web relationships may affect the entire web **(SM.2.a & SM.2.b).**
    c. **Potential midpoint performance:** Students is able to describe interdependent relationships **(SM.3.b)** but is only able to represent positive feedback loops (thereby always concluding that there is a lack of resilience in the ecosystem) **(SM.2.d.i)**.
4. **Systems and System Models –** Task re-contextualizes the ecosystem changes in terms of risks to change (i.e. an unintended consequence of an ecosystem service). Students are asked to make an argument which contrasts potential benefits of the ecosystem service against potential negative consequences (both of action and inaction).
    a. **Successful Performance:** Student is able to describe mediating factors that contribute or mitigate risk, including feedback loops, entity mediators, and relational mediators in order to predict outcomes **(SM.4.a & SM.4.b)**. Argument for/against ecosystem service should be grounded in that prediction **(SM.4.e)**.
        i. **Alternative:** Student is able to critique a provided argument/prediction on the basis of relational mediators **(SM.4.e)**.
    b. **Unsuccessful Performance:** Student fails to describe how system features or entities may act to mediate other system relationships (i.e. fails to describe how

population growth affects overall resource availability) **(SM.3.d)**. They make inaccurate or inappropriate predictions or fail to argue how model elements exist to support that prediction **(SM.2.e)**.

5. **Structure and Function –** Task provides a description of several potential engineering solutions each designed to provide an ecosystem service in light of the needs of a local community. Students are asked to model the criteria and constraints of the problem definition, highlighting relevant aspects of the local environment. (Models may take the form of illustrations, concept maps, linguistic descriptions, etc.)
    - **a. Successful Performance:** Student is able to identify relevant ecosystem structures and describe how the engineering solution fits into those pre-existing structures. **(SF.1.a & SF.1.d)**
    - **b. Unsuccessful Performance:** Student overemphasizes familiar structures or functions **(SF.1.c).** Student represents only structures or only function **(SF.1.b).**
6. **Structure and Function –** Task provides further details regarding the proposed engineering solution for an ecosystem service. Students are asked to draw connections between the structure/function relationships that are common to several proposed designs.
    - **a. Successful Performance:** Student successfully maps a particular engineered structure to the functional goal of the design **(SF.2.a).** They are able to identify several instances in which a given structure-function pairing appears in many different proposed solutions. **(SF.2.b)**
    - **b. Unsuccessful Performance:** Student fails to form a structure-function pairing or fails to see that same relationship in other designs **(SF.1.a).**
7. **Structure and Function –** Task provides additional context in which a local community is seeking to evaluate the proposed solutions in order to implement one that best suits their needs. Students are asked to build on the connections from the bridge task to identify sources of evidence that could be used to assess and select the best design.
    - **a. Successful Performance:** Student is able to use commonalities in structure-function relationships across designs to identify potential sources of evidence **(SF.3.c).** Student is able to identify evidence that manifests at a different level than the structure-

	function relationship or across levels of the structure-function relationship **(SF.3.b).** **b. Unsuccessful Performance:** Student fails to develop an argument for how evidence collected supports an assessment of the proposed design **(SF.2.c).** Student fails to contextualize relationships in terms of the criteria and constraints bounding the problem **(SF.2.d).** 8. **Structure and Function –** Task provides an argument for or against the implementation of a particular ecosystem service via the construction of some human-engineered structure. Students are asked to use a model of the functioning of the solution to critique the proposed implementation/construction strategy. **a. Successful Performance:** Student effectively uses the model to identify mediating relationships that run counter to the functioning of the engineering design **(SF.4.b)** and proposes amendments to the implementation that may act to mitigate unintended consequences **(SF.4.c).** **b. Unsuccessful Performance:** Student fails to present an account of the functioning of the design solution whereby mediators transform a many-to-one relationship into a many-to-many relationship **(SF.3.a).**
Common Misconceptions*	
Additional Assessment Boundaries	

Not an exhaustive list

Appendix B: Hypothesized Learning Progressions

Systems and Systems Models

	Level 1 →	**Level 2** →	**Level 3** →	**Level 4** →	**Level 5**
SM1. Aggregate level phenomenon	**SM1.1.A.** Inaccurately identify phenomena of interest (i.e., perturbed state, equilibrium) **SM1.1.B.** Inaccurately describe phenomenon using observable features and personal experience.	**SM1.2.A.** Accurately identify relevant phenomenon of interest. May inaccurately identify other phenomena. **SM1.2.B.** Accurately describe the phenomenon using observable features and personal experience only. May be inconsistent in describing phenomenon.	**SM1.3.A.** Accurately identify relevant phenomenon of interest only. **SM1.3.B.** Accurately describe the phenomenon using observable features and one hidden mechanism. May be inconsistent in describing phenomenon.	**SM1.4.A.** Already at highest **SM1.4.B.** Accurately describe the phenomenon using observable features and several hidden mechanisms. May be inconsistent in describing the mechanisms.	**SM1.5.A.** Already at highest **SM1.5.B.** Accurately describe the phenomenon using observable features and hidden mechanisms.

SM2. Components	**SM2.1.A.** Identify irrelevant and/or inaccurate components related to the phenomenon **SM2.1.B.** No identification of sub-components **SM2.1.C.** No description of components as being static and/or changing.	**SM2.2.A.** Accurately identify some components relevant to the phenomenon. May identify irrelevant components. **SM2.2.B.** Identify sub-components but may be irrelevant and/or inaccurate. **SM2.2.C.** Describe components as being static or unchanging (initial conditions).	**SM2.3.A.** Accurately identify all components relevant to the phenomenon. **SM2.3.B.** Accurately identify all sub-component	**SM2.4.A.** Already at highest **SM2.4.B.** Already at highest **SM2.4.C.** Describe components (or initial conditions) as changing due to external and internal factors. Description include short-term and long-term changes and may be inaccurate.	**SM2.5.A.** Already at highest **SM2.5.B.** Already at highest **SM2.5.C.** Accurately describe components (or initial conditions) as changing due to external and internal factors. Description include short-term and long-term changes.

			SM2.3.C. Describe components (or initial conditions) as changing due to external factors. Description is limited to short-term change and may be inaccurate.		
SM3. Relationships between components			**SM3.3.A.** Describe causal relationships between components	**SM3.4.A.** Accurately describe causal relationships between	**SM3.5.A.** Already at highest **SM3.5.B.** Accurately mention bidirectional

		and sub-components but may be irrelevant and/or inaccurate. **SM3.3.B.** Identify unidirectional between components and sub-components but may be irrelevant and/or inaccurate **SM3.3.C.** Identify and one to one relationships but may be irrelevant and/or inaccurate **SM3.3.D.** Identify linear relationship between components and sub-components but may be irrelevant	components and sub-components. **SM3.4.B.** Accuratel y identify unidirectional between components and sub-components. **SM3.4.C.** Identify one to one relationships and one-to-many relationships but may be irrelevant and/or inaccurate. **SM3.4.D.** Accuratel y identify linear relationships between components and sub-components	and/or cyclical relationship between components and sub- components. **SM3.5.C.** Accurately identify one to one relationships and one-to-many relationships **SM3.5.D.** Accurately identify linear and non-linear relationships between components and sub- components

		and/or inaccurate		
SM4. Mechanisms (processes/ relationship) **(e.g., emergence, feedback loops, adaptation, iteration, randomness).**		**SM4.3.A.** Identify one central mechanism that explains observed phenomena. May be inaccurate. **SM4.3.B.** Identify mechanisms at one level only (e.g., local only and no aggregate). May be inaccurate. **SM4.3.C.** Does not identify random nature of mechanisms at the local or individual level **SM4.3.D.** Describe mechanisms as static, even as the components change.	**SM4.4.A.** Identify multiple mechanisms that explains observed phenomena. May be inaccurate and/or continue to identify one central mechanism. **SM4.4.B.** Identify mechanisms at multiple levels (e.g., local and aggregate). May be inaccurate. **SM4.4.C.** Identify random nature of mechanisms at the local or individual level. May be inaccurate.	**SM4.5.A.** Accurately identify multiple mechanisms that interact together to give rise to observed phenomena. **SM4.5.B.** Accurately identify mechanisms at local and aggregate levels. **SM4.5.C.** Accurately identify random individual/local mechanisms that interact across multiple levels in the system. **SM4.5.D.** Accurately describe mechanisms and components as changing dynamically in response to one another.

		SM4.3.E. Does not make any predictions about system level behaviors based on mechanisms	**SM4.4.D.** Describe mechanisms and components as changing dynamically in response to one another. May be inaccurate. **SM4.4.E.** Predictions about system level behaviors from mechanisms are linear in scope (e.g., single mechanisms as impacting system)	**SM4.5.E.** Predictions about system level behaviors from mechanisms are dynamic. (e.g., multiple interacting behaviors as impacting system)
SM5. Properties of System		**SM5.3.A.** Limited or unclear description of relevant boundaries **SM5.3.B.** System may be inappropriately categorized as open or closed	**SM5.4.A.** Detailed description of the boundaries of the system. May be inaccurate. **SM5.4.B.** System categorized as open or closed to external inputs. May	**SM5.5.A.** Accurate description of the boundaries of the system **SM5.5.B.** System accurately categorized as open or closed to external inputs. **SM5.5.C.** Understanding that

		to external inputs. **SM5.3.C.** Inaccurate understanding of equilibrium as simply meaning positive balance	include inaccuracies **SM5.4.C.** Limited understanding that mathematical equilibrium includes stable & positive population only	mathematical equilibrium includes both unstable and stable, positive and extinction of populations (i.e., chaos is the norm)

Structure and Function

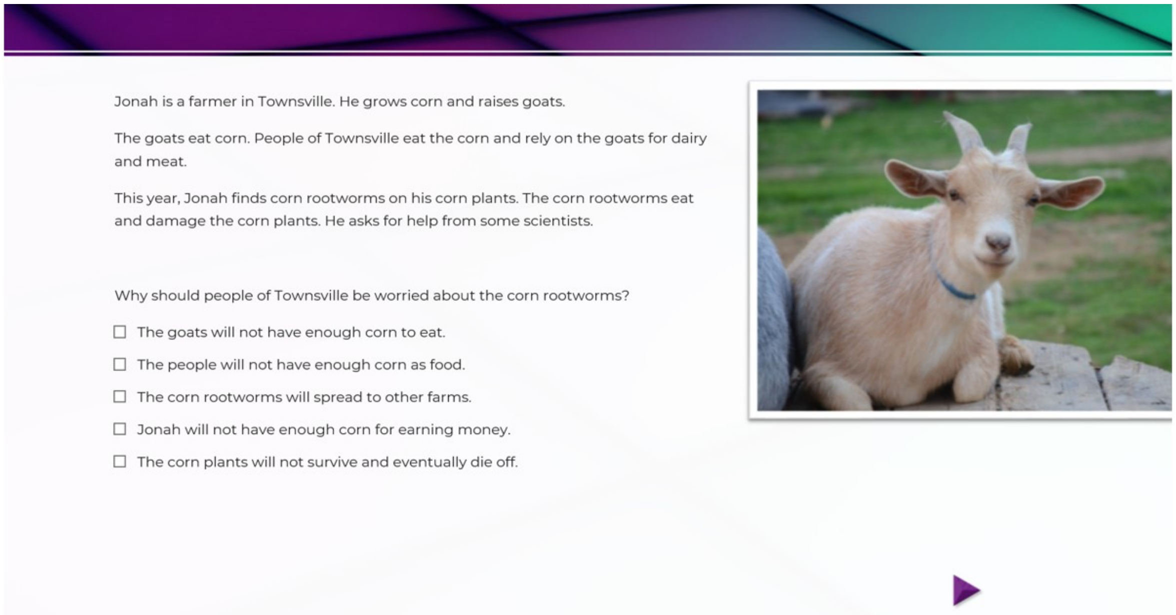
	Level 1 →	Level 2 →	Level 3 →	Level 4 →	Level 5
SF1. Structure (Components)	**SF1.1.A.** Identify a single structure but not all of its components. Description is based on familiarity or visual similarity. May be inaccurate. **SF1.1.B.** Describe structures as consisting of smaller components. May be inaccurate.	**SF1.2.A.** Accurately identify a single structure and its sub-structures **SF1.2.B.** Accurately describe structures as consisting of smaller components.	**SF1.3.A.** Accurately identify multiple structures and including sub-structures **SF1.3.B.** Already at highest	**SF1.4.A.** Already at highest **SF1.4.B.** Already at highest	**SF1.5.A.** Already at highest **SF1.5.B.** Already at highest

SF2. Function: Purpose of the structure	**SF2.1.A.** Describe function by inferring from observable conditions or surface details, but details are limited or significantly inaccurate.	**SF2.2.A.** Describe function by inferring from observable conditions or surface details, but details are limited.	**SF2.3.A.** Describe individual functioning of structures and sub-level structures but does not describe overall function as a product of sub-level conditions.	**SF2.4.A.** Describes the multiple functions of structures and sub-level structures. Infers function as a product of sub-level conditions. May be inaccurate.	**SF2.5.A.** Accurately describes the multiple functions of structures and sub-level structures. Accurately infers function as a product of sub-level conditions.
SF3. Behavior (Mechanisms): Attributes or specific states derived from the structure Examples: cell wall or membrane		**SF3.2.A.** Does not describe properties/behavior of structure **SF3.2.B.** Does not describe conditions under which the properties of the structure support its function **SF3.2.C.** Does not describe relationships between structural	**SF3.3.A.** Limited description of properties or behavior that are inferred from macro level structures only. **SF3.3.B.** Limited description of conditions under which the properties of the structure supports its function **SF3.3.C.** Limited or no description	**SF3.4.A.** Infer properties or behavior from articulating relationship among micro-macro structures and their function. May be inaccurate **SF3.4.B.** Causal description of the conditions under which the properties of the structure supports its function. May be inaccurate	**SF3.5.A.** Accurately infer properties or behavior from articulating relationship between (sub)-structures and their function. **SF3.5.B.** Accurate causal description of the conditions under which the properties supports its function. **SF3.5.C.** Describe causal relationships between structural

		components and overall function.	of causal relationships between structural components and overall function on one level only. May be inaccurate.	**SF3.4.C.** Describe causal relationships between structural components and overall function that are parallel and occur at multiple levels. May be inaccurate.	components and overall function that are parallel and integrated across multiple levels in relation to overall function of system

Appendix C: Ecosystem Task Items with Design Rationales

QUESTION 1: Why worry about the rootworm invasion?



Jonah is a farmer in Townsville. He grows corn and raises goats.

The goats eat corn. People of Townsville eat the corn and rely on the goats for dairy and meat.

This year, Jonah finds corn rootworms on his corn plants. The corn rootworms eat and damage the corn plants. He asks for help from some scientists.

Why should people of Townsville be worried about the corn rootworms?

□ The goats will not have enough corn to eat.

□ The people will not have enough corn as food.

□ The corn rootworms will spread to other farms.

□ Jonah will not have enough corn for earning money.

□ The corn plants will not survive and eventually die off.

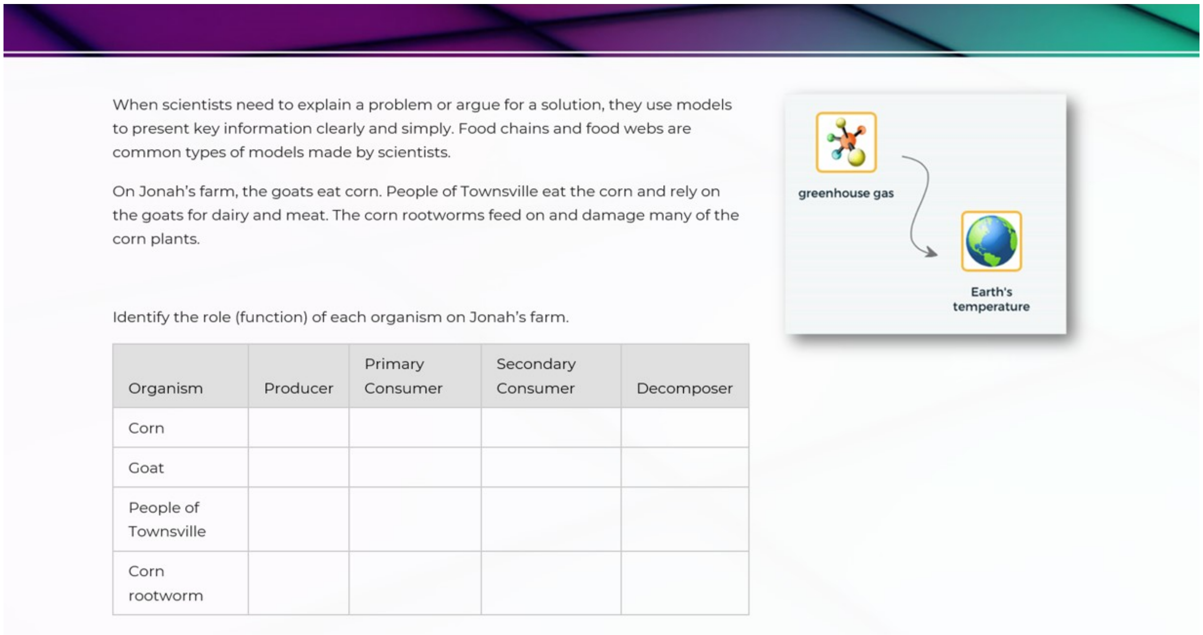
Key: Unkeyed. This item was redesigned to ask students to rank the concerns according to their salience for the people of Farmville.

Question 1: Design Rationale

- **What is the assessment item asking?** This item introduces students to the relevant entities that will be at play over the course of the whole activity while probing their understanding that disturbances to one element in a system can affect other system components.
- **What information is important?** It is important to pay attention to respondents' ability to identify likely outcomes of the rootworm invasion, particularly that *multiple* negative outcomes may occur.
- **What is the rationale associated with each possible answer? Note that rationales also exist for distractor responses.**
  1. **The goats will not have enough corn to eat. –** The corn rootworms (CRWs) directly decreases the amount of corn available. Goats eat the corn, so a reduction in corn decreases the amount they can eat.
  2. **The people will not have enough corn as food. –** The CRWs directly decreases the amount of corn available. People in Townsville eat the corn, so a reduction in corn decreases the amount they can eat.
  3. **The corn rootworms will spread to other farms. –** The CRWs are mobile and can move from farm to farm (**Note:** check respondent's thoughts about central mechanism described in the story).

4. **Jonah will not have enough corn for earning money. –** The CRWs directly decreases the amount of corn available. Jonah sells excess corn to support his family, so a reduction in corn decreases the amount he can sell.
5. **The corn plants will not survive and eventually die off. –** The CRWs consume the corn, if left unchecked, the CRWs will consume all the corn (**Note:** check respondent's thoughts about human ability to prevent collapse).

QUESTION 2: Differentiating the role of organisms on the farm



When scientists need to explain a problem or argue for a solution, they use models to present key information clearly and simply. Food chains and food webs are common types of models made by scientists.

On Jonah's farm, the goats eat corn. People of Townsville eat the corn and rely on the goats for dairy and meat. The corn rootworms feed on and damage many of the corn plants.

Identify the role (function) of each organism on Jonah's farm.

Organism	Producer	Primary Consumer	Secondary Consumer	Decomposer
Corn				
Goat				
People of Townsville				
Corn rootworm				

Key: Corn – Producer, Goat – Primary Consumer, Corn rootworm – Primary Consumer
Human – Primary Consumer *and* Secondary Consumer

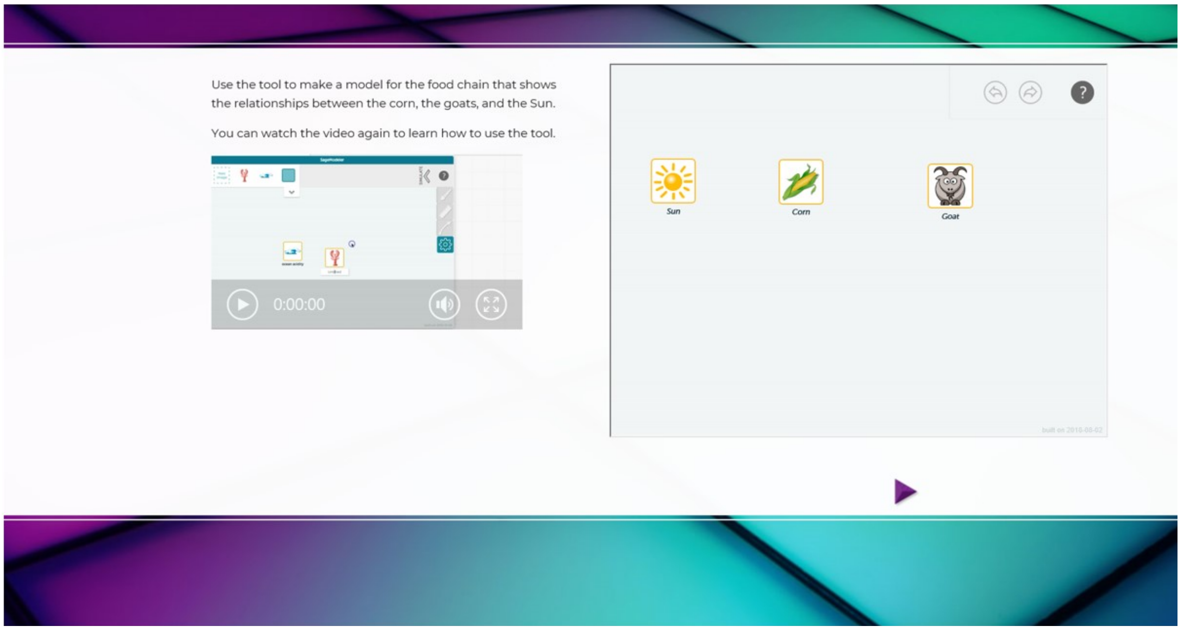
Item Design Rationale

- **What is the assessment item asking?** This item presents a variety of trophic roles organisms can take on within an ecosystem and asks respondents to correctly associate story elements (corn, goats, people, and corn rootworms [CRWs]) with those roles.
- **What information is important?** It is important to pay attention to respondents' ability to differentiate primary and secondary consumption, identify which kind of organisms are considered producers, and to see CRWs as invasive consumers rather than key decomposers.
- **What is the rationale associated with each possible answer? Note that rationales also exist for some select distractor responses. Further rationales may need to be investigated with interview probes.**

7. **The corn are producers –** Corn produces edible biomass by converting matter (carbon in the air and water from air/soil) by using energy from the sun.
8. **The goats are primary consumers. –** The goat directly consumes the corn.

9. **The people of Townsville are both primary consumers *and* secondary consumers. –** Humans consume corn both by directly eating it and by eating organisms (the goats) that eat the corn.
10. **The people of Townsville are one of either primary *or* secondary consumers -** Humans consume corn *either* by directly eating it *or* by eating organisms (the goats) that eat the corn. (**Note:** check respondent's reading of the story and whether they noted that both consumptions occur).
11. **The corn rootworms are primary consumers. –** The CRWs directly consume the corn.
12. **The corn rootworms are decomposers. –** The CRWs cause the corn to die (decompose).

QUESTION 3: Creating the baseline energy-transfer model



Use the tool to make a model for the food chain that shows the relationships between the corn, the goats, and the Sun.

You can watch the video again to learn how to use the tool.

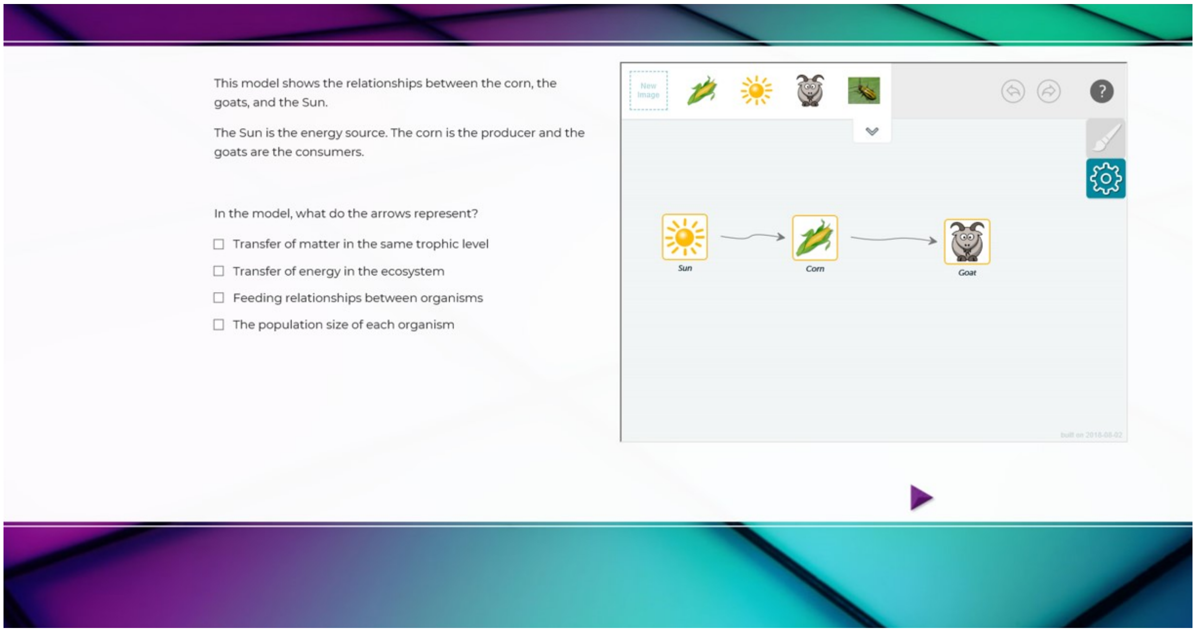
Key: Sun -> Corn; Corn -> Goat

Item Design Rationale

- **What is the assessment item asking?** This item provides a selection of elements that ought to be incorporated via the SageModeler interface.
- **What information is important?** It is important to pay attention to respondents' ability to incorporate the appropriate entities, draw arrows in appropriate direction, the order of element incorporation, and any errors made along the way.
- **What is the rationale associated with each possible answer? Note that rationales also exist for some select distractor responses. Further rationales may need to be investigated with interview probes.**

1. **Sun -> Corn -> Goat –** The sun produces the energy that supports the production of plant (corn) matter that is ultimately consumed by the goat.
2. **Goat -> Corn -> Sun –** The goats derive energy from the corn which derive energy from the sun (**Note:** check on respondent's understanding on the meaning of arrows and their directionality).
3. **Sun not in model or not connected to other model elements –** Food webs only show the flow of matter, sunlight is not matter and, therefore, should not be included in the model (**Note:** check on respondents reading of the task prompt).
4. **Goat connected via the sun –** The goats live during the day (when the sun is out); they would die if the sun was gone (reasoning may vary); the sun provides necessary thermal energy to the goats.

QUESTION 4: What do model elements represent?



This model shows the relationships between the corn, the goats, and the Sun.

The Sun is the energy source. The corn is the producer and the goats are the consumers.

In the model, what do the arrows represent?

- □ Transfer of matter in the same trophic level
- □ Transfer of energy in the ecosystem
- □ Feeding relationships between organisms
- □ The population size of each organism

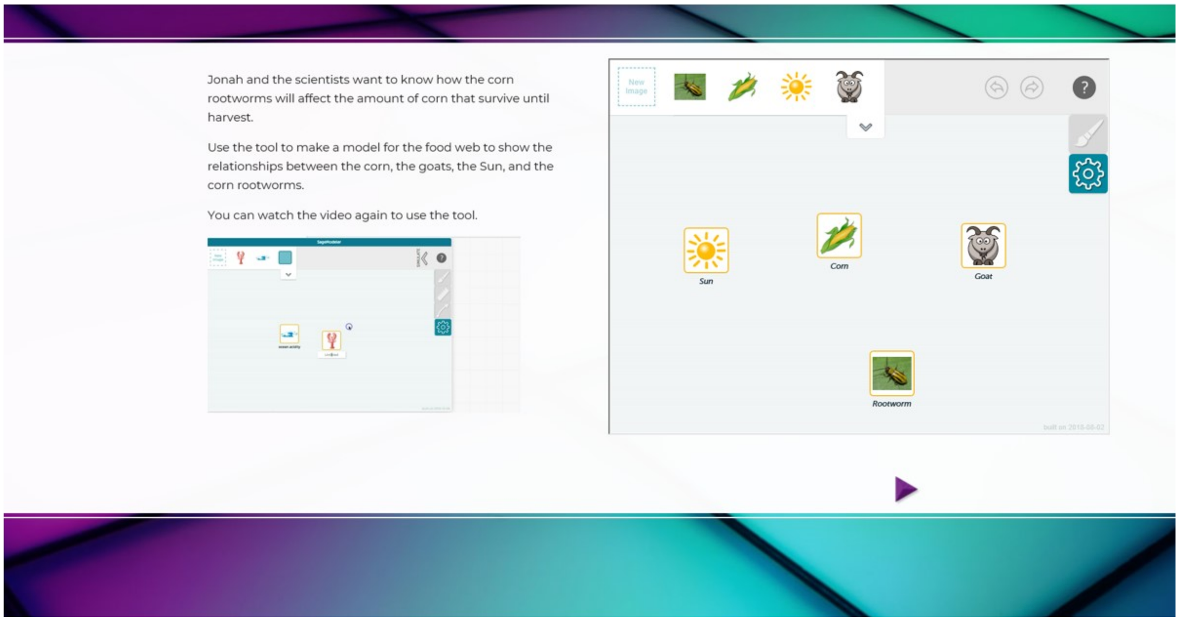
Key: 2

<u>Item Design Rationale</u>

- **What is the assessment item asking?** This item levels students from the past model-creation task and asks them to reflect on the information about the ecosystem that the model is attempting to convey.
- **What information is important?** It is important to pay attention to respondents' ability to identify the mechanisms by which model elements serve a role within the ecosystem (as represented by the arrows in the model).
- **What is the rationale associated with each possible answer? Note that rationales also exist for distractor responses.**

1. **Transfer of matter in the same trophic level. –** All the organisms are in the same level (horizontally) therefore the matter is flowing within that single level (**Note:** check respondent's understanding of what "trophic levels" are).
2. **Transfer of energy in the ecosystem. –** Energy in the ecosystem originating from the Sun moves from element at the back of the arrow to the element at the head of the arrow.
3. **Feeding relationships between organisms. –** The goats on Jonah's farm survive by feeding on the corn.
4. **The population size of each organism –** Within the ecosystem there is a sun, a corn, and a goat. (**Note:** check how students would model ecosystem if the farm contained many corn stalks or many goats).

QUESTION 5: Integrating the corn rootworm into the model



Jonah and the scientists want to know how the corn rootworms will affect the amount of corn that survive until harvest.

Use the tool to make a model for the food web to show the relationships between the corn, the goats, the Sun, and the corn rootworms.

You can watch the video again to use the tool.

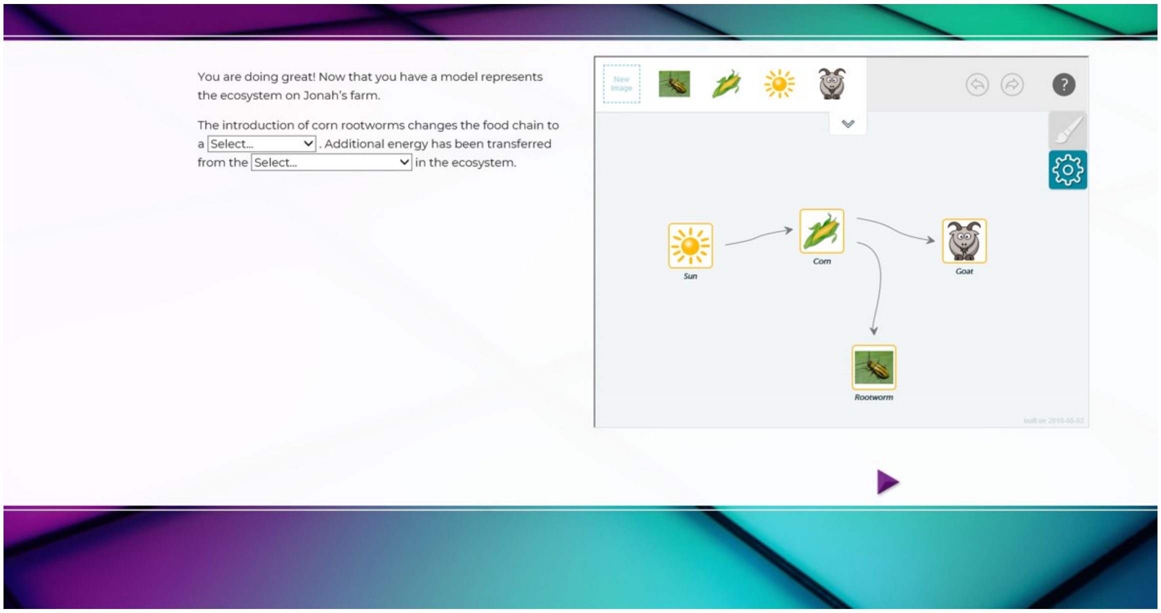
Key: Sun -> Corn; Corn -> Goat; Corn -> Rootworm

<u>Item Design Rationale</u>

- **What is the assessment item asking?** This item provides the corn rootworm as a new element to be incorporated into the previous model via the SageModeler interface.
- **What information is important?** It is important to pay attention to respondents' ability to incorporate the appropriate entities, draw arrows in appropriate direction, the order of element incorporation, and any errors made along the way.
- **What is the rationale associated with each possible answer? Note that rationales also exist for some select distractor responses. Further rationales may need to be investigated with interview probes.**
    1. **Corn -> Corn Rootworm –** Corn is a food/energy source for the rootworms.

2. **Corn Rootworm-> Corn –** The rootworms derive energy from the corn which derive energy from the sun (**Note:** check on respondent's understanding on the meaning of arrows and their directionality).
3. **Corn Rootworm connected to the goats –** The CRWs compete with the goats and cause the goat population to change (**Note:** check on respondents reading of the task prompt).

QUESTION 6: What do model elements represent?



You are doing great! Now that you have a model represents the ecosystem on Jonah's farm.

The introduction of corn rootworms changes the food chain to a Select... . Additional energy has been transferred from the Select... in the ecosystem.

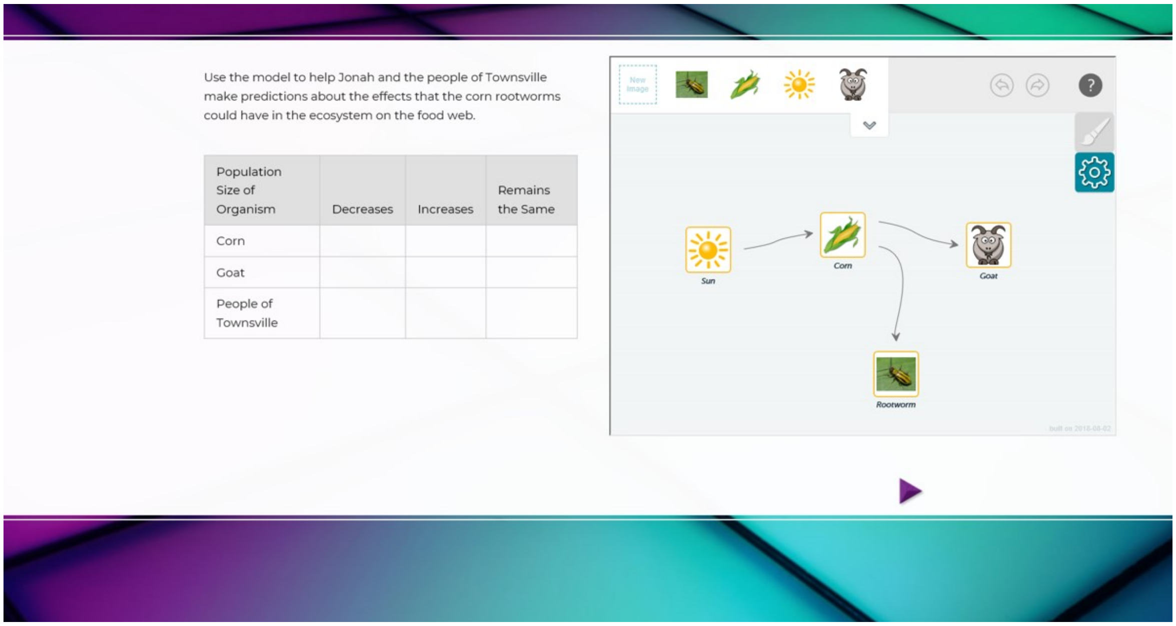
Key: 1- new food chain; **food web**
2- **producers to consumers**; consumers to producers

<u>Item Design Rationale</u>

- **What is the assessment item asking?** This item levels students from the past model-creation task and asks them to reflect on the changes to the information contained within the model as a function of the incorporation of the CRW.
- **What information is important?** It is important to pay attention to respondents' ability to relate the new role of the CRW within the ecosystem to a change in where available energy in the systems goes.
- **What is the rationale associated with each possible answer? Note that rationales also exist for distractor responses.**
  1. **New food chain. –** Previously the model showed 1 food chain (sun->corn->goat) and now it also shows another food chain (sun->corn->CRW) as well (**Note:** check student reasoning about the difference between multiple food chains and a single food web).
  2. **Food web. –** The integration of the CRWs means we have created a model in which many organisms serve as the producer for or consumer of many other organisms.

3. **Producers to consumers –** The CRWs consume the corn and derive energy produced from its growth process (photosynthesis).
4. **Consumers to producers. –** When the CRWs die, their bodies provide nutrients to the soil that corn needs to grow (i.e. "circle of life" naïve conception; **Note:** check on respondent's understanding on the meaning of arrows and their directionality).

QUESTION 7: Inferring the consequences of the invasion



Use the model to help Jonah and the people of Townsville make predictions about the effects that the corn rootworms could have in the ecosystem on the food web.

Population Size of Organism	Decreases	Increases	Remains the Same
Corn			
Goat			
People of Townsville			

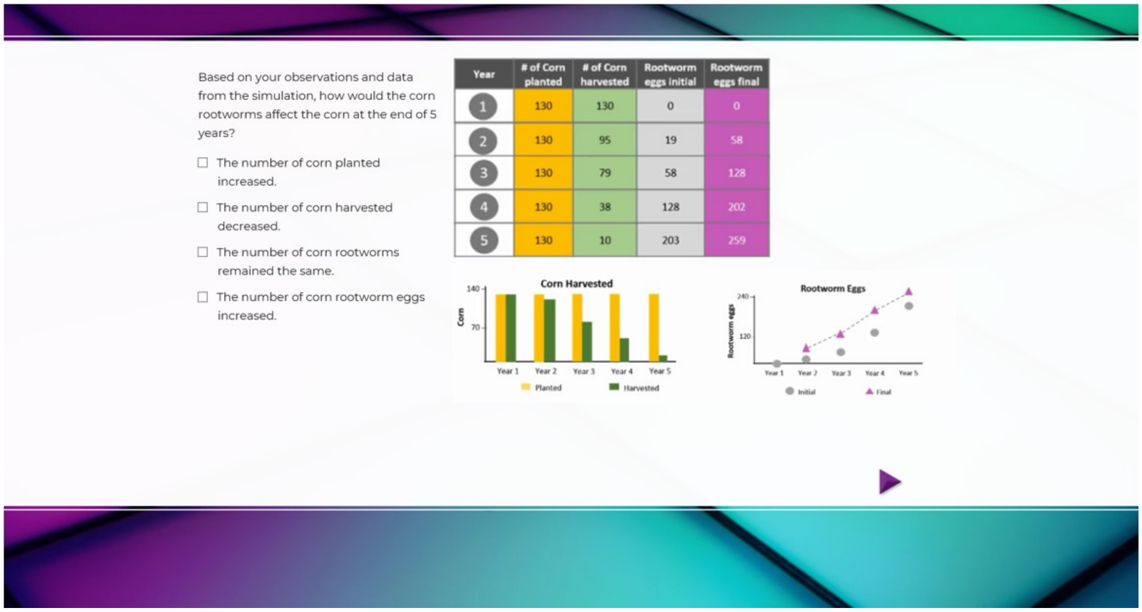
Key: Corn – decrease
Goat – decrease (if corn is the only food available in the farm)
People – remain the same

Item Design Rationale

- **What is the assessment item asking?** This item asks respondents to make predictions about changes to populations (both seen and unseen) living in the ecosystem as a result of the CRW invasion.
- **What information is important?** It is important to pay attention to respondents' ability to identify likely outcomes of the rootworm invasion, particularly that *multiple* negative outcomes may occur. They should be able to back these predictions with a statement about evidence available in the model.
- **What is the rationale associated with each possible answer? Note that rationales also exist for some select distractor responses. Further rationales may need to be investigated with interview probes.**
    1. **Corn population decreases. –** The corn rootworms (CRWs) consume the corn, directly reducing corn population numbers.

2. **Corn population increases –** When the CRWs die, their bodies provide nutrients to the soil that corn needs to grow (i.e. “circle of life” naïve conception; **Note:** check on student conceptions about timeline displayed in the model – that the food web is a snapshot in time).
3. **Corn population stays the same –** The corn draws its energy from the sun, so new consumers don’t affect how much food the corn gets (or by extension the corn population supported by the ecosystem; **Note:** check respondent’s view of the connection between energy acquisition and matter consumption depicted in the model).
4. **Goat population decreases –** The invasion of the CRWs means there is a new competitor for the corn, decreasing the amount of energy available to support the current goat population, and resulting in a goat population decrease.
5. **Goat population stays the same –** While the CRWs consume corn, goats are mobile and can consume other plants with the aid of farmer Jonah (**Note:** check respondent’s thoughts about central mechanism described in the story).
6. **Population of people decreases. –** The CRWs directly decreases the amount of corn available. Jonah and his family eat the corn, so a reduction in corn decreases the amount he and his family can consume (**Note:** check respondent’s thoughts about human ability to prevent collapse).
7. **Population of people increases** – Human populations increase because, on average, more children are born than die (**Note:** check respondent’s ability to connect claim made about population in information contained within the model)
8. **Population of people stays the same. –** Unlike the goats, humans can readily acquire energy from new sources and are not reliant on eating corn (hence why they were not necessary to include within the food chain/web).

QUESTION 8: Understanding the consequences of the invasion



Based on your observations and data from the simulation, how would the corn rootworms affect the corn at the end of 5 years?

☐ The number of corn planted increased.

☐ The number of corn harvested decreased.

☐ The number of corn rootworms remained the same.

☐ The number of corn rootworm eggs increased.

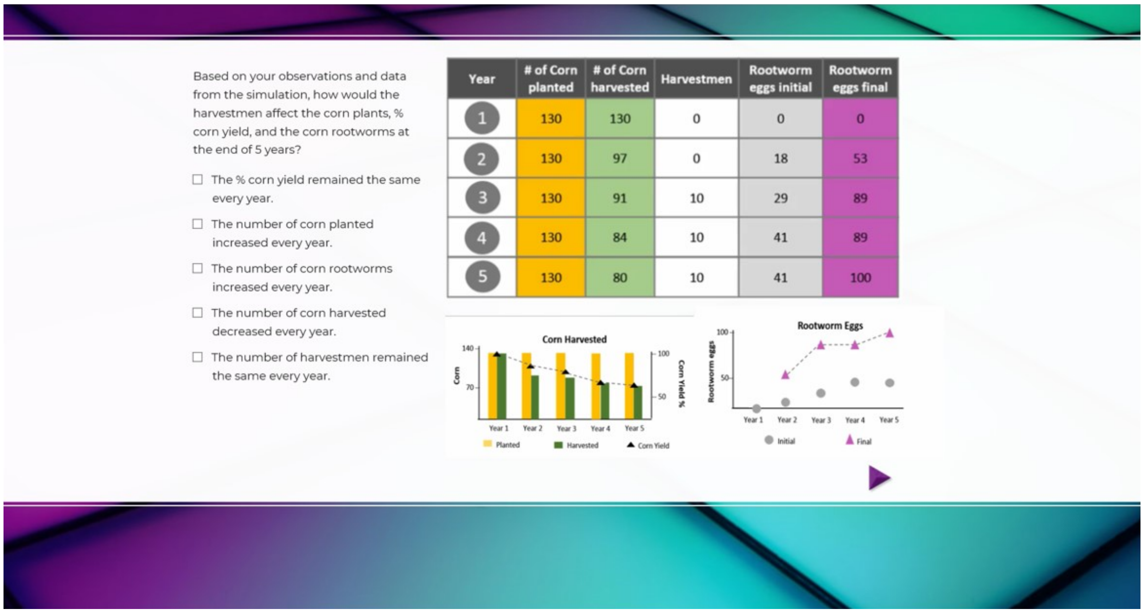
Year	# of Corn planted	# of Corn harvested	Rootworm eggs initial	Rootworm eggs final
1	130	130	0	0
2	130	95	19	58
3	130	79	58	128
4	130	38	128	202
5	130	10	203	259

Key: 2 & 4

<u>Item Design Rationale</u>

- **What is the assessment item asking?** This item provides a variety of data displays from the results of the CRW-invasion simulation and asks students to reason about the consequences of not implementing any control strategies.
- **What information is important?** It is important to pay attention to respondents' ability to link claims made about the impacts of the CRW invasion with evidence from the data displays and reasoning about the mechanisms at play during the invasion.
- **What is the rationale associated with each possible answer? Note that rationales also exist for distractor responses.**
  1. **The number of corn planted increased. –** Since the corn rootworms (CRWs) directly decreases the amount of corn that survives, the farmer Jonah should plant more so that enough survives the season (**Note:** check respondent's ability to connect claim made about population in evidence contained in the data display)
  2. **The number of corn harvested decreased. –** Each year the number of corn that survives until harvest decreases per the data in the green column.
  3. **The corn rootworms remained the same. –** The corn can only support a fixed number of CRWs, so their population will be the same over time (**Note:** check respondent's ability to connect claim made about population in evidence contained in the data display).
  4. The number of corn rootworm eggs increased. – As CRWs survive the season, they are able to lay more than 1 egg per rootworm. This causes the amount of CRW eggs to rapidly increase.

QUESTION 9: Determining the efficacy of harvestmen predators pt.1



Based on your observations and data from the simulation, how would the harvestmen affect the corn plants, % corn yield, and the corn rootworms at the end of 5 years?

□ The % corn yield remained the same every year.

□ The number of corn planted increased every year.

□ The number of corn rootworms increased every year.

□ The number of corn harvested decreased every year.

□ The number of harvestmen remained the same every year.

Year	# of Corn planted	# of Corn harvested	Harvestmen	Rootworm eggs initial	Rootworm eggs final
1	130	130	0	0	0
2	130	97	0	18	53
3	130	91	10	29	89
4	130	84	10	41	89
5	130	80	10	41	100

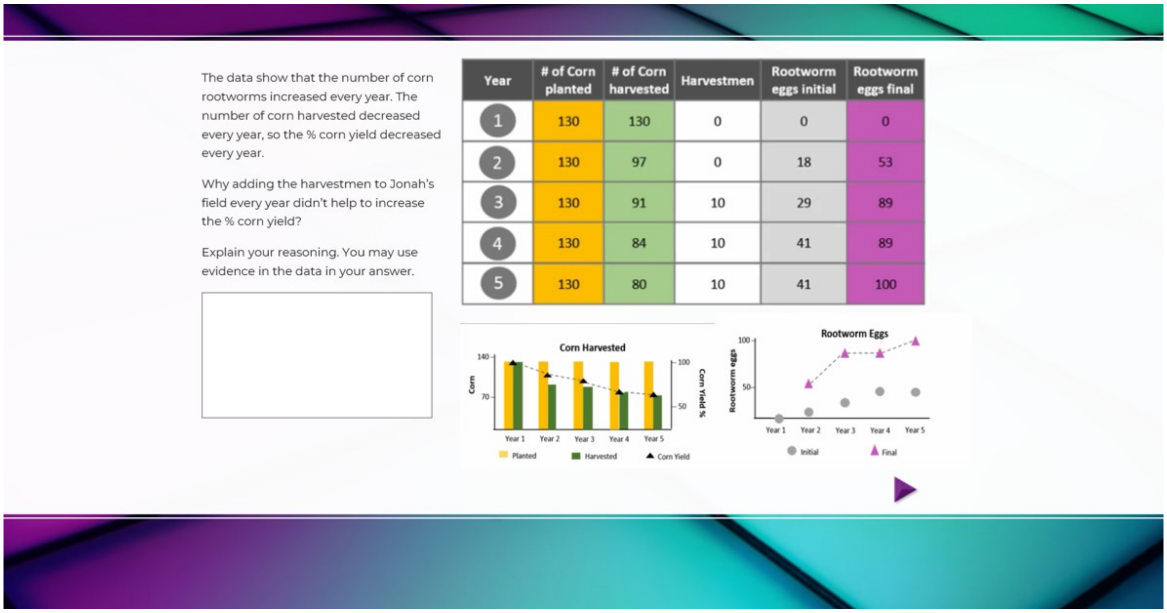
Key: 3, 4, & 5

Item Design Rationale

- **What is the assessment item asking?** This item provides a variety of data displays from the results of the CRW-invasion simulation when the harvestmen-based control strategy was implemented and asks students to reason about the changes that occurred to farm populations over time.
- **What information is important?** It is important to pay attention to respondents' ability to link claims made about the impacts of the CRW invasion with evidence from the data displays and reasoning about the mechanisms at play during the invasion.
- **What is the rationale associated with each possible answer? Note that rationales also exist for distractor responses.**
    1. **The % corn yield remained the same every year. –** Implementing the control strategy decreases the growth rate of the corn rootworms (CRWs) population, improving corn yield percentage (**Note:** check respondent's ability to connect claim made about population in evidence contained in the data display).
    2. **The number of corn planted increased every year–** Since the corn rootworms (CRWs) directly decreases the amount of corn that survives, the farmer Jonah should plant more so that enough survives the season (**Note:** check respondent's ability to connect claim made about population in evidence contained in the data display).
    3. **The number of corn rootworms increased every year. –** Despite implementing the control strategy, the CRW population (as depicted in the pink column) continued to increase.
    4. **The number of corn harvested decreased every year. – –** Each year the number of corn that survives until harvest decreases per the data in the green column.

5. **The number of harvestmen remained the same every year. –** The harvestmen are not native organisms to the farm and can only survive while there are rootworms available to eat. During the winter, they die off and Jonah must release a new set of 10 each spring.

QUESTION 10: Determining the efficacy of harvestmen predators pt.2



The data show that the number of corn rootworms increased every year. The number of corn harvested decreased every year, so the % corn yield decreased every year.

Why adding the harvestmen to Jonah's field every year didn't help to increase the % corn yield?

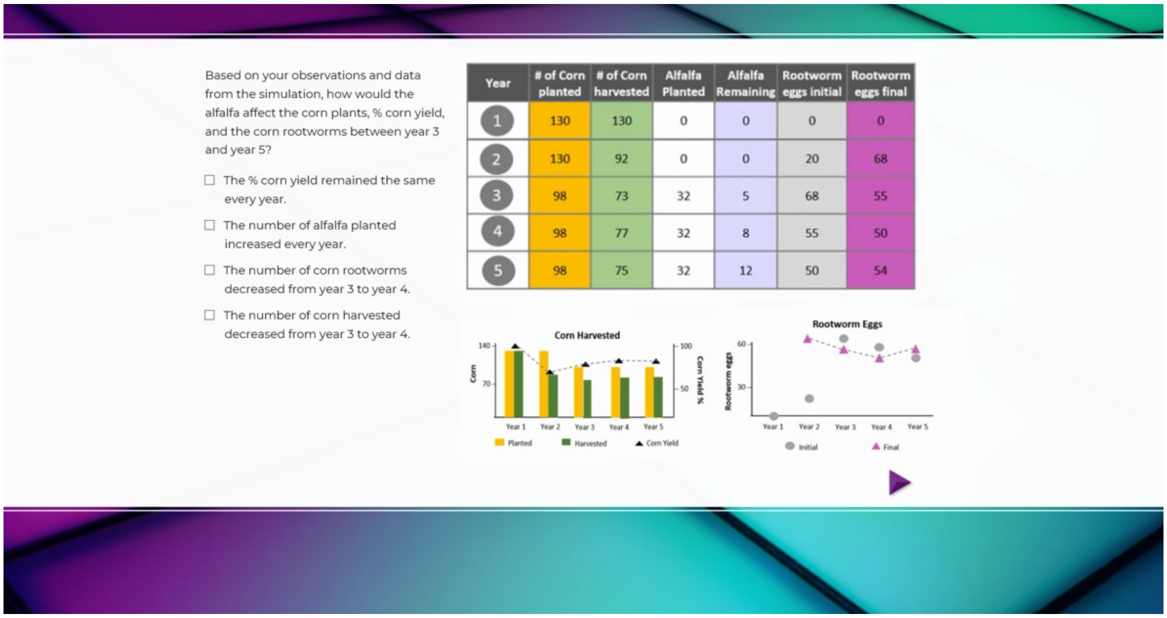
Explain your reasoning. You may use evidence in the data in your answer.

Year	# of Corn planted	# of Corn harvested	Harvestmen	Rootworm eggs initial	Rootworm eggs final
1	130	130	0	0	0
2	130	97	0	18	53
3	130	91	10	29	89
4	130	84	10	41	89
5	130	80	10	41	100

Item Design Rationale

- **What is the assessment item asking?** This item levels students so they recognize relevant patterns in population numbers following the implementation of the harvestmen-based control strategy and asks them to provide some reasoning for why the strategy was ineffective.
- **What information is important?** It is important to pay attention to respondents' ability to the criteria students generate for determining what counts as effective solutions (students might note that the harvestmen strategy is still better than nothing). What elements of the system (and the simulation) do students focus on? Do they tie back their responses to the not-simulated goat and human populations?
- What is the rationale associated with each possible answer? Note that rationales also exist for some select naïve responses. Further rationales may need to be investigated with interview probes.
    1. Answer focuses on insufficient number of harvestmen added each year.
    2. Answer focuses on the strategy still being more effective than doing nothing at all.
    3. Answer focuses on the available amount of energy to support the CRWs.
    4. Answer focuses on visual elements from the video (i.e. "the harvestmen didn't interact with the bugs enough")

QUESTION 11: Determining the efficacy of the trap crop pt.1



Based on your observations and data from the simulation, how would the alfalfa affect the corn plants, % corn yield, and the corn rootworms between year 3 and year 5?

☐ The % corn yield remained the same every year.

☐ The number of alfalfa planted increased every year.

☐ The number of corn rootworms decreased from year 3 to year 4.

☐ The number of corn harvested decreased from year 3 to year 4.

Year	# of Corn planted	# of Corn harvested	Alfalfa Planted	Alfalfa Remaining	Rootworm eggs initial	Rootworm eggs final
1	130	130	0	0	0	0
2	130	92	0	0	20	68
3	98	73	32	5	68	55
4	98	77	32	8	55	50
5	98	75	32	12	50	54

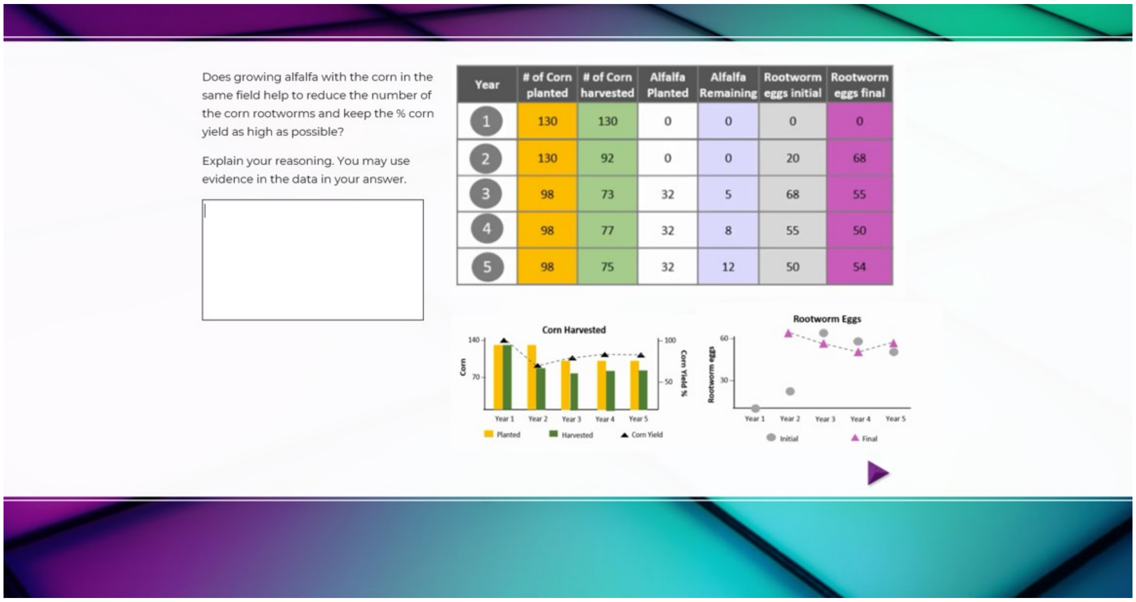
Key: 1 & 3

Item Design Rationale

- **What is the assessment item asking?** This item provides a variety of data displays from the results of the CRW-invasion simulation when the trap-crop-based control strategy was implemented and asks students to reason about the changes that occurred to farm populations over time.
- **What information is important?** It is important to pay attention to respondents' ability to link claims made about the impacts of the CRW invasion with evidence from the data displays and reasoning about the mechanisms at play during the invasion.
- **What is the rationale associated with each possible answer? Note that rationales also exist for distractor responses.**
    1. **The % corn yield remained the same every year. –** Implementing the trap crop distracts the CRWs with less nutritious non-corn alternatives, allowing the corn to grow to maturity before destruction by the CRWs (as depicted in the green column) while also limiting the CRW population.
    2. **The number of alfalfa-trap-crop planted increased every year. –** By planting lots of alfalfa, but not harvesting it for the goats or humans, its population can increase (**Note:** check that student can identify relevant data columns).
    3. **The number of corn rootworms decreased from year 3 to year 4. –** The CRWs do not receive enough nutrition from the trap crop, inhibiting them from laying enough eggs to replace the incident CRW population.
    4. **The number of corn harvested decreased from year 3 to year 4. –** The CRWs still continue to consume the corn, even if they also consume the alfalfa. Coupled with the decrease in corn planted by replacing some land with the non-corn trap

crop, means that corn to harvest may decrease (**Note:** check that student can identify relevant data columns in relevant years).

QUESTION 12: Determining the efficacy of trap crop pt.2



Does growing alfalfa with the corn in the same field help to reduce the number of the corn rootworms and keep the % corn yield as high as possible?

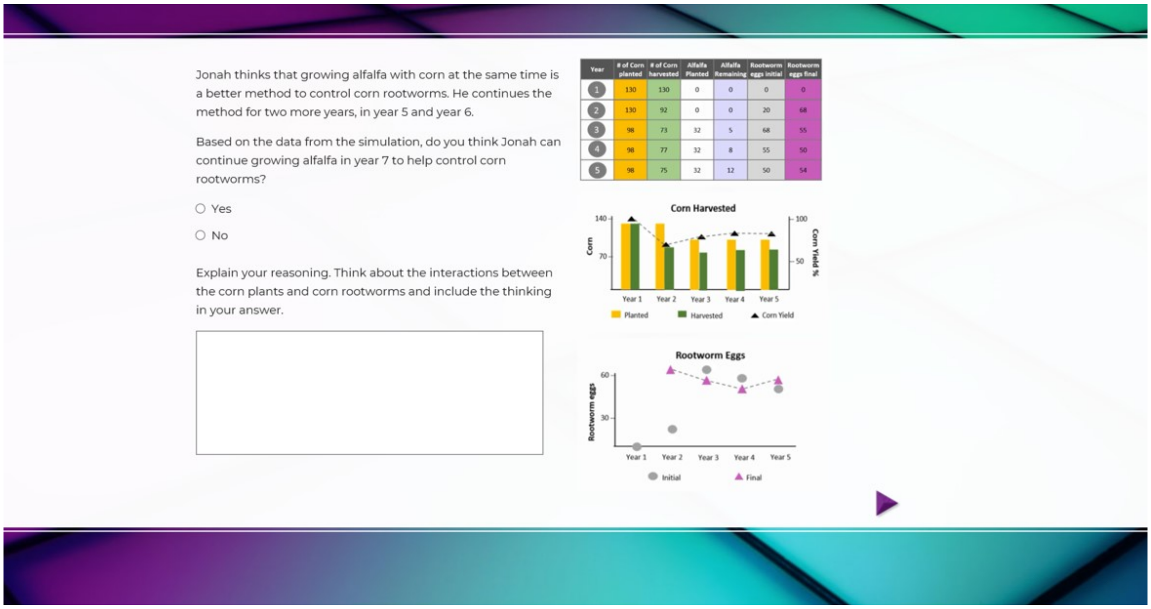
Explain your reasoning. You may use evidence in the data in your answer.

Year	# of Corn planted	# of Corn harvested	Alfalfa Planted	Alfalfa Remaining	Rootworm eggs initial	Rootworm eggs final
1	130	130	0	0	0	0
2	130	92	0	0	20	68
3	98	73	32	5	68	55
4	98	77	32	8	55	50
5	98	75	32	12	50	54

<u>**Item Design Rationale**</u>

- **What is the assessment item asking?** This asks students to generate an argument about the efficacy of the trap-crop-based control strategy and asks them to provide some reasoning for why the strategy was or was not effective.
- **What information is important?** It is important to pay attention to respondents' ability to the criteria students generate for determining what counts as effective solutions (students might note that the trap crop strategy entails an immediate reduction in the amount of corn planted/harvested). What elements of the system (and the simulation) do students focus on? Do they tie back their responses to the not-simulated goat and human populations?
- **What is the rationale associated with each possible answer? Note that rationales also exist for some select naïve responses. Further rationales may need to be investigated with interview probes.**
    1. **Answer focuses on short term reduction, concluding the strategy was ineffective**
    2. **Answer conceded short term reduction but also long-term stabilization.**
    3. **Answer focuses on the strategy still being more effective than doing nothing at all.**
    4. **Answer focuses on the available amount of energy to support the CRWs.**
    5. **Answer focuses on visual elements from the video (i.e. "the alfalfa takes up too much space")**

QUESTION 13: Drawing a conclusion about the strategies



Jonah thinks that growing alfalfa with corn at the same time is a better method to control corn rootworms. He continues the method for two more years, in year 5 and year 6.

Based on the data from the simulation, do you think Jonah can continue growing alfalfa in year 7 to help control corn rootworms?

○ Yes

○ No

Explain your reasoning. Think about the interactions between the corn plants and corn rootworms and include the thinking in your answer.

Year	# of Corn planted	# of Corn harvested	Alfalfa Planted	Alfalfa Remaining	Rootworm eggs initial	Rootworm eggs final
1	130	130	0	0	0	0
2	130	92	0	0	20	68
3	98	73	32	5	68	55
4	98	77	32	8	55	50
5	98	75	32	12	50	54

<u>Item Design Rationale</u>

- **What is the assessment item asking? What is the assessment item asking?** This item levels students so they recognize relevant patterns in population numbers following the implementation of the trap-crop-based control strategy and asks them to provide some reasoning for why the strategy was effective (particularly several years after implementation began).
- **What information is important?** It is important to pay attention to respondents' ability to extend the reasoning and argument from Q12.
- **What is the rationale associated with each possible answer? Note that rationales also exist for some select naïve responses. Further rationales may need to be investigated with interview probes.**
  6. **Answer focuses on short term reduction, concluding the strategy was ineffective**
  7. **Answer conceded short term reduction but also long-term stabilization.**
  8. **Answer focuses on the strategy still being more effective than doing nothing at all.**
  9. **Answer focuses on the available amount of energy to support the CRWs.**
  10. **Answer focuses on visual elements from the video (i.e. "the alfalfa takes up too much space")**

Appendix D: Post-Task Interview Protocol

Item 2 (optional questions- ask only if time allows)

- How does identifying the role of CORN, GOATS and CORN ROOTWORM help scientists think about the impacts of the corn rootworm?

Item 3 (Required)

- Why did you add element X (sun, corn, goat) into the model? How did you connect it to element Y?
- Why did you leave element X (sun, corn, goat) out of the model? How did you connect it to element Y?
- What do the arrows in the model represent? How did you decide to place them in the direction you did (indicate direction based on notes)?
- Relevant SageModeler usability questions.
  - How easy was using with the modeling tool?
  - Was the tutorial video helpful to teach you what to do?
  - Was it clear how to fix a mistake (if made)?
  - Any suggestions for improvement?

ITEM 5 (Required)

- How did you decide to connect the rootworms to element Y? (Alternative: Why did you chose not to connect the rootworms to any of the other model elements?)
- How did adding the rootworm to the model help the scientists think about the impacts of the invasion on the farm ecosystem?

ITEM7 (Required)

- How might the invasion of the corn rootworms have the stated effect? (Interviewer should remind the students of the responses they selected)
- Did you use the model to make your hypothesis? If so, how was the model helpful?

VIDEO: Examining the Impact of the Rootworm Invasion (Required)

- Was the video useful to help you understand the relationships in the system?
- Based on the video, what was the purpose of the simulation? What did you learn from the video?

ITEM 8 (Required)

- Usability questions regarding the data table and graphs.
  - What did different columns in the data table represent?
  - For each graph, what patterns did you see? What is the relationship between the graph and the data table?
  - What did we mean by yield percentage? Why is yield percentage a useful measure of Jonah's Farm?

VIDEO: Enacting the Harvestmen Control Strategy (Required)

- Was the video useful to help you understand how Harvestmen Control Strategy works?
- Based on the video, what was the purpose of the simulation? What did you learn from the video?

ITEM 10 (Required)

- Why were only 10 harvestmen released per year on the farm?
- What other data might have been helpful in drawing a conclusion about the efficacy of the harvestmen-based control strategy?

VIDEO: Enacting the Alfalfa Trap Crop Control Strategy (Required)

- Was the video useful to help you understand how the Alfalfa Trap Crop Control Strategy works?
- Based on the video, what was the purpose of the simulation? What did you learn from the video?

ITEM 12 (Required)

- Why did planting the Alfalfa decrease the amount of corn planted?
- What other data might have been helpful in drawing a conclusion about the efficacy of the trap-crop-based control strategy?

ITEM 13 (Required)

- What information about the efficacy of the strategy was added by letting the simulation run for several more years?
- How did you use the videos of the simulated control strategies to help you respond to relevant questions?

Overall Task Feedback Questions

- Did you feel like you had all the information needed to engage with the task? What other information might have been helpful?
- Were the experiences of this task similar to tasks you had already engaged in within your science classroom?
- Do you think the farm scenario was meaningful to you? If not, what types of scenarios will be more engaging?
- Were you hoping that you could have direct interactions with the simulation shown in the video? If you were able to interact with the simulation shown in the video, what tests would you like to run?

[End of Semi-Structured Protocol]

[If time, interview asks for task feedback or about interesting observations made during think aloud]

Appendix E: DAT-CROSS: Background Questionnaire

Before we get started, please answer the following questions.

1. What grade are you currently in?
   ○ 6th grade
   ○ 8th grade
   ○ 10th grade

2. How old are you? _*(text entry)*_

3. What is your gender?
   ○ Male
   ○ Female
   ○ Other__________
   ○ Prefer not to answer

4. Which of the following best describes you? (Select all that apply.)

   □ African American

   □ White

   □ Asian

   □ Hispanic

   □ Pacific Islander

   □ Native American

   □ Mixed Race

   □ Other_______________

   □ Prefer not to answer

5. Have you received any instruction about **systems**? Explain and use examples if necessary

6. Have you received any instruction about **system models**? Explain and use examples if necessary

7. Have you received any instruction about **structure and function**?
8. Do you have any experience using online or virtual simulations in school?

Appendix F: Data Collection Matrix

Data Collection	Analysis Methods
• Background information questionnaire (See Appendix E) • Assessment items that probe students to be more explicit in their CCC reasoning but are not expected to fit into an actual teaching setting. ○ Mixture of Multiple Choice/Multiple Select and Free Response Items that examine a CCC across DCIs from several domains. (See Appendix C) • In-situ think aloud with students that provide them greater opportunity to vocalize their reasoning. This will allow us to form some validity in our assessment rubric. (See Appendix H) • Post-task reflective interviews that aloud students to discuss challenges working with the task, amend their prior thinking, and recommend changes. (See Appendix D).	• Rubric for questionnaire that can distinguish between differing levels of sophistication in CCC reasoning. • Coded interviews to ensure that the rubric aligns to the ways students are thinking about the assessment task. • Using R to generate Rasch model to confirm construct leveling in students CCC reasoning. • Structural equation to determine if different constructs are divergent.

Appendix G: Etic Coding Scheme for Interview Data

The etic scheme for coding draws from prior literature into the nature of CCCs (Rivet et al., 2016; Weiser et al., 2017) as well as into known avenues of progression in system thinking (Breslyn et al., 2016; Gunckel et al., 2012; Jin & Anderson, 2012: Mohan et al., 2009; Songer et al., 2009).

- CCCs as Lenses – The role of the CCC is to highlight salient features that may not be immediately obvious due to scale, scope, or size.
- CCCs as Bridges – The role of the CCC is to draw connections between two entities, facilitating transfer of understanding.
- CCCs as Levers – The role of the CCC is to combine several, related ideas, entities, or representations in order to make understanding take on a new form towards a particular goal.
- CCCs as Rules – The role of the CCC is to validate a representation's utility in explaining or predicting the natural world.
- System Phenomena – As students build sophistication in their thinking around systems, they are better able to describe phenomena as a system of many simultaneous interactions.
- System Components – As students build sophistication in their thinking around systems, they are better able to break down the components of a system into their constituent, dynamic parts.
- System Relationships – As students build sophistication in their thinking around systems, they are better able to describe the relationships between previously identified components of the system.
- System Boundaries – As students build sophistication in their thinking around systems, they are better able to define the boundaries of the system and track the flows of inputs and outputs across those boundaries.

Appendix H: Rubric for Assessing Constructed Responses

Overall Guidance on Using Rubric

- Each CR item should be scored from 0 to 3 holistically. That is, while guidance is provided to assess the items in terms of the claim, evidence, and reasoning of their argument, scores should correspond to the overall quality of the performance.
- At each score level for each item in this rubric you will see a conceptual definition (in black), guidance for what elements must appear in order to be scored at least as high as the current level (in red), and an example as provided by a student (in blue).
- Due to limitations on the sophistication of the task, it is not reasonable to expect students to display a broad range of claims within their arguments. Score levels 2 and 3 may have similar criteria for the quality of argumentative claims made by students in their overall response. The distinction between a score of 2 or of 3 rests on the sophistication of the evidence and reasoning provided as well as the lack incorrect or inaccurate statements.

Question 10: Determining the efficacy of harvestmen predators pt.2

Score Level	Scoring Criteria
3 (This level broadly aligns to level 4 elements on the learning progression)	Student response demonstrates complex understanding of systems and system modeling within the three main parts of their argument: the claim, the evidence cited, and their reasoning. • Claim (about system phenomena, SM1): response shows that the student can accurately describe efficacy of the control mechanism using at least one unseen/hidden/underlying mechanism (**SM1.3.B**). ○ Student responses at this level should claim that the failure of the harvestmen was a result of at least one of the following aspects of the system: ▪ There was a finite number of harvestmen released, which was insufficient to consume the pre-existing number of rootworms. ▪ There was not enough time in a season for the harvestmen to consume enough rootworms. ○ Additional claims comparing the efficacy of the harvestmen to the baseline (in which no strategy is implemented) are permitted even though they are not part of the question prompt. • Evidence (from system components or relationships, SM2/3): response cite from changes to two or more system components that are accurately related to the use of the relevant control strategy and the degree to which those components changed (**SM2.4.C+SM3.4.A**). ○ Student must cite evidence related to all three of the number of harvestmen, the number of rootworms, and the amount of corn that survives the season.

	o Response should include a statement of *rate* (i.e. how fast the rootworms are procreating and/or consuming corn). o Student may also include a statement about *rootworm survival* – that some rootworms survive the season without being eaten by the harvestmen. o Additional evidence from the baseline (in which no strategy is implemented) can be cited without penalty. • Reasoning (from system relationships and mechanisms, SM3/4): response provides a relevant mechanism related to a causal chain between the harvestmen, rootworms, and the amount of corn harvested (**SM4.5.D**). All three components must be addressed. Reasoning provided must be relevant to the claims students make and the evidence they have cited. o Student response must make reference to the following relationships: ▪ The harvestmen consume some of the rootworms, but some rootworms survive the season (student may but do not have to reference laying eggs). ▪ If sufficiently many rootworms survive the season (to lay eggs for next year), then their population can continue to grow even as the harvestmen attempt to consume them. ▪ As the rootworm population grows, the amount of corn that survives until the harvest decreases. o Response should include a statement of *rate attribution* – attributing the rate (or changes to the rate) of one change in relevant components (the corn yield and/or the rootworm population) to the behavior of other system components. **From Datusability40:** There might not have been enough harvestmen to keep the population of corn rootworms under control, so even though the population of corn rootworms did not go up as rapidly as before, as shown in the graph data, the population went up nevertheless, which also means that the number of corn harvested still went down.
2 (This level broadly aligns to level 3 elements on the learning progression)	Student response demonstrates acceptable understanding of systems and system modeling within the three main parts of their argument: the claim, the evidence cited, and their reasoning. The student response includes some incorrect or inappropriate elements but is still sophisticated. • Claim (about system phenomena, SM1): response shows that the student can accurately describe efficacy of the control mechanism using at least one unseen/hidden/underlying mechanism (**SM1.3.B**). o Student responses at this level should claim that the failure of the harvestmen was a result of at least one of the following aspects of the system: ▪ There was a finite number of harvestmen released, which was insufficient to consume the pre-existing number of rootworms. ▪ There was not enough time in a season for the harvestmen to consume enough rootworms.

	○ Additional claims comparing the efficacy of the harvestmen to the baseline (in which no strategy is implemented) are permitted even though they are not part of the question prompt. • Evidence (from system components or relationships, SM2/3): response cite from changes to two or more system components but does not provide a relationship between those evidences and a relevant control mechanism (**SM2.3.C**). ○ Student must cite evidence related to at least two of the number of harvestmen, the number of rootworms, and the amount of corn that survives the season. ○ Student may also include a statement about *rootworm survival* – that some rootworms survive the season without being eaten by the harvestmen. ○ Additional evidence from the baseline (in which no strategy is implemented) can be cited without penalty. ○ Some stated evidence can be inaccurate or inappropriate, provided that sufficiently much accurate/appropriate evidence is also cited in support of the claim. • Reasoning (from system relationships and mechanisms, SM3/4): response provides a relevant mechanism related to a causal chain between the harvestmen, rootworms, and the amount of corn harvested. All three components must be addressed. Response includes a mixture of both appropriate and inappropriate system properties (**SM3.4.A+SM4.4.D)**. ○ Student response must make some reference to the following relationships: ▪ The harvestmen consume some of the rootworms, but some rootworms survive the season (student may but do not have to reference laying eggs). ▪ If sufficiently many rootworms survive the season (to lay eggs for next year), then their population can continue to grow even as the harvestmen attempt to consume them. ▪ As the rootworm population grows, the amount of corn that survives until the harvest decreases. ○ Response includes references to appropriate relationships but fails to accurately attribute changes in the number of rootworms or amount of corn harvested to those relationships. **From Datusability42:** I think adding the harvestmen to the field for 5 years didnt {sic} help because {sic} if you look the harvestmen population only hit 10 in year 3 while the rootworm population was starting at year 2 with an initial population of 18 and a final population of 53. This shows that the population of rootworms was always greater than the population of harvestmen from the start making it harder for the harvestmen to kill off any rootworms.
1 (This level broadly aligns to level 2 elements on	Student response demonstrates insufficient understanding of systems and system modeling within the three main parts of their argument: the claim, the evidence cited, and their reasoning. Response may be broadly inaccurate, but still articulate the existence of multiple system components (rootworms, rootworm eggs, corn, harvestmen, etc.) that may have been relevant during the simulation.

the learning progression)	• Claim (about system phenomena, SM1): response shows that the student inaccurately describes efficacy of the control mechanism **(SM1.2.B)**. ○ Student response directly answers question but does not refer to any aspects/features/behavior of the harvestmen. ▪ May include statements like “Adding harvestmen to the field did not help increase the corn yield percentage” • Evidence (from system components or relationships, SM2/3): response cites from changes to only a single system component (**SM2.2.A**). ○ Response only cites from changes in the amount of corn harvested as indicative of the efficacy of the harvestmen **OR r**esponse only cites from changes in the population of rootworms as indicative of the efficacy of the harvestmen. ○ Student may refer to *rootworm survival* but only as evidence of lack of efficacy, not as evidence in support of a relationship (i.e. does not connect survival to future growth of rootworm population). • Reasoning (from system relationships and mechanisms, SM3/4): response provides a relevant mechanism related to a single relationship between any two of the harvestmen, rootworms, and the amount of corn harvested but doesn’t address all components (**SM3.3.B)**. ○ Response only describes a relationship between the harvestmen and the rootworms **OR** response only describes a relationship between the rootworms and the corn. **From Datusability9:** The number of corn did not increase when the harvestmen were put into the field. This is becuse {sic} there were still many rootworm eggs in the end of the year, and they can't stop the rootworms completaly {sic}.
0 (This level broadly aligns to level 1 elements on the learning progression)	Student response demonstrates little or no understanding of systems and their response is missing argumentative elements. They may provide a simple description of the video or restate information that was previously provided to them Or Student response is off topic or inappropriate. This score may also be assigned for any response that fails to meet minimum expectations for a level 1 score. • Student response indirectly answers question but does not include a “because” statement **From Datusability7:** In the video simulater {sic} the havestmen {sic} ate both the rootworms anf {sic} the corn so the corn harvested still decreased.

Question 12: Determining the efficacy of trap crop pt.2

Score Level	Scoring Criteria
3 (This level broadly	Student response demonstrates complex understanding of systems and system modeling within the three main parts of their argument: the claim, the evidence cited, and their reasoning.

aligns to level 4 elements on the learning progression)	• Claim (about system phenomena, SM1): response shows that the student can accurately describe efficacy of the control mechanism using at least one unseen/hidden/underlying mechanism (**SM1.3.B**). o Student responses at this level should claim that the success of the alfalfa was a result of at least one of the following aspects of the system: ▪ The rootworms ate alfalfa instead of eating corn. ▪ Eating alfalfa in some way prevented the rootworms from laying eggs o Additional claims comparing the efficacy of the alfalfa to the baseline (in which no strategy is implemented) or to the harvestmen method are permitted even though they are not part of the question prompt. • Evidence (from system components or relationships, SM2/3): response cite from changes to two or more system components that are accurately related to the use of the relevant control strategy and the degree to which those components changed (**SM2.4.C+SM3.4.A**). o Student must cite evidence related to all three of the presence, the number of rootworms, and the amount of corn that survives the season. o Response should include a statement of *rate* (i.e. how fast the rootworms are procreating and/or consuming corn). o Student may also include a statement about rootworm eggs. o Additional evidence from the baseline (in which no strategy is implemented) can be cited without penalty. • Reasoning (from system relationships and mechanisms, SM3/4): response provides a relevant mechanism related to a causal chain between the alfalfa, rootworms, and the amount of corn harvested. All three components must be addressed. Reasoning provided must be relevant to the claims students make and the evidence they have cited (**SM4.4.B+SM2.5C**). o Student response must make reference to all three of the following relationships: ▪ The rootworms consume alfalfa in addition to consuming corn. ▪ Consuming alfalfa in some way prevents corn rootworms from surviving or procreating. ▪ As the rootworms consume the alfalfa instead of the corn, the corn yield improves either because the alfalfa is being eaten in place of the corn or because consuming the alfalfa decreases the rootworm population (and thereby decreases the loss of corn). o Response should include a statement of *rate attribution* – attributing the rate (or changes to the rate) of one change in relevant components (the corn yield and/or the rootworm population) to the behavior of other system components. **From Datusability12:** I think that it does, and it is more effective than method one, this is because even though the start was rough, the % did start to go up. Also there was a higher harvest when you look at the # of corn planted to the # of corn

	harvested. The rootworm eggs also started to decrease. I think this was because they didn't have enough time to plant eggs because they were eating the alfalfa instead of the corn, which is how they grow. In the simulation, the rootworms were becoming adults in the fall innstead {sic} of the summer, which meant that they couldn't all lay their eggs, making the problem less of an issue.
2 (This level broadly aligns to level 3 elements on the learning progression)	Student response demonstrates acceptable understanding of systems and system modeling within the three main parts of their argument: the claim, the evidence cited, and their reasoning. The student response includes some incorrect or inappropriate elements but is still sophisticated. • Claim (about system phenomena, SM1): response shows that the student can accurately describe efficacy of the control mechanism using at least one unseen/hidden/underlying mechanism (**SM1.3.B**). ○ Student responses at this level should claim that the success of the alfalfa was a result of at least one of the following aspects of the system: ▪ The rootworms ate alfalfa instead of eating corn. ▪ Eating alfalfa in some way prevented the rootworms from laying eggs ○ Additional claims comparing the efficacy of the alfalfa to the baseline (in which no strategy is implemented) or to the harvestmen method are permitted even though they are not part of the question prompt. • Evidence (from system components or relationships, SM2/3): response cite from changes to two or more system components but does not provide a relationship between those evidences and a relevant control mechanism (**SM2.3.C**). ○ Student must cite evidence related to all three of the presence, the number of rootworms, and the amount of corn that survives the season. ○ Response should include a statement of *rate* (i.e. how fast the rootworms are procreating and/or consuming corn). ○ Student may also include a statement about rootworm eggs. ○ Additional evidence from the baseline (in which no strategy is implemented) or to the harvestmen method can be cited without penalty. ○ Some stated evidence can be inaccurate or inappropriate, provided that sufficiently much accurate/appropriate evidence is also cited in support of the claim. • Reasoning (from system relationships and mechanisms, SM3/4): response provides a relevant mechanism related to a causal chain between the alfalfa, rootworms, and the amount of corn harvested. All three components must be addressed. Response includes a mixture of both appropriate and inappropriate system properties (**SM3.4.A+SM4.4.D)**. ○ Student response must make some reference to at least one of the following relationships: ▪ The rootworms consume alfalfa in addition to consuming corn.

	▪ Consuming alfalfa in some way prevents corn rootworms from surviving or procreating. ▪ As the rootworms consume the alfalfa instead of the corn, the corn yield improves either because the alfalfa is being eaten in place of the corn or because consuming the alfalfa decreases the rootworm population (and thereby decreases the loss of corn). ○ Response includes references to appropriate relationships but fails to accurately attribute changes in the number of rootworms or amount of corn harvested to those relationships. **From Datusability34:** Growing alfalfa helped keep the % of the corn yield as high as possible because when the alfalfa was near to the rootworms, the rootworms died, which decreased the amount of corn they could eat, because their was less of them.
1 (This level broadly aligns to level 2 elements on the learning progression)	Student response demonstrates insufficient understanding of systems and system modeling within the three main parts of their argument: the claim, the evidence cited, and their reasoning. Response may be broadly inaccurate, but still articulate the existence of multiple system components (rootworms, rootworm eggs, corn, alfalfa, etc.) that may have been relevant during the simulation. • Claim (about system phenomena, SM1): response addresses only one of the "reduce the corn rootworm" or "keep the corn yield high" components of the question. Response shows that the student inaccurately describes efficacy of the control mechanism **(SM1.2.B)**. ○ Student response directly answers question but does not refer to any aspects/features/behavior of the alfalfa. ▪ May include statements like "Planting alfalfa in the field did not help increase the corn yield percentage" • Evidence (from system components or relationships, SM2/3): response cites from changes to only a single system component (**SM2.2.A**). ○ Response only cites from changes in the amount of corn harvested as indicative of the efficacy of the alfalfa **OR r**esponse only cites from changes in the population of rootworms as indicative of the efficacy of the alfalfa. ○ Student may refer to the trade-off between land used to plant corn and land used to plant alfalfa, but only as evidence of lack of efficacy, not as evidence in support of a relationship (i.e. does not connect survival to future growth of rootworm population). • Reasoning (from system relationships and mechanisms, SM3/4): response provides a relevant mechanism related to a single relationship between any two of the alfalfa, rootworms, and the amount of corn harvested but doesn't address all components (**SM3.3.B)**. ○ Response only describes a relationship between the alfalfa and the rootworms **OR** response only describes a relationship between the rootworms and the corn. **From Datusability41:** The alfalfa planted in the corn field does help the corn yield as high as it can be. And even in year 3-4 there was an increase of corn harvested.

	However there is the downside of not planting as much corn as possible but this method seemed to work better than the first one.
0 (This level broadly aligns to level 1 elements on the learning progression)	Student response demonstrates little or no understanding of systems and their response is missing argumentative elements. They may provide a simple description of the video or restate information that was previously provided to them Or Student response is off topic or inappropriate. This score may also be assigned for any response that fails to meet minimum expectations for a level 1 score. **From Datusability9:** Yes, becuse {sic} the results produce a more suggesting number that the corn is harvesting more when there is Alfalfa in it. This is becuse {sic} the Alfafa {sic} is stronger and can get rid of more rootworms.

Question 13: Drawing a conclusion about the strategies

Score Level	Scoring Criteria
3 (This level broadly aligns to level 4 elements on the learning progression)	Student response demonstrates complex understanding of systems and system modeling within the three main parts of their argument: the claim, the evidence cited, and their reasoning. • Claim (about system phenomena, SM1): response shows that the student can accurately describe efficacy of the control mechanism using at least one unseen/hidden/underlying mechanism (**SM1.3.B**). o Student responses at this level should claim that the success of the alfalfa will or will not continue into future years because of at least one of the following aspects of the system: ▪ The alfalfa was successful in reducing the rootworm population in years 3-5 ▪ The corn rootworm population increased when alfalfa was not planted o Response should include a statement of *prediction* in which the student asserts that the strategy will/will not continue to be effective in the future. o Additional claims comparing the efficacy of the alfalfa to the baseline (in which no strategy is implemented) or to the harvestmen method are permitted even though they are not part of the question prompt. • Evidence (from system components or relationships, SM2/3): response cite from changes to two or more system components that are accurately related to the use of the relevant control strategy and the degree to which those components changed (**SM2.4.C**). o Student must cite evidence related to all three of the presence alfalfa, the number of rootworms (or rootworm eggs), and the amount of corn that survives the season.

	○ Response may also include a statement regarding balance or equilibrium (i.e. “held the rootworms at bay”). ○ Student may also include a statement about rootworm eggs. ○ Additional evidence from the baseline (in which no strategy is implemented) or to the harvestmen method can be cited without penalty. • Reasoning (from system relationships and mechanisms, SM3/4): response provides a relevant dynamism (ability to change in the future) to the consequences of a causal relationship between the alfalfa, rootworms, and the amount of corn harvested. All three components must be addressed. Reasoning provided must be relevant to the claims students make and the evidence they have cited (**SM4.5.D+SM4.5.E**). ○ Student response must make reference to both of the following relationships: ▪ The rootworms population grows in the absence of a control strategy ▪ Consuming alfalfa in some way prevents corn rootworms from surviving or procreating. ○ Response must include a statement of *predictability* in which the student asserts that evidence from prior years can be used to estimate the future state of the system. ○ Response may include a statement of *dynamism* in which the student asserts that aspects of the system (like number of rootworms or amount of corn surviving to harvest) are subject to change in response to changes to the system (i.e. the introduction of a control strategy). **From Datusability6:** Jonah can continue to plant alfalfa in year 7. He can continue that because it helps keeping the corn healthy and keeping the corn rootworms as low as possible. In the data table it shows that from year 1 to year 2 the corn decreased by a lot when he didnt {sic} use the alfalfa. But when he used it, it actually increased. In the other data table of the rootworm eggs it shows that from year 1 to year 2 the eggs increaseed {sic} when he didnt {sic} plant alfalfa. But from year 3 to year 5 the rootworms decreased. In conclusion. Jonah should use the alfalfa in year 7.
2 (This level broadly aligns to level 3 elements on the learning progression)	Student response demonstrates complex understanding of systems and system modeling within the three main parts of their argument: the claim, the evidence cited, and their reasoning. • Claim (about system phenomena, SM1): response shows that the student can accurately describe efficacy of the control mechanism using at least one unseen/hidden/underlying mechanism (**SM1.3.B**). ○ Student responses at this level should claim that the success of the alfalfa will or will not continue into future years because of at least one of the following aspects of the system: ▪ The alfalfa was successful in reducing the rootworm population in years 3-5 ▪ The corn rootworm population increased when alfalfa was not planted

	○ Response should include a statement of *prediction* in which the student asserts that the strategy will/will not continue to be effective in the future. ○ Additional claims comparing the efficacy of the alfalfa to the baseline (in which no strategy is implemented) or to the harvestmen method are permitted even though they are not part of the question prompt. • Evidence (from system components or relationships, SM2/3): response cite from changes to two or more system components but does not provide a relationship between those evidences and a relevant control mechanism (**SM2.3.C**). ○ Student must cite evidence related to all three of the presence of alfalfa, the number of rootworms (or rootworm eggs), and the amount of corn that survives the season. ○ Response may also include a statement regarding balance or equilibrium (i.e. "held the rootworms at bay"). ○ Student may also include a statement about rootworm eggs. ○ Additional evidence from the baseline (in which no strategy is implemented) or to the harvestmen method can be cited without penalty. ○ Some stated evidence can be inaccurate or inappropriate, provided that sufficiently much accurate/appropriate evidence is also cited in support of the claim. • Reasoning (from system relationships and mechanisms, SM3/4): response provides a relevant dynamism (ability to change in the future) to the consequences of a causal relationship between the alfalfa, rootworms, and the amount of corn harvested. All three components must be addressed. Response includes a mixture of both appropriate and inappropriate potential relationship dynamics (**SM3.4.B+SM4.4.E)**. ○ Student response makes reference to only one of the following relationships: ▪ The rootworms population grows in the absence of a control strategy ▪ Consuming alfalfa in some way prevents corn rootworms from surviving or procreating. ○ Response may include a statement of *dynamism* in which the student asserts that aspects of the system (like number of rootworms or amount of corn surviving to harvest) are subject to change in response to changes to the system (i.e. the introduction of a control strategy). **From Datusability13:** The rootworms decreased in the amount of time the alfalfa was being used. This method would probably continue to work because the egg count has already gone down, and when the egg count goes down, more corn can be harvested. There will be less rootworms to eat the corn.
1 (This level broadly aligns to	Student response demonstrates insufficient understanding of systems and system modeling(?) within the three main parts of an argument: claim, evidence, and reasoning.

level 2 elements on the learning progression)	• Claim (about system phenomena, SM1): response includes a claim that only addresses the state of the alfalfa. Response shows that the student inaccurately describes efficacy of the control mechanism **(SM1.2.B)**. ○ Response should include a statement of *prediction* in which the student asserts that the strategy will/will not continue to be effective in the future but does not refer to any aspects/features/behavior of the alfalfa. ▪ May include statements like "Planting alfalfa in the field did not help increase the corn yield percentage" • Evidence (from system components or relationships, SM2/3): response cites from changes to only a single system component (**SM2.2.A**). ○ Response only cites from changes in the amount of corn harvested as indicative of the efficacy of the alfalfa **OR r**esponse only cites from changes in the population of rootworms as indicative of the efficacy of the alfalfa. ○ Student may refer to the trade-off between land used to plant corn and land used to plant alfalfa, but only as evidence of lack of efficacy, not as evidence in support of a relationship (i.e. does not connect survival to future growth of rootworm population). ○ Response may cite directly from the graphs without connection to a causal relationship or mechanism. • Reasoning (from system relationships and mechanisms, SM3/4): response provides a relevant mechanism related to a single relationship between any two of the alfalfa, rootworms, and the amount of corn harvested but doesn't address all components (**SM3.3.B)**. ○ Response only describes a relationship between the alfalfa and the rootworms **OR** response only describes a relationship between the rootworms and the corn. ○ Student overemphasizes slight variation in years 4, 5, and 6 to assert that the control strategy is no longer effective. ○ Response may include an *inaccurate* statement of *dynamism* in which the student asserts that aspects of the system (like number of rootworms or amount of corn surviving to harvest) *will not change* despite changes to other aspects of the system (like the cessation of control strategy implementation). **From Datusability40:** I think that growing alfalfa in year 7 to help control corn rootworms will not help, because although the number of corn rootworms did decrease when the alfalfa was planted in years 3 and 4, the numbers began to increase again in year 5, and if it continues at that same rate, planting alfalfa in years 6 and 7 will not continue to help.
0 (This level broadly aligns to level 1 elements on	Student response demonstrates little or no understanding of systems and their response is missing argumentative elements. They may provide a simple description of the video or restate information that was previously provided to them Or Student response is off topic or inappropriate.

the learning progression)	This score may also be assigned for any response that fails to meet minimum expectations for a level 1 score. **From Datusability7**: On the second chart rootworms goes up and down so he cannot predict what will happen next year and for that reason he cannot control the amount of rootworms

Appendix I: R Script for Structural Equation and Rasch Model

```
library(psych)
library(GPArotation)
library(eRm)
library(mixRasch)
library(dplyr)
library(haven)
library(lavaan)
library(QuantPsyc)
library(MASS)
library(Hmisc)
library(readxl)
library(semPlot)

#Read in data
DATXecosystemUsabilityData <- read_excel("Dissertation Stuff/DATX Score
Analysis.xlsx")
ItemScoresOnly<-
dplyr::select(DATXecosystemUsabilityData,Q2_SUM:Q5,Q6_SUM,Q7_SUM,Q8:Q13)

#checking for unidimensionality
scree(ItemScoresOnly)
pca(ItemScoresOnly, nfactors = 2,residuals=TRUE)

#sem cfa
model <- '
#measurement model
StructureFunction =~  Q8 + Q7_SUM + Q6_SUM + Q5 + Q4 + Q3 + Q2_SUM
SystemModels =~ Q13 + Q12 + Q11 + Q10 + Q9
#regressions
StructureFunction ~~ SystemModels
StructureFunction ~ Grade
SystemModels ~ Grade
'
fit<-lavaan::sem(model, data = DATXecosystemUsabilityData)
summary(fit, standardized=TRUE)
semPlot::semPaths(fit)
semPlot::semPaths(fit,what="path",whatLabels = "par")

#Cronbach Alpha
psych::alpha(ItemScoresOnly)
psych::alpha(dplyr::select(DATXecosystemUsabilityData,Q2_SUM:Q5,Q6_SUM,Q7_SUM
,-Q8))
psych::alpha(dplyr::select(DATXecosystemUsabilityData,Q9:Q13))

#Rasch modeling-Combined
RaschCombined<-
eRm::PCM(dplyr::select(DATXecosystemUsabilityData,Q2_SUM:Q5,Q6_SUM,Q7_SUM,Q8:
Q13))
#Average Difficulty and Threshold position
thresholds(RaschCombined)
#Item Category Curves
plotICC(RaschCombined, mplot = TRUE, legpos = FALSE, ask = FALSE)
#Person Item Map
plotPImap(RaschCombined,sorted=TRUE,main="Person-Item Map of Combined
Behavior Items")
```

```
#item-fit statistics
itemfit(person.parameter(RaschCombined))
######Examining Fit Statistics
RaschCombinedPersons<-person.parameter(RaschCombined)
RaschCombinedPersonResid<-residuals(RaschCombinedPersons)
scree(RaschCombinedPersonResid) #some indication that unidimensionality
assumption fails
CombinedTheta<-RaschCombinedPersons$theta.table

####Repeat with SF Separated
RaschSF<-
eRm::PCM(dplyr::select(DATXecosystemUsabilityData,Q2_SUM:Q5,Q6_SUM,Q7_SUM,Q8)
)
#Average Difficulty and Threshold position
thresholds(RaschSF)
#Item Category Curves
plotICC(RaschSF, mplot = TRUE, legpos = FALSE, ask = FALSE)
#Person Item Map
plotPImap(RaschSF,sorted=TRUE,main="Person-Item Map of Structure Function
Items")
#item-fit statistics
itemfit(person.parameter(RaschSF))
######Examining Fit Statistics
RaschSFPersons<-person.parameter(RaschSF)
RaschSFPersonResid<-residuals(RaschSFPersons)
scree(RaschSFPersonResid) #some indication that unidimensionality assumption
fails
SFTheta<-RaschSFPersons$theta.table

####Repeat with SSM Separated
RaschSSM<-eRm::PCM(dplyr::select(DATXecosystemUsabilityData,Q9:Q13))
#Average Difficulty and Threshold position
thresholds(RaschSSM)
#Item Category Curves
plotICC(RaschSSM, mplot = TRUE, legpos = FALSE, ask = FALSE)
#Person Item Map
plotPImap(RaschSSM,sorted=TRUE,main="Person-Item Map of Systems and System
Model Items")
#item-fit statistics
itemfit(person.parameter(RaschSSM))
######Examining Fit Statistics
RaschSSMPersons<-person.parameter(RaschSSM)
RaschSSMPersonResid<-residuals(RaschSSMPersons)
scree(RaschSSMPersonResid) #some indication that unidimensionality assumption
fails
SSMTheta<-RaschSSMPersons$theta.table

#Wilcoxon Signed Rank Test for difference between Thetas
wilcox.test(SFTheta$`Person Parameter`,SSMTheta$`PersonParameter`,paired =
FALSE)
```

Appendix J: Definitions of Some Key Terms

- Cognitive Labs (or Cog Labs or Usability Tests) – In assessment design, a critical step in the development of a complex instrument (such as those used in the DAT-CROSS projects) are guided interviews (typically think-alouds), surveys, and item-response analyses that examine the degree to which the instrument is effectively engaging respondents in the kinds of cognition relevant to the measured construct without posing unnecessary challenges that may systematically prohibit the subject from acting toward their true ability. Drawing on the focus of diminishing non-construct relevant barriers to participants' use of the instrument, the process may also be referred to as 'usability testing' (Zaharias & Poylymenakou, 2009).
- Models – In teaching parlance, modelling is an instructional behavior (i.e. the teacher models the behavior she wishes the student to enact). However, models in this paper refer to scientific models – means of representing science knowledge by mapping abstract science concepts onto pictures, words, or other structures. Models serve many purposes in science and some philosophers argue that all science knowledge can be thought of in terms of the various representations scientists use to explain, describe, and predict the natural world. More on models in the NGSS can be found in Kracjik and Merritt (2012).
- Performance Expectations – The standards in the NGSS are not a mere listing of important ideas in science (though this dimension does appear as DCIs) because a central theme of the NGSS is that what students do or do not know and only be evaluated in terms of the operationalization of that knowledge. Performance expectations in the NGSS detail ways students can operationalize their science knowledge via some activity.
- Practices – In this paper, "practices" refer to the suite of activities authentic to the scientific community outlined in the Science and Engineering Practices (SEPs) dimension of the *Framework* (NRC, 2012). These activities represent the various ways that scientists engage with their science knowledge, and feature as a dominant part of the language of each performance expectation of the NGSS.
- Scripts – The foci of investigation in Engestrom's (2000) activity theory analytical framework are the developed schema for how people engage in various activities that make up both everyday and professional life. Engestrom calls these schema *scripts*. Much like a script for an actor, activity scripts are constructed prior to any specific instance of an activity in order to inform how to proceed as the activity occurs. Just as the production of a play can be disrupted in such a way as to push an actor off-script, so too can disruptions in the learning setting push a student off script. While there can be value in disrupting a script for the purpose of constructivist learning, such disruptions are an impediment to effective measurement of students' current states of understanding.

- Storyboards – The development of rich performance assessments entails the movement through a design process that starts at a very general level (domain modeling), moves to include specifics of what performances might be possible (task models), and culminates as a sequence of stimuli (including written prompts, videos, and images) and response collections. This series of stimuli and response, the last design stage before a task become operational, comprises a document called the "storyboard" (NRC, 2014) which details the specifics of what students will interact with as they engage in the task.