UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
The Affordances of Item Response Theory: The Case of a Complex Learning Progression about Evolution

Permalink
https://escholarship.org/uc/item/9jb8w632

Author
Cardace, Amy Elizabeth

Publication Date
2018

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

The Affordances of Item Response Theory:
The Case of a Complex Learning Progression about Evolution

By

Amy Cardace

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Education

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark Wilson, Chair
Professor Kathleen Metz
Professor Steven Raphael

Fall 2018

Abstract

The Affordances of Item Response Theory:
The Case of a Complex Learning Progression about Evolution

by

Amy Cardace

Doctor of Philosophy of Education

University of California, Berkeley

Professor Mark Wilson, Chair

The National Research Council (NRC) has published multiple reports in recent years that describe potential improvements to elementary science instruction, and call for higher standards for the development of related assessments. To meet this charge, the *Development of the Conceptual Underpinnings of Evolution (CUE)* project was funded by the National Science Foundation and produced the data analyzed with item response theory in this dissertation.

The CUE project team designed curricula for introducing concepts of micro-evolution to second- and third-grade students to engage young learners in reasoning about complex scientific phenomena. The team studied student capacities for understanding concepts in micro-evolution, while investigating the interplay of cognitive development, learning, and instruction. The team also developed one-on-one interview assessments, based on a learning progression that describes children's pathways to more sophisticated types of reasoning about micro-evolution. The design of this progression was a novel attempt to model early learning in this context, informed by cognitive development theory, classroom experiences, and assessment activities in an iterative process. Throughout this work, the CUE team examined the complexity of children's conceptualization of the concepts under study, especially with respect to the differentiation and order of conceptual levels within the learning progression and their multidimensional nature.

This dissertation examines the affordances of Item Response Theory (IRT) in investigating and validating the hypothetical learning progression and representing the range of student thinking, while enabling the quantitative assessment of learning gains. The empirical findings address practical questions about item function, multidimensionality, open-ended items, and pre-post effects, and ultimately support the use of this complex learning progression to model student thinking. This dissertation presents an uncommon case where IRT is applied to interview data, and complementary outcome measures were chosen to best represent the complex phenomenon under study. This work provides a model for strengthening validity arguments by examining assumptions about how items, responses, and coding schemes relate to the theoretical construct, and how the scores they yield inform what we know about student reasoning and learning gains.

Acknowledgments

This dissertation would not have been possible without the support and guidance I received from my academic advisors, Professor Mark Wilson and Professor Kathleen Metz, and my parents, Dominick and Lynne Cardace. I am also very thankful to Andrew Galpern for his advice and generous help with logistics.

This research was supported by a grant from the National Science Foundation for the *Development of the Conceptual Underpinnings of Evolution in Second and Third Graders* project awarded to Dr. Kathleen Metz.

Table of Contents

Abstract 1
List of Tables ii
List of Figures iii
Chapter 1. Introduction 1
Chapter 2. Methods 6
Chapter 3. Results: Validity and Reliability 12
Chapter 4. Results: Pre-Post Change 35
Chapter 5. Discussion and Conclusions 49
References 52
Appendix A. Interview Protocol Excerpt for Cricket Item 54
Appendix B. Coding Book Excerpt with Examples for Cheetah Item 56

List of Tables

Table 1. Baseline Learning Progression: Building Explanation of the Fit between Organisms and Their Environment (CUE codebook, 6/6/2010)

Table 2. Task Administration by Cohort

Table 3. Spearman Rho Interrater Estimates for Summer School Codes

Table 4. Summer School Interrater Statistics by Item

Table 5. Interrater Percent Agreement and Spearman Coefficients for Prime Coders

Table 6. Disagreement Cases by Levels and Warrants

Table 7. Interrater Data for First 20% of Final Dataset

Table 8. Number of Items in Each Task by Dimension

Table 9. Unidimensional and Multidimensional Results

Table 10. Items in Multidimensional Analysis with Reversed Response Categories

Table 11. Summary Statistics for Sum of Squares Discrepancy Indices (DI) by Dimension

Table 12. Examples of Cased with High Discrepancy Indices for State Dimension (D1)

Table 13. Multidimensional Version of the Learning Progression

Table 14. Latent Regression of All Cases with External Variables

Table 15. Percent of Students Scoring Proficient or Above by Site

Table 16. Students in Dataset by Cohort

Table 17. Mean Scores for Returners by Cohort over Time

Table 18. Latent Regression of All Cases Controlling for Variables of Import

Table 19. Latent Regression of Novice Cases Controlling for Variables of Import

List of Figures

Figure 1. The Learning Progression Perspective as a Bridge between Instruction and Assessment
Figure 2. BEAR Assessment System Model in the CUE Project Setting

Figure 3. The Cumulative Probability Curve Plot for Brassica 2

Figure 4. The Cumulative Probability Curve Plot for Butterfly

Figure 5. The Cumulative Probability Curve Plot for Cheetah 1

Figure 6. Map of student ability distribution and threshold locations for each item (UD)

Figure 7. Scatterplot of EAP scores on Dimension 1 (State) and Dimension 2 (Process)

Figure 8. Multidimensional Map of Student Ability Levels and Item Threshold Locations after DDA

Figure 9. Distribution of Regular School Teacher Rating Variable

Figure 10. Scatterplot of EAP Estimates and Regular School Teacher Ratings at Pre-Interview

Figure 11. Notation of Research Design

Figure 12. Pre-Post Change in Mean Scaled Scores for Novice Students

Figure 13. Wright Map for Dimension 2 (Process)

Figure 14. Pre-Post Change in Mean Scaled Scores for Return Students

Figure 15. Scatterplot of Max Score and EAP Score by Pre-Post

Figure 16. Pre-Post Change in Maximum Scores for Novice Students

Figure 17. Pre-Post Change in Maximum Scores for Return Students

Chapter 1. Introduction

The National Research Council (NRC) has published several reports in recent years that describe potential changes to science instruction, and call for higher standards for the development of related assessments. The 2007 publication, *Taking Science to School*, reviewed current research and practical teaching challenges involved in science education for elementary and middle school students (National Research Council, 2007). The report describes advances in our knowledge about how children learn and how instruction can be improved to better support this learning. The authors delineate four strands of scientific literacy that incorporate content knowledge and active scientific inquiry: (1) know, use, and interpret scientific explanations of the natural world; (2) generate and evaluate scientific evidence and, explanations; (3) understand the nature and development of scientific knowledge; (4) participate productively in scientific practices and discourse (National Research Council, 2007, p.51). They argue for interweaving the strands during instruction to maximize learning. Discussing the need for reform, the authors point out that "in contrast to the commonly held and outmoded view that young children are concrete and simplistic thinkers, the research evidence now shows that their thinking is surprisingly sophisticated" (National Research Council, 2007, p. 335). Accordingly, examining children's capabilities was a central goal of the Development of the Conceptual Underpinnings of Evolution (CUE) project. The project's framework integrated curriculum and assessment design with insight from cognition research.

More recently, the 2012 NRC publication, *A Framework for K-12 Science Education*, embraced this concept in their vision for science education where students "actively engage in scientific and engineering practices and apply crosscutting concepts to deepen their understanding of each field's disciplinary core ideas" (NRC, 2012, p. 2). The proposed instructional focus on the triad of scientific practices, crosscutting concepts, and core disciplinary ideas holds the promise to dramatically change the way science is taught in public schools and would theoretically achieve a higher standard for all students. After this movement evolved further with the introduction of the Next Generation Science Standards (NGSS), the National Research Council issued a report asserting the need for "a thorough rethinking of each element of science education, including science assessment" (NRC, 2014, p. ix). That report specifically proposed particular approaches to assessment, focusing on the construct-centered approach applied in this dissertation.

Research questions

I consider the reform principles noted above in the practical setting of curriculum design, implementation, and evaluation. My dissertation examines the learning progression and structured interview assessment developed as part of the CUE project. It provides an example of learning progression-based assessment to measure elementary student conceptions in the natural sciences. These data can help us address questions at the forefront of science education reform today.

How should researchers establish validity for assessments based on learning progressions? While there is an emerging consensus that assessments should be closely related to a construct theory, there are various ways of doing so. My research describes one strategy for developing a new construct to describe conceptual development in a classroom setting, and

discusses how both the design process and related quantitative analyses can contribute to validity arguments. The CUE example presents a case in which student interview data is transformed into ordered codes that are directly linked the learning progression under study. I describe this process and the quantitative analyses that informed theory development and ensured validity.

What analyses meet current standards of validity for the CUE learning progression? Current assessment standards for validity highlight multiple types of analyses that are essential for establishing validity for new constructs. In this case, the data required for these analyses include interrater codes, student responses, pre- and post-test scores, and student demographics such as school grade and academic proficiency as rated by teachers. My study incorporates these data into a validity argument designed to provide a model for upholding the most recent Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, National Council for Measurement in Education, 2014).

What can CUE outcome measures reveal about pre-post change in students' understanding? Two related issues arise in this study: what outcome measures provide useful information about pre-post change, and what is the most accurate way to estimate pre-post change in scaled scores. I consider the usefulness of both maximum level achieved and scaled score for assessing students' conceptual growth, and describe pre-post trends based on both measures. I also provide an example of latent regression in order to accurately account for the measurement error inherent to studying the relationships between latent and independent variables.

Theoretical framework

The educational reform movement described above relied on advances in cognitive science to reframe what optimal instruction looks like. Central to this reframing is a critical review of mainstream conceptions regarding developmental appropriateness and how those conceptions have restricted the development of optimal instructional activities. Multiple papers cited in *Taking Science to School* (National Research Council, 2007), including ones authored by the CUE project's principal investigator, reflected this perspective (Metz, 1995, 1997). Essentially, typical curricula often depend on theories of learning that consider children's potential conceptual development as limited by discrete cognitive stages. This way of thinking imposes unnatural constraints in pursuit of what are commonly termed "developmentally appropriate" learning environments.

Metz (1995) presents evidence that this interpretation of learning theory is problematic, as highlighted by unpacking the logic behind recent elementary science education standards in relation to foundational theories in cognitive development. Specifically, in the case of elementary science education, a common curricular strategy is to limit instruction to concrete activities such as observing, ordering, categorizing, and making inferences. This strategy does not reflect a comprehensive understanding of Piaget's work, which included evidence in support of "the potential for fuller scientific inquiry at this age and younger" (Metz, 1995, p.96). Further, more recent theoretical perspectives suggest that teaching scientific inquiry effectively requires opportunities for authentic inquiry in collaborative settings, even for elementary school students. The problematic "build-blocks curricular structure" (Metz, 1995, p. 118), in which the scientific enterprise is fragmented and decontextualized, ignores evidence of children's greater conceptual flexibility, and the dependence of that conceptual flexibility on the instructional context.

In light of this evidence, as well as others, *Taking Science to School* (National Research Council, 2007, p.35) pointed out that "a key question for instruction is thus how to adapt the instructional goals to the existing knowledge and skills of the learners, as well as how to choose instructional techniques that will be most effective." In a context where elementary science curricula have historically underestimated what children can do, the dynamic ways in which instructional contexts affect learning, and the importance of providing opportunities for authentic scientific inquiry, the practical steps for achieving reform are complicated. *A Framework for K-12 Science Education* describes the need for more evidence to support specific reform plans in the future, and identifies teachers as the "the linchpin in any effort to change K-12 science education" (NRC, 2012, p. 255).

Researchers have accordingly presented guidance about how to redesign curricula to optimize learning, and considered the everyday experience of teachers in this process. In an earlier research project related to the CUE curriculum, Metz (2004) describes animal behavior and botany curricula that were developed according to specific instructional design principles to address the pedagogical challenges noted above, and administered by elementary teachers in the public school system. The design principles serve as guidance for curriculum developers to achieve a more optimal level of instructional activities by encouraging the contextualization of science activities to foster motivation and understanding, supporting the appropriation of related knowledge, proving opportunities for collaboration, and inviting creativity in the design of research projects.

Building upon the theoretical and empirical foundations noted above, the CUE project provided an opportunity to study the implementation of a sophisticated elementary science curriculum, the validity of an assessment designed for the new curriculum, and the conceptual change it inspired among students. Ideally, the implementation of more dynamic curricula should include the design of more rigorous assessment, primarily through integration with the underlying learning theory. To this end, The CUE project designed instruction and assessment with reference to a common learning progression theory to ensure coherence between the two.

In *Taking Science to School* (National Research Council, 2007), the authors describe contemporary research in cognitive development and assert that student conceptions emerge from "a complex interplay among maturation, experience, and instruction" (p. 3). They also describe learning progressions in particular as a method for developing constructs related to conceptual growth in educational settings. Specifically, the report calls for learning progressions to model "successively more sophisticated ways of thinking about a topic that can follow and build on one another as children learn about and investigate a topic over a broad span of time" (p. 214). The authors also note that learning progressions can play a critical role in aligning curriculum, instruction, and assessment to create more effective classrooms. Along these lines, I rely on the learning progression approach to align instruction and assessment activities within the CUE project, as shown in Figure 1 below.

Metz (2009) theorizes about the role of learning progressions, highlighting how they can model the development of knowledge, inform instructional design, and accordingly "have substantial impact on how students' understandings emerge" (p.17). Viewing student learning through the lens of a defined learning progression also introduces a risk of underestimating the variation that truly exists, whereas contextual factors always affect the particular pathways of conceptual development. Along these lines, Lehrer and Schauble (2009) suggest caution in the utilization and interpretation of constructs in learning environments due to their traditionally unidimensional and unidirectional nature (p.734). One strategy for addressing this concern is to

model the relationships between constructs to better describe reality, as suggested in Wilson (2009). Not only can multiple constructs be defined as dimensions within a common learning progression, but analyzing these dimensions can show linkages between levels across dimensions (Wilson, 2012).

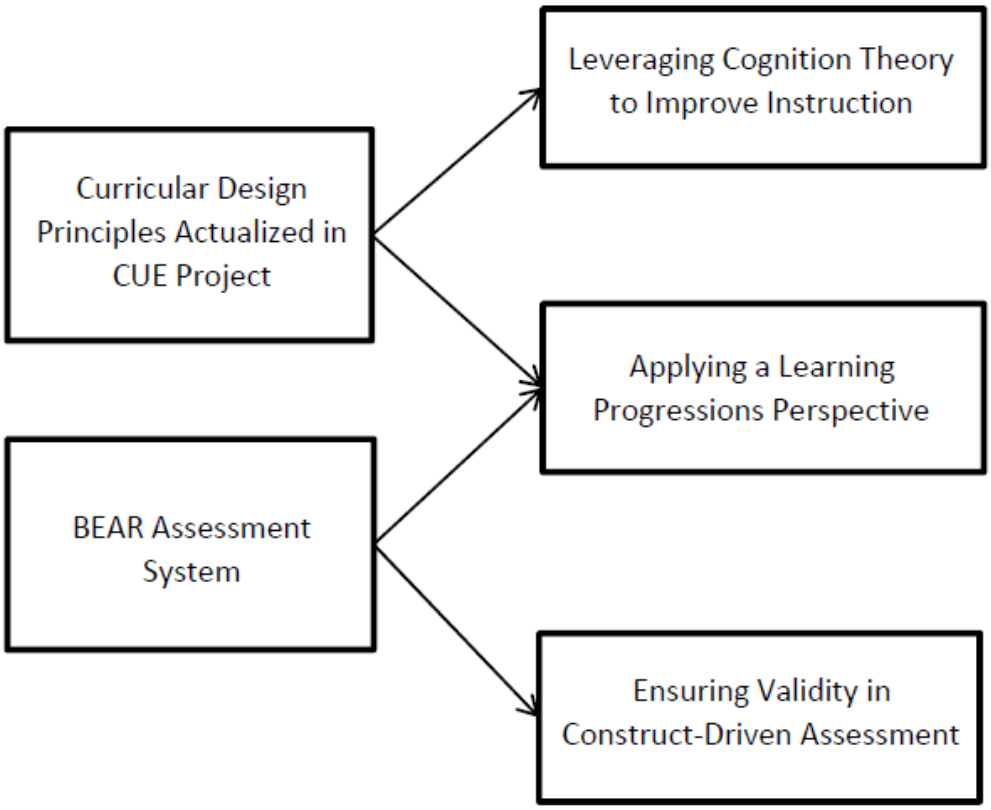


Figure 1. The Learning Progression Perspective as a Bridge between Instruction and Assessment

An additional critique of learning progression research proposes that reliance on empirical data to define the theory can create a bias toward confirmation and limit the range of possible interpretations. One strategy for promoting rigor given this concern is the researchers' responsibility to "seek disconfirmation" (Shavelson, 2009, p. 10). The BEAR Assessment System (BAS; Wilson and Sloane, 2000; Wilson, 2005) provides one model for managing the balance between designing and questioning new learning progression theories. BAS defines four "building blocks" that provide a foundation for developing constructs for measurement. First, the construct map defines the construct under study and describes how each level within that construct differentiate from lower to higher levels. We assume that the latent construct can be manifested in some way that we can assess, and that there are higher and lower values ordered in some progression-like form. The corresponding set of items, the second building block, is designed to map directly to levels within the proposed construct, ideally with particular responses that signal a proposed location on the construct map. The third building block is the outcome space, where analysts' decisions about how to code responses must directly relate to levels within the proposed construct. Ultimately these codes must be given values, or scored, in

some way that relates back to the construct via a measurement model, the fourth building block (see Figure 2).

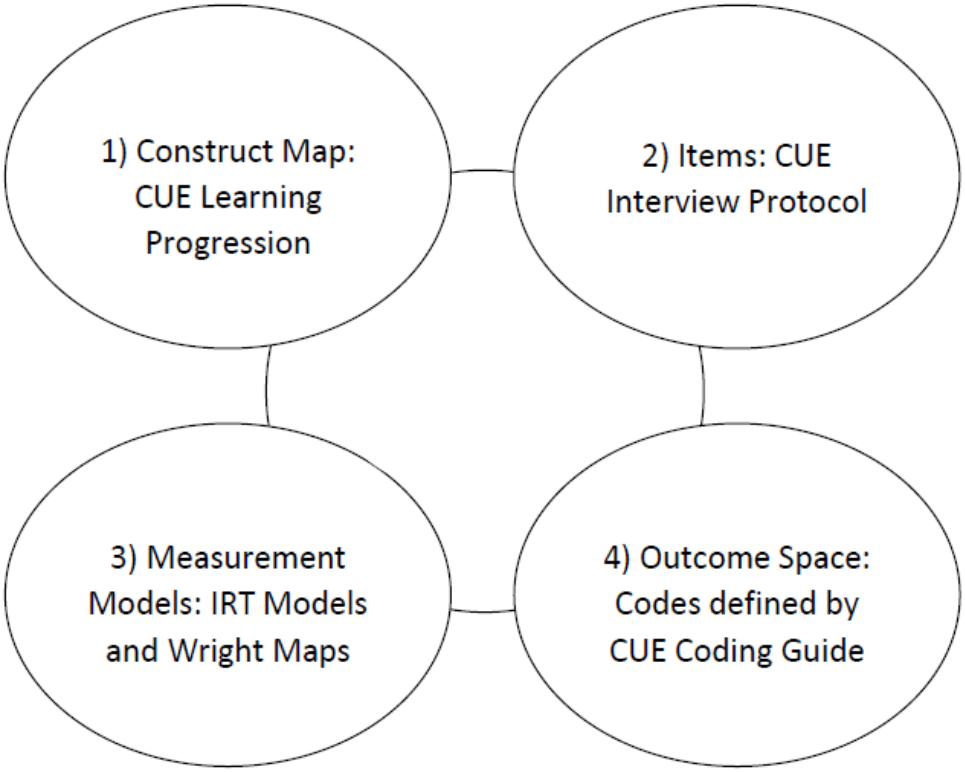


Figure 2. BEAR Assessment System Model in the CUE Project Setting

I adopt the BAS approach described in Wilson (2009) and utilize item response theory (IRT) to bridge the philosophical and analytical gap between item-focused (e.g., Guttman) and instrument-focused (e.g., Classical Test Theory) approaches (Wilson, 2009, p. 89). IRT methods used here are based on the Rasch model, which enables analysis at both the item and instrument levels. Beyond summative approaches to calculating student scores, IRT models the probability of student responses. This model assumes that each student has a latent location that can be estimated on the construct's scale, and that each item has a corresponding estimate of item difficulty. The transformation of both students' estimated scores and items' estimated difficulties onto a logit scale enables the production of a Wright map (Wilson, 2009, p. 90) in which the interaction of the two becomes evident. This approach allows me to directly address questions of internal consistency, item functioning, and multidimensionality as noted above, and also enables latent regression for analyzing pre-post effects while taking measurement error into account.

Chapter 2. Methods

One of the CUE project objectives was to strengthen the learning progressions framework with quantitative analyses that leverage item response theory (IRT) to maximize the information we can glean from student responses. As noted above, the BEAR Assessment System (Wilson and Sloane, 2000; Wilson, 2005) helps to ensure validity in this process by systematically investigating the relationships between (1) the underlying theory of a hierarchical construct, (2) the items that can differentiate levels within that construct, (3) the scoring rules that link responses to the construct, and (4) the measurement model that can provide evidence that the response patterns function as the construct theory proposed.

CUE learning progression

In this case, the first step in this validation process is the research-based development of the learning progression. The learning progression theory developed by CUE researchers describes the fit between organisms and their environment, and the way organisms adapt to maintain that fit. In the context of microevolution, students involved in the curriculum are challenged to consider how organisms adapt given environmental conditions. Table 1 shows an initial hypothesis of the learning progression before multiple dimensions were defined.

The levels in Table 1 describe student conceptions along the path to a sophisticated understanding of the topic. They were developed by the research team with data from classroom activities and interview responses and informed by the developmental research literature. The research team considered whether these levels were exhaustive, ordered, and on a single dimension throughout the project, and the learning progression theory evolved accordingly. During this process, the team used the IRT analyses to provide empirical feedback about multiple relationships between item responses and conceptual levels, the hierarchical nature of the learning progression, and issues of multidimensionality. Ultimately, the unidimensional learning progression theory evolved to differentiate two dimensions, discussed in detail in Chapter 3.

CUE curriculum, sample, and data

The original learning progression provided a foundation for the CUE curriculum, which was designed to embody five tenets of design-based inquiry as defined by Brown (1992), and described in more detail in the related manuscript (Metz, Cardace, Berson, Ly, Wong, Sisk-Hilton, Metz, & Wilson, under review). These two curricula were taught to students in second and third grades at two urban elementary schools in Northern California schools. Both schools had large percentages of students designated as socioeconomically disadvantaged: 65% in the regular school year site, and 42% in the summer school site. (California Department of Education, 2017. Cohort 1 included four classes of summer school students; Cohort 2 included eight classrooms of students who participated in the program during their regular academic year. At each site, half of the classes were taught in Year I and half in Year II. Over thirty percent of participants in YII were returning students who also participated in the curriculum in YI. Each student who completed the curriculum was interviewed both at the beginning and end of instruction.

Table 1
Baseline Learning Progression: Building Explanation of the Fit Between Organisms and Their Environment (CUE codebook, 6/6/2010)

Learning Progression Level	Description
LP 8	This process of Natural Selection leads to organisms that are adapted to where they live.
LP 7	Over many generations inherited traits that help an organism's chance to survive and reproduce where they live become more common. Those that hurt its chances become less common.
LP 6	Inherited traits' survival value can make a different on their relative frequency in the next (offspring) generation.
LP 5B	Traits' survival value can make a difference on their relative frequency in the population (same generation).
LP 5A	Individuals with a trait that gives them an advantage are more likely to live long enough to produce offspring.
LP 4	Individuals of the same species are not identical. Individual organisms' trait(s) can confer risk or advantage for the individual's survival or not make a difference for where they live.
LP 3	Structures of organisms enable particular functions of behaviors and/or enable the organisms to live under particular environmental conditions.
LP 2	Organisms live in places where they can get their needs met.
LP 1.5	Organisms live where they live because of what's there or what they can do there.
LP 1	Negligible understanding of fit between organisms and their environment.

Note: LP5A and LP5B are assumed to be equivalent but conceptually different.

The research team designed an interview protocol intended for assessment and evaluation purposes. Some items were based on other cases of microevolution from biology research literature and previous studies of related science learning. Additional items were designed specifically to target both the curricular concepts being taught by the research team, and also novel examples to study students' ability to transfer concepts across contexts. Items were

designed and grouped intentionally to cover both plant and animal examples of microevolution, with various degrees of transfer from the classroom instruction. For instance, one segment of the curriculum focused on the survival value of different wing types among crickets. An item specifically about the cricket example was included as a direct transfer item for students who were taught the Animal Behavior module and paired with a nearly identical item about the Brassica rapa plant that was based on the Botany module (see Appendix A). Additional items were designed to elicit student responses about related phenomena that were not directly addressed in the curriculum (e.g., Desert/Rainforest, Moth, Butterfly), and others were added to mirror or elaborate on items in the literature mentioned above (e.g., Guppy, Cheetah). In total, there were 10 tasks, each with one to three items, yielding 18 items. Each of the two cohorts was administered a subset of these ten tasks according to Table 2.

Table 2
Task administration by cohort

Cohort	Guppy	Brassica	Cricket	Otter	Kelp	Moth	Cheetah	Butter-fly	Desert Rain-forest	Desert Plants
2009	X	X	X	X	X	X	X			
2010	X	X	X				X	X	X	X

Note: There were multiple items within some tasks, denoted by numerical codes such as Guppy 1, Guppy 2, or Guppy 3.

There were seven common items that enable comparison between the cohorts (Guppy 1, Guppy 2, Guppy 3, Brassica 1, Cricket 1, Cheetah 1, and Cheetah 2). All items were administered as structured interviews lasting between thirty and forty-five minutes by trained interviewers who videotaped each interview. Our dataset contains only a subset of the interviews due to staffing and budget constraints: the entire Cohort 1 was coded yielding pre and post scores for 65 students, and a subset of Cohort 2 was coded yielding scores for 70 pre and 123 post interviews.

Coding

The research team implemented consensus coding for a majority of our sample. We double-coded items and any disagreements were investigated by the Principal Investigator with discussions to follow until consensus was reached. Throughout this process, the coding guide was examined and revised for clarity, including examples of actual student responses at each level of the learning progression. (See Coding Guide excerpt in Appendix B.) Due to the complexity of the learning progression and the open-ended nature of the interview items, double-coding was critical for the majority of the data we coded, and the consensus-seeking process aided in the thorough development of the coding guide, and elaboration of the learning progression. A small subset of YII data was single-coded after inter-rater correlations exceeded 80% agreement. These data included Butterfly, Rainforest/Desert, and Guppy items for only a small set of students. Following calls for more rigorous qualitative data analysis via more transparent reporting about coding practices (Hammer, D., & Berlans, L., 2014), I present

analyses across items and raters at different time points, as well as examples of particular interview transcriptions that caused disagreement and how they were resolved.

Measurement Models

The development of the interview protocol and coding guide progressed iteratively as the CUE team integrated empirical data to improve items, clarify the coding rules, and elaborate on the underlying learning progression theory. Our analytical process follows BAS (Wilson and Sloane, 2000; Wilson, 2005) as described above, and utilizes IRT in the design of measurement models. Accordingly, I examine a partial credit unidimensional IRT model estimated in the ACER ConQuest program (Wu, Adams, & Wilson, 1998), and present statistics related to internal consistency, item fit, and item step order.

This IRT model analyzes response patterns to estimate both item difficulties and student proficiencies simultaneously, reporting both student score and item difficulty estimates on a logit scale to enable comparison as shown in Figure 6 in Chapter 3. The partial credit item response model (Wilson, 2005; Wright and Masters, 1982) denoted in Equation 1 extends Rasch and rating scale models to estimate different step parameters within each item; these steps do not need to be equal.

Equation 1

$$\ln \frac{P(x_{ip} = j|\theta_p)}{P(x_{ip} = j-1|\theta_p)} = \theta_p - \delta_{ij}$$

Equation 1 describes the log odds of person (*n*) achieving the adjacent item response category (*j*) on item (*i*). Coefficient θ_n represents the ability estimate for student *n*, and δ_{ij} represents the difficulty estimate for level *j* within item *i*.

Given the evidence of potential multidimensionality, discussed in more detail below, I also examine more complex models to analyze model fit. To do so, I apply the Multidimensional Random Coefficients Multinomial Logit Model (Adams, Wilson, & Wang, 1997; Briggs & Wilson, 2003). I analyze the CUE data in a "between-item" multidimensional partial credit model where each item is assigned to one of the two dimensions (Adams, Wilson, Wang (1997); Schwartz (2011)). This design allows for a Wright map comparison where each set of step parameters within an item is used to locate steps on the logit scale. In this model, responses to categories (*j*) within items (*i*) are analyzed to estimate item and step parameters, as well as student scores (θ) for each person (*p*) on each dimension (*d*), shown in Equation 2 below.

Equation 2

$$\ln \frac{P(x_{ip} = j|\theta_{pd})}{P(x_{ip} = j-1|\theta_{pd})} = \theta_{pd} - \delta_{ij}$$

I analyze response data based on the two theorized dimensions ("State" and "Process") and calculate item difficulties, student abilities, and AIC statistics accordingly.

Delta Dimensional Alignment

I also locate item difficulties across the two dimensions on the same scale to show how they relate and inform the learning progression theory. I calibrate the item difficulties across dimensions using the Delta Dimensional Alignment approach (Schwartz & Ayers, 2011) which enables a revision of the original learning progression theory to show how the levels in each dimension aligned. In order to examine the LP levels across dimensions more closely, ACER ConQuest produces maps that juxtapose the distribution of student abilities and the degree of difficulty for each response category within each item. Because we cannot assume that both dimensions will have a mean of zero, as is necessary in the multidimensional IRT model estimated in this case, an additional transformation of item difficulty and threshold parameters is necessary in order to compare dimensions.

The Delta Dimensional Alignment (DDA) technique (Schwartz & Ayers, 2011) enables this transformation by recalculating the item-related parameters, using the unidimensional mean and variance for each dimension. Conceptually, this technique assumes the unidimensional model estimated the average item difficulty for each dimension's item set reasonably well, but it did not take into account the multidimensional relationships when estimating the specific item difficulties. Accordingly, the unidimensional average for each dimension's item set acts as a reference point around which the multidimensional item difficulties can be scaled.

Equation 3

$$\text{Transformed Item Difficulty } \delta_{id} = \delta_{i\ multi} \left(\frac{\sigma_{uni}}{\sigma_{multi}} \right) + \mu_{uni}$$

Equation 4

$$\text{Transformed Step Parameter } \tau_{ikd} = \tau_{ik\ multi} \left(\frac{\sigma_{uni}}{\sigma_{multi}} \right)$$

These transformed item and step parameters can then be used as starting values to anchor the multidimensional estimation and produce Wright maps that facilitate comparison across dimensions.

In this case, the DDA technique was critical for aligning the dimensions in order to consider how LP levels from different dimensions compare in terms of item difficulty. For example, one theoretical issue discussed during coding conversations was whether the levels describing structure-function relationships (originally LP3) which were assigned to the State dimension in the multidimensional analysis would correspond with the levels about trait variation and survival value (originally LP4) on the Process dimension. A map of student abilities and item threshold locations that were transformed using the DDA technique can help researchers address this question.

External validity analyses

I also describe analyses to provide further evidence for the structured interview, according to contemporary assessment standards. I examine the CUE assessment in terms of each type of validity evidence as defined in *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council for Measurement in Education, 2014): test content; response processes; internal structure; external variables; and consequences. I utilize the BAS and IRT models as noted above to show how construct development, response processes, and internal structure relate to the assessment's validity. I also consider external variables by analyzing the correlations between student estimates and independent measures that would likely have positive correlations:

1) School: the summer school site performed might higher in state standardized tests, so I expect it to also have higher scores on our instrument at pre-test.
2) Grade level: I expect third grade students to perform better on our instrument at pretest given the likelihood that they have had more exposure to related course content and life experience.
3) Teacher rating of student proficiency: I expect teacher ratings of proficiency will be positively correlated with our measure at pretest.

Pre-Post Change

My final set of analyses considers pre-post change in students' scaled and maximum scores. I use both visual displays and statistics from latent regression analyses to describe trends in student scores and consider the impact of the instructional intervention. In addition to summary statistics and graphical displays created in Stata 14 and Microsoft Excel, I created a latent regression model in ACER ConQuest (Wu, Adams, & Wilson, 1998). The latent regression in essentially a multi-level model in which the item response data is analyzed to estimate item difficulties and student scores, while simultaneously estimating fixed effect coefficients for independent variables, such as pre-post in this example. The first level model was shown in Equation 2 above. The second level model, shown in Equation 5 below, models the student scaled scores (θ_{op}) as the dependent variable in a linear regression. In this case, external validity variables noted above, as well as control variables of import, are included as independent variables to estimate their relative relationships with student scores.

Equation 5

$$\theta_{op} = \beta_1 \boldsymbol{prepost}_{op} + \beta_2 \boldsymbol{year}_{op} + \beta_3 (\boldsymbol{prepost\ x\ year})_{op} + \beta_4 \boldsymbol{novice}_{op} + \beta_5 \boldsymbol{grade}_{op} + \beta_6 \boldsymbol{teacher\ rating}_{op} + \beta_7 \boldsymbol{school}_{op} + \zeta_{1p}$$

In basic terms, θ_{op} is influenced by a pre-post effect described by the magnitude of β_1, with a random effect at the person level, described by ζ_{1p}. The coefficients and their interpretations are described in more detail in Chapter 4.

Chapter 3. Results: Validity and Reliability

Item responses were coded by the CUE project team according to an elaborate coding guide, with examples of responses at different levels (see excerpt in Appendix A). Our coding processes evolved over the course of the project. Initially, all responses were coded by at least two coders, the team met regularly to review difficult cases, and interrater agreement was analyzed at multiple time points. As coders gained experience and the coding guide integrated the diversity of typical responses, dual coding continued with coders primarily working individually. When coders did not agree on a code after deliberation, the project lead would code the case and review the coding guide and related criteria with the coding team. The coding guide was elaborated as needed to address all necessary issues.

Early in the project eight coders participated in initial coding and coding book development. Table 3 shows the Spearman Rho correlations for the initial scores of those coders, who individually scored between 54 and 390 items at that point in the project. The majority of these correlations, 14 of 25, were at least 0.80 before reviewing and later agreeing on a consensus code. Six of the 25 were less than 0.70. Only five coders continued to analyze data as the project continued, referred to by the letters A, D, N, S, and U, and initial correlations among those five were particularly strong (see shaded cells in Table 3 below).

Table 3

Spearman Rho Interrater Estimates for Summer School Codes

Coder	A	D	E	H	N	S	T	U
A	1.00 (242)							
D	0.50 (3)	1.00 (54)						
E	0.66 (44)	.	1.00 (383)					
H	0.89 (71)	0.93 (9)	0.73 (71)	1.00 (390)				
K	.	.	0.85 (11)	.				
N	0.97 (33)	0.82(6)	0.69 (102)	0.73 (50)	1.00 (352)			
S	0.92 (5)	.	0.93 (15)	0.92 (17)	0.90 (32)	1.00 (115)		
T	0.74 (22)	0.90 (25)	0.54 (71)	0.75 (52)	0.62 (15)	.	1.00 (206)	
U	0.81 (33)	1.00 (7)	0.80 (45)	0.91 (79)	0.76 (75)	0.88 (48)	0.07 (27)	1.00 (319)

Note: The project PI reviewed any codes where initial coders did not achieve consensus which yielded 15 cases that were scrutinized in detail with a final code confirmed by the PI. Shading highlights correlations for the coders who continued coding long-term.

Interrater data for initial rounds of coding also revealed variation by item. Spearman Rho correlations varied between 0.52 and 1.00 across the 15 items that had adequate data, as shown in Table 4. The Kelp (1.00) and Butterfly (0.92) items were particularly strong, and the Desert/Rainforest set of three items was also at the upper end with an average of 0.75. These data were essential for considering the potential for single coding as the project continued.

Table 4
Summer School Interrater Statistics by Item

Item	Agree	Total	Percent Agree	Spearman
br1	47	62	76%	0.70
br2	17	31	55%	0.66
but	53	62	85%	0.92
ch1	47	93	51%	0.65
ch2	43	79	54%	0.58
ch_u	71	103	69%	0.81
cr1	42	57	74%	0.73
cr2	16	29	55%	0.66
dr1	48	63	76%	0.64
dr2	46	63	73%	0.78
dr3	37	63	59%	0.82
gu1	83	108	77%	0.52
gu2	93	111	84%	0.77
gu3	86	112	77%	0.80
kel	36	36	100%	1.00
mo1	9	11	82%	NA
mo2	8	8	100%	NA
ott	33	36	92%	0.82
Total	815	1127	72%	0.82

Note: Missing codes were included in agreement percentages to account for all agreements, but not in correlations.

Later in the project, we relied on three primary coders. The project standard was to establish eighty percent agreement before allowing single coding. To this end, we considered interrater disagreements in more detail, analyzing the percent of items with disagreement, the degree of disagreement, the LP levels involved, and how these factors varied by particular items and coding pairs over time. Table 5 shows 14 sets of double-codes among the three long-term coders for comparison, categorized by item and pre-post administration. In seven of these coding sets, at least one correlation did not meet the 0.80 threshold despite high percentages of agreement (denoted with * in the Table 5 below). In order to better understand this phenomenon,

I also analyzed the degree of disagreement by including the number of response categories, or steps, between the two disagreeing codes.

Table 5
Interrater Percent Agreement and Spearman Coefficients for Prime Coders

						Degree of disagreement				
		#Total	#Agree	%Agree	Corr.	1-step	2-step	3-step	4-step	5-step
BR1 POST	E_N	16	13	81%	0.56*	3	0	0	0	0
	N_U	9	7	78%	0.96	2	0	0	0	0
	E_U	9	7	78%	0.91	1	1	0	0	0
BR2 POST	E_N	16	7	44%	0.36	5	1	1	1	1
	N_U	8	5	63%	0.52	1	1	0	0	1
	E_U	10	4	40%	0.91	3	3	0	0	0
CR1 POST	E_N	28	23	82%	0.49*	2	1	2	0	0
	N_U	13	12	92%	0.84	1	0	0	0	0
	E_U	16	13	81%	0.63*	3	0	0	0	0
CR2 POST	E_N	29	17	59%	0.64	7	4	1	0	0
	N_U	13	11	85%	0.93	2	0	0	0	0
	E_U	15	12	80%	0.87	3	0	0	0	0
BR1 PRE	E_N	15	12	80%	0.73*	1	1	0	1	0
	N_U	19	13	68%	0.62	2	2	2	0	0
BR2 PRE	E_N	10	7	70%	0.40	1	0	0	2	0
	N_U	13	10	77%	0.83	2	1	0	0	0
GU1 POST	E_N	16	14	88%	0.93	2	0	0	0	0
	N_U	10	9	90%	0.95	1	0	0	0	0
GU2 POST	E_N	15	14	93%	0.97	1	0	0	0	0
	N_U	9	8	89%	0.69*	0	0	1	0	0
GU3 POST	E_N	16	12	75%	0.78	2	2	0	0	0
	N_U	9	8	89%	0.94	1	0	0	0	0
GU1 PRE	E_N	17	16	94%	NC	0	1	0	0	0
	N_U	19	18	95%	NC	1	0	0	0	0
GU2 PRE	E_N	17	17	100%	NC	0	0	0	0	0
	N_U	19	19	100%	NC	0	0	0	0	0
GU3 PRE	E_N	16	16	100%	NC	0	0	0	0	0
	N_U	19	19	100%	NC	0	0	0	0	0
CH D1	E_N	11	11	100%	NC	0	0	0	0	0
	N_U	7	7	100%	NC	0	0	0	0	0
CH D2	E_N	27	25	93%	0.96	1	1	0	0	0
	N_U	27	23	85%	0.94	4	0	0	0	0

Note: Codes were scaled so that one step is the distance between any given LP level and the nearest reported LP level for that item even if they are not numerically consecutive. Subcode disagreements were omitted. NC denotes "not calculated" usually due to no variation between codes. Shading highlights cases where agreement and correlations achieve 80% or higher. * denotes cases where the correlation is low despite over 80% agreement.

The initial interrater analyses showed that coders had more discrepancies when coding Cricket #2 and Brassica #2. As suggested by Hammer and Berlans (2014), I describe the nature of those disagreements and the mode of reconciliation in more detail. When any disagreement occurred, the original coders reexamined the case, as well as the other coder's warrant, or reasoning, for the original code. If disagreement persisted, the project team reviewed the case as a group and discussed, making edits to the coding guide as needed. Later in the project, when the coding guide was fully developed and disagreements were few, the project PI considered any disagreements and reported back to the group about the most appropriate code. In most cases across all items, disagreements were resolved after the initial coders elaborated on their warrants and recoded. The majority of disagreements were different by only one coding level.

However, there were also cases of disagreement across multiple learning progression levels. To examine these in more detail, I focused on the Brassica item, because the initial interrater results showed higher rates and more varied types of disagreement when compared to other items. The two Brassica items showed disagreements that varied by as many as 5 levels. The cases presented in Table 6 show the general types of resolutions we experienced during the project. These cases are all in response to the Brassica Rapa task, where students are told about the different stem types, some hairy and some smooth, and asked to reason about how that variation would affect which were more likely to be eaten by caterpillars (see Appendix A). Many students generated explanations about one type of stem being more or less preferable to caterpillars, and spoke about how that preference would affect the plants' future generations.

In Case A, both coders disagreed about whether the response achieved only LP4B (survival value), or LP5B (changing distributions within that generation) as well. Upon review, the coders agreed that the student's response showed evidence of "differential caterpillar preference" but no changes in population distributions, and the code was ultimately assigned as LP4B. Similarly, in Case B, coders disagreed about LP4B and LP5B levels, but ultimately decided that there was sufficient evidence for a LP5B code because the student "is taking away smooth ones from the original count" to show a change in distribution. In this case, noting how the student used the manipulatives was critical in understanding her reasoning. In Case C, coders agreed to interpret the student's description of a changing distribution as referring to the next generation, which was sufficient for a LP6 code as opposed to LP5B where the shifting distribution is only evident within the same generation.

Cases D through G show wider disagreements across multiple categories. In Case D, one coder revisited an original interpretation of the response to be signaling survival value, but ultimately decided it was too "cryptic" and agreed to a lesser code of LP1. Case E presents an example where coders began with LP1 and LP5B codes and ultimately agreed that LP4B was most appropriate. Case F presents an example where coders considered a student's previous response within the task when coding an item later in the task. Rather than LP1 based on a partial answer, the coders agreed that survival value was proven within the task and the student's elaboration on this item was actually evidence of LP5B.

Case G also shows the complexity involved. In this case, the coders did not reach agreement after debating whether a student's earlier response made their interrupted response to this item adequate for LP5B. The PI was engaged in this disagreement due to the lack of resolution and agreed with the proposal to code the item as LP1 (negligible understanding). Cases like these informed the refinement of the coding book. For instance, at one point we clarified that "if S[tudent] articulates a more abbreviated explanation to a second item within the same task, unless reason to interpret otherwise _-- in words or actions (10/25/11) --- treat

abbreviation as an avoidance of repetition of that just said [operating in accordance with Grice's Maxims of Discourse], and assign code of prior item" (CUE Research Team, 2011).

Table 6
Disagreement Cases by Levels and Warrants

Case	Item	Levels Involved	Consensus Warrant	Coders' Finding	PI Finding	Final Code
A	Br1	P4B v P5B	Differential caterpillar preference, but no differential reproduction or survival in terms of distribution.	P4B		P4B
B	Br1	P4B v P5B	I think it's 5b because she is taking away smooth ones from the original count. She didn't add any hairy ones because she's showing that the smooth ones are being eaten from that very same field. In that way, it IS a shift in the distribution. (She's showing gen 1, but reduced by # of smooth ones.)	P5B		P5B
C	Br1	P5B v P6	Understands the survival advantage, "and then it grew as a brassica rapa plant" = talking about the next generation, and "there's more hairy ones" indicates the shift in distribution.	P6		P6
D	Br1	P1 v P4B	Kinda thought that he mentioned the survival advantage of having a lot of seeds, but it's cryptic, so I agree [to downcode].	P1OTHER		P1OTHER
E	Br1	P1 v P5B	I want to knock it down to 4B now, because I'm not sure she understands how the advantage affects the distribution.	P4B		P4B
F	Br2	P1 v P5B	Looking at BR1 I think she understands survival advantage and states that they will be spikey because the (others) will die…to me still LP5B	P5B		P5B
G	Br2	P1 v P5B v Missing	CODER 1: Change to 5B because of Brassica 1, I'm scoring this way because the interview was interrupted, and it seemed like they didn't go DOWN in understanding, but we didn't hear the full idea; CODER 2: I think this is LP1UNCODEABLE (due to interruption) or LP1OTHER because his reason is based on a pattern of alternating generations.	P5B	YES: 1 Other	P1OTHER

After extensive double-coding, 1612 item codes were assigned, comprising over 65% of what would ultimately be 2412 total item codes over the course of the project. Before approving single coding, the project team analyzed a subset of responses for the four items left to be coded (Desert/Rainforest 1, 2, 3, and Butterfly). For this subset of items, agreement ranged between 85% and 100%, and the coders initiated single coding for these items, yielding 274 additional codes. Finally, the same experienced coders assigned codes for an additional set of students to improve our dataset, yielding 526 additional codes for Desert Plant, Cheetah, Cricket, and Guppy items.

Table 7
Interrater Data for First 20% of Final Dataset

		Total	#Disagreements	%Agreement	Type of Disagreement
E_N	DR1	22	2	91%	LP1U-2; LP2-3B
E_N	DR2	22	0	100%	
E_N	DR3	22	1	95%	LP1U-2
U_N	BUT	20	3	85%	LP1U-5B; LP1U-5B; LP1-4B

Unidimensional Model

The initial set of psychometric analyses to study the item set's item functioning and reliability use a partial credit unidimensional IRT model estimated using the ACER ConQuest program (Wu, Adams, & Wilson, 1998). This model includes all items that were designed to measure the unidimensional learning progression theorized at the beginning of the CUE project, previously discussed in Chapter 2 and illustrated in Table 1. Several quality indices indicate that the item set functions reasonably well as a unidimensional construct. The separation reliability statistics range from 0.75 using WLE estimates to 0.79 using EAP estimates. All 18 items have weighted mean square fit statistics that fall within the expected range (0.75-1.33). The two items with the lowest fit indices were Brassica2 (0.79) and Cricket1 (0.81), suggesting possible local dependence which typically occurs when items are strongly correlated. In this case, two Brassica items and two Cricket items were based directly on the curricular content as non-transfer items; they ask students questions about contexts discussed in class. In addition, the overarching Brassica and Cricket tasks were intentionally structured to be as comparable as possible.

To illustrate how these fit statistics inform us about item functioning in practice, following are three Cumulative Probability Curve plots for items with low, moderate, and high weighted mean square indices. Each item has five response categories and four corresponding thresholds. The plots show variability in how the probability curves differentiate from one another, and how closely the actual data map to the probability curves. For Brassica2, an example of a low weighted mean square fit, we see frequent gaps between the actual data points and the probabilistic curve, suggesting relatively weak fit (see Figure 3). In this case we also see multiple points where actual data for different levels overlap, suggesting that the response categories within the item are not clearly differentiated.

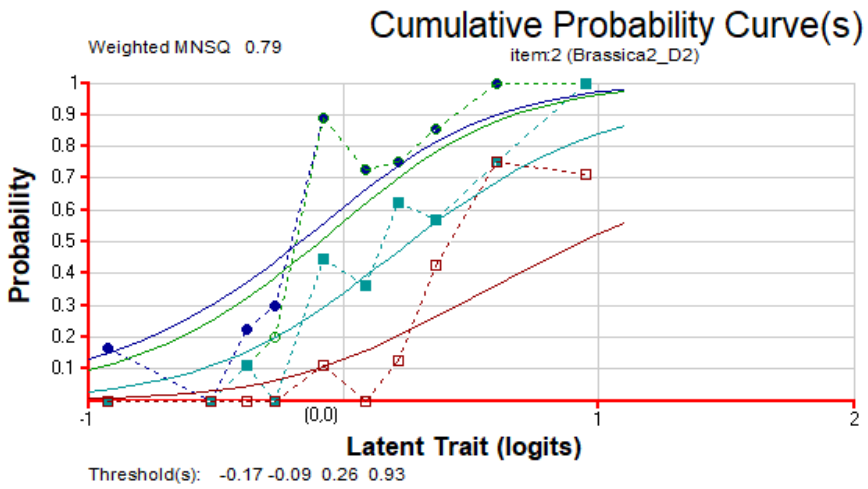


Figure 3. Cumulative Probability Curve Plot for Brassica2

The Butterfly item shows an example of a strong mean square fit index (0.98), where the actual data points map more closely to the probabilistic curve. The consistent overlap between the dark blue and green levels suggests they are not differentiated from one another well. Yet, the lowest probability levels (red and aqua) do show little overlap and are clearly differentiated from the highest levels (dark blue and green).

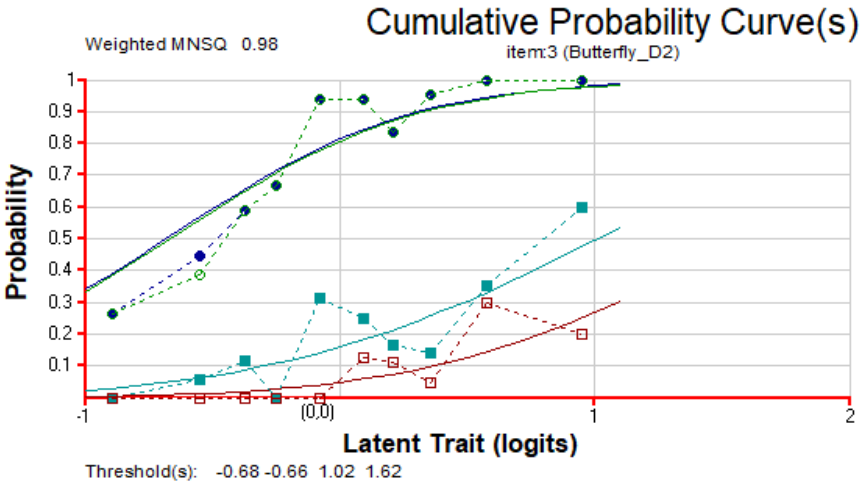


Figure 4. Cumulative Probability Curve Plot for Butterfly

In the case of Cheetah1, with a relatively high weighted mean square fit (1.21), we see different trends in the way the actual data varies from the predicted curves as the latent trait varies. For instance, the light green category overlaps with a higher category at the low end of

the latent trait scale, but deviates closer to the lower response categories at the higher end of the student score distribution.

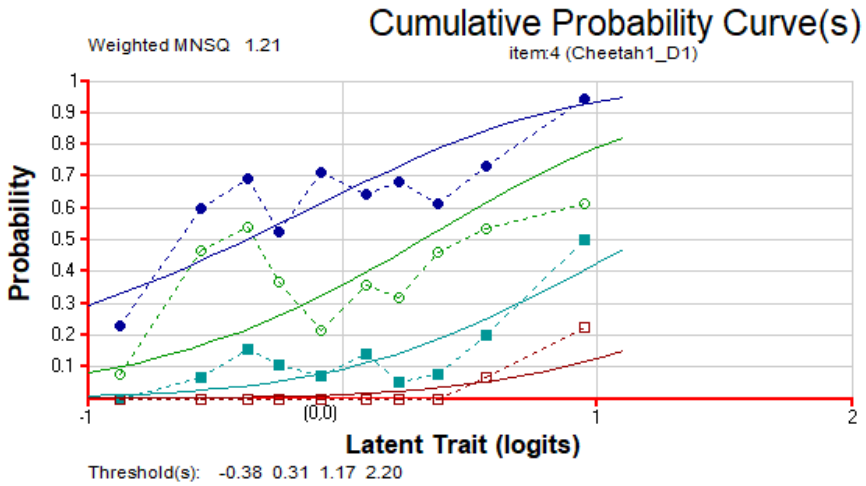


Figure 5. Cumulative Probability Curve Plot for Cheetah1

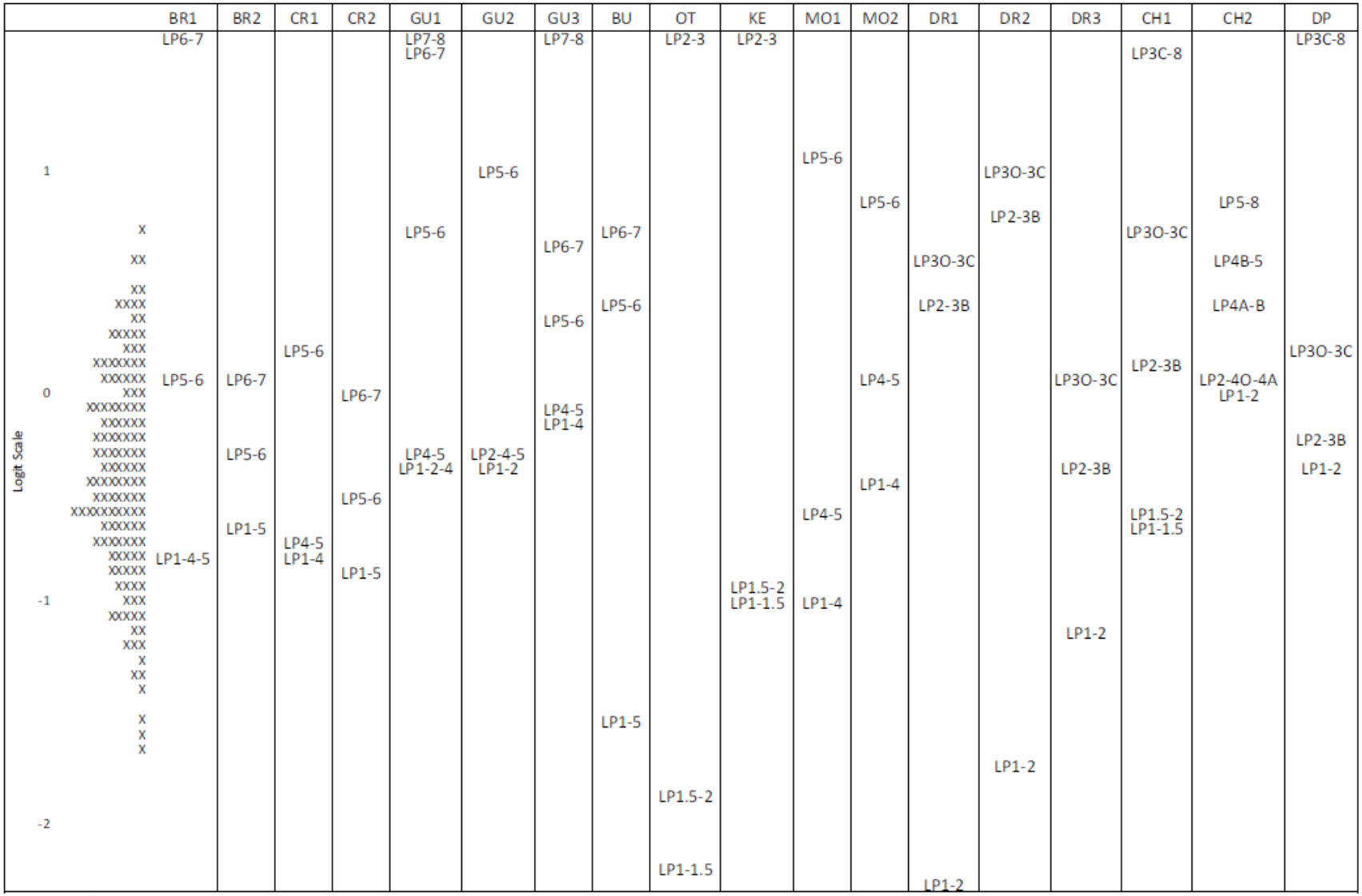
These three examples illustrate how the plots can vary, even within the acceptable range of weighted mean square fit indices (-0.75 - 1.33). The illustrations highlight the usefulness of considering how response categories within an item differentiate, especially in relation to the underlying learning progression theory. For instance, when examining the mean location of each response group in terms of logits, Brassica2 and Cheetah1 both show some irregular patterns. For Brassica2, the location of category 1 is below that of category 0, but there were only 3 respondents in category 1. Despite this reversal, the other response categories with larger sample sizes ranked logically, and the overall weighted mean square index did fall within the range noted above.

Cheetah1 is more problematic, with the mean location of response category 1 out of place, even with 38 respondents in the group. In conjunction with the borderline weighted mean square index, these data call into question the usefulness of the item. Its importance when considering the item set as a whole becomes more apparent when examining the Wright map in Figure 6.

This Wright map shows a vertical distribution of student ability levels transformed into logits by the IRT analysis; each “X” represents approximately 1 student. The numbers to the right of the distribution show the equivalent locations of each item threshold. This display allows analysts to compare items, identify which items are particularly useful for differentiating more or less sophisticated responses, and consider whether the items adequately cover the range of student abilities. Each threshold location shows the point at which a student at that level would be 50% likely to score above or below that threshold. For instance, for the first item (BR1), the LP1-4-5 threshold is located near -0.75 logits with 5 students scoring at that level. This placement suggests that students above this point would be likely to score at or above LP5 on this item, and those below would be more likely to score LP1. LP4 did not have a differentiated threshold in this case, showing that when students achieved LP4, they were equally likely to achieve LP5.

We see that all items have multiple response categories; 11 items have 4 response categories and 7 items have between 5 and 7 categories. The locations of thresholds between these response categories stretch above and below the student ability distribution. This wide range suggests that the item set has good coverage of the student abilities in the given sample, despite measuring a complex topic among a very young group of students. In general, the items function well together as a unidimensional construct with strong reliability statistics.

Further, the Wright map shows affordances of items Brassica2 and Cheetah1. Brassica2 functions with Brassica1 just as Cricket2 functions with Cricket1, and the two item subsets are very well aligned. These data suggest that both may be useful for showing students' understanding of direct course content, and that they do act in comparable ways as intended. The Wright map also highlights the very few items that offer students an opportunity to show mastery at the high end of the distribution; Cheetah1 is one of only five items where students might achieve the highest score of LP8 with evidence of reasoning about natural selection. For these reasons, I would not suggest omitting any of these items from future interview protocols solely on the basis of these borderline psychometric properties.



Note: Each 'X' represents 0.9 cases.

Figure 6. Map of student ability distribution and threshold locations for each item in the unidimensional model. This figure shows how the Wright map aligns logit scores for item difficulties and student scores.

Multidimensional Model

The unidimensional analysis also reveals a conceptual divergence in student responses between those addressing the good fit between organisms and their environment (the "State" dimension) and those describing how adaptation enables this fit (the "Process" dimension). This conceptual divergence was also noted separately by the teaching team before they received empirical feedback. One of the earliest empirical signs of this divergence is evident in the Wright map, where we see that LP3C appears consistently more difficult than LP4, and generally equivalent to higher LP levels; the LP30-3C thresholds are located near LP5-6 thresholds on other items. Upon further examination, the research team noticed that LP3 responses about an organism's physical structures and their related functions were conceptually different from the LP4-7 responses which focused on how survival value affects changing distributions of traits over time and generations. Reviewing student responses and considering the levels to which they relate, we were able to differentiate between reasoning about State and Process across the learning progression.

At the lowest level, some common responses address organisms' **state** of being in their environment (e.g., organisms live where they belong) while others address the **process** by which that state occurs (e.g., organisms adjust to where they live). More sophisticated concepts such as structure-function relationships on the State dimension, and survival advantage on the Process dimension dominate the middle of the progression. Student responses expose a wide conceptual range and inspired the addition of multiple sublevels. Some *Structure-Function* responses (LP3) are more sophisticated in terms of conceptual understanding than *Trait Variation* responses (LP4) and they are often of a qualitatively different nature. Common LP3 responses typically explain the fit between organisms and their environment, and common LP4 responses typically explain the mechanism of change.

Most items elicit responses that generally map onto either the State or Process dimension. For instance, the Otter task (OT) includes items about a picture of an otter eating in the water and questions whether otters could live in other environments. Student responses typically address the otters' physical characteristics and needs, thus mapping onto the State dimension. In comparison, the Cricket task (CR1 & CR2) questions how a sample of crickets would change over generations given variation in chirping, its reproductive value, and a change in the environment. This task evokes explanations about reproduction, survival, and shifting distributions of the chirping trait, which map clearly to the Process dimension. The Cheetah item (CH1 & CH2) is unique in that it described a trait change over time ("cheetahs today are faster than those who lived many years ago") and asks students to explain the change. Concepts of structure-function relationships AND survival value are both central to the item, and student responses did address both dimensions.

Accordingly, the multidimensional model assigns each item to one of the two dimensions, except for the Cheetah item which coders analyzed to provide scores on both dimensions (see Table 8).

Table 8
Number of Items in Each Task by Dimension

	OT,# items	KE,# items	DR,# items	DP,# items	CH,# items	GU,# items	BR,# items	CR,# items	MO,# items	BU, # items
D1. State	1	1	3	1	1					
D2. Process					1	3	2	2	2	1

Note: The Cheetah (CH) codes were based on a common interview item that was coded to reflect a score for each dimension.

The MRCML analysis described in Chapter 2 provides information about the reliability for each dimension and the correlation and covariance between the dimensions, as well as item and person-specific estimates. The person separation reliability statistics vary depending on estimation method as shown in Table 9.

Table 9
Unidimensional and Multidimensional Results

	Unidimensional	Multidimensional	
		"State" Dimension	"Process" Dimension
MLE reliability	0.76	0.37	0.62
WLE reliability	0.75	0.30	0.70
EAP/PV reliability	0.79	0.62	0.80
Variance	0.34	0.39	0.47
Correlation of dimensions	NA	0.71	
Covariance of dimensions	NA	0.31	
Deviance	5916.02	5885.03	
Parameters	76	78	
AIC	6068.02	6041.03	

As expected, the EAP/PV method results in higher reliability indices[1], with reliabilities for the entire item set (0.79) and the Process dimension alone (0.80) matching quite closely. The State dimension dropped to 0.62, likely because it included only 7 items as opposed to 11 on the Process dimension. For the multidimensional model, all mean square fit indices fall within the expected range.

The responses within each item are generally ordered as expected. A review of the average estimated locations for each response category identify three items in which response

[1] The EAP/PV estimates use the entire response vector, including all responses for items on both dimensions, compared to likelihood estimates which only use the responses related to each dimension.

categories are reversed, where there are more than 10 respondents in each category. There were three items on the State dimensions with a misaligned response category location for a category with more than 10 respondents: Cheetah 1; Desert Plant; and Desert Rainforest 1. Table 10 shows the nature of these response category reversals. In each case, the reversal related to LP levels that were likely too fine-grained to clearly differentiate. All involved codes that were so infrequent across the entire data set that they were essentially collapsed in future analyses (LP1.5, LP2, and LP3 subcodes O, A, B, or C). In general, empirical data for the other items reflect the expected hierarchy of the progression and support the multidimensional theory.

Table 10
Items in Multidimensional Analysis with Reversed Response Categories

Item	Reversed Category 1 (n)	Reversed Category 2 (n)
Cheetah 1	LP1.5 (38)	LP2 (39)
Desert Plant	LP1 (61)	LP2 (19)
Desert Rainforest 1	LP3OAB (12)	LP3C (17)

Note: Response categories wither fewer than 10 responses were omitted from this analysis due to their inherently high variability.

Of note, the three items with internal response category reversals all fall on the State dimension which also has weak reliability statistics ranging from 0.30 (WLE) to 0.62 (EAP). The two dimensions are strongly correlated (0.71) but not perfectly so, supporting the theory that they are related but unique. Possibly, with further development of the State dimension items, the two dimensions would be more clearly differentiated and yield improved statistics. ‘

Overall, AIC estimates suggest that model fit hardly improves between the unidimensional and the multidimensional models, with a decrease from only 6068.02 to 6041.03. A likelihood ratio test can examine specifically whether the change in model fit is statistically significant. In this case, the log likelihoods were calculated as -2958.01 (unidimensional) and -2942.51 (multidimensional), yielding a likelihood ratio test statistic of 31 and a two-tailed chi-square p-value of 1.855e—07. These findings suggest that the change in model fit, while minor, is statistically significant.

Considering how multidimensional analyses will affect interpretation is also relevant for model choice. As shown in Briggs and Wilson (2003), student scores on highly correlated dimensions can also vary in educationally meaningful ways. A pairwise analysis of EAP scores on the two dimensions yielded a correlation of 0.87 and the scatterplot in Figure 7 below. Not only does the scatterplot show a generally positive relationship, but it also illustrates cases where EAP scores on D1 vary from D2 in ways that would complicate interpretation. For instance, D1 and D2 scores vary by over 0.50 logit for 17 cases (7%).

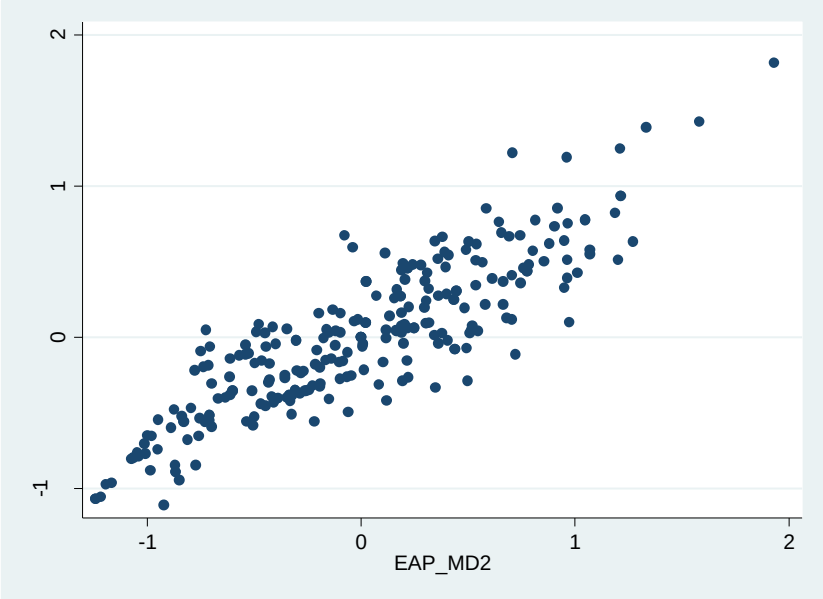


Figure 7. Scatterplot of EAP scores on State Dimension (D1) and Process Dimension (D2)

Pairwise correlations between the unidimensional scores and those from the State and Process dimensions equal 0.93 and 0.99 respectively. To consider how these relationships might affect the interpretation of scores, a discrepancy index is useful. This index is based on the squared difference between one dimension's score and the overall unidimensional score (Briggs and Wilson, 2003). The State dimension discrepancy index showed greater variation than that for the Process dimension, consistent with the correlations above.

Table 11
Summary Statistics for Sum of Squares Discrepancy Indices (DI) by Dimension

Variable	Mean	Std. Dev.	Min	Max	Obs
State Dimension DI	0.04	0.05	0.00	0.31	258
Process Dimension DI	0.01	0.02	0.00	0.12	258

The max discrepancy index was 0.31 compared to 0.12 for the Process dimension, and there were thirteen cases where the index was greater than 0.14 (two standard deviations above the mean). Among these cases, the difference between EAP scores on the two dimensions can show notably different locations. The majority of these cases had differences of more than 0.50 logits between the scores, as noted by * in Table 12. If we ignore the multidimensionality, we are essentially assuming that a single estimate can adequately describe students' thinking. Yet, in actuality, some students' responses showed varying degrees of sophistication depending on the dimension under study.

Table 12
Examples of Cases with High Discrepancy Indices for State Dimension (D1)

DI (State)	EAP (Uni)	EAP (State)	EAP (Process)	Δ>0.5
0.31	0.44	-0.11	0.72	*
0.30	0.12	0.67	-0.08	*
0.29	0.26	-0.29	0.50	*
0.28	0.63	0.10	0.97	*
0.24	-0.43	0.05	-0.72	*
0.22	0.12	0.60	-0.04	*
0.22	0.14	-0.33	0.35	*
0.16	0.53	0.13	0.68	*
0.15	0.83	1.22	0.71	*
0.15	-0.45	-0.06	-0.71	*
0.15	0.31	-0.07	0.49	*
0.14	-0.29	0.09	-0.48	
0.14	-0.04	-0.42	0.12	

These data support my finding that while the two dimensions do not dramatically improve the item set's psychometric fit at this point in development, they do in fact contribute to an interpretation of student conceptual growth that is more useful in practice than the unidimensional version. The multidimensional model also matches the evolving learning progression theory under development more closely than the original unidimensional model. Given the usefulness of the multidimensional model, equating the item threshold locations across dimensions is important for establishing whether the response categories in the State dimension align with those in the Process dimension as expected. To this end, I standardize the thresholds across dimensions using Delta Dimensional Alignment (Schwartz & Ayers, 2011). Figure 8 shows the two resulting distributions of student ability levels, each based on responses to one dimension's item set. The State dimension, on the left, yields a more narrow distribution that stretches across approximately 3 logits, slightly wider than that for the Process dimension. Next to each distribution are columns for each item related to that dimension. Within each item's column, the thresholds between response categories are identified with the learning progression levels. Each threshold marks the point at which students at that logit level on the distribution would have a fifty percent chance of responding either below the threshold, or at or above the threshold.

For instance, on the State dimension, the lowest thresholds for Kelp (KE) differentiate between responses that express *Negligible Understanding* (S1), the idea that organisms match well with certain environmental characteristics (S1.5), or that organisms need certain things to survive in an environment (S2). The Kelp threshold "S1-1.5-2" is located near 0 logits, and shows where students would typically be 50% likely to articulate conceptions either at S1 or at S2 or higher; S1.5 was typically very infrequent and accordingly did not always show a uniquely differentiated threshold. Students located beneath this threshold would most likely reason at the S1 level, and those above would likely reason at S2 or higher levels. Using this logic, the threshold map shows the likely response category for a student at any particular ability level for each item.

In the State dimension, student responses fall into eight distinct levels, increasing in sophistication from negligible understanding at the bottom of the learning progression to an understanding of natural selection at the top (see Table 13). Only a few students are able to fully explain the structure-function-environment relationship, as shown by the relatively high location of thresholds (shaded medium green in the threshold map below). The few students with ability levels near thresholds are approximately 50% likely to provide a complete structure-function-environment response. Students with ability levels below these thresholds are more likely to provide a less sophisticated response. The cells with light green shading illustrate the thresholds where students move from *Needs/Environmental Fit* (S2) responses to *Structure-Environment*, *Structure-Function*, or *Function-Environment* responses (S3). In general, these thresholds are consistently aligned within the State dimension, except for Otter and Kelp items (labeled as OTT and KEL in Figure 8), where student responses do not typically include the more sophisticated concepts that emerge in other State items. This difference was likely because more sophisticated students can respond to the Otter and Kelp items with a reasonably complete answer without needing to address structure-function-environment concepts.

The Process dimension follows a similar progression from showing negligible understanding (P1), to identifying how environmental factors relate to organisms' needs (P2), and how traits vary (P3). At higher levels, students reason about survival value (P4) and how that value affects reproductions (P5a) and shifting distributions of the trait within the generation (P5b). This line of thinking leads to considering how the prevalence of a trait with survival value changes over one (P6) or multiple generations (P7), ultimately informing an integrated understanding of natural selection at the highest level (S5+P7 or NS).

The logit scale facilitates examination of which levels on the Process dimension align with those on the State dimension. A major concept on the Process dimensions was traits' survival value, which is represented by achieving the P4 threshold (shaded in light blue). Ten of eleven Process items have threshold estimates for this level that fall between 0.0 and 1.0 logits. Within this section of the logit scale (0.0-1.0), the most common State dimension threshold is that for S2-3, representing reasoning about elementary concepts of structure-function relationships (shaded in light green). This parallel suggests that students who show conceptual awareness about survival value are also likely to do so about elementary structure-function relationships.

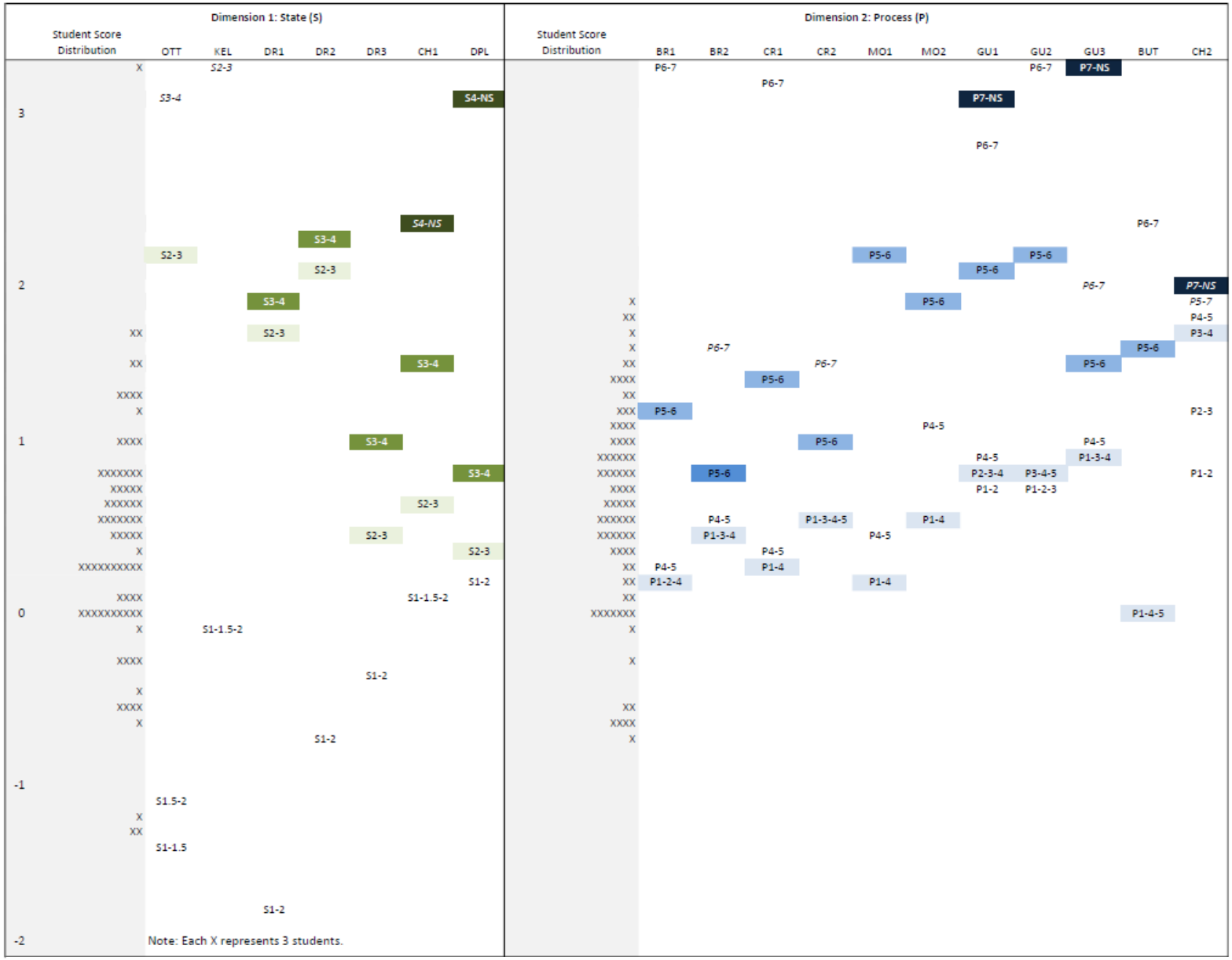


Figure 8. Multidimensional Map of Student Ability Levels and Item Threshold Locations after DDA

There is a similar relationship between more sophisticated responses on the two dimensions; all P5-6 thresholds (shaded medium blue) fell between 0.75 and 2.25 logits, a range that also included 5 of the 6 State thresholds at the S3-4 level (shaded in medium green). These data suggest that students who are able to effectively address the shifting prevalence of traits in a future generation depending on their survival value (P6), are also likely to provide *Structure-Function-Environment* (S4) responses on the State dimension. At the highest level, the Process and State dimensions merge where students can articulate an integrated understanding of natural selection. The S4-NS and P7-NS thresholds (shaded in dark green and blue) are all located above 2 logits.

Implications for Learning Progression Development

Given these empirical data, the dimensions are likely aligned as shown in the multidimensional learning progression (Table 13). Decisions about where to align the conceptual levels across dimensions were based on the map of item thresholds in Figure 8.

Common to both dimensions is the pattern of increasingly sophisticated conceptual frames at higher points on the Wright map, culminating at the highest level in an understanding of natural selection that incorporates concepts from both dimensions. In the State dimension, students at the lower levels are able to reason about what organisms need to survive and consider the surrounding environment. As they begin to think in terms of structure-function relationships they can describe physical structures of the organism, such as body parts, that affect its survival. The data include a range of responses about structure, function, and environmental conditions, with more sophisticated responses with reasoning about all three in concert. The progression in the Process Dimension follows a similar pattern at the lower levels. Students can identify traits and acknowledge that those traits vary within a population at a given time. As they progress, students can reason about the survival advantage of particular traits and consider the role of inheritance when reasoning about the proliferation or decline of particular traits across generations. Table 13 shows how these dimensions align, with an integrated top level at S4 + P7 to represent Natural Selection.

Table 13
Multidimensional Version of the Learning Progression

"State" Dimension	"Process" Dimension
S4+P7 (NS) This process of Natural Selection leads to organisms that are adapted to where they live (natural selection explains fit).	
	P7 Over many generations in the population inherited traits that help an organism's chance to survive and reproduce become more common and those that hurt its chances, less common (Process of natural selection).
S4 Some structure of the organism-kind enables a particular function or behavior that enables it to live under particular environmental conditions (structure-function-environment).	P6 In the offspring generation of the population, inherited traits that help its chances to survive and reproduce tend to become more common and those that hurt its chances, less common (one-generation natural selection).
S3c Some function of the organism/kind enables it to live under particular environmental conditions (function-environment). S3b Some structure of the organism/kind enables them to live in particular environmental conditions. (structure-environment). S3a Some structure of the organism/kind enables particular function or behavior (structure-function).	P5b Among organisms of the same kind living together, traits that hurt chances to survive tend to be less common within a generation (differential mortality population effects). P5a Individuals with a trait that gives them an advantage are more likely to live long enough to produce offspring (trait conveys reproductive advantage).
S2 Organisms live in places where they can get their needs met (needs/environmental fit). S1.5 Organisms live where they live because of what's there or what they can do there (organisms-environment association).	P4 Even individuals of the same kind living together at the same point in time are not identical. Different traits of the individual organism can hurt or help its chances to survive under particular environmental conditions (within-population trait variation). P3b Individuals of the same kind have not been identical over time. Different traits of the individual organism can hurt or help its chances to survive under particular environmental conditions (over-time trait change). P3a Individuals of the same kind at different points/stages of their life are not identical. Those differences can help or hurt its chances to survive under particular environmental conditions (over-lifecycle trait change). P2 Organisms survive when their needs are met (differential needs/environmental fit).
PS1 Negligible understanding of fit between organisms and their environment (negligible understanding).	

Note: Horizontal lines illustrate bands of approx. aligned conceptual levels.

External Variables

In order to rigorously defend the validity of using this assessment in classroom contexts, additional analyses examine whether the assessment relates to external variables as expected. Logically, if the assessment truly measures the sophistication of student reasoning, relationships with other variables should be measurable based on the following assumptions:

1. Student reasoning about topics addressed in typical instruction or everyday life should become more sophisticated with increased years of schooling on average, yielding a positive effect of being in a higher **grade**.
2. Students with higher levels of general classroom achievement should respond with higher levels of sophistication on average, yielding a positive correlation between student estimates and **teacher ratings**.
3. A **school** with higher mean proficiency levels in related topics should also have higher mean scaled scores, compared to a lower proficiency school.
4. Composite student estimates should increase with related instruction on average, yielding positive **pre-post** change.

In order to study these relationships while most appropriately accounting for measurement error, the latent regression described below estimates both the composite scores and the regression coefficients. I modeled the latent regression based on a partial credit item response model to calculate student estimates for the Process Dimension, and included multiple regression coefficients:

1) Pre-Post (0=pre, 1=post);
2) Year II (0=Y1, 1=YII);
3) The interaction of Pre-Post and Year to show whether the rate of improvement varied by year;
4) Novice as opposed to a small subset of returning students who experienced the instructional intervention twice (0=returner, 1=novice);
5) 3rd Grade where students were in either second or third grades at the time of instruction (0=second; 1=third);
6) Teacher-rating provided by Regular School instructors to differentiate students at "High", "Above Average", "Average", "Below Average" achievement levels[2]; and
7) Regular School was also included to identify any differences between schools that affect interpretation (0=summer school, 1=regular school).

[2] The "low" achievement level category was omitted in this analysis because of irregular functioning as illustrated in Figure 10 below. Also, because we only have teacher Rating for regular school students, including the lowest level in the latent regression model dramatically changed the Regular School coefficient when compared to other models.

Table 14
Latent Regression of All Cases with External Variables

Variable	Coefficients	Standard Error	T	Significance
Pre-Post	0.80	0.09	9.08	*
Regular School	-0.25	0.11	-2.19	**
Novice	-0.51	0.09	-5.60	*
3rd Grade	0.22	0.07	3.08	*
Year II	0.06	0.09	0.67	
Pre-Post x Year	0.05	0.12	0.40	
Teacher Rating	0.00	0.02	0.07	

Note: * denotes statistical significance at α=0.001. ** denotes statistical significance at α=0.005.

The latent regression results presented in Table 14 address the external variables related to the assessment validity assumptions noted above:

1. Grade is statistically significant with a positive effect of 0.22 logits attributed to being in third grade rather than second, as hypothesized.
2. Teacher rating is not statistically significant, with a coefficient of zero.
3. Regular School is statistically significant, with a coefficient of -0.25. This negative effect corresponds with the gap in proficiency levels between the schools; according to the California Standards Test Scores, students at the summer school site were notably more proficient in both Mathematics and English Language Arts (see Table 15; California Dept. of Education (2018)).
4. Pre-Post was statistically significant with a coefficient of 0.82, notably larger than the other independent variables which ranged from -0.51 to 0.22. This coefficient describes an average increase of 0.82 logits in student scaled scores over the course of the CUE curriculum which was described in more detail in Chapter 2.

Table 15
Percent of Students Scoring Proficient or Above in 2009-2010 by Site

Site	Math		ELA	
	2nd Grade	3rd Grade	2nd Grade	3rd Grade
Summer School	82	84	78	78
Regular School	56	74	52	49

Considering that the distribution of student estimates on the Process dimension largely fell between -1 and 2 logits, as shown in Figure 8, the pre-post effect describes an average increase of over a quarter of the overall range. In terms of the learning progression, this change would be equivalent to a conceptual shift from negligible understanding to a grasp of survival value, or an increase from understanding survival value to reasoning about shifts in the prevalence of a trait over a generation. Further discussion of the cohort mean scores and the empirical findings about the nature of these conceptual shifts is discussed in more detail in Chapter 4 below.

The grade level effect also suggests meaningful differences, as third graders placed nearly a quarter-logit above their second-grade peers on average. The teacher rating effect is nearly independent in terms of the learning progression distribution and did not show a strong correlation. Interestingly, these teacher ratings were distributed fairly evenly across all response categories (see Figure 9).

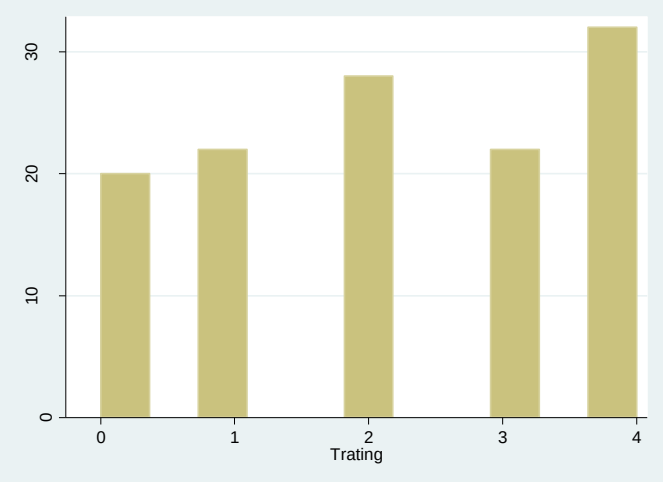


Figure 9. Frequency Distribution of Regular School Teacher Rating Variable

Examining the scatterplot between student EAP estimates and teacher ratings, the expected positive relationship appears to hold for students with higher Teacher ratings and higher EAP scores (see Figure 10).

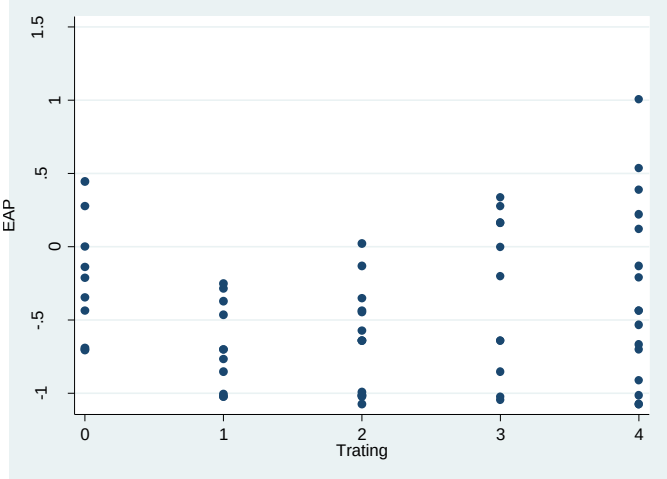


Figure 10. Scatterplot of EAP Estimates and Regular School Teacher Ratings at Pre-Interview

The analyses above describe an assessment with reasonably strong internal consistency that relates to a defined construct theory in meaningful ways. Each analysis provides support for the CUE assessment as a reliable and valid measure of the content taught in the CUE curriculum. Composite student estimates should increase with instruction on average, yielding positive pre-post change. These gains are significant with a relatively large coefficient, even after controlling for factors that would typically relate to student learning in this context, such as grade, school site, and teacher ratings. This evidence supports the hypothesis that the instructional intervention did yield meaningful conceptual growth for students, which will be discussed in more detail in Chapter 4 below.

Chapter 4. Results: Pre-Post Change

The research design for evaluating the CUE curriculum as an intervention was informed by best practices in quasi-experimental methods (Shadish, Cook, and Campbell, 2002), working within the practical limitations that are present when introducing a new curriculum within schools. I examined changes in outcome measures based on the pre-post intervention plan over two years at two sites. Because a subset of students participated in both Year I and Year II, there were six non-random (NR) cohorts with cycles of observation-intervention-observation. Figure 11 shows the research design, where | denotes independent samples between years. Summer school students were invited and then self-selected into the program, and regular school students in particular classrooms were taught the curriculum by their regular teachers. The summer cohort began before the regular school cohort and one of two curriculum modules were taught each year. Students were administered pre- and post-interviews before and after each intervention session.

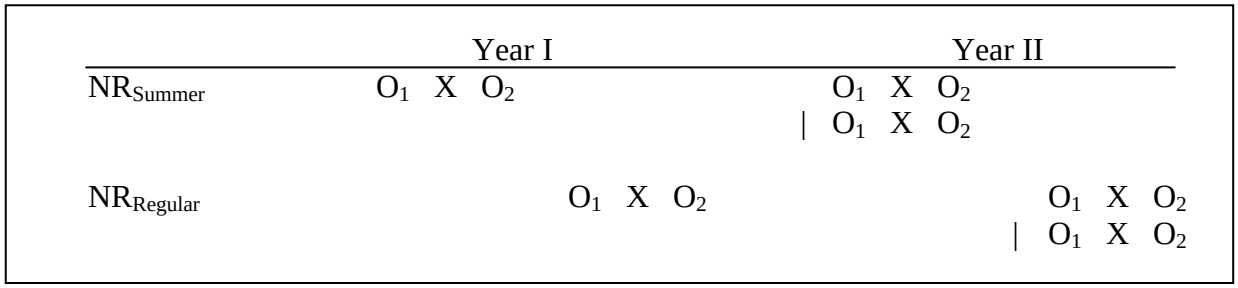


Figure 11. Notation of Research Design

This research design leverages the existence of multiple cohorts to test whether measurable conceptual changes occur during the course of the curriculum in similar ways across groups, thereby bolstering the strength of findings through replication.

The CUE research team examined student conceptual change related to the curricular intervention using two main student outcome measures. The first measure is a student scaled estimate based on a set of interview questions which were scored to serve as items and analyzed using a multi-dimensional IRT model. The second measure is a student maximum score, calculated as a student's most sophisticated response to any item. The two measures serve different functions, allowing analyses of both overall improvements in student reasoning, as well as more detailed pre-post shifts between particular "best" performance levels within the learning progression.

For the purpose of pre-post analyses, I included all data for the IRT analyses to extract as much information as possible for item and student score estimation, but only compared the cohort means for students who had both pre- and post-scores for the aggregated charts. These data include students from two sites who fall into one of two categories: "novices" who engaged with the curriculum for the first time in either YI or YII, and "returners" who repeated participation with a different module of the curriculum in YII, as previously described in Chapter 2.

The dataset for comparing cohort means includes 103 students in second and third grades grades with matched pre-post interview codes; 54 were from a summer school program taught by

members of the research team, and 49 were from regular school classrooms in which teachers adopted the curriculum. As noted above, approximately thirty percent of the YII cohort was made up of returning students. Our dataset included a subset of those, because not all cases had valid pre- and post-interviews, and not all data was coded (see Table 16). These data on returning students enable study of learning retention and the potential for ongoing conceptual growth over time.

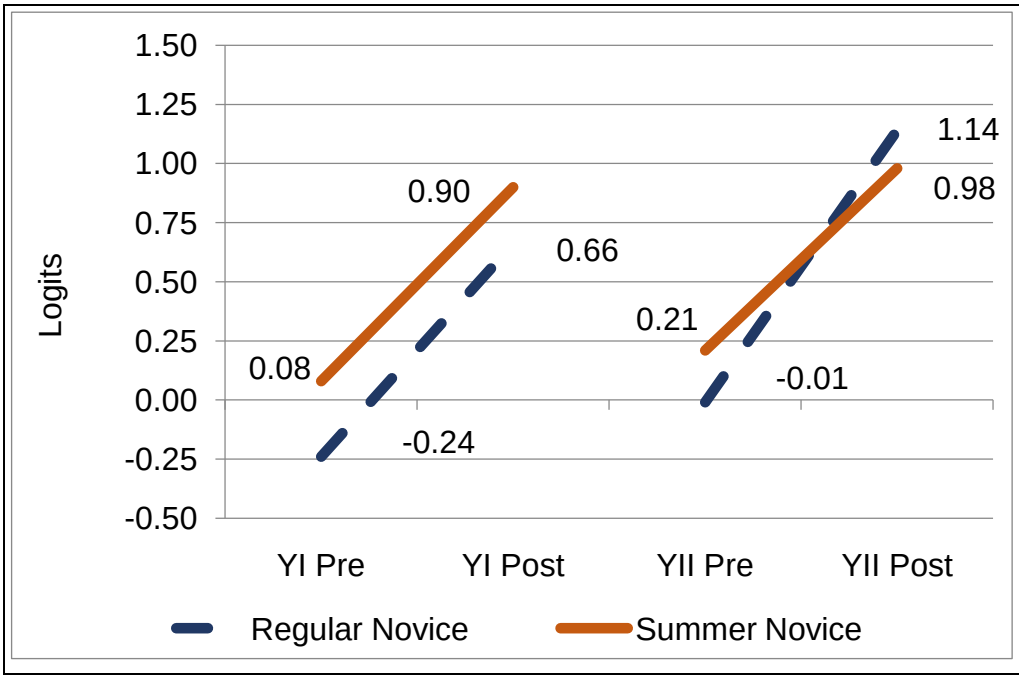
Table 16
Students in Dataset by Cohort

Site	YI	YII	
	New students	New students	Return students
Summer school	n=32	n=22	n=12
Regular school	n=27	n=22	n=11

Scaled Scores

To calculate scaled scores, Item Response Theory (IRT) suited both the project's philosophical approach and the complex instrument design. The use of probabilistic estimation to calculate a range of likely student scores addressed concerns about the impossibility of assigning any student a valid discrete conceptual score. This philosophical impossibility is rooted in the many reasons why students' raw responses cannot fully represent the complexity of their thinking, and our belief that knowledge itself is inherently varied across situational factors and constantly subject to change. Further, the item set varies between YI and YII as shown in Table 2 (see Chapter 2 above). This complexity can be handled by the probabilistic nature of item response models which yield student scores that are not instrument- or sample-dependent. For instance, the probability of a student choosing the correct response for any given item depends on the distance between that student's ability level and the difficulty of that item (Wilson, 2005, p. 94). Considering our intent to investigate pre-post change, IRT also enables latent regression (Adams, Wilson, & Wang, 1997) in which the item difficulties and student scores can be modeled while also estimating the effects of independent variables, such as the effects of school site, year, or pre-post administration as described in more detail below.

The CUE items had multiple response categories which required a partial-credit IRT model to estimate measurement statistics and EAP (expected a posteriori) scores. This model produced student scaled estimates for the Process dimension that enable aggregation by cohort to examine mean changes. Novice students at the two sites showed similar conceptual gains, ranging between a low of 0.77 logits for summer students in Year II to a high of 1.13 logits for regular school students in Year II (see Figure 12).



Cohort	Mean Logit Change
YI Regular	0.90
YI Summer	0.82
YII Regular	1.15
YII Summer	0.77

Figure 12. Pre-post Change in Mean Scaled Scores for Novice Students

This degree of change can be interpreted based on other information we gained from the development process described above about the interpretation of the logit scores. Referencing the Wright map (see Figure 13), we can see that the estimated distribution of student scores on the Process dimension ranges from -1.5 to 1.5 with a fairly normal shape (mean=0.00; s.d.=0.61). In that context, a 0.75-logit shift would be equivalent to crossing multiple conceptual levels according to our LP. For instance, a student with a mean estimate at -0.5 would typically respond below P4 on most items, suggesting a lack of understanding that traits vary and can help or hurt organisms. From this conceptual starting point, a 0.75-logit increase to 0.25 yields a position on the Wright map aligned with conceptual understanding above P5 for most items, suggesting an understanding that trait differences can affect the distribution of that trait in a population. Accordingly, the size of gains noted above is both large and educationally meaningful in terms of students' conceptual growth.

It is also useful to note that the 0.75-logit shift is meaningful in terms of the typical conceptual variation between students in grades 2 and 3. Examining scaled estimates for novice students at both sites, the average difference between second- and third-graders was 0.12 logits at the Summer School, 0.19 logits at the Regular School, and 0.16 logits when averaged across both cohorts. Accordingly, a change of 0.75 logits would reflect over four times typical student growth would be between second and third grades absent the intervention.

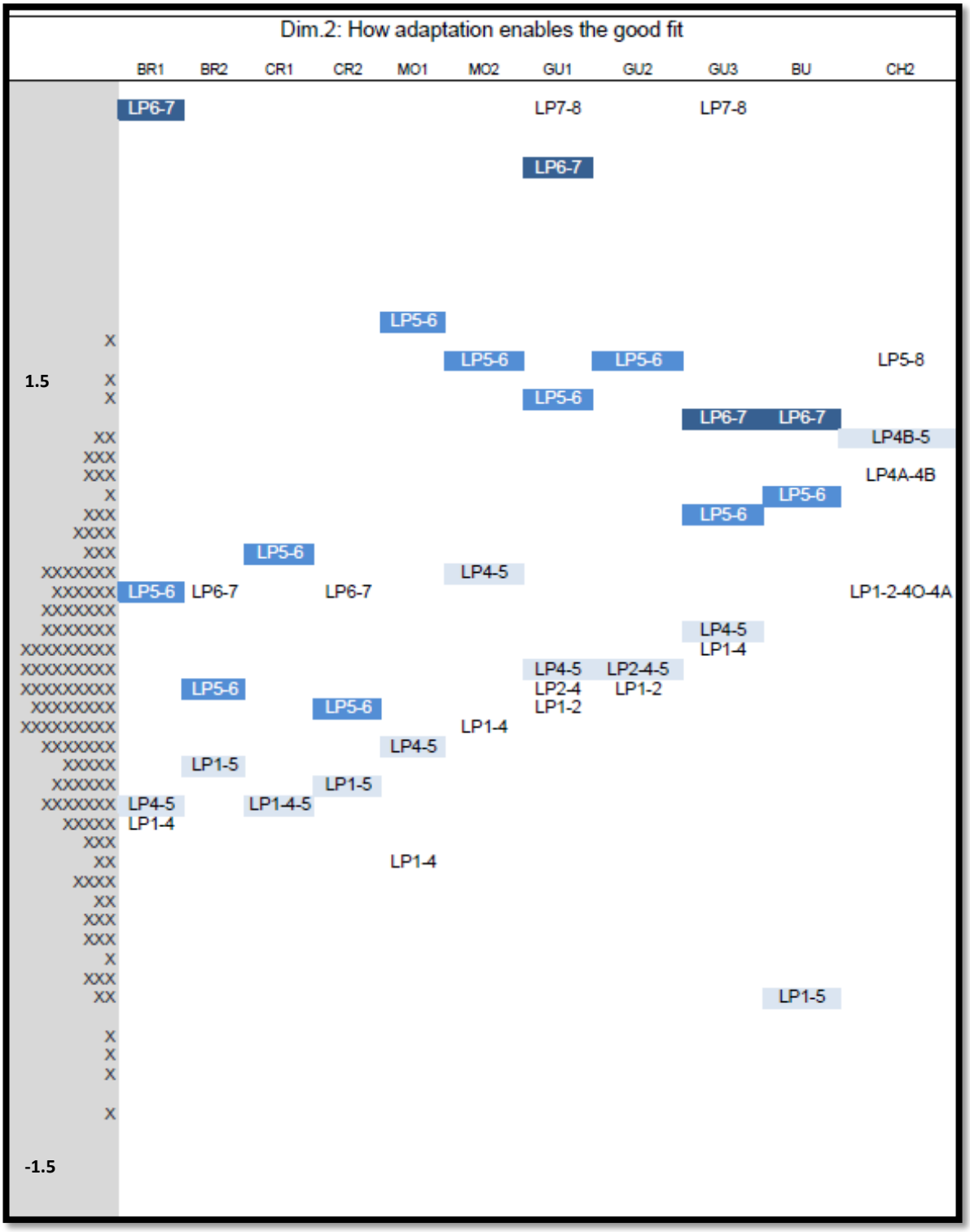


Figure 13. Wright Map for Process Dimension (D2)

The subset of students who repeated the curriculum in Year II presents an opportunity to consider how their growth changed over the four time points: YI Pre, YI Post, YII Pre, and YII Post (see Figure 14). YI gains persisted despite absence of instruction for periods ranging from 3 months for the regular school students, to 9 months for the summer school students.

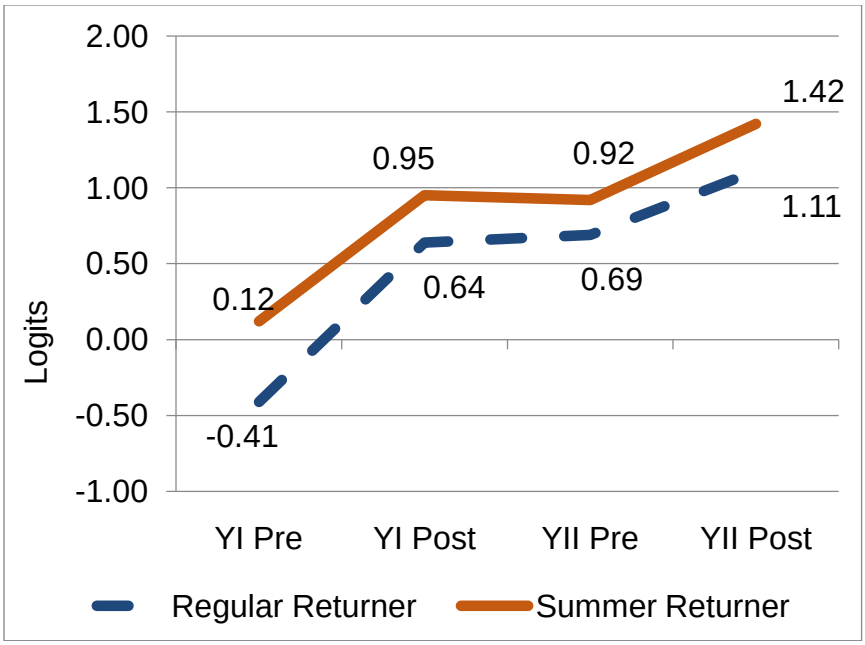


Figure 14. Pre-Post Change in Mean Scaled Scores for Return Students

Further, continued growth is evident in YII where students gained in meaningful ways (0.50 in the summer school and 0.42 in the regular school), although less dramatically than in YI (0.83 in summer school and 1.05 in regular school).

Table 17
Mean scores for returners by cohort over time

	Summer School		Regular School	
	Mean Student Estimate (logits)	Change	Mean Student Estimate (logits)	Change
YI Pre	0.12		-0.41	
YI Post	0.95	0.83	0.64	1.05
YII Pre	0.92	-0.03	0.69	0.05
YII Post	1.42	0.50	1.11	0.42

Latent regression enables a more precise examination of pre-post changes and the effects of other variables. This multi-level model simultaneously estimates scaled scores and regression coefficients to most accurately account for measurement error. To address our research questions about whether student scores improved between pre- and post-administrations, I designed a latent regression including effects for school site, novice-returner status, grade level, year of implementation, pre-post, and an interaction variable to show whether the rate of pre-post change varied by year. The results are shown in Table 18 below.

Pre-post denotes whether the score was for a students' pre or post assessment (0=pre, 1=post);
Regular School denotes the school site (0=summer, 1= regular);
Novice denotes whether student is new to curriculum (0=returning student, 1=novice);
3rd Grade denotes student school grade during intervention (0=second, 1=third);
Year II denotes which year (0=YI, 1=YII);
Pre-post x Year denotes an interaction between pre-post and year to study whether the rate of pre-post change varied by year.

Table 18
Latent Regression of All Cases Controlling for Variables of Import

CONSTANT (0.00)	Coeff	s.e.	T	Significance
Pre-Post	0.83	0.10	8.44	*
Regular School	-0.27	0.07	3.79	*
Novice	-0.48	0.10	4.70	*
3rd grade	0.22	0.07	3.08	*
Year II	0.14	0.10	1.40	
Pre-Post x Year	-0.02	0.14	0.13	

*Significant at α=0.01.

The results show an average pre-post change of 0.83 logits (standard error=0.10), an effect that was statistically significant where $\alpha = 0.01$. The following effects were also statistically significant:

- Summer school students scored 0.27 logits higher than their regular school peers. This effect is consistent with school differences as described above in Table 15.
- Returning students' pre-scores were 0.48 logits higher than their novice peers on average. Because all returning students were also 3rd graders, this effect may be due to either returning to the course OR being one grade older. (See novice-only results in Table 19.)

- Third grade students scored 0.22 logits higher than second grade students. Because a portion of the third-grade students included all repeaters, this effect may be due to the increased exposure for that group (See novice-only results in Table 19.)

As noted above, all returners in YII were also 3rd grade students. In order to differentiate between the change in scores that is due to grade level versus returner status, we can examine the mean grade effects for novices only. Given that comparisons within the novice cohort can provide a better estimate of the grade effect, we can use that effect to reason about how grade and returner status contribute to the YII effect.

Compared with the previous model, the novice-only latent regression results (see Table 19) yield nearly identical coefficients and standard errors for Pre-Post, School, and Grade. These effects are statistically significant at α=0.01. The effect of 3rd grade level increases from 0.22 in Table 18 to 0.26 in Table 19, suggesting that the inclusion of repeaters above created a minor negative bias.

Table 19
Latent Regression of Novice Cases Controlling for Variables of Import

CONSTANT (0.00)	Coeff	s.e.	T	Significance
Pre-Post	0.84	0.10	8.57	*
Regular School	-0.27	0.08	3.50	*
3rd grade	0.26	0.08	3.21	*
Year II	-0.06	0.11	.58	
Pre-Post x Year	0.31	0.15	2.09	**

*Significant at α=0.01. **Significant at α=0.05.

The effects for Year II and the interaction effect (Pre-Post x Year) changed in meaningful ways. The Year II effect declined from 0.14 in Table 18 to -0.06 in Table 19, suggesting that the inclusion of repeaters created a positive bias, which we would expect due to their greater exposure to the intervention. However, this effect is not statistically significant. The interaction variable, Pre-Post x Year, measures the change in pre-post gains between YI and YII. This effect is now positive and statistically significant at α=0.05 (coefficient=031; standard error=0.15), which supports our theory that the improved curriculum and more experienced teachers in YII would have a greater impact. Because we included only novices in this model, we can rule out the impact of returning students in boosting YII scores.

Omitting the returning students in YII enables more clear interpretation of the effects of Year and Pre-Post x Year. These statistics provide further evidence that the intervention did create statistically significant pre-post gains, that school scores reflect what we know about proficiency

levels between schools, and that the rate of pre-post gain in YII was in fact statistically significant when compared to YI.

Comparing the coefficients from these latent regression models also highlights how the inclusion of returning students in YII affects other variables. The change in coefficients for YII declined, because the returner students boosted the mean score for YII. After omitting this positive bias in YII mean scores, the interaction variable Pre-Post x Year shows a coefficient of larger magnitude than School or Grade, as well as statistical significance as noted above. In Figure 14 we can see that the mean returner scores were relatively stable between intervention periods and increased less in YII. Accordingly, the latent regression findings here reiterate that trend, and provide evidence that the returner effect was meaningful enough to mask the statistical significance related to the rate of pre-post change.

In addition, differences between the two latent regressions show that the effect of being in 3rd Grade does not change by omitting returners. This finding offers additional evidence about the impact of grade level regardless of returner status or other variables, and also supports our interpretation that typical conceptual development between Grade 2 and Grade 3 approximates a 0.26-logit gain on our scale.

The analyses above provide evidence that the intervention did in fact affect student learning, and that student learning was related to other variables as we would expect. They also affirm the ability of the coded interview to measure conceptual levels with adequate power to study conceptual change.

Maximum Scores

Less common in pre-post analysis, I also analyzed the maximum learning progression (LP) level students achieved across all items. These maximum scores were particularly useful given our reliance in interview data. There are multiple reasons why examining the maximum scores might reveal meaningful information given this research context:

1) In the conversational context of interviews, people typically display a tendency to offer less-elaborate explanations after fully articulating their ideas earlier (Grice, 1975). If students conform to Grice's maxim in this case, the sophistication of their responses would decrease after an initial high.

2) A primary intention of the CUE curriculum was to identify students' **capacity** for reasoning about micro-evolution given optimal instruction. Accordingly, the maximum score represents student's optimal reasoning. A main criticism of this assumption relates to the role of error; if one student's response to one item is unexpectedly higher than their responses on the other items, it is most likely due to error. Yet, this argument is complicated due to both the known conversational tendencies noted above, and also multiple known sources of variation in item difficulties across items within the CUE interview protocol. Some items were designed to measure concepts that were explicitly taught in the course, and others intended to challenge students to apply concepts to other examples with varying degrees of conceptual distance (e.g., extrapolating from one animal type to another, or from an example in the animal family to one in the plant family). The interview method used in this case helps to minimize the potential positive bias of giving credit where it is not truly due, given that all codes were based on entirely student-generated responses.

The maximum scores did relate to the EAP scores as show in Figure 15. The maximum scores are more closely related to the EAP scores on the pretest, where Spearman's rho=0.81. On the post-test there was greater variability and Spearman's rho=0.64. The scatter plots show that maximum scores and EAP scores were more closely related at the lower end of the scale. As students achieved more sophisticated conceptual levels, the variation between EAP and maximum scores increased. The EAP scores take into account variation within a student's response pattern across all items, and they can accordingly differentiate better between students with similar max scores.

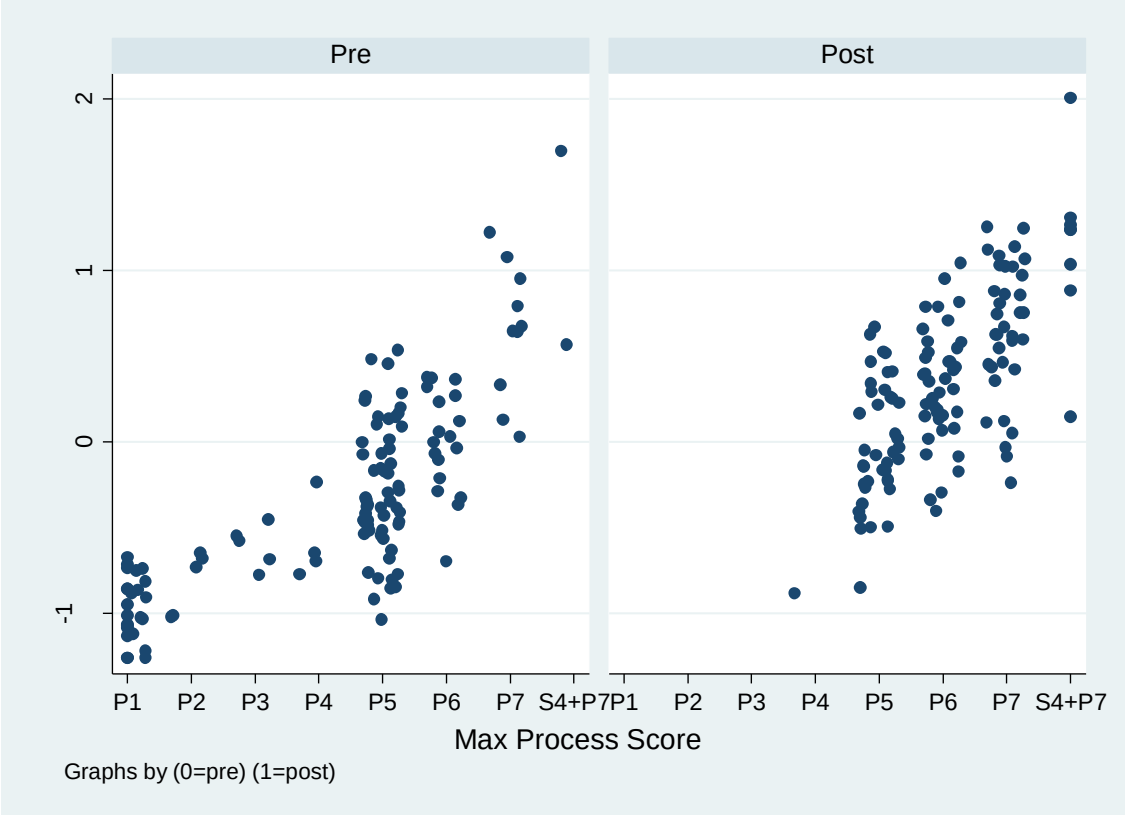


Figure 15. Scatterplot of Max Score and EAP Score by Pre-Post

Although the maximum scores do not take students' response patterns into account, they are uniquely capable to address questions about student capabilities. We use the maximum score as evidence of students' optimal reasoning, and our analyses provide evidence about how that optimal capacity varies over time. As shown in Figure 16 below, the maximum scores for novices spread over the range of learning progression levels, varying from P1 to P7 in the pre-interview, and from P4 to P7+S4 in the post-interview. The pre-post distributions of maximum scores for novices illustrate a shift in conceptual thinking, with insight to general conceptual benchmarks. The distributions consistently shift upward, and some students showing comprehension of the highest-level concepts (P7+S4) at post-interview.

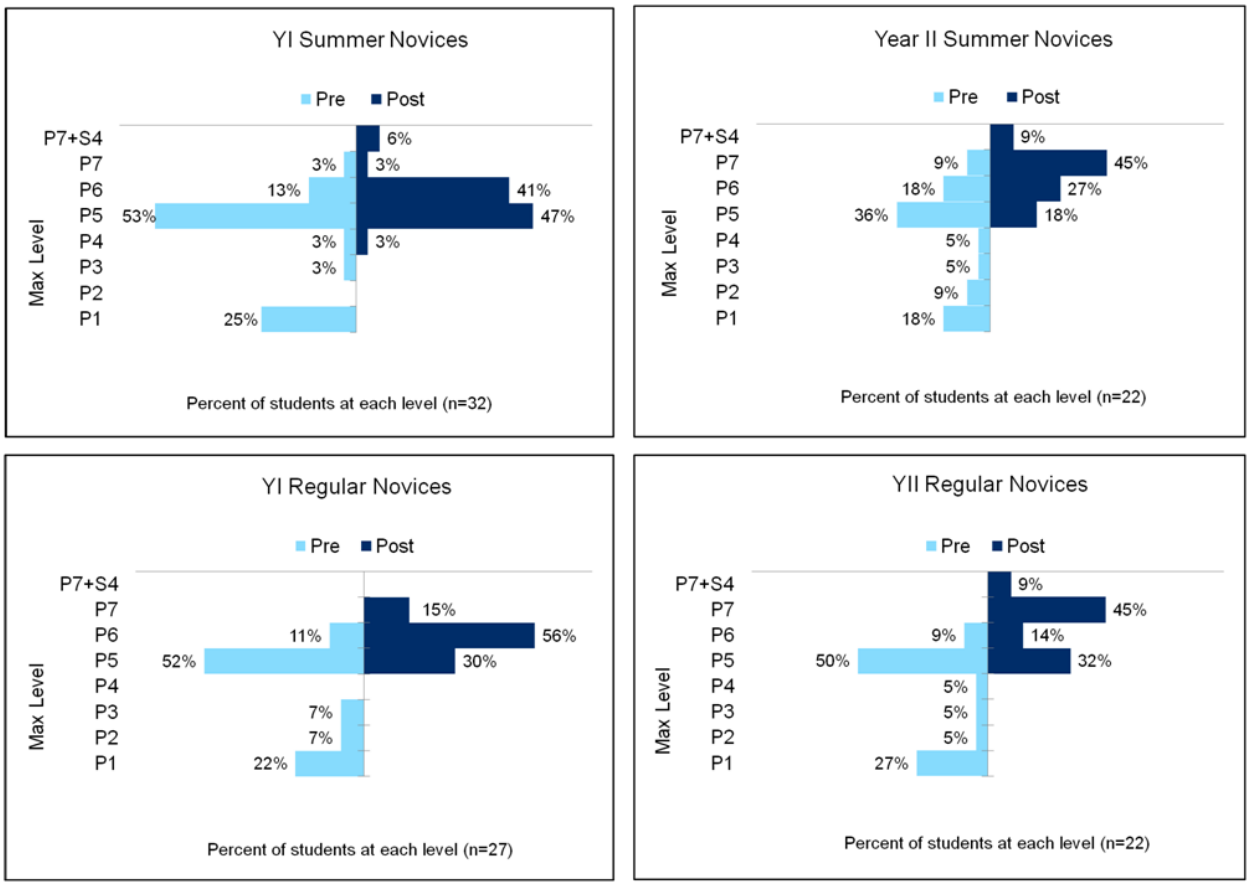


Figure 16. Pre-Post Change in Maximum Scores for Novice Students

To reach the top conceptual level suggests an understanding of a basic model of the natural selection process in the context of microevolution, by integrating lower-level concepts of survival value, and population shifts over generations, with the impact of environmental changes in micro-evolution contexts. The following excerpt from one student's post-interview response to the Guppy3 item presents a useful example of this conceptual level:

Interviewer: How are these guppies different from the ones in the pool 3 generations earlier?
Student 226: Well there were 4 colorful ones and now there is non-colorful ones.
Interviewer: How did that happen?
Student 226: Well the guppies ate the fish.
Interviewer: The guppies ate the fish?
Student 226: No, the fish ate the guppies. And the colorful guppies couldn't survive because these guys (points to predator fish) were maybe too quick for them, and these guys knew were these guys were because they were colorful. And I think before the predator fish came, the ones that had a survival advantage were the colorful ones cause they can a [sic] mates more easily.
Interviewer: Just tell me again, what do you have here then.
Student 226: Well there's lots and lots of the see-through ones because maybe one of the

14 see-through ones got eaten by these guys (points to predator fish) but it shouldn't be that much.

In this section of the interview, we hear the student reasoning about the survival value of coloration, and how that value changes with an environmental shift, the predator's arrival. Fish with bright colors may have previously had the benefit of attracting mates, but now have the burden of attracting predators, according to this student's reasoning. The student refers to both survival advantage and reproductive advantage, and the changing numbers in the population that result over time with a clear attribution of these changes to the environmental stimulus.

Examining these maximum score examples highlights the conceptual differences between responses, and shows how we might attribute conceptual locations to responses at different learning progression levels. This lens enables a pre-post examination of student scores that directly connects to the learning progression under study. Considering this pre-post change by year and program site, Figure 16 above shows consistent upward shifts in student's reasoning across cohorts.

While there are variations between cohorts, we see some common patterns. In each case, a notable portion of students showed evidence of reasoning at the P1 level, described as Negligible Understanding of micro-evolutionary concepts. In Year I, the percent of novices at this level ranged from 22% among regular school students, to 25% among summer school students. In Year II, those numbers were 27% and 18% respectively. In each cohort, the most frequently reported maximum score in the pre-interview was P5, capturing between thirty-six and fifty-three percent of total scores depending on the cohort. The P5 conceptual level requires evidence of understanding that survival value will affect the distribution of organisms with related traits in the current population and/or that survival value will affect the likelihood that an organism will reproduce. The two concepts provide a requisite foundation for advancing to the P6 level where students understand that survival value will affect the distribution of organisms with related traits in the future generation.

Of note, nearly all students expressed at least P5 reasoning in the post-interviews. Only 1 of 103 students did not reach P5 in the post-interview. In comparison, 57 of 103 scored maximum levels below P5 on the pre-interview. At post-interview, over 50% of students in each cohort achieved P6, the conceptual level where students are able to reason about how the prevalence of a related trait in the next generation's population will shift in response to environmental stimuli. While this conceptual level appears attainable at post-interview in YI, P7 was quite rare. Only 9% of summer students and 15% of regular students achieved this level or higher in YI. The conceptual challenge between P6 and P7, where student can reason about how population shifts will continue over larger periods of time, requires consideration of generations. The CUE team identified the trouble students had with this concept during YI and implemented instructional supports accordingly for YII. This curricular change might explain the larger upward shifts in YII, where over half of all students achieved at least P7.

The data described above do not include students who returned in YII and therefore experienced the curriculum twice. While the number of these "returners" with matched pre-post scores was low, only 12 summer and 11 regular students, the shifts in maximum scores are notable, as shown in Table 17. Not only did students who scored on the lower end of the learning progression on the pre-interview progress to at least P5 on the post in YI, all returners maintained that conceptual level through the YII pre-interview. Both regular and summer student cohorts experienced a period with no instruction between YI and YII interventions that lasted

approximately three months and nine months respectively. In all cases, YI gains persisted. Further, students continued to show increasing conceptual sophistication in YII, even after YI gains left them at the higher end of the learning progression. There is evidence of continued learning gains where the distributions of maximum scores continued shifting upward at both sites in YII. These data support the scaled score findings described above, and provide additional detail about how conceptual gains relate to the learning progression and the instructional activities. They also serve to validate the interview and coding scheme as a measure of students' conceptual growth, given that the learning gains evident in the classroom mirrored the shifts reported here in both scaled and maximum score measures.

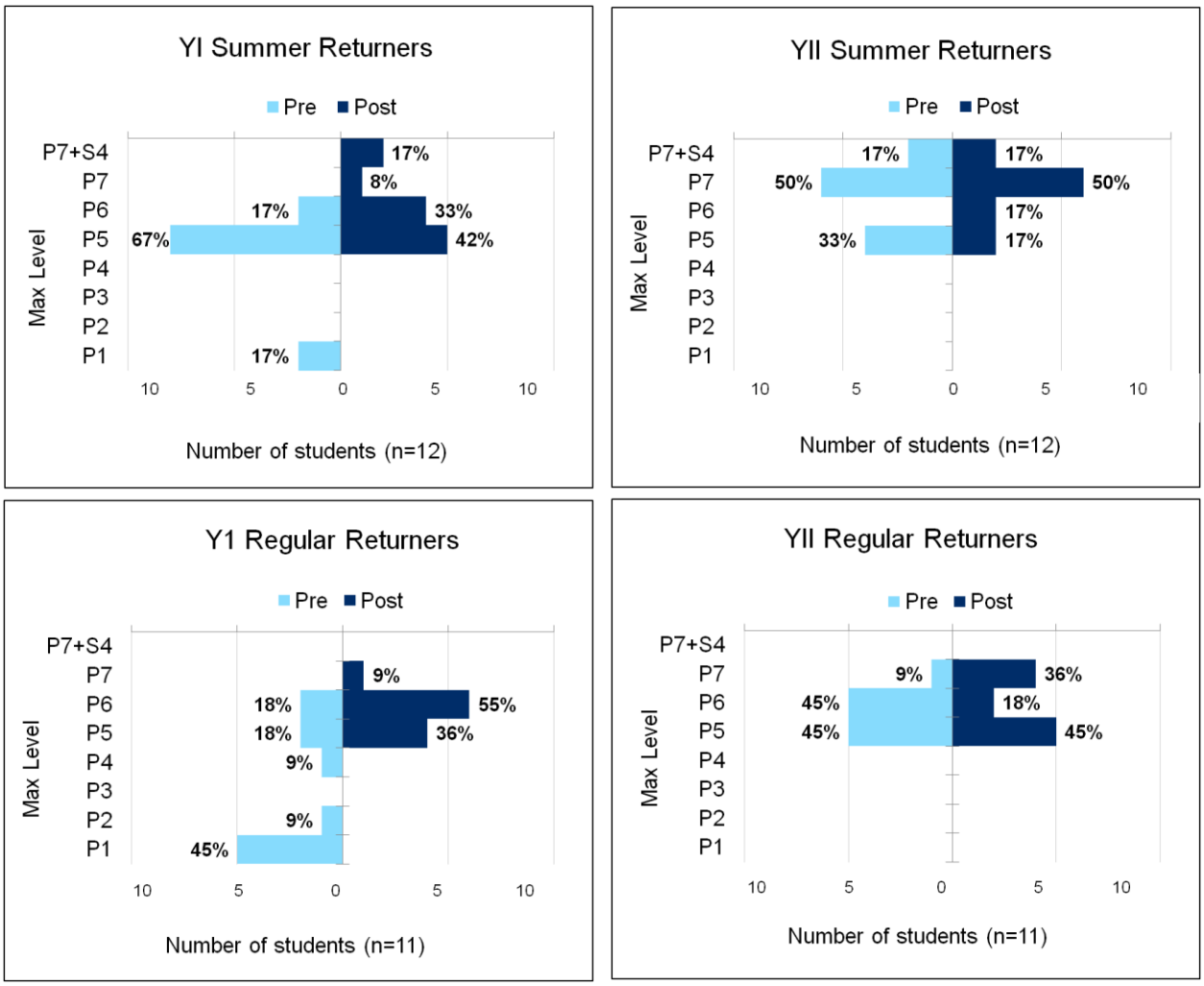


Figure 17. Pre-Post Change in Maximum Scores for Return Students

The pre-post analyses of both measures provide evidence in support of the underlying assumptions of the CUE project. The instructional intervention can affect student learning in both optimal (summer school) and practical (regular school) educational settings. This learning was measurable in terms of a complex learning progression coded both for a probabilistic model to calculate scaled scores and a maximum score analysis to show shifts in terms of conceptual

levels. Both the measures and intervention proved to be robust; we were able to calculate student scores with strong reliability and fit statistics that show gains across the entire range of learning progression level, and these conceptual gains were strong enough to persist over periods with no instruction.

Chapter 5. Discussion and Conclusions

This study provided evidence for the utility of applying item response theory in a novel context to promote validity and interpretability of measurement findings. Compared to closed-ended items which are commonly used to create student scaled scores, the interview format introduced increased variability in responses. This variability presented a challenge to aggregating responses for pre-post analysis in a meaningful way. The Bear Assessment System (BAS) approach (Wilson, 2009) applied in this dissertation served to optimize the connections between the learning progression, interview items, student responses, coding guide, and scoring scheme to ensure validity.

As illustrated in Figure 1, the learning progression framework served as a bridge between assessment and instruction over the course of the CUE project. During design of the learning progression, IRT analyses were used to test theories of how items related to learning progression levels, and how items functioned according to unidimensional or multidimensional theories. The related IRT statistics discussed above provided evidence about how student conceptions at particular progression levels might not advance in the sequential form of the original learning progression (e.g., LP3 was generally more difficult than LP4). The item fit analyses supported the CUE team's emerging theory about the complexity involved, and signaled the value of testing both uni- and multidimensional models. Ultimately, the triangulation of information from different analyses yielded a sophisticated version of the learning progression that best suits the CUE project's understanding of children's conceptual development in micro-evolution under particular learning conditions.

This study relied on triangulation. First, it incorporated expert knowledge, classroom observations, and the empirical study of student responses to produce a nuanced understanding of the learning progression theory. Later, it presented analyses of the item set's internal functioning, the relation of scaled scores to other variables, and the identification of pre-post changes to form a comprehensive validity argument. Analyzing new measures at both the instrument- and item-level enabled closer investigation of how the measure relates to both the empirical data and the design theory, especially around issues of adequate item spread (Do the items adequately cover the range of student responses?) and item difficulty (Does the likelihood of students achieving sophisticated levels on an item correspond with the underlying theory?). Making these connections in an empirical way strengthened the validity argument according to the contemporary standards described in Chapter 2.

The initial validity analyses and the later pre-post analyses provided support for one another. The use of IRT to analyze reliability and validity during the design phase of the project enabled us to interpret student conceptual levels. The pre-post findings relied on that earlier IRT work, as well as the application of IRT-dependent latent regression. This approach produced both dependable scaled scores for pre-post comparison, and regression coefficients that account for measurement and prediction error to describe how important variables affected student scores. Further, the clear upward shifts in students' mean scaled scores after exposure to the curricular intervention provide additional evidence that the methods for translating interview responses into scaled scores are in fact able to measure conceptual levels in this domain.

In this case, additional measures were useful for investigating conceptual change according to the CUE project's foundational research interest in student capacity. The CUE team also used maximum scores to address this topic, providing an example for researchers with similar perspectives. Using both scaled and maximum scores in this dissertation provided

additional detail about the nature of students' conceptual change. Also, the similar pre-post shifts that both measures showed serve to reinforce their validity for use in this context.

The research activities described above also stimulated collaboration on new projects for future development. There were multiple points at which the research team's methodological roots in both qualitative and quantitative methods challenged us to analyze and present data in novel ways. Related to pre-post change in maximum scores, the comparative horizontal bar charts were designed to most clearly reflect the students' distribution across progression levels at different points in time. Regarding the quantitative measurement analyses, the research team discussed the potential usefulness of augmenting the Wright map to visually represent the score distributions that the IRT analyses estimate for each student. The probabilistic nature of the IRT scaled scores is critical for accurately representing cognitive development theories that model student thinking as highly complex and affected by many contextual variables[3].

Our findings also relate to recent papers about the role of learning progressions in classroom assessment. Gotwals (2018) discusses learning progressions in terms of the importance of scope and grain size. These characteristics affect the potential usefulness of learning progressions for formative assessment. The CUE learning progression focuses on a distinct set of concepts in a particular domain, and describes conceptual growth across nuanced levels that are capable of shifting during periods of instruction. These affordances make the CUE progression potentially "likely to support inferences that teachers could use to support student learning" (Gotwals, 2018, p. 158). In practice, the CUE learning progression shows promise for addressing the key questions Gotwals suggests for teachers and students to consider (Gotwals, 2018, p. 159), which relate to the current state of understanding, the possible direction of conceptual growth, and how to close the gap and support student learning.

The CUE research team also considered future applications of the CUE learning progression to make related assessment more accessible. Given the high level of resources involved in interviewing large numbers of student and analyzing their responses, a written version would be more useful. Yet, among this young population, there are serious limitations to how well students can express themselves in writing. With these issues in mind, I designed a written in-class assessment that maps directly to the learning progression. As part of the CUE project, 65 students responded to this written assessment, and I will analyze the data in future analyses.

Is it possible to differentiate between levels with that type of assessment in this context? How well do the written scores correlate with the interview-based scores analyzed here? These research questions will inform the extent to which a written alternative might be useful in practical settings. Developing useful classroom assessment materials based on the CUE progression could also benefit student learning "by illuminating student thinking and by providing insights about instructional next steps" (Shepard, 2018). Applying our theoretical and empirical work to practical tools that might contribute to classroom teaching would further develop the "horizontal coherence" that the CUE project aimed to model (NRC, 2001). Ideally, cognition, instruction, assessment, and teacher learning are all based on a common learning theory. To this end, future analyses may provide a useful case study about how alternative assessment forms can represent learning progressions for both formative and summative purposes.

[3] Our early work on this topic will be presented at the *National Council on Measurement in Education* annual meeting in April 2019.

The findings presented above described a collaborative effort to integrate instruction and assessment based on the learning progression. The application of IRT played a major role in providing interpretable empirical feedback, with enough detail to inform the development of sub-levels within the progression. These analyses provided evidence that the learning progression accurately reflects student thinking and that the intervention did affect conceptual growth. Beyond the scope of this project, our work offers the potential to model effective analytical strategies for collaborative projects based on learning progressions.

References

Adams, R., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21 (1), 1-23.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Briggs, D., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, 4 (1), 87-100.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in evaluating complex interventions in classroom settings. *The Journal of the Learning Sciences*. 2 (2), 141-178.

California Department of Education. (2018). *DataQuest [Database]*. Retrieved from https://www.caschooldashboard.org.

Catley, K., Lehrer, R., & Reiser, B. (2005). *Tracing a prospective learning progression for developing understanding of evolution*. [Paper commissioned by the National Academies Committee on Test Design for K-12 Science Achievement]. Retrieved from http://www7.nationalacademies.org/bota/evolution.pdf.

CUE Research Team. (2011). *Coding Book 11-9-11*.

Gotwals, A. W. (2018). Where are we now? Learning progressions and formative assessment. *Applied Measurement in Education*, *31*(2), 157-164.

Hammer, D., & Berland, L. K. (2014). Confusing claims for data: A critique of common practices for presenting qualitative research on learning. *Journal of the Learning Sciences*, *23*(1), 37-46.

Lehrer, R., & Schauble, L. (2009). Images of learning, images of progress. *Journal of Research in Science Teaching*, *46* (6), 731-735.

Metz, K.E. (1995). Reassessment of developmental constraints on children's science instruction. *Review of Educational Research*. 65 (2), 93-127.

Metz, K.E. (1997). On the complex relation between cognitive developmental research and children's science curricula. *Review of Educational Research*. 67 (1), 151-163. [Reprinted in Smith, P.K. & Pellegrini, A. (2000). *Major writings in the Psychology of Education.* Vol. 111: The school curriculum. London: Routledge.]

Metz, K.E. (2000). Young children's inquiry in biology: Building the knowledge bases to empower independent inquiry. In J. Minstrell and E.H. vanZee (Eds.), *Inquiry into inquiry learning and teaching in science* (pp. 371-404). Washington, DC: American Association for the Advancement of Science.

Metz, K. E. (2006). The knowledge building enterprises in science and elementary school science classrooms. In *Scientific inquiry and nature of science* (pp. 105-130). Springer Netherlands.

Metz, K. E. (2008). Narrowing the gulf between the practices of science and the elementary school classroom. *Elementary School Journal*. 109 (2) 138-161.

Metz, K. E., Cardace, A. E., Berson, E., Ly, U., Wong, N., Sisk-Hilton, S., Metz, S. E., Wilson, M. (2017). *A learning progressions approach to the challenge of teaching evolution: Developing second and third graders' understanding of natural selection*. Under review.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.
National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. National Academies Press.
National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.
Quinn, H., Schweingruber, H., & Keller, T. (Eds.). (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.
Schwartz, R., & Ayers, E. (2011). *Delta dimensional alignment: Comparing performances across dimensions of the learning progression for assessing data modeling and statistical reasoning.* (Unpublished research paper). University of California, Berkeley, CA.
Shepard, L. A. (2018). Learning progressions as tools for assessment and learning. *Applied Measurement in Education, 31*(2), 165-174.
Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.
Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal for Research in Science Teaching*, 46 (6), 716-730.
Wilson, M. (2012). Responding to a challenge that learning progressions pose to measurement practice: hypothesized links between dimensions of the outcome progression. *Learning progressions in science: Current challenges and future directions.* Rotterdam: Sense Publishers.
Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied measurement in education*, 13 (2), 181-208.
Wu, M.L., Adams, R.J., & Wilson, M. (1998). *ACER ConQuest* [computer program]. Hawthorn, Australia: ACER.

Appendix A. Interview Protocol Excerpt for Brassica Item

2A. HAIRY BRASSICA PLANTS (for kids in botany class *only*)

Materials

Part I:

- Image of hairless Brassica
- Image of hairy Brassica
- Field with 16 hairless Brassica & 4 hairy Brassica
- Caterpillar icons
- Loose set of 16 hairy & 16 non-hairy.
- Template of field

Introduction:

A scientist went to a field to study a kind of plant called a Brassica. [Show image of field with 16 hairless Brassica & 4 hairy Brassica]

[S] he noticed something really interesting about the Brassicas in these fields. Almost all of them had very smooth stems: they looked like this. [Point to some smooth-stemmed Brassica]. But a small number of them had hairy stems: They looked like this. [Point to hairy-stemmed Brassica].

Environmental Press: Introduction of the Caterpillars

One day the scientist noticed some caterpillars come into the field! [Add two sets of caterpillars to the image]

The caterpillars found the Brassicas and began to eat them!

So which Brassicas do you think the caterpillars will be more likely to eat? [Wait for response]

If S thinks they'll be more likely to eat the smooth ones:
Yes
If S thinks they'll be more likely to eat the hairy one
Actually:

the ones with the smooth stem! The hairs on these Brassicas [point to the a couple hairy stemmed plants] protect the plant from the caterpillars, so they are less likely to be eaten.

2) Prediction of Generation 2 With icons

a) Request for prediction

The scientist had to leave the field with the Brassica plants and the caterpillars.
A year later, [s]he want back to see what the Brassica plants looked like.

The scientist knew the *same* plants wouldn't be living anymore, because Brassica plants don't live that long. You are one generation after your parents. The seeds from a plant

grow up into the next generation of plants. Since [s]he went back a year later, [s]he knew [s]he'd see the next generation of Brassica plants, the children of the Brassica's [s]he saw when [s]he first got there.

What do you think the Brassicas looked like one generation later, after the caterpillars got there? Use whatever you need of this stuff to show what you think the Brassica plants looked one generation after the caterpillars got there.
[Provide set of 16 hairy & 16 non-hairy Brassicas; the field template with two caterpillar icons pasted on]

b) Explaining prediction, using icons

What do you think the Brassicas looked like one generation later, after the caterpillars got there? Use what you did here to explain to me what you think will happen.
If the child does not quantify icons,

I want to make sure I understand your idea. What do you have here?. If S wants to adjust numbers as [s]he is in the process of quantifying, allow. The issue here is child's intention, not necessarily an accurate counting thereof. For example, if the child intended to put out 13 hairy plants & 3 non-hairy, that what is important, not the fact that they actually put out 12 & 4]

If just one template

Will they all look like this?

3) Explanation of Mechanism of the change

So when the scientist first came, the Brassicas looked like that [point to original state]. And you think that the Brassicas in the next generation will look like that. [point to S's arrangement of icons].

How will that happen?

>>>>>PROBE<<<<<

Tell me more! I'm really interested in your ideas!
Tell me more! I want to make sure I understand your idea.

Appendix B. Coding Book Excerpt with Examples for Cheetah Item

EXAMPLAR LP CHEETAH RESPONSES

LP Level	**Examples from Cheetah Task** **Explaining change: How come cheetahs today are so much faster than their ancestors?**
1. Negligible understanding of the fit between organisms & their environments	"Because the ones who lived thousands of years ago, they got older and older on their birthdays... Sometimes when people or things grow, they can't run that fast… When they get old they can't run that fast anymore" (224BR, 2010, Pre).
	"They probably communicate and teach themselves" (036PA, 2009, Pre).
	"The parents were born to run fast and their ancestors' parents weren't born to run fast" (024AH, Post, 2009).
	"they all kept on trying so they got their prey easily … They just did it and they started to run faster" (202JA, 2010, Post)
2. Get needs met: Organisms live where they live or do what they do to get their needs met/ *extent to which they get what they need*	"They caught more food, that gave them more energy to run more faster… (008AN, 2009, Post)
	" There wasn't just that much food and they couldn't run as fast… There's more food now" (210MI, 2010-Pre)
	"They still get energy from their food, and more and more energy, and more and more air, and more and more water makes them really run fast… (010MI, 2009, Post).
3a Some structure(s) of the organism-kind enables *particular function or behaviors.*	"Because they have changed, for many years because probably their legs got longer, well kind of long. … Actually their body became longer and longer to jump, to run faster 'cause it kind of liked jump, jumps so then they'll get … Cause they've grown. They might have grown, they grow one part of their body and if it gets bigger and bigger it'll allow them to run faster and faster." (204AN, 2010, Pre)
	"Because they have stronger muscular legs than their ancestors" (007LE, 2009, Post)
4 Different traits of the individual can hurt or help its chances to survive [under particular environmental conditions]	"… the ones that run about 20 miles, they can't even catch a gazelle, so they had less prey. They might have died because they had nothing to eat but they can drink water from other places. But some of them might have survived, because they can blend in with the grass and jump up on them and attack them" (217MY, 2010, Post)
	"Probably the ancestors didn't have a lot of energy to chase after their prey. That's why they had to go 20mph and couldn't try to run as fast as their great grandchildren. The ones that can run 60 mph as fast as a car they have a better survival advantage to live than their ancestors had, because they can find more meat and gazelles and other animals that have meat" (215GR, 2010, Post)
	FASTER… CATCH THEIR PREY…. 226 JE

5A Those individuals with a trait that gives them an advantage are more likely to live long enough to produce offspring	No Cheetah exemplars for 5A were found in the interviews.
5B Among organisms of the same kind living together, traits that help their chance to survive tend to become more common: Same generation	No Cheetah exemplars for 5B were found in the interviews.
6 In the next (offspring) generation of the population, inherited traits that help its chances to survive & reproduce tend to become more common & those that hurt its chances, less common.	No Cheetah exemplars for 6 were found in the interviews.
7 Over many generations in the population, inherited traits that help an organism's chance to survive and reproduce become more common & those that hurt its chances, less common.	"Probably because it reproduced and it had one who can really run fast and that one would reproduce and reproduce and then the ones today can run fast… Their ancestors when they reproduce they would actually they would sometimes have a baby who *would,* who would run fast and it can reproduce and reproduce and then the ones today can run fast" (234FA, 2010, Post)
	"Because there might have been some faster ones, like ones that are a teensy-weensy bit faster than these (points to picture of cheetahs). And their babies keep getting faster and faster and faster. That's what Paul Williams said with brassica. He selected fastest ones, and they kept getting faster and faster. Like if they were 20mph, he got, there were 32 days to flower, and he got them to 15 days. And he's seen ones that flower in 13 days. (Interviewer probes: But there wasn't a scientist picking out the cheetahs that ran really fast) "That's artificial selection but in this case, it's natural selection. That means the earth picks out the fastest one, and they have babies, and they keep getting more and more and more .. they keep getting more and more and more and more and fast. From 20 to 65 miles per hour)" (014RU, 2009, Post)

	S:Are there any fast cheetahs? Interviewer: Back in the day? Um…I don't know. S:Cause that would make it way easier. Interviewer: Explain why, what are you thinking? S: Cause like, if there was one cheetah, or actually two would be better, two fast running cheetahs and they mated and then they had a baby, that baby would be a fast running cheetah. And then if that baby mated another cheetah, that would be so it would keep on going until there were like, a lot of cheetahs. (005EL, 2009, Post)
8 Natural selection leads to organisms that are adapted to where they live.	"It's kind of like the sparrows, they adapt to where they're living, like this gazelle, it can run up to 40 or 50 mph maybe, I'm not sure, so the cheetahs that can run 20mph, it couldn't catch its prey, if it couldn't catch its prey, it would die out because it couldn't catch its prey and it wouldn't have food. (Interviewer probes) There could be some fast ones in a cheetah herd and if a fast one, like the sparrow, if the fast one married a fast one, it would have a fast child. If that fast one married a fast one, they would have an even faster one, and if that even faster one married an even faster one, they would have an even fasterest cheetah. And so eventually there was a really fast-running cheetah and since that cheetah could survive more easilier because they can run more fast, that kind of cheetah would be adapted to live in there and so there will be more and more of that and eventually the slow-running cheetah would die out." (214RU, 2010, Pre)