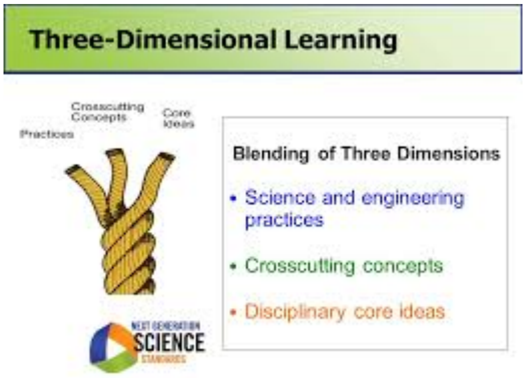
INTRODUCTION

Background on the project and NGSS based assessments that led to the investigation of using machine learning scoring

Carbon TIME (Transformations in Matter and Energy) is a design-based implementation research project (Cobb et al, 2003; Fishman et al, 2013; Mohan, Chen, and Anderson, 2009) that includes curriculum and assessments focused on transformations of matter and energy in living systems and Earth systems at multiple scales, from atomic-molecular to global. Its foundation is learning progression research analyzing the development of student practices as they investigate and explain these systems (Jin and Anderson, 2012; Jin, Zhan, & Anderson, 2012; Mohan, Chen, & Anderson, 2009). As design-based research it focused on both how students learn and how to support that learning within the ecology of educational systems. (Fishman et al, 2013). Additionally, Fishman et al. (2013) stress that design based educational research is iterative with cycles of invention and revision based on evidence gathered in learning systems.

The current reform movement in science education encompasses important changes in science education standards. Science education standards have changed a great deal over the last several decades. The National Science Education Standards (NRC, 1996) and similar standards of the time (NAEP, PISA, etc) made a move towards including science inquiry (and sometimes the nature of science) as a separate but important component. Assessments based on these standards often included specific question targets, such as point totals, for these domains in the blueprint (Ohio Graduation Test Blueprint, 2006; Florida Department of Education, 2019). This was an important first step away from teaching and assessing science primarily as a body of facts, towards an emphasis on science as a sense-making endeavor that used common practices to make claims about what happens in the natural and designed world and why. However, these inquiry practices were often taught and assessed in isolation from science content. The Next Generation Science Standards (NGSS, NGSS Lead States, 2013) and NRC Framework (NRC, 2013) shifted from teaching inquiry and content as separate but equal to an integrated three-dimensional framework, including (a) science and engineering practices, (b) crosscutting concepts, and (c) disciplinary core ideas.

The NRC Framework (2012) stated that: “To support students’ meaningful learning in science and engineering, all dimensions need to be integrated into standards, curriculum, instruction, and assessment.” (pg 2). In fact, this notion is so prevalent, that the image of the three strands as intertwined as shown in Figure 1, not only in instruction but also in assessment is common.



Three-Dimensional Learning

Blending of Three Dimensions

- Science and engineering practices
- Crosscutting concepts
- Disciplinary core ideas

Figure 1 NGSS as three dimensional

Need for Three-Dimensional Assessments

The shift to a three-dimensional framework requires a new kind of assessment. Traditional science assessment, particularly at scale, relied heavily on forced-choice tasks. As stated in the Framework: “Assessments of this type can measure some kinds of conceptual knowledge, and they can also provide a snapshot of some practices. But they do not adequately measure other kinds of achievements, such as the formulation of scientific explanations or communication of scientific understanding…or engaging in scientific argumentation.” (pg 212).

So, new assessments will require item types that have not been frequently used in large scale assessment such as constructed response, composite items, and performance-based tasks. To adequately assess standards based on the NRC Framework and NGSS, three-dimensional science learning, constructed response or composite items that combine some forced choice with constructed response portions are necessary to measure some of the key science and engineering practices such as scientific argument and communicating scientific ideas. While forced-choice items do present evidence of some of these skills and practices, they are unable to adequately asses all of the knowledge, skills, and abilities within the targeted practices (NRC, 2014).

Thus the assessment system should elicit and evaluate student performances that integrate all strands together to give the student a holistic score that is based on complex integrated performances. The more that we assess students integrating the three strands together, the stronger the claims that we will be able to make about three-dimensional learning.

Needs and Challenges for Machine Scoring

However, human scoring of constructed response items or composite items can be time consuming and costly (Williamson, Bejar, and Mislvey, 2006; Mao et al, 2018). “In short, from a pragmatic point of view a key goal in any assessment is to maximize construct representation at the least cost. Human scoring might be the way to achieve that goal under some circumstances, but not in others.” (Bejar, Williamson, and Mislevy, 2006. Pg 53) Given the desire of stakeholders, such as states, to keep costs down for assessments of NGSS (Gorin and Mislevy, 2013; Toch, 2006), a more cost effective solution is necessary. The NRC report on Assessing NGSS (2014) suggests that technology may supply answers to some of these problems through the use of simulations, tech-enhanced items, and other emerging technologies. The most likely emerging technology for scoring constructed response questions is using machine learning to classify students’ responses along a learning progression. (Gambrell, Thomas, Meisner, and Bolender, 2016; Thomas, 2017; Thomas, Kim, & Draney, 2018).

Assessing NGSS poses an especially difficult problem. Generic machine learning scoring engines such as those used to score argumentative essays regardless of context (Attali, Bridgeman, & Trapani, 2010; Shermis, 2015; Shaw, Meisner, & Bolender, 2019) cannot be used for these assessments because whether or not the claim is scientifically correct is at least as important as the structure of the argument. So, machine scoring techniques need to make separate scoring models for every task or scenario.

The NRC Report on Assessing NGSS (Pellegrino, et al., 2014) made several recommendations that directly relate to the design of the curriculum, assessments, and machine scoring systems related to *Carbon TIME* (NRC, 2014). Recommendations included concerns about: carefully designed tasks with evidence collected to support claims; tasks that demonstrate students engaging in all three practices that are aligned to the Framework; assessments that use technological innovations for administration and scoring; assessments should be able to identify students along a learning progression; and that systems be built that consider the constraints of cost.

Conclusions 2-1, 2-3, 2-4, 2-4, 4-1a, and 4-2 deal with item development, design, and deployment. One of the key facets is an understanding that the Performance Expectations represent one example of how to validly assess the combination of three strands around a DCI. Another set (2-2, 4-1b, 4-1c) of conclusions revolve around identifying where students are along a learning progression.

Iterative Design of Assessment in *Carbon TIME*

The *Carbon TIME* project began developing assessments of students' integrated performances in 2007 and went through yearly revision cycles through 2018. The initial learning progression work focused primarily on students' explanation practices (Jin & Anderson, 2012; Mohan, Chen, & Anderson, 2009). Starting in 2015, the assessments were based on learning progressions in three strands, focused on explanation practices, investigation practices, and reasoning about large-scale systems (Covitt & Anderson, 2018). The first phase of the project (2007-14) relied on human scoring of students' constructed responses. Machine learning and machine scoring were central to the second phase (2014-18). We briefly describe the first phase below, then focus on the second phase for the remainder of this article.

Phase 1: Iterative development based on human scoring. In order to tap complex constructs that involve all three dimensions of science learning and instruction, (science and engineering practices, crosscutting concepts, and disciplinary core ideas) composite items were developed prior to any attempts at machine scoring. The most successful assessment tasks used a mix of forced choice and constructed response components to elicit evidence of what students know and can do in relation to the learning targets. [Jin and Anderson, 2012; Mohan, Chen, & Anderson, 2009].

Each task included one or more constructed response parts that were scored by human coders using a rubric that (a) assigned student responses to a learning progression level and (b) identified the specific elements in the student response that provided evidence for the learning progression level code. These indicator codes could theoretically provide valuable formative insight for teachers and curriculum developers. If several students had the same sub-code, a teacher could use that information to make choices in instruction or remediation that would address that particular alternate conception or reasoning that could be addressed.

However, because of limitations of human scoring in cost and time, only a portion of the student data was analyzed, and not until after the students had finished the course. So, the value of the sub-codes for formative assessment was lost; however, the patterns found in the data for the case study teachers was able to be used to revise the curriculum and support materials (Doherty, et al, 2015).

Phase 1 was based on the BEAR assessment model: an iterative process which includes editing items, rubric construction and alignment, coding, IRT, and developing a shared understanding of the reasoning that students at each level are relying on to produce their answers (Yao, Berson, Ayers, Choi, & Wilson, 2010). Reliability checks in human coding led to revisions of items and rubrics. Items and rubrics that had unusual IRT statistics or bad item fit were identified by the psychometric staff and revised or discarded. However, because only a few hundred student responses to each item were being scored, the power of psychometric analysis was not being fully utilized. As design based research a new invention and iterative cycle would be necessary to scale up the assessment process.

Thus the first phase of *Carbon TIME* project addressed several of the issues identified in the NRC Report on Assessing NGSS (Pellegrino, et al., 2014). However, several recommendations (6-3, 7-7) deal with the use of emerging technologies to address issues with scalability and cost. Human scoring for authentic three-dimensional assessment tasks was too costly and slow to meet the design and implementation goals.

Phase 2: Iterative development including machine learning and machine scoring. During the second phase of the project (2014-18), machine learning (ML) and machine scoring were included in the development cycles, as illustrated in Figure 2 below. The blue boxes in Figure 2 depict steps in the Phase 1 development cycles; the green boxes depict ML steps added during Phase 2.

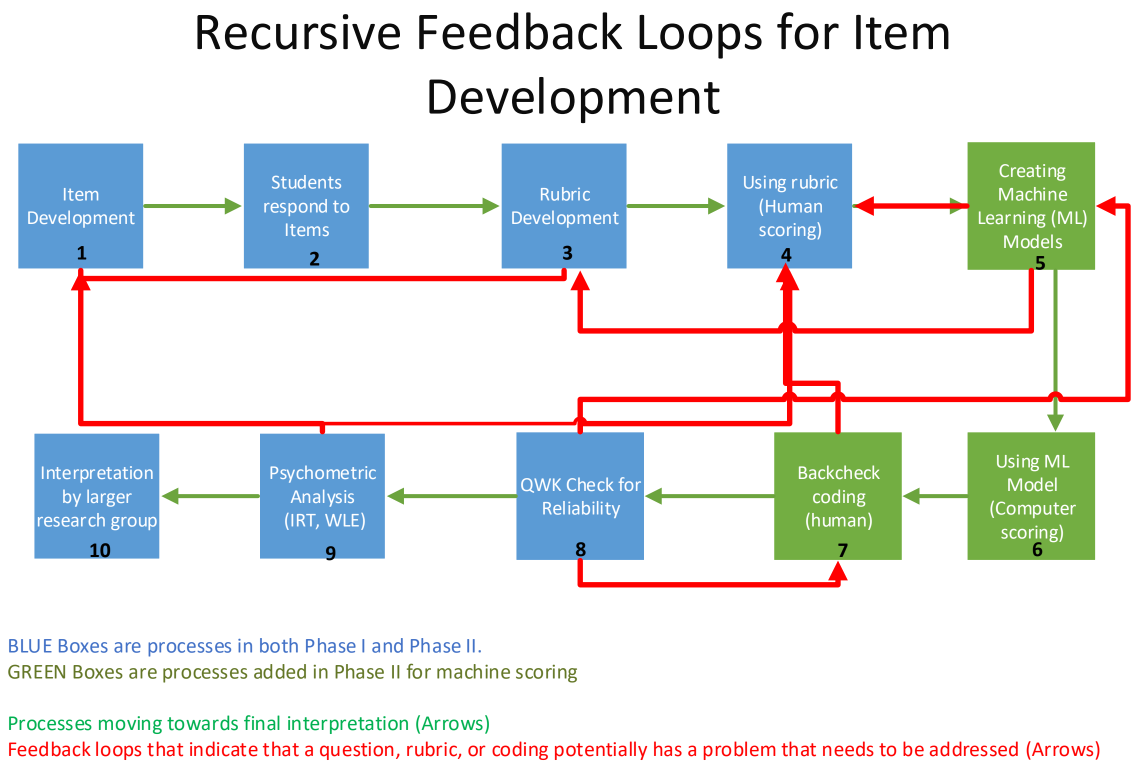

Recursive Feedback Loops for Item Development
Item Development 1
Students respond to Items 2
Rubric Development 3
Using rubric (Human scoring) 4
Creating Machine Learning (ML) Models 5
Using ML Model (Computer scoring) 6
Backcheck coding (human) 7
QWK Check for Reliability 8
Psychometric Analysis (IRT, WLE) 9
Interpretation by larger research group 10
BLUE Boxes are processes in both Phase I and Phase II.
GREEN Boxes are processes added in Phase II for machine scoring
Processes moving towards final interpretation (Arrows)
Feedback loops that indicate that a question, rubric, or coding potentially has a problem that needs to be addressed (Arrows)

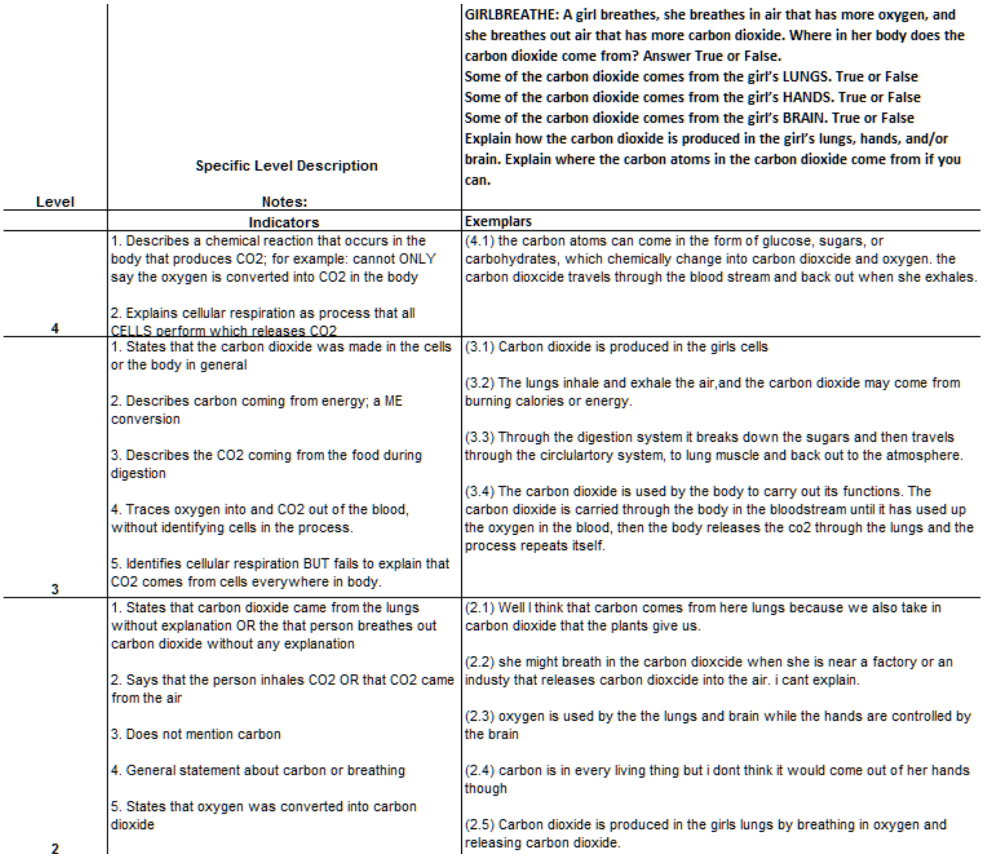
***Figure 2** Feedback loops in item development for Phases 1 and 2 of Carbon TIME*

The paper will now focus on three aspects of the ML scoring:

1. Creating ML models and insights gained about the scoring process
2. Using ML scoring as a driver to improve the assessment system: item development, rubric development, human scoring, and ML scoring
3. The scalability of the methodology as well as its limitations

<u>Creating ML models and insights gained about the scoring process</u>

The work on developing three-dimensional items began in Phase 1, as described above. Data from student responses were used to validate and revise the learning progression framework (Jin & Anderson, 2012; Mohan, Chen, & Anderson, 2009). The student responses were used to create rubrics for human coders to score student responses based on one or more constructed responses to a prompt. Most items also included one of more forced choice (either true/false or multiple choice) components as well. Usually, the rubric did not use student forced-choice responses as part of the classification rubric. (See Figure 3)



Level	Specific Level Description Notes:	**GIRLBREATHE: A girl breathes, she breathes in air that has more oxygen, and she breathes out air that has more carbon dioxide. Where in her body does the carbon dioxide come from? Answer True or False.** **Some of the carbon dioxide comes from the girl's LUNGS. True or False** **Some of the carbon dioxide comes from the girl's HANDS. True or False** **Some of the carbon dioxide comes from the girl's BRAIN. True or False** **Explain how the carbon dioxide is produced in the girl's lungs, hands, and/or brain. Explain where the carbon atoms in the carbon dioxide come from if you can.**
	Indicators	**Exemplars**
4	1. Describes a chemical reaction that occurs in the body that produces CO2; for example: cannot ONLY say the oxygen is converted into CO2 in the body 2. Explains cellular respiration as process that all CELLS perform which releases CO2	(4.1) the carbon atoms can come in the form of glucose, sugars, or carbohydrates, which chemically change into carbon dioxcide and oxygen. the carbon dioxcide travels through the blood stream and back out when she exhales.
3	1. States that the carbon dioxide was made in the cells or the body in general 2. Describes carbon coming from energy; a ME conversion 3. Describes the CO2 coming from the food during digestion 4. Traces oxygen into and CO2 out of the blood, without identifying cells in the process. 5. Identifies cellular respiration BUT fails to explain that CO2 comes from cells everywhere in body.	(3.1) Carbon dioxide is produced in the girls cells (3.2) The lungs inhale and exhale the air,and the carbon dioxide may come from burning calories or energy. (3.3) Through the digestion system it breaks down the sugars and then travels through the circlulartory system, to lung muscle and back out to the atmosphere. (3.4) The carbon dioxide is used by the body to carry out its functions. The carbon dioxide is carried through the body in the bloodstream until it has used up the oxygen in the blood, then the body releases the co2 through the lungs and the process repeats itself.
2	1. States that carbon dioxide came from the lungs without explanation OR the that person breathes out carbon dioxide without any explanation 2. Says that the person inhales CO2 OR that CO2 came from the air 3. Does not mention carbon 4. General statement about carbon or breathing 5. States that oxygen was converted into carbon dioxide	(2.1) Well I think that carbon comes from here lungs because we also take in carbon dioxide that the plants give us. (2.2) she might breath in the carbon dioxcide when she is near a factory or an industy that releases carbon dioxcide into the air. i cant explain. (2.3) oxygen is used by the the lungs and brain while the hands are controlled by the brain (2.4) carbon is in every living thing but i dont think it would come out of her hands though (2.5) Carbon dioxide is produced in the girls lungs by breathing in oxygen and releasing carbon dioxide.

Figure 3 *Human scoring rubric for GIRLBREATHE*

During Phase 2, we developed assessments around three strands, each with its own learning progression framework. (For more detailed descriptions of these frameworks, see Covitt & Anderson, 2018.) As Table 1 below shows, each strand focused on different sets of scientific practices and disciplinary core ideas, with all sharing a common set of crosscutting concepts.

Table 1: Learning Progression Frameworks

Framework	***Practices***	***Core Ideas***	***Crosscutting Concepts***
Macroscopic explanation (carbon)	Explanation, using models	Carbon-transforming processes (combustion, photosynthesis, cellular respiration, digestion, biosynthesis) at multiple scales	Conservation, flows, cycles, of matter and energy Systems and system models Scale
Macroscopic inquiry	Asking questions, analyzing data, arguments from evidence		
Large-scale systems	Data & model interpretation, explanation, prediction	Ecosystem & global carbon cycling & energy flow, climate change	

Step 5: Creating machine learning models from Figure 2. We used the Open Source machine learning engine LightSide Researcher's Workbench (Mayfield, Adamson, and Rose, 2014), to code student responses. The engine extracted text features (n-grams, stem words, parts of speech, etc.), forced choice selections as an additional feature, as well as stretchy patterns for the data feature set. The human code was treated as a nominal category

and the engine was used to create models using a variety of techniques (decision trees, logistic regressions, and Bayes nets).

Our goal was to classify student responses to the correct learning progression level to meet the industry accuracy standard of a quadratic weighted kappa (QWK) of at least .7 (Gambrell et al, 2016), comparing machine-scored and human-scored responses. For many items we were not successful in achieving this level of accuracy using only the learning progression level codes in the training set.

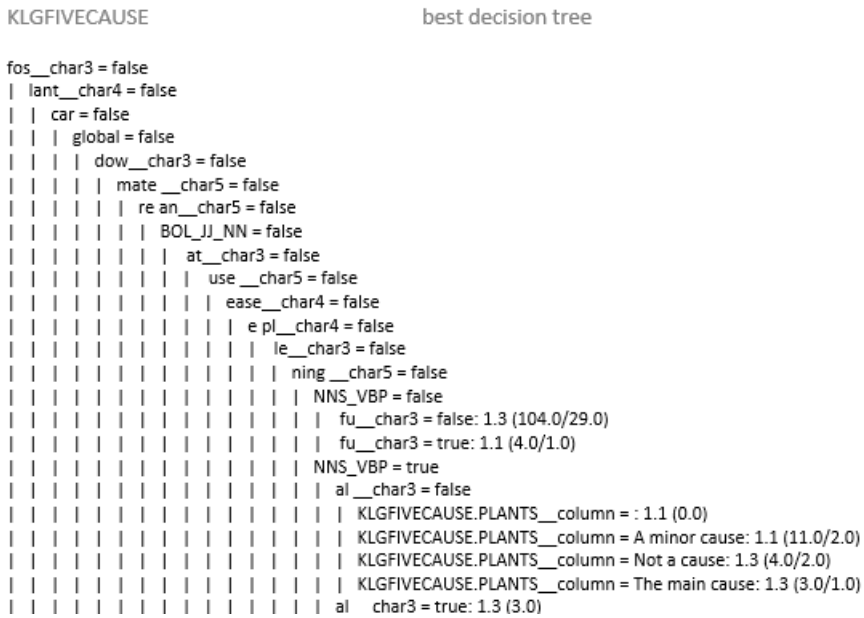
As described in Figure 3 above, the scoring rubrics identified overall learning progression levels: Level 2, Causal events with hidden mechanisms; Level 3, Incomplete understanding of matter and energy cycles (school based science understanding); Level 4, Qualitative model-based account of processes in systems (Mohan, Chen, and Anderson, 2009). The rubrics for human scoring identified criteria not only for levels of the learning progression but also specific indicators for each level.

The indicators represent typical patterns of student responses that fall within the broader learning progression level. Within a given level, there is no hierarchy of indicators. Indicator 2.3 is no closer to the student attaining learning progression Level 3 than indicators 2.1 or 2.2. However, students at learning progression Level 3 are closer to the upper anchor target understanding than those at Level 2. The rubrics classify students along the learning progression into both large bins (overall levels) and smaller more actionable bins (indicators).

The engine was then trained to code student responses at the more specific indicator code. The QWK at the learning progression level was then calculated and the engine was able to score more accurately using the smaller grain size. By using smaller bins of answers, the engine was able to create working models with much smaller sample sizes than typical for ML scoring. Usually 75 responses per indicator was sufficient and a point of diminishing returns was reached beyond 125 responses. One item (BIOMASSPYRAMID) was able to make a sustainable model with only 13 responses in the highest-scoring indicator. An important finding here was that human coded responses were needed to create a ML model. If the bins show too much inconsistency, such as those at the LP level, the ML engine will be confused by a feature that appears to be equally likely to occur in more than one scoring bin, even when linked to other features through stretchy patterns.

If an acceptable model could not be built, the confusion matrix of samples that the computer scoring did not match the original code was created and sent to an expert coder. The expert would determine if the original human coder or the computer were correct in assigning the indicator and level code. Occasionally, new indicators were created at this step for responses that did not clearly fit into the original rubric categories. This backchecked set was then used to retrain the engine. If this did not work, additional human responses could be added to the training set and the ML training could continue. If multiple attempts could not create a working model, then the item was sent back to the assessment developers to determine how if the item should be revised or replaced. The failure of the ML engine to create a working model was treated as a symptom of a larger assessment problem rather than the problem itself.

ML Scoring capturing 3-dimensional learning that human coders could not. The other interesting aspect of this Phase was that the ML engine frequently used forced-choice responses for classifying student responses even when the forced-choice was not included as part of the rubric. The ML engine was finding patterns in large groups of responses that humans could not find because of the immensity of the data set. Generally, the forced-choice responses were weighted lower in logistic regressions or appeared lower in a decision tree (Figure 4)

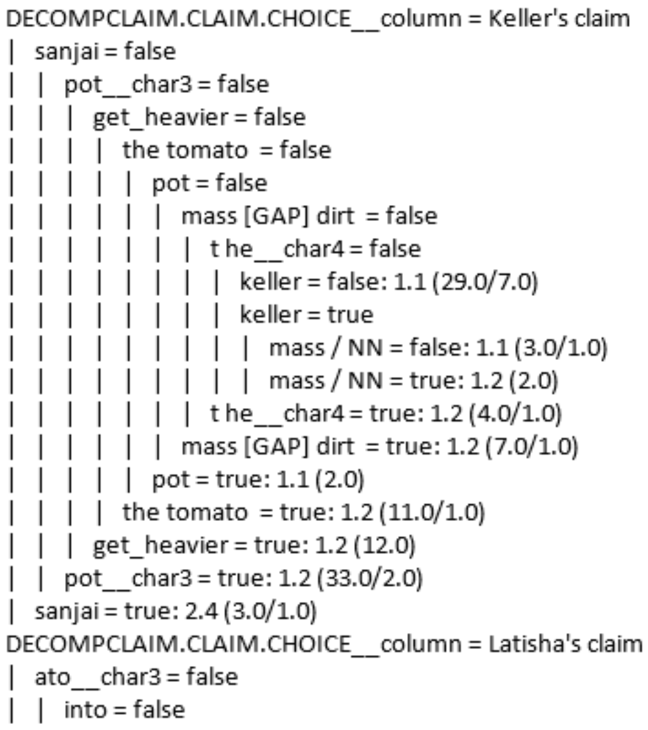


```
KLGFIVECAUSE                                        best decision tree

fos__char3 = false
|  lant__char4 = false
|  |  car = false
|  |  |  global = false
|  |  |  |  dow__char3 = false
|  |  |  |  |  mate __char5 = false
|  |  |  |  |  |  re an__char5 = false
|  |  |  |  |  |  |  BOL_JJ_NN = false
|  |  |  |  |  |  |  |  at__char3 = false
|  |  |  |  |  |  |  |  |  use __char5 = false
|  |  |  |  |  |  |  |  |  |  ease__char4 = false
|  |  |  |  |  |  |  |  |  |  |  e pl__char4 = false
|  |  |  |  |  |  |  |  |  |  |  |  le__char3 = false
|  |  |  |  |  |  |  |  |  |  |  |  |  ning __char5 = false
|  |  |  |  |  |  |  |  |  |  |  |  |  |  NNS_VBP = false
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  fu__char3 = false: 1.3 (104.0/29.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  fu__char3 = true: 1.1 (4.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  NNS_VBP = true
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  al __char3 = false
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  KLGFIVECAUSE.PLANTS__column = : 1.1 (0.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  KLGFIVECAUSE.PLANTS__column = A minor cause: 1.1 (11.0/2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  KLGFIVECAUSE.PLANTS__column = Not a cause: 1.3 (4.0/2.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  KLGFIVECAUSE.PLANTS__column = The main cause: 1.3 (3.0/1.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  al    char3 = true: 1.3 (3.0)
```

Figure 4 *Example of decision tree for item with no use of forced-choice in rubric*

However, items in which the forced-choice was part of the human rubric, the ML engine usually put the forced-choice at the top of the decision tree (Figure 5). Although the best operational scoring algorithms are rarely decision trees, they present an efficient method to show stakeholders what is driving the machine scoring of student responses. In operational scoring, forced-choice data facilitates reliably locating students along the learning progression that human coders could not have used (Thomas et al, 2018) as examining dozens or hundreds of responses was required to find the pattern, which is far beyond the working memory capacity of even the most expert coders.



```
DECOMPCLAIM.CLAIM.CHOICE__column = Keller's claim
|  sanjai = false
|  |  pot__char3 = false
|  |  |  get_heavier = false
|  |  |  |  the tomato  = false
|  |  |  |  |  pot = false
|  |  |  |  |  |  mass [GAP] dirt  = false
|  |  |  |  |  |  |  t he__char4 = false
|  |  |  |  |  |  |  |  keller = false: 1.1 (29.0/7.0)
|  |  |  |  |  |  |  |  keller = true
|  |  |  |  |  |  |  |  |  mass / NN = false: 1.1 (3.0/1.0)
|  |  |  |  |  |  |  |  |  mass / NN = true: 1.2 (2.0)
|  |  |  |  |  |  |  t he__char4 = true: 1.2 (4.0/1.0)
|  |  |  |  |  |  mass [GAP] dirt  = true: 1.2 (7.0/1.0)
|  |  |  |  |  pot = true: 1.1 (2.0)
|  |  |  |  the tomato  = true: 1.2 (11.0/1.0)
|  |  |  get_heavier = true: 1.2 (12.0)
|  |  pot__char3 = true: 1.2 (33.0/2.0)
|  sanjai = true: 2.4 (3.0/1.0)
DECOMPCLAIM.CLAIM.CHOICE__column = Latisha's claim
|  ato__char3 = false
|  |  into = false
```

Figure 5 *Example of decision tree in which forced-choice is part of original rubric*

Since the forced-choice tend to target disciplinary core ideas more heavily than the constructed-response portions of question, the incorporation of both aspects increases the integrated nature of the classification of the student along the learning progression. By combining scientific explanation (a science practice) with matter and energy tracing to explain cause and effect relationships (crosscutting concepts) with appropriate biology content (disciplinary core ideas) to holistically score the student for both the learning progression level and specific indicator, the ML engine is able to provide strong evidence of what a student knows and can do.

Finally, there were surprising findings that the ML engine found different patterns between students who used "CO_2" and those who used "carbon dioxide" in their responses. (Gambrell et al, 2016). Human coders who know science, particularly science teachers or pre-service science teachers, instinctively treat these two as equivalent, because from an instructional or linguistic view, they are. However, the ML engine found that using the shorter "CO_2" was more common among students at the higher levels of the progression than those who used "carbon dioxide" as it appears in the text of the assessment task. This is an example of the ML engine finding patterns that human coders could not because experts automatically chunk the equivalent forms of carbon dioxide to reduce mental load while scoring items. Although the exact relationship between the use of CO_2 and higher levels of the learning progression is not clear, the continued reliability of the model based on the QWK attributes to the robustness and consistency of this pattern.

Another insightful aspect of ML scoring is that the engine can look for "stretchy patterns" of words, phrases, or letters across white spaces. For example, the ML engine found that the letter pattern "to en" was strongly associated with sublevel 3.1 on several items. Closer analysis revealed that this letter pattern is indicative of a matter-energy conversion misconception. Students at this learning progression level will use phrases like "the bread turned into energy" or "changes mass to energy." Conversely, the ML engine on another item had "to che" as a strong indicator of a level 4 response. Students who trace energy, tend to name the type of energy: chemical energy, solar energy, etc. So the stretchy pattern "to che" indicated that students who understand photosynthesis trace the energy to chemical energy stored in chemicals like glucose or high energy bonds.

One of the advantages of these stretchy patterns is that they can find patterns even if several words or letter patterns intervene. Also, since the ML engine was using the feature "Stem N-words," the engine was able to identify patterns even if the word was not spelled correctly. This is important since spelling is not a construct that is being assessed in this science context; however, research shows that human coders often score tasks with spelling errors lower, even when spelling is not relevant to the construct being assessed. (Chase, 1983) Since computers do not experience fatigue, frustration, or associate responses to specific students, the ML engine does not score with contrast effects, halo or racial bias, or fatigue effects. (Fleming, 1999) or other aspects of the Pygmalion effect (Sprouse and Webb, 1994).

Since multiple model-building algorithms, logistic regression, decision trees, and Bayes nets, are available in the software, the best available model, as measured by QWK, was used regardless of the type. During model development, the number of features can be also be altered to maximize the QWK of the final model. The best models generally use between 200 and 600 features of the thousands of features extracted and available for categorization. The flexibility of changing the number of features and model types for operational scoring allows the most salient features to drive the scoring of student responses rather than predetermined constraints under which some other ML engines operate.

Using ML scoring as a driver to improve the assessment system: item development, rubric development, human scoring, and ML scoring

Recursive feedback loop to item development. One aspect that became clear was that training the machine learning engine could be used as a quality check on items. The *Carbon TIME* project already had a method for evaluating and revising assessment items prior to using machine scoring, the feedback loop from rubric development back to item development. The machine scoring gave empirical evidence, a failed QWK score, that there was a flaw in the item that should be addressed. As Figure 2 shows, the ML allowed for multiple feedback loops at a wider variety of stages (boxes five through eight). This feedback enabled item refinement that created over 50 machine scorable tasks over several iterations. Since the assessment triangle (Pellegrino, Chudowsky, and Glaser, 2001) requires quality observations in order to make valid claims, creating quality items is of great importance.

As discussed previously, models that failed at Step 5 (Figure 2) had methods to improve rubrics, human coding, and machine scoring. If training proved impossible, it was an indication that there were flaws in the initial human coding. If the training set still had very few of a particular indicator, additional responses might be coded to develop a more robust and representative training set. Often, additional posttest responses were required to supply enough responses at the highest level to the ML engine.

Other researchers who have struggled with machine scoring of NGSS based items, may not have had flawed ML scoring engines but rather were using flawed items or human scoring processes. It is important to realize that for most large scale testing programs, many pre-tested items fail to meet psychometric parameters (Schmeiser and Welch, 2006); so it is not surprising that many items developed to assess NGSS would fail as well.

Computer scoring and backcheck human scoring (Steps 6-8). In addition to the feedback loop at the creation of ML models, additional feedback became possible to strengthen the item pool. Stratified random samples of ML-scored student responses were back-checked by human coders to verify that the operational scoring maintained at least a 0.7 QWK. If operational scores did not meet that threshold, the ML engine was retrained and the entire batch of student responses were rescored. If new classes of answers were discovered in the larger answer pool, the rubric could be revised to add a new indicator to capture those responses that had not been present in the original training set. Finally, IRT analysis was completed (Nine) which could identify items that had poor fit, aberrant difficulties, or other psychometric flaws (Thomas, Kin, and Draney, 2018; Thomas et al, 2019). These items could then be revised or replaced to further improve the item pool.

Additional recursive feedback loops. The ML scoring process served as a control lever for the entire development process. If the scoring engine was unable to create a model with an acceptable accuracy, then there was an issue with the item that should be addressed. The item might contain unclear wording. The item might allow students to substitute an easier question (Kahneman, 2011) rather than answer the intended target. The rubric might not include a code for a set of responses that did not fit any of the indicators. Human coders might not be consistently applying the rubric. The feedback loops improved item quality and consequently the claims that could be made about what students know and can do. (Figure 2) Consequently, over the four years of *Carbon TIME* Phase 2, numerous items were revised or replaced based on information derived from the various feedback loops.

Additionally, since each unit test and each full year test had multiple items designed to elicit evidence of student understanding, problematic items could be eliminated from the analysis pool. The following year, a revised item or entirely new item could be created to elicit the practices, crosscutting concepts, and disciplinary core ideas targeted by the problematic item. Consequently, the assessment pool was strengthened by multiple iterations of revisions.

Finally, since there were multiple assessment items associated with each of the six *Carbon TIME* units, a wider variation of the targeted science practices (including explanation, argumentation, and data analysis) could be integrated with the primary crosscutting concepts (scale, tracing matter and energy, systems and system models) so that stronger inferences could be made about what students know and can do.

Another aspect of this process that insured great reliability was that every student response could be re-scored by the ML engine with little additional time or cost. Unlike human coding, removing the old codes and rescoring hundreds or thousands of responses could be done in seconds or minutes rather than weeks and months. So, if the initial model was flawed, those ML codes could be thrown out and replaced with codes that meet accuracy criteria, further insuring that the final identification of every student response along the learning progression is as accurate as possible given measurement uncertainty. This facet is important not only because of the responsiveness of the process but also because of the contrast with human scoring techniques which would make such a change extremely costly and time consuming if not impossible.[1]

Step 9: Psychometric analyses. The large data sets generated each year from machine scoring all student responses to all items have provided opportunities to generate a variety of evidence to support the validity of claims about student learning. Each year, each item has had an IRT model created with thresholds for each level transition calculated as well as verifying that the item fit function is within acceptable parameters. (Adams and Khoo, 1996). These results are reported in more detail elsewhere (Doherty, et al., 2015; Draney & Bathia, in preparation). In general, they show that the tests are three-dimensional, with the dimensions corresponding to the three learning progression frameworks summarized in Table 1, above: Macroscopic explanation, macroscopic inquiry, and large-scale reasoning.

Wright maps: IRT thresholds for the transitions between Level 2, 3, and 4 were identified for each item. We conducted multiple psychometric analyses including item fit, student fit, and overall test reliability. Wright maps (Figure 6 below), sources of evidence specifically mentioned in the NRC (2014) report about assessing NGSS, provide further validity evidence that the items, unit tests, and full year tests demonstrate both a reliable measurement of students' location along the learning progression and the growth of students understanding from pretest to posttest at both the unit and school year level.

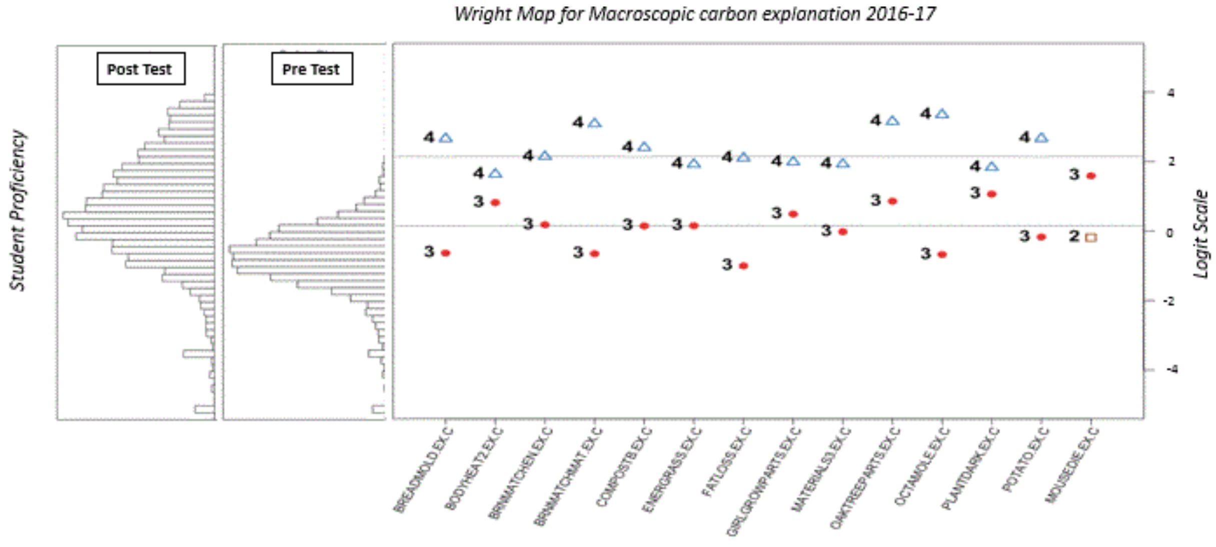


Figure 6 *Wright Map for all items 16-17 student data.*

The similarity in item functioning and Wright maps from year to year based on the ML scored items provide further evidence that the assessments are validly assessing the 3-dimensional learning targets based on over 1.8 million student responses. Additionally, items that have thresholds that are vastly different from others cause a feedback loop back to item development to determine if the item is targeting the construct as intended. Each additional feedback loop helps to increase the confidence that the items are eliciting evidence of what students know and can do from which researchers and teachers can draw valid conclusions about teaching and learning.

Reliability of forced-choice and constructed-response components. Since one of the targeted practices in the *Carbon TIME* curriculum is scientific explanation, it would be difficult to have valid assessment without student constructed response from a construct validity, face validity, or assessment blueprint perspective. Since most of

[1] The author was told at an AP Workshop that if a new acceptable answer is discovered through the course of grading and added to the rubric, previously scored responses are not checked to see if they would be scored differently based on the adjustments to the rubric.

the items on the assessment were composite items, including both constructed response and forced-choice components, we were able to compare the reliability of weighted likelihood estimates of student proficiency based on the separate components and the items as a whole. Figure 7 shows that using only forced-choice data there is a great deal of error variance that may cause many students to be misclassified along the learning progression. However, adding the information from ML constructed-response codes greatly improves the reliability of the proficiency estimates. Beginning with the constructed-response items, and then adding the forced-choice as an additional data points reduces the spread further and gives greater confidence in assigning the learning progression level to an individual student.(Thomas et al, 2019).

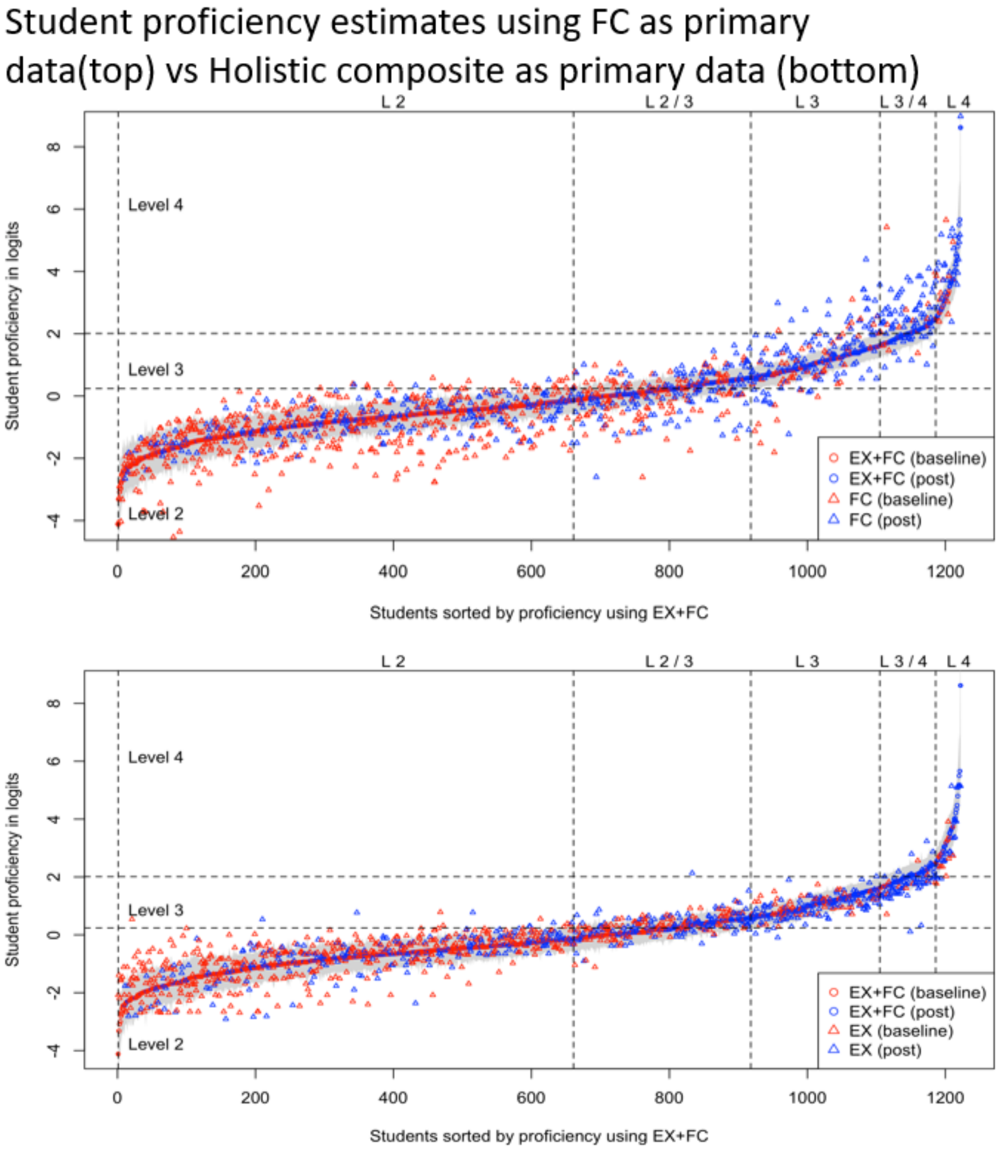


Figure 7. Comparison of using FC or Holistically scored (EX) items as primary data for assigning student ability level.

Figure 7 provides additional insights to the larger group (Step 10). Students in the baseline group (red) are students who had just finished a traditional curriculum in biology for teachers who would be joining the *Carbon TIME* project for the following school year. Students in the post-test group (blue) had the same teacher but experienced at least three units of the *Carbon TIME* curriculum. So, students who received the curriculum demonstrated higher levels of understanding on the LP of the carbon dimension. These sorts of conclusions will be more closely examined in the final section on scalability.

The scalability of the methodology as well as its limitations

One unresolved problem with reformed standards such as NGSS that stress the importance of three-dimensional performances is how to assess student proficiencies at scale. As Pellegrino et al. pointed out (NRC, 2014), assessing NGSS will need to use emerging technologies including machine scoring. During Phase 1 of the Carbon TIME project we only scored a sample of student responses from participating teachers. However, after initiating ML scoring of the items, we were able to score every response from every student.

Table 3, below, shows the scale of the ML project. From the pilot using the 2015-16 data until the end of the Carbon TIME project in the 2018-19 school year, over 1.8 million student responses were scored by the ML engine. The table also shows that there were fluctuations in items from year to year. As previously discussed, these revisions and replacements were guided by the feedback loops that resulted from efforts to use ML to score these student responses.

Table 2. Student responses scored using ML by school year.

School year	Responses scored	Unique items scored	Assessments scored
15-16	175,265	33	27,981
16-17	532,825	39	61,475
17-18	693,086	41	66,335
18-19	409,266	39	42,117
Total	1,810,442	57	197,908

Given that expert human coders are only able to score 100 responses per hour, we have been able save over 18,000 labor hours. Using an average cost of $18 per labor hour this represents a savings of over $325,000. In addition to these savings, the ability to build working models on smaller training sets than most ML scoring engines also improves the scalability as smaller pre-test and calibrating samples are needed. In reality, since most testing programs require some, if not all, student responses to be double scored the savings would be even larger for operational scoring.

In addition to the sheer volume of student responses scored, the speed with which they could be scored also impacted the process. For example, the 2017-18 data set would take over 6,900 labor hours to human score. This would take four full-time coders over 9 months (an entire school year) to code. The entire data set was machine-scored within a few weeks of the final data collection. This speed enabled the rest of the team to complete IRT, HLM, and other analyses in time to inform decisions about changes to curriculum, student worksheets and labs, assessment items, and professional development for the next school year. Importantly, the rapidity of ML scoring allows for matching student responses for pre-post gain scores, HLM, and other analyses to be completed in a facile way. Finally, since most IRT models work better with large data sets, the vastness of the set allows greater confidence in the IRT findings. The larger group was able to look at post-test results of the two groups (Figure 8). Using traditional instruction only 14% of students move beyond Force Dynamic reasoning (Level 2) while 66% of Carbon TIME students achieve at least Level 3. 25% of Carbon TIME students reach the upper anchor, Level 4, of the learning progression while only 0.3% reach that level with traditional instruction.

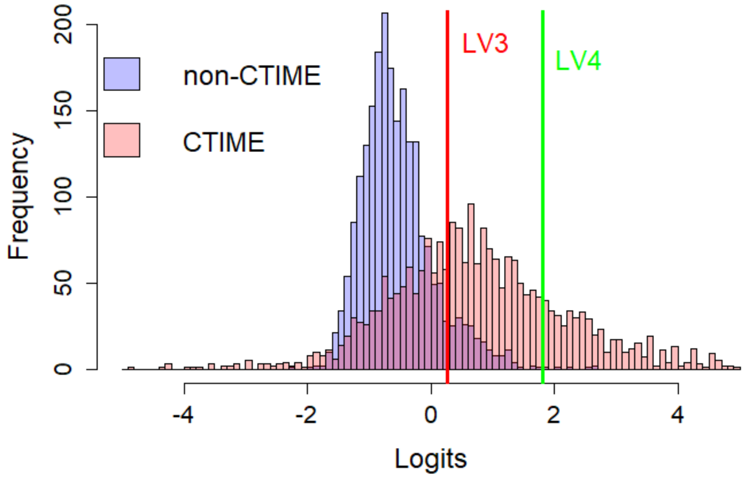


Figure 8. Comparison of post-test scores of students in traditional and Carbon TIME curricula.

Moreover, the larger group was able to examine trends in student learning when looking at matched students pre-test and post-test results (Figure 9).

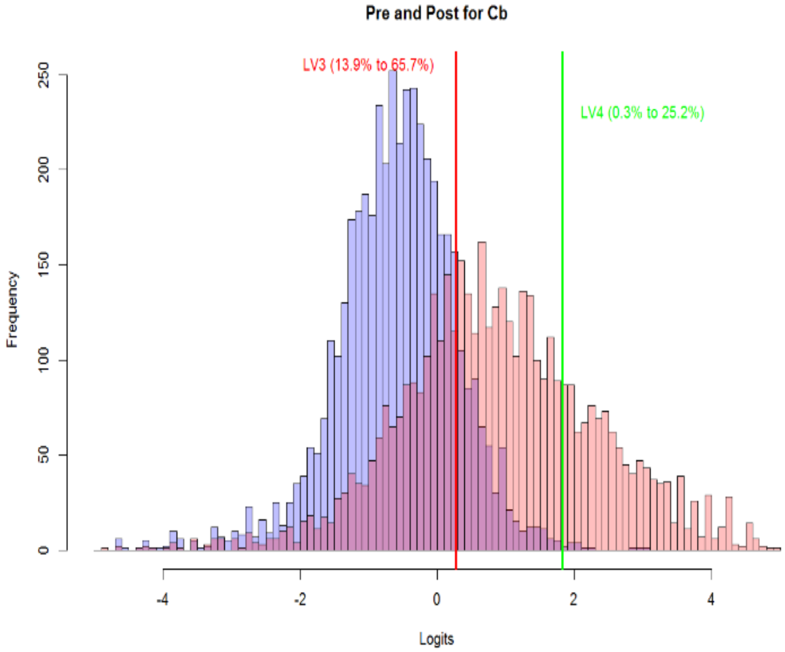


Figure 9. Pre and post-test scores from 2017-18 data. Pretest results in blue and posttest results in pink.

The ML scores could be used to evaluate the effectiveness of the curriculum and PD across all three frameworks. The impacts of changes to PD, curricula, etc. can be examined by the larger research group based on the student responses of all students from all teachers.

This data set has supported research claims about students, teachers, curriculum, and professional development. (Lin et al, 2018; Covitt et al, 2018; Parker et al, 2018; Edwards, Scott, and Anderson, 2018). We have strong evidence that the instructional materials are bringing about student learning because we are looking at literally thousands of students with hundreds of thousands of responses that involve integrating all three dimensions of the Framework: students are authentically being assessed on making sense of phenomena using science practices, cross cutting concepts, and disciplinary knowledge.

Student learning in *Carbon TIME* was compared to pre-test data (Figure 7) as well as to end of course assessment of traditional biology curriculum. Teachers planning to use *Carbon TIME* materials the following year gave the end of course assessment to students in their classes. These baseline results represent the three dimensional understanding of students using traditional coursework. Figure 8 (Lin, Frank, and Anderson, 2019) shows that not only did students in *Carbon TIME* achieve greater understanding, the baseline students using traditional curriculum do not show much growth beyond pretest results for other students. Although matching pre/post test data is not available for baseline students, similar disparities between baseline and *Carbon TIME* data across several years and dozens of teachers suggest that traditional curricula does not lead to three-dimensional understanding of carbon cycling.

Using traditional instruction only 14% of students move beyond Force Dynamic reasoning (Level 2) while 66% of *Carbon TIME* students achieve at least Level 3. 25% of *Carbon TIME* students reach the upper anchor, Level 4, of the learning progression while only 0.3% reach that level with traditional instruction.

Since every response of every student can be evaluated using ML, the gain scores of teachers can be compared. (Lin et al., 2019). Figure 10 below shows that gain scores can be compared so that the relative effectiveness of teachers can be compared. Furthermore, large matching ML-scored data sets have been used to create models that estimate the importance of *Carbon TIME*, teacher effects, and school FRL percentage.

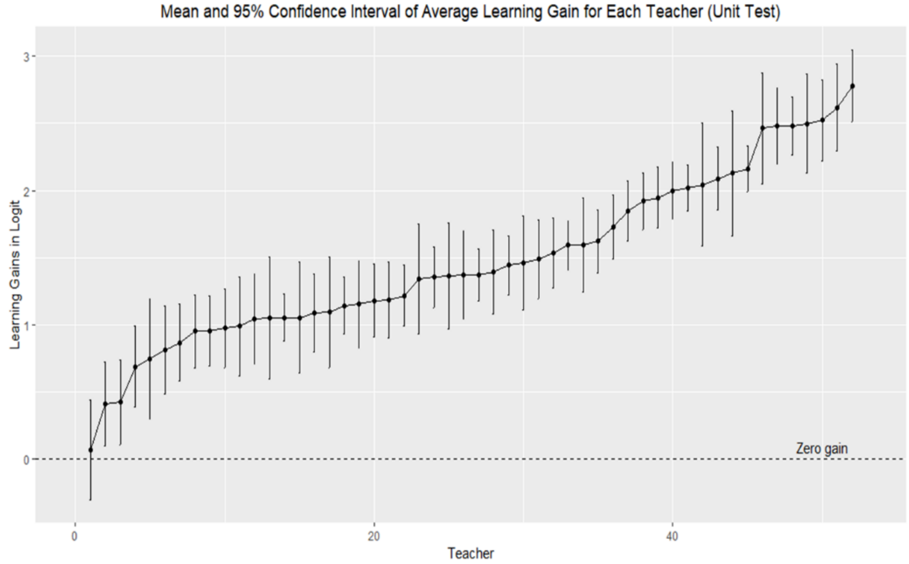


Figure 10 Average gain scores by teacher

In order to create reliable models of constructs important such as student growth, value added, etc. to stake holders at the district, state, and national level, massive sets of matched data are required. Limitations to this methodology do exist. Specifically, items must have rubrics created at a grain size fine enough that the ML engine can consistently identify differences between patterns of responses. Training sets must be gathered and coded by human experts in order to train the engine. Samples of student responses must be human coded to insure that the scoring consistently operates at or above industry standards for reliability. These stages all require time, money, and expert coders. Assessment tasks that are not scorable by the ML engine must be examined for flaws in the process along the feedback loops. A willingness to throw away items that do not work and replace those items with better ones can cause delays and increased costs. However, the same delays and costs would be incurred for any item that failed to meet the blueprint parameters during pre-testing. Since most IRT modeling of 2PL or 3PL require samples of 500 or 1,000 responses to converge (Stone & Yumoto, 2004; Kean & Reilly, 2014) requiring a similar number of responses to train the ML engine should not add too large a burden to the calibration and pre-testing phase.

Conclusions

The NGSS and related state standards pose new problems for assessment, scoring, and reporting. Many of the targeted three-dimensional performances will require some form of constructed response to validly assess what students know and can do. However, other testing programs that are using constructed-response and other extended essay formats are experiencing delays in scoring that are not compliant with ESSA. (Education Dive, 2019) As states move to assessing these new standards, they will need to find a way to assess NGSS while maintaining control over costs related to item and test construction, delivery, scoring, and reporting. At the same time, ESSA places greater demands for validity evidence through peer review and other processes.

The ML scoring technique used to evaluate student performances in *Carbon TIME* provides insight into how curriculum, assessment, scoring, and reporting may be integrated to facilitate teaching, learning, and assessment. We have demonstrated over a multi-year project that three-dimensional performance items can be assessed reliably at scale. The data from those assessments can be used to generate evidence of student learning and to evaluate the effectiveness of programs.

There are lessons to be learned. Items that are problematic for humans to score will likely be problematic for the ML engine to score as well. Consequently, items may need to be revised or replaced as they would be in other testing programs if they do not meet statistical criteria related to the blueprint. Some items may be better at eliciting information at specific portions of the learning progression (Briggs and Alonzo, 2012), so multiple items that elicit information along the learning progression will be needed. Additionally, since items target different facets of the same set of science and engineering practices, disciplinary core ideas, and crosscutting concepts multiple items may be needed to make valid claims about student learning and understanding.

Design-based research is an iterative process and using ML scoring allowed for additional feedback loops (Figure 2) based on large data sets to drive conclusions on the entire design: professional development, curriculum, assessment, and learning. This project showed that nearly 2 million student responses could be scored over several school years.

But to fully assess NGSS, many more tasks assessing other performance expectations will need to be developed, pre-tested, and used to train ML engines. More learning progressions with meaningful indicators will be needed as well. This ML technology may provide a way to assess at scale, but experts from teaching, learning, and assessment will be needed to produce assessment tasks, curricular materials, and professional development so that the ML scores can be used both for summative and formative assessment. The investment of time and money in this development process could make monitoring assessments affordable to deliver and score at scale. However, it took data from many school districts and thousands of students to generate using experts in item design, rubric development, machine learning, and psychometrics. As the NRC (2014) report suggests these scalable monitoring assessments should represent only a part of a system of assessments, at a variety of scales, to validly assess all facets of NGSS.

ACKNOWLEDGEMENT This paper is adapted from an earlier version co-authored with Andy Anderson, Qinyun Lin, and Kenneth Frank (MSU) as well as Karen Draney and Shruti Bathia (BEAR). When submitted for publication rather than a presentation all co-authors will receive the credit that they richly deserve. Karen and Shruti's work is

primarily contained in another paper in this symposium. Qinyun, Andy, and Kenneth are presenting in another session at NARST.

References

Adams, R., & Khoo, S. T. (1996). Quest : the interactive test analysis system. Retrieved from http://works.bepress.com/ray_adams/36

Bejar, I. I., Williamson, D. M., & Mislevy, R. J. (2006). Human scoring. *Automated scoring of complex tasks in computer-based testing*, 49-81. Retrieved from https://www.researchgate.net/profile/Isaac_Bejar/publication/236167620_Automated_Scoring_of_Complex_Tasks_in_Computer-Based_Testing/links/0deec53b4446cd0d8e000000/Automated-Scoring-of-Complex-Tasks-in-Computer-Based-Testing.pdf on May 13, 2019.

Briggs, D. C., & Alonzo, A. C. (2012). The psychometric modeling of ordered multiple-choice item responses for diagnostic assessment with a learning progression. In *Learning progressions in science* (pp. 293-316). Brill Sense.

Chase, C. I. (1983) Essay test scores and reading difficulty. *Journal of Educational Measurement*, *20* (3) ,293-297.

Covitt, B. A., & Anderson, C. W. (2018). Assessing scientific genres of explanation, argument, and prediction. In A. L. Bailey, C. Maher, & L. Wilkinson (Eds.) *Language, literacy, and learning in the STEM disciplines: How language counts for English learners,* pp. 206-230*.* New York, NY: Routledge.

Doherty, J. H., Draney, K., Shin, H. J., Kim, J., & Anderson, C. W. (2015). Validation of a learning progression-based monitoring assessment. Manuscript submitted for publication. Retrieved from https://carbontime.bscs.org/sites/default/files/educator_resources/CarbonTIMEAssessmentValidation.pdf on 9/30/2019.

Draney, K., and Bathia, S. (in preparation). Assessment frameworks: Improving the evidence we collect about student cognition. To be presented at the annual meeting of the National Association for Research in Science Teaching, Portland, Oregon, 2020.

Education Dive (2019). Growing pains, grievances for new ESSA school report cards. Retrieved from https://www.educationdive.com/news/growing-pains-grievances-for-new-essa-school-report-cards/544313/ on 10/1/19.

Edwards, K. D., Scott, E. E., & Anderson, C. W. (2018, March). Designing curriculum to support student engagement in inquiry practices. Presented at the annual meeting of the National Association for Research in Science Teaching, Atlanta. https://carbontime.bscs.org/sites/default/files/research/conference-presentations/1_Edwards_Inquiry_poster.pptx retrieved on 10/1/19.

Fishman, B. J., Penuel, W. R., Allen, A. R., Cheng, B. H., & Sabelli, N. O. R. A. (2013). Design-based implementation research: An emerging model for transforming the relationship of research and practice. *National society for the study of education*, *112*(2), 136-156.

Fleming, N. D. (1999). Biases in marking students' written work: quality. *Assessment matters in higher education: choosing and using diverse approaches*, 83-92.

Florida Department of Education (2019). Statewide Assessment Program Information Guide, 2019-2020. Retrieved from http://www.fldoe.org/core/fileparse.php/5663/urlt/swapig.pdf on 9/30/2019.

Gambrell J, Thomas Jay, Meisner R, Bolender Brad. Machine Learning Analysis of Student Responses to Carbon-TIME Learning Progression Items. Presented at AERA; 2016 April 12; Washington DC, USA.

Gorin, J. S., & Mislevy, R. J. (2013, September). Inherent measurement challenges in the next generation science standards for both formative and summative assessment. In *Invitational research symposium on science* *assessment*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.800.5350&rep=rep1&type=pdf on May 13, 2019.

Jin, H., and Anderson, C. W. (2012). A learning progression for energy in socio-ecological systems. *Journal of Research in Science Teaching*, 49(9), 1149–1180

Jin, H., Zhan, L., & Anderson, C. W. (2013). Developing a fine-grained learning progression framework for carbon-transforming processes. *International Journal of Science Education*, *35*(10), 1663-1697.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kean, J., & Reilly, J. (2014). Item response theory. *Handbook for clinical research: Design, statistics and implementation*, 195-198.
Li, H., Gobert, J. D., & Dickler, R. (2017). Automated Assessment for Scientific Explanations in On-line Science Inquiry. In *EDM*. Retrieved from http://educationaldatamining.org/EDM2017/proc_files/papers/paper_55.pdf on May 13, 2019.

Lin, Q., Kim, J., Holste, E., Bathia, S., Draney, K., & Frank, K. A. (2018). Teacher effectiveness and their *Carbon TIME* practices and knowledge. Presented at the annual meeting of the National Association for Research in Science Teaching, Atlanta. https://carbontime.bscs.org/sites/default/files/research/conference-presentations/4_Lin_HLM_poster.pptx retrieved on 10/1/19.

Lin, Qinyun, Ken Frank, and Charles W. Anderson. What factors affect students' learning? Paper presented at NARST Conference in Baltimore, MD; April 2, 2019.

Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H. S., & Pallant, A. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment*, *23*(2), 121-138.

Mayfield, E., Adamson, D., & Rosé, C. P. (2014). LightSide researcher's workbench user manual. http://ankara.lti.cs.cmu.edu/side/LightSide_Researchers_Manual.pdf *Retrieved November*, *12*, 2015.

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

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

National Research Council. (2005). *Systems for State Science Assessment.* Washington D.C.: National Academies Press.

National Research Council. (1996). *National science education standards*. National Academies Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8.* Washington, DC: National Academies Press.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: National Academies Press.

National Researcher Council (2014). *Developing Assessments for the Next Generation Science Standards*. Edited by Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S.National Academies Press. 500 Fifth Street NW, Washington, DC 20001.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. National Academies Press.

Ohio Graduation Test, Blueprint for Science (2006). http://education.ohio.gov/getattachment/Topics/Testing/Achievement-Tests/Blueprints-for-Ohio-Graduation-Tests/OGT-Science-Test-Blueprint.pdf.aspx retrieved 09/30/2019.

Parker, J., Kohn, C., Covitt, B., Lee, M. & Anderson, C. W. (2018, March). Curriculum materials supporting 3-dimensional learning about the global carbon cycle. Presented at the annual meeting of the National Association for Research in Science Teaching, Atlanta. Poster: https://carbontime.bscs.org/sites/default/files/research/conference-presentations/2_Parker_large_scale_poster.pptx retrieved 10/05/2019.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. National Academy Press, 2102 Constitutions Avenue, NW, Lockbox 285, Washington, DC 20055.

Schmeiser, C.B., & Welch, C.J. (2006). Test development. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed.) (pp. 307-354). Westport, CT: American Council on Education and Praeger Publishers.

Shaw, D., Meisner, R., & Bolender, B. (2019) Prompt-Agnostic Automated Essay Scoring Project. ACT Research Report.

Shermis, M. D. (2015). THE ROLE OF MACHINE SCORING IN SUMMATIVE AND FORMATIVE ASSESSMENT. The Next Generation of Testing: Common Core Standards, Smarter? Balanced, PARCC, and the Nationwide Testing Movement, 83.

Sprouse, J. L., & Webb, J. E. (1994). *The Pygmalion Effect and Its Influence on the Grading and Gender Assignment on Spelling and Essay Assessments*.

Stone, M., & Yumoto, F. (2004). The effect of sample size for estimating Rasch/IRT parameters with dichotomous items. *Journal of applied measurement*.

Thomas, Jay. Using Open Source Machine Learning to Holistically Score 3-Dimensional Science Composite Items based on NGSS. (2017) Poster Presented at Education Technology and Computational Psychometrics Symposium in Iowa City, IA, USA; November 15, 2017.

Thomas, J., Kim, J., & Draney, K. (2018). Machine scoring and IRT analysis. Presented at the annual meeting of the National Association for Research in Science Teaching, Atlanta. Poster: https://carbontime.bscs.org/sites/default/files/research/conference-presentations/3_Thomas_Scoring_poster.pptx retrieved 10/1/19.

Thomas, Jay, JinHo Kim, and Karen Draney (2018). Measuring progress toward measuring three-dimensional learning at scale: Machine Scoring and IRT Analysis. Poster presented at ETCPS, Iowa City, Iowa, USA; October 3, 2018.

Thomas, Jay, Ellen Holste, Karen Draney, Shruti Bathia, and Charles W. Anderson (2019). Developing Automated Scoring for Large-scale Assessments of Three-dimensional learning. Paper presented at NARST Conference in Baltimore, MD; April 2, 2019.

Williamson, D. M., Bejar, I. I., & Mislevy, R. J. (2006). Automated scoring of complex tasks in computer-based testing: An introduction. *Automated scoring of complex tasks in computer-based testing*, 1-13. Retrieved from https://www.researchgate.net/profile/Isaac_Bejar/publication/236167620_Automated_Scoring_of_Complex_Tasks_in_Computer-Based_Testing/links/0deec53b4446cd0d8e000000/Automated-Scoring-of-Complex-Tasks-in-Computer-Based-Testing.pdf on May 13, 2019.

Yao, S. Y., Berson, E., Ayers, E., Choi, S., & Wilson, M. (2010, April). *The Qualitative Inner-Loop of the BEAR Assessment System*. Presented at the International Objective Measurement Workshop, University of Colorado, Boulder.