**Supplementary Materials**

This document includes supplementary materials to provide a more comprehensive understanding and context for our paper.

1. A full sample item and an example of a second item, accompanied by text explaining how the first item is related to a relevant NGSS performance expectation and to each dimension of the NRC framework (Item Types and Example Items);
2. Additional Figures and Tables for the paper (Figure S1; Table S1, S2, S3);
3. Methods Supplement that describes how we applied hierarchical linear models to decompose variance in student learning (Methods\_Supplement);
4. Supplemental material on how we applied machine-learning scoring to assess three-dimensional science learning (ML\_Supplement);
5. Abstracts of our related qualitative research that (1) focuses on science classroom discourse and its connection to student learning (2) connects teacher perceptions to classroom discourse and student learning outcomes (QualitativeResearchAbstracts\_Supplement);
6. Supplemental material that walks through each step of the sample exclusion procedure, providing rationale, as well as information regarding how samples and results change with each step (Sensitivity\_Supplement).

Table of Contents

[Item Types and Example Items 3](#_Toc80047861)

[Item Types 3](#_Toc80047862)

[The FATLOSS Item 4](#_Toc80047863)

[FATLOSS Item Text 5](#_Toc80047864)

[Human Scoring Rubric for Students’ Written Explanations 6](#_Toc80047865)

[Machine scoring 8](#_Toc80047866)

[Connections to NGSS and the NRC Framework 8](#_Toc80047867)

[PLANTCLAIM Item 9](#_Toc80047868)

[Figure S1 13](#_Toc80047869)

[Table S1 14](#_Toc80047870)

[Table S2 15](#_Toc80047871)

[Table S3 16](#_Toc80047872)

[Methods Supplement 17](#_Toc80047873)

[Abstract: Effects of Teaching Strategies on Students’ Three-dimensional Learning in a Design-based Implementation Research Project 22](#_Toc80047874)

[Abstract: Connecting Teachers’ Commitments and Perceptions to Classroom Discourse and Student Learning Outcomes 24](#_Toc80047875)

[Sensitivity Analysis Supplement 26](#_Toc80047876)

# Item Types and Example Items

This supplemental material contains a single item that is included in one form of the full pretest/posttest and in the *Animals* unit pretest and posttest. The full item set is available on the *Carbon TIME* website: <https://carbontime.bscs.org/assessment-links> . In this supplemental material we (a) summarize the types of items included in the assessments; (b) provide an example explanation item (FATLOSS) with human scoring rubrics, data on machine scoring, and related NGSS performance expectations, practices, disciplinary core ideas, and crosscutting concepts; and (c) an example inquiry item: PLANTCLAIM.

## Item Types

The item types in assessments are based on our learning progression work, in which we have developed cognitive models for three related learning progressions. In the table below, these three item types are represented by the rows and the NGSS dimensions by the columns; note that each item is three-dimensional. (The full list of DCIs for each item type is available on the *Carbon TIME* website.) The full pretests and posttests include items representing all three item types. The unit posttests that we used for our HLM analyses contained only items representing the first two item types: macroscopic explanation and inquiry. These items are based on the same DCIs and CCCs, but engage students in different scientific practices.

|  |  |  |  |
| --- | --- | --- | --- |
| ***Framework*** | ***Practices*** | ***Disciplinary Core Ideas*** | ***Crosscutting Concepts*** |
| ***Macroscopic explanation*** | 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 |

## The FATLOSS Item

This section includes the following:

1. The FATLOSS item
2. Human scoring rubric for students’ written explanations
3. A summary of item statistics for machine scoring
4. Connections to NGSS and the NRC *Framework*

This item is structured like most items on the assessment in that it contains (a) a series of forced-choice responses and (b) one or more constructed response follow-up questions in which students explain their reasoning. The version below is a teacher’s version in which blue italics provide guidance about correct responses and how the learning progression framework can be used to interpret students’ explanations.

Like all the items on the test, this item is the result of a 10-year iterative development process that included multiple revisions in the wording of the item and scoring rubrics, explained in Thomas, et al. (2019); Thomas, (2019). (We found, for example, that wording the question to ask about atoms rather than more generic wording—"What happened to the man’s fat?”—improved the quality of Level 4 responses but did not affect Level 2 or Level 3 responses.)

### FATLOSS Item Text

Fat is mostly made of molecules such as stearic acid: C18H36O2.

**a) Decide and circle whether each of the following statements is true or false about what happens to the atoms in a man’s fat when he exercises and loses weight.**

***True*** False Some of the atoms in the man’s fat are incorporated into CARBON DIOXIDE in the air.

True ***False*** Some of the atoms in the man’s fat are converted into ENERGY that he uses when he exercises.

True ***False*** Some of the atoms in the man’s fat are BURNED UP AND DISAPPEAR.

True ***False*** Some of the atoms in the man’s fat are converted into HEAT.

***True*** False Some of the atoms in the man’s fat are incorporated into WATER VAPOR in the air.

**b) Explain the pattern in your answers. What happens to the atoms in the fat of a person who loses weight?**

***Level 4 explanations recognize that the fat molecules are oxidized in a chemical change. For example:***

* ***They may explain that the fat molecules are being used for cellular respiration, meaning that the products are CO2 and H2O.***
* ***They may invoke conservation laws and “follow the rules.”***
  + ***Atoms cannot be burned up or disappear.***
  + ***Atoms cannot change into forms of energy.***
  + ***There must be products of the matter change that include the atoms in stearic acid: C, H, and O.***

*Level 3 explanations may suggest that fat molecules are converted into energy/heat.*

*Level 2 explanations may suggest that the fat is “burned up” and disappears.*

### Human Scoring Rubric for Students’ Written Explanations

The table below contains the rubric used for human scoring of students’ explanations, including (a) specific indicators for each learning progression level and (b) examples of student responses at that level.

| ***Level*** | ***Indicators*** | ***Example Responses*** |
| --- | --- | --- |
| 4 | Indicators: 1. Traces atoms from fat molecules to CO2 and water vapor (although mention of water vapor not necessary)   2. Traces atoms from fat molecules to other molecules AND chooses True for some atoms incorporated into CO2 | (4.1) A person gets energy by breaking up food monomers (this is called cellular respiration). Fat molecules in a body can turn into glucose. The glucose then breaks up into CO2 and H2O and H2O gets released (for example sweating) and the CO2 gets breathed out.  (4.2) The atoms of fat our converted into other places and substances but not destroyed (FC: CO2, T; H2O, T) |
| 3 | Indicators:  1. States that fat is converted into heat or energy (M/E conversion); may mention exercise as the process through which this happens    2. States that fat is converted into sweat or other molecules/substances without addressing CO2 as an accompanying byproduct   3. States one of the laws of conservation of energy OR conservation of matter  4. States that ATOMS exit the body (saying "fat" leaves the body is not enough) | (3.1) When a man is losing weight some of the atoms of the mans fat turns into sweat which is the water vapor. Also when someone is hot from running it is because the atoms turn into heat and energy.   (3.2) The fat burns from a person sweating when they are exercising, so while they are exercising they are burning the calories, so they begin to sweat, and that sweat loses their weight.   (3.3) The atoms of the fat cannot just disappear and go away. They have to go into something else or be converted into something else.   (3.4) They exit the body. |
| 2 | Indicators: 1. States that the atoms “burn up”, are used up, or converted with no other information  2. States that atoms/fat completely disappear OR that fat leaves the body  3. States that the atoms shrink/get larger  4. Just includes everything from the prompt in a list form without an explanation  5. ONLY explanation is that does NOT become CO2 or H2O  6. Speaks in general terms of the processes associated with weight loss (e.g., exercising, sweating) without recognizing the molecular underpinnings of these processes  7. Does not correctly describe a physiological process i.e. fat turns to muscle without any of the L3 or L4 indicators | (2.1) They will eventually start to burn up.   (2.2) the atoms disappear  (2.3) They start to shrink.  (2.4) what happens is thatoms burn up and dissapear, some of them are converted into energy, and some them are incorporated into water vapor in the atmosphere.   (2.5) I don't believe atoms in the man's fat is turned into carbon dioxide in the air.   (2.6) the person that losses the weight will be able to get more weight.   (2.7) turns to muscel |
| 0 | Unintelligible, nonsense, not related to question |  |
| 7 | I don’t know, I guessed, ? OR provided multiple choice with I don't know, I guessed, ? or similar |  |
| 8 | Choice with no explanation |  |

### Machine scoring

A sample of 2,116 human-scored responses was used as a training sample for machine learning. If machine scoring was not sufficiently reliable with human scoring (quadratic weighted kappa > 0.7), then the rubric was revised and an additional sample of responses was human- and machine-scored. Items that did not meet the criterion of quadratic weighted kappa > 0.7 were discarded (see Gambrell, Thomas, Meisner, and Bolender, 2016; Thomas, Kim, and Draney, 2018). The final model for this item had an initial QWK of 0.764 for the training set. A stratified random sample of ML scored responses was backchecked by expert human coders to confirm the continued reliability of the scoring engine. The backchecked sample had a QWK of 0.740 which was still above the accepted threshold for using the ML scoring. (Thomas, Holste, Draney, Batthia, and Anderson, 2019; Thomas, 2019). In total, this item was scored 84,487 times across the four years of machine scoring.

### Connections to NGSS and the NRC Framework

No single item can comprehensively assess an NGSS performance expectation or a dimension of the NRC *Framework.*  FATLOSS does, however, include elements that assess the following:

**NGSS Performance expectation: HS-LS1-7.** Use a model to illustrate that cellular respiration is a chemical process whereby the bonds of food molecules and oxygen molecules are broken and formed resulting in a net transfer of energy.

**Science and engineering practices:**

2. (Developing and) using models

6. Constructing explanations (and designing solutions)

**Crosscutting concepts**

4. Systems and system models

5. Energy and matter: flows, cycles, and conservation

**Disciplinary core ideas**

LS1: C: Energy and matter flow in organisms:

* As matter and energy flow through different organizational levels of living systems, chemical elements are recombined in different ways to form different products.
* As a result of these chemical reactions, energy is transferred from one system of interacting molecules to another and energy is released to the surrounding environment and to maintain body temperature. Cellular respiration is a chemical process whereby the bonds of food molecules and oxygen molecules are broken and new compounds are formed that can transport energy to muscles.

## PLANTCLAIM Item

The item below is an inquiry-focused item, with subparts addressing the practices of asking questions, analyzing data, arguments from evidence. The version below is a teacher’s version in which blue italics provide guidance about correct responses and how the learning progression framework can be used to interpret students’ explanations.

A class is investigating how plants grow. The teacher asks the students, “Where does most of the mass of a plant come from?”

a. Three students shared their ideas about what happened. Do you agree or disagree with each student’s claim?

|  |  |  |
| --- | --- | --- |
| ***Agree*** | ***Disagree*** | Mike: "I think a growing plant gains most of its mass from nutrients in the soil." |
| ***Agree*** | ***Disagree*** | Lucia: "I think a plant gains most of its mass from gases in the air." |
| Agree | ***Disagree*** | Oscar: “I think a plant gains most of its mass from the sunlight.” |

b. Provide an explanation. Why did you agree or disagree with each student’s claim?

***Level 4 responses disagree with Oscar because matter cannot be converted into energy in chemical and physical changes; agrees or disagrees with Lucia that air/gases can provide mass to plants; agrees or disagrees with Mike because soil nutrients and/or water provide mass to plants.*** *Level 3 responses may include: i) sunlight is a source of matter for plants (agrees with Oscar) AND/OR ii) disagrees with Lucia's claim that air could account for the tree's increased mass. Level 1 of Level 2 responses may only reason about 1 or 2 of the claims in a force-dynamic way; e.g., i) the plant needs light, soil, water to grow, ii) air/gas cannot provide mass to the plant.*

c. The class does an experiment to investigate how plants grow. They started by selecting six **identical** plants. Three of those plants were grown in regular soil. The other three plants had extra soil nutrients added to the soil in their pots. The class put all six plants under **identical** conditions (i.e., the same light conditions, the same watering conditions) and let them grow for one month. At the end of the month, the class weighed each of the six plants and recorded their weights in the table below. They also recorded the weight of the soil nutrients added to three of the pots.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Plant** | **Initial weight (g)** | **Added soil nutrients (g)** | **Final weight (g)** | **Plant growth (g)** |
| 1 | 30 | 0 | 50 | 20 |
| 2 | 31 | 0 | 52 | 21 |
| 3 | 29 | 0 | 48 | 19 |
| **Average** | **30** | **0** | **50** | **20** |
| 4 | 30 | 3 | 68 | 38 |
| 5 | 31 | 3 | 62 | 31 |
| 6 | 28 | 3 | 65 | 37 |
| **Average** | **30** | **3** | **65** | **35** |

Which claim do you think is best supported by the data? (Circle one choice.)

1. Mike’s claim
2. ***Lucia’s claim***
3. Oscar’s claim

Explain how the patterns in the data support the claim that you chose.

***Level 4 responses recognize that there is an unaccounted for difference between the decrease in soil mass and the increase in plant mass and use this mass discrepancy as evidence to support Lucia’s claim.*** *Level 3 responses identifiy all matter pools, or recognizs missing pools, but agree with Oscar’s claim that violates principles (Matter to Energy conversion), agree with Mike’s claim which is inconsistent with the data, or agree with Lucia’s claim but has flaws in their reasoining. Level 1 and 2 responses choose Mike’s claim because the plant gained mass while the soil lost mass. OR select claims based on force-dynamic reasoning without using mass data (may provide an explanation about food use or soil nutrients for rationale.*

d. What additional evidence would you collect to help show that the claim you chose is the best claim?

***Level 3 responses propose questions that target limitations in the data (recognize there is an unaccounted for matter pool, i.e., gas); they focus on matter tracing and are constrained by principles such as matter to energy conversion.*** *Level 2 responses propose evidence that partially address limitations in the data. Level 1 responses identify aspects of the system that students are curious about independent of the data, they critique the experimental design, or do not recognize that additional evidence needs to be collected.*

References

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

Thomas, Jay, JinHo Kim, and Karen Draney. Measuring progress toward measuring three-dimensional learning at scale: Machine Scoring and IRT Analysis. NARST; 2018 March 12; Atlanta, GA, USA.

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

Thomas, Jay, Karen Draney, and Shruti Bathia. Using machine learning to make assessment of NGSS based three-dimensional science scalable. Paper accepted for AERA Conference in San Francisco, CA: April, 2020.

# Figure S1

*Mean and 95% Confidence Interval of Value-added Measures for 133 Teachers and Their 245 Classrooms* Chart

Description automatically generated

*Note:* We calculated confidence interval based on standard error of the HLM model. However, the estimation of standard errors in VAM is controversial because the standard errors can be quite sensitive to how one conceptualizes the level of analysis and different sources of dependencies among students (Bibler et al., 2014).

# Table S1

*Reliability and Person Fit*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2015–2016 | | 2016–2017 | | 2017–2018 | | 2018–2019 | |
|  | Reliability | % Person misfit | Reliability | % Person misfit | Reliability | % Person misfit | Reliability | % Person misfit |
| Full Test | 0.793 | 1.1% | 0.786 | 13.0% | 0.879 | 15.7% | 0.874 | 15.9% |
| Systems & Scale | 0.701 | 13.7% | 0.838 | 11.4% | 0.828 | 13.7% | 0.827 | 13.7% |
| Animals | 0.736 | 1.4% | 0.812 | 12.4% | 0.845 | 13.4% | 0.841 | 13.2% |
| Plants | 0.677 | 2.7% | 0.839 | 19.1% | 0.872 | 13.6% | 0.863 | 14.8% |

*Note*: Person misfit is defined as percentage of students with weighted fit statistics larger than 1.33. According to Adams & Khoo (1996), the acceptable range of fit values is from 0.75 to 1.33.

# Table S2

*Comparing Learning Gains after Different Units*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | N | Mean | Std. Dev. | Min | Max |
| S&S gain | 10,832 | 1.897 | 1.601 | –3.675 | 10.751 |
| Animals gain | 10,832 | 1.953 | 1.297 | –3.839 | 9.073 |
| Plants gain | 10,832 | 2.129 | 1.345 | –3.509 | 8.141 |
| Paired t test between S&S and Animal: difference = 0.056, *SE* = 0.014, *p* < 0.001. | | | | | |
| Paired t test between Animal and Plant: difference = 0.176, *SE* = 0.011, *p* < 0.001. | | | | | |
| Paired t test between S&S and Plant: difference = 0.232, *SE* = 0.014, *p* < 0.001. | | | | | |

# Table S3

*Collinearity between Teachers’ Gaining Experience and Improvement in the Project Over Time*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Notes.* Number of teachers are presented in each cell. | | Teachers’ gaining experience | | | |
| Y1 | Y2 | Y3 | Y4 |
| Improvement in the project | 2015–2016 | 26 |  |  |  |
| 2016–2017 | 56 | 19 |  |  |
| 2017–2018 | 40 | 42 | 9 |  |
| 2018–2019 |  | 31 | 15 | 7 |

# Methods Supplement

This supplemental material presents how we decomposed the variance in students’ learning gain scores into three factors: school factors, students’ prior knowledge, and teachers’ identities. Specifically, we compared the change in variance explained when including different variables into the model. That said, however, because of the potential collinearity between variables, we cannot cleanly distinguish the contribution of each factor in explaining the variance in students’ learning gains.

Figure 1 presents the potential overlap among the three factors. Each circle represents the variance each factor explains in students’ learning gains (i.e., the outcome variable). The figure is based on conventional representations of overlap in variances. But the Figure does not explicitly show the overlap with the outcome. This will be explained below.

On the left-hand side are three overlapping classroom-level factors, including average pretest, school factors, and teachers, and each circle overlaps. On the right-hand side is a student-level factor—individual students’ deviance from their classroom’s average pretest scores. As explained in the paper, the student-level group centered pretest is orthogonal to the included classroom-level factors; thus, there is no overlap between the three classroom-level factors and the one student-level factor. To make the following analyses easier to follow, we use different characters to represent each factor and potential relationships between and among factors. A represents the variance in the outcome that is uniquely explained by Average pretest. B represents the variance in the outcome that is uniquely explained by School factors. Meanwhile, C and D represent the collinearity between students’ classroom average pretest scores and school factors in explaining the variance in student learning gains. That is, we cannot statistically distinguish if students’ classroom average pretest scores or school factors contribute to C and D, which represent some of the variance in the outcome variable.

Accordingly, the variance of student learning gains explained by students’ classroom average pretest scores equals the sum of A, C, D, and E. Similarly, the variance in the outcome variable explained by school factors equals the sum of B, C, D, and F.

**Figure 1**

*Venn Diagram for Variance Decomposition*

A close up of a mans face

Description automatically generated

Next, we present the analysis step-by-step. We started with the most complicated model–M1 (this is the same as the Model 2 presented in Table 2 within the main text). In this model, the residual variance at the classroom-level is 0.419 (i.e., var(r00j)), and the residual variance at the student-level is around 0.909 (i.e., var(eoij)).

M1

We choose this model because the classroom-level variance (i.e., var(r00j)) is a clean measure of teacher effect, capturing what teachers can explain in students’ learning gains after controlling for all other factors. In other words, G in Figure 1 is equivalent to 0.419.

We then removed students’ classroom average pretest scores from the most complicated model M1, as shown in the following model (M2).

M2

Theoretically, the residual variance at the classroom-level should increase given the exclusion of a predictor variable. However, M2’s results were almost identical to those of M1, suggesting that the omission of students’ classroom average pretest scores produce little change. This result also means that the summation of A and E in Figure 1 is almost equal to zero. Here, the difference in residual variance between M1 and M2 does not include C and D because any variance explained by school factors were excluded from var(r100j).

Next we removed school factors from the most complicated model M1, as presented below (M3).

M3

As in M2, excluding school factors should only affect residual variance at the second level. Compared to M1, the residual variance at level 2 increased from 0.419 (M1) to 0.477 (M3). The difference, which is around 0.06, represents the sum of B and F. Here, the difference in residual variance between M1 and M3 does not include C and D because var(r200j) excludes the variance explained by students’ average classroom pretest scores.

Next, we removed both school factors and classroom average pretest from M1, as shown below (M4).

M4

Again, excluding school factors and classroom average pretest should only affect variance at the second level. Now the residual variance in M4 increased from 0.419 (M1) to 0.478 (M4). The difference, which is about 0.06, represents everything explained by either school factors or average pretest. Any overlap between these two factors and teachers are also included. That is, A + E + C + D + B + F = 0.06.

So far, we have presented the variance decomposition at the second level, from which we obtain A + E = 0, C + D = 0.001, B + F = 0.06, and G = 0.419. Next we looked at the circle on the right-hand side, which represents variance in the outcome variable explained by the student-level predictor. Specifically, we removed the student-level pretest from M1, as presented below (M5). The residual variance at the student-level increased from 0.909 (M1) to 1.053 (M5). The difference, which is about 0.144, captures H in Figure 1. In other words, this is the amount of variance explained by individual’s deviance from the classroom average pretest at the student-level.

M5

Table 1 presents a summary of the variance decomposition result. The middle column shows the amount of variance in the outcome variable explained by corresponding factors listed in the left column. The right column used a common denominator so that we obtain a percentage of the variance explained by the factors in each row. Here we used the total variance in students’ learning gain scores (1.49) since students’ learning gain is our outcome variable.

**Table 1**

*Summary of the Variance Decomposition Results*

|  |  |  |
| --- | --- | --- |
| H | 0.14 | 9.6% |
| A + E | 0.00 | 0% |
| C + D | 0.00 | 0.1% |
| B + F | 0.06 | 3.9% |
| G | 0.41 | 27.7% |
| A + E + C + D + B + F | 0.06 | 4.0% |
| A + E + C + D + B + F + G | 0.47 | 31.4% |
| A + E + C + D + B + F + G + H | 0.58 | 39.0% |

Finally, we combined the information from Table 1 and Figure 1 to calculate a range of variance in the outcome variable that can be accounted for by each factor: school factors, students’ prior knowledge measured by students’ pretest, and teachers, as summarized by Table 2.

**Table 2**

*Range of Variance Explained by Each Factor*

|  |  |  |
| --- | --- | --- |
|  | Min | Max |
| School factor | B + F = 3.9% | C + D + B + F = 4.0% |
| Students’ prior knowledge | A + E + H = 9.6% | A + E + C + D + H = 9.7% |
| Teachers | G = 27.7% | A + E + C + D + B + F + G = 31.4% |

# Abstract: Effects of Teaching Strategies on Students’ Three-dimensional Learning in a Design-based Implementation Research Project

This article reports findings from classroom videos, interviews, and student work in a design-based implementation research (DBIR) project that aims to support student learning at scale, with learning defined as achieving three-dimensional learning goals in the Next Generation Science Standards (NGSS). The curriculum, assessments, and professional development in this project helped most participating teachers make substantial changes in their classroom practices that led to improvements in students’ three-dimensional learning, but the effects of participation in the project were uneven. Our study compares classrooms where students were more and less successful in achieving three-dimensional performance expectations.

This article uses value-added measures from quantitative analyses of student learning data to support qualitative analyses of the activities of teachers and students in individual classrooms. We draw on six case studies of teachers and students that include video records, student work, and interviews with teachers and students. All six teachers and their students completed at least three instructional units designed by the project. These units posed instructional challenges for the teachers. In particular, all of the units included sequences of tasks that were high in intellectual demand: They were designed to engage students in three-dimensional target performances that involved making sense of phenomena.

Our analyses describe how high-learning gains and low-learning gains classrooms were alike and different in their responses to this challenge.

**1. How all classrooms were similar.** All the classrooms were alike in that engaging students in making sense of phenomena was not teachers’ only concern; the teachers also responded to obligations that came from their students, parents, colleagues, and administrators. In particular, this meant that teachers made sure that students who attended class regularly and completed their assigned work received passing grades—a basic obligation of secondary school teaching. Thus all teachers confronted a tension between this basic obligation and the high intellectual demands of three-dimensional sensemaking.

**2. Differences between higher- and lower-learning gains classrooms.** Teachers responded to this tension in different ways. Teachers in lower-learning gains classrooms tended to accomplish this goal by simplifying three-dimensional tasks, reducing both their difficulty and their potential for three-dimensional learning. Teachers in higher-learning gains classrooms modeled, coached, and scaffolded their students’ three-dimensional performances. These teachers used many of the strategies for scaffolding students’ three-dimensional engagement with phenomena described in the current design-based research. We also note two additional dimensions of their practice that are not as widely discussed in current science education literature; teachers with higher learning gains (a) used a variety of apprenticeship strategies to model, coach, and scaffold their students’ three-dimensional performances and (b) held students accountable for canonically correct three-dimensional writing performances.

This article addresses a challenge: How can the current standards-based reform effort in science education lead to system-wide effects on classroom teaching and students’ three-dimensional learning? The evidence in this study leads to some ideas about investments that will be necessary (though probably not sufficient) to meet this challenge, including investments in curriculum, classroom and large-scale assessment, professional development, and school professional communities.

# Abstract: Connecting Teachers’ Commitments and Perceptions to Classroom Discourse and Student Learning Outcomes

In this article we analyze data from interviews with 68 teachers who participated in the project focusing on their commitments and perceptions of teaching and learning. We describe patterns in teachers’ interview responses that were associated with differences in their students’ learning gains. In particular, we examine patterns in teachers’ commitments and perceptions with respect to (a) their professional identities, (b) their students’ potential and accomplishments, (c) the project*’s* utility in meeting their goals, and (d) extent of support from and alignment with other professionals in their local context.

**Comparing interviews of teachers whose students had higher and lower learning gains.** A small minority of teachers reported discouragement or negative perceptions of their students’ success, the project units, or their relationships with their professional colleagues. Students in these teachers’ classrooms generally had lower learning gains. Most teachers, including teachers with higher and lower learning gains, reported more positive perceptions. There were, however, important differences in how teachers with higher and lower learning gains described teaching and learning in their classrooms.

The higher-learning gains teachers talked specifically about their commitments to students’ three-dimensional learning and provided examples of students’ success in three-dimensional performances. They described ways that they used the project materials to scaffold students’ three-dimensional performances. They sometimes acknowledged limitations in students’ initial knowledge and practice (“raw talent” in the words of one high-learning gains teacher) and students’ complaints about working too hard, but they portrayed students as ultimately willing and able to achieve difficult performances.

The lower-learning gains teachers described their classrooms in less specific terms (“constructivist,” “student-centered”), focused on how they made their teaching fun or interesting for their students, or focused on one-dimensional goals for learning content or skills. They described their students’ interest in investigations and hands-on activities while indicating that students were less interested in sustained work toward three-dimensional performances. They modified the project units to add science content; reduce what they saw as repetitive activities; or simplify tasks (generally reducing them from three to one-dimensionality) so that students would be more successful.

**School professional communities.** Many teachers reported considerable professional autonomy, so that school professional communities were not particularly salient to their classroom teaching and learning. A few higher-learning gains teachers described school professional communities that supported three-dimensional teaching and learning. A few lower-learning gains teachers described their instruction being constrained by one-dimensional shared assessments or pacing guides.

**Summary.** The most important differences among teachers concerned their understanding and commitment to three-dimensional learning goals and their beliefs about students’ interests and abilities. Higher-learning gains teachers described strategies for supporting students’ three-dimensional learning. Lower-learning gains teachers described ways to motivate students with pace and novelty while pursuing one-dimensional learning goals.

# Sensitivity Analysis Supplement

This supplemental material presents how we filtered the data step by step. For each step in the exclusion procedure, we provide rationale for why we implemented the exclusion, how much data were excluded, as well as how hierarchical linear model estimates, and corresponding inferences changed. Most of our conclusions stay the same across all steps of the exclusion procedure (see Table 2 below). Below we provide more detail.

We started with a sample with all students with full pretest and at least one unit posttest (Sample 1). It would be impossible to calculate a gain score without both a before and after score. Although we could use a psychometric algorithm to predict the missing score based on other student data, we felt it was more appropriate to base our calculations and conclusions on the data that we observed rather than estimated or simulated. Then we excluded data following three steps as below:

1. Starting from Sample 1, we excluded students with pretest or posttest lower than –5, which generates Sample 2. In order to achieve a score this low, multiple and often the majority of answers on the test would have been recorded as missing, off-topic, “I don’t know”, or unresponsive and coded as a zero rather than a score on the learning progression. This indicates a lack of meaningful effort which was also confirmed by response time data. This was also a way to eliminate incomplete tests as items that were left unanswered either purposefully skipped or never reached were included in the zero score category.
2. Starting from Sample 2, we excluded teachers with fewer than 15 students’ data in one year, which generates Sample 3. Teachers with fewer than 15 students would lack the sample size needed for reliable calculation of the value added scores. It would also indicate that less than one full class had usable data.
3. Starting from Sample 3, we excluded students with only one unit data, meaning only students with at least two units’ data were included. This generates the final sample (Sample 4). Students who completed only one unit test had several issues that were problematic. Some were transient, either by switching teachers within a school out of the project or by completely leaving the school; these students did not receive all the instruction and would not fairly measure the gains of students who did participate in the program. They may have never completed the work, so that it would be impossible to fully ascertain what they knew and could do about the unit. Having only one data point for post analysis would weaken the inferences that we could make.

Table 1 below presents how sample size changes with each step of exclusion, from which we can see the largest sample change happened in the last step when we excluded students with only 1 unit data.

**Table 1**

*Sample Size Change in Each Step of Sample Filtering*

|  |  |  |
| --- | --- | --- |
|  | Number of students | Number of Teacher\_Year |
| Sample 1: all students with full pretest and at least one unit posttest | 19,571 | 285 |
| Sample 2: exclude students with pretest or posttest lower than –5 from Sample 1 | 19,025 | 283 |
| Sample 3: exclude teachers with fewer than 15 students’ data from Sample 2 | 18,913 | 272 |
| Sample 4: exclude students with only one unit data from Sample 3 | 16,195 | 245 |

Figures 1–3 below compared the distribution of pretest, posttest, and gain scores for sample 1 and 2, where the difference is due to the exclusion of low effort students. Importantly, Figure 3 shows that we are not only removing students with low gain scores but also those with high gain scores, because both cases can be due to lack of effort (in either pretest or posttest).

**Figure 1**

*Comparing Distribution of Full Pretest in Sample 1 and 2*



**Figure 2**

*Comparing Distribution of Unit Posttest in Sample 1 and 2*



**Figure 3**

*Comparing Distribution of Gain Scores in Sample 1 and 2*



In Table 2 below, we report how the hierarchical model (using Model 2 in the main text as an example since this is the main model) results changed for each step of sample exclusion. The only change in statical inference comparing Sample 1, 2, 3 versus Sample 4 is the inference regarding the average pretest: the average pretest is negatively associated with students’ learning outcome based on Sample 1 while this significant association essentially went to zero (less than its standard error) once low score students were excluded. We investigated that this negative relationship applies to both cases: exclusion of students with extremely low *pretest* only, as well as exclusion of students with extremely low *posttest* only. This indicates that the negative association is not driven by any particular exclusion of pretest or posttest, adding further support for the argument that this significant negative association is invalid. From another perspective, this also means the effect size we reported in the paper using Sample 4 is conservative. The other major change reflected in Table 2 is that Sample 1 has the largest residual variance, which is consistent with Figure 3 above, indicating that the total variance in the gain score outcome shrunk from Sample 1 to Sample 2.

**Table 2**

*Parameter Estimates for Two-level Hierarchical Linear Models 2 Predicting Students’ Gain Scores*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Outcome: Gain score | Sample 1: Model 2 | Sample 2: Model 2 | Sample 3: Model 2 | Sample 4: Model 2 |
| Deviation from class average pretest. | – 0.671\*\*\* | – 0.504\*\*\* | – 0.504\*\*\* | – 0.507\*\*\* |
| (0.008) | (0.010) | (0.010) | (0.010) |
| Class average pretest. | – 0.684\*\*\* | – 0.024 | – 0.033 | – 0.135 |
| (0.099) | (0.128) | (0.130) | (0.138) |
| Percent of free and reduced lunch. | – 1.492\*\*\* | – 0.993\*\*\* | – 0.969\*\*\* | – 1.077\*\*\* |
| (0.395) | (0.275) | (0.282) | (0.293) |
| Percent of marginalized students of color. | – 0.197 | – 0.374 | – 0.495 | – 0.544 |
| (0.414) | (0.283) | (0.292) | (0.304) |
| Constant | 1.154\*\*\* | 2.059\*\*\* | 2.079\*\*\* | 2.028\*\*\* |
| (0.174) | (0.186) | (0.188) | (0.195) |
| Observations | 19,571 | 19,025 | 18,913 | 16,195 |
| Number of Teacher\_Year | 285 | 283 | 272 | 245 |
| Level 2 residual variance | 1.011 | 0.431 | 0.429 | 0.419 |
| Level 1 residual variance | 1.240 | 0.993 | 0.996 | 0.909 |

*Notes.* Standard error in parentheses. \*\*\* *p* < 0.001, \*\* *p* < 0.01, \* *p* < 0.05.

We also evaluated how robust our inference is with respect to potential bias using the Robustness of Inference to Replacement (RIR) (Authors, 2000; Authors, 2013). Specifically, RIR describes how different the sample would need to be to change our inferences. We report the results for the two most important predictors below in Table 3. The RIR number can be interpreted as how many students need to be replaced with students for whom the effect is zero to invalidate the inference at a 0.05 level (i.e., change the inference from statistically significant to not statistically significant). For example, to invalidate the inference that percentage of free and reduced lunch is negatively associated with students’ learning gains based on Sample 4, we need to replace 7559 students (47%), which is much larger than the sample difference from Sample 1 to Sample 4 (19751-16195=3556, 18%). This indicates that our inference is robust with respect to our filtering of the sample. The inference regarding the deviation from class average pretest is even more robust, for which the RIR indicates that one need to replace more than 96% of the students to invalidate the inference.

**Table 3**

*Robustness of Inference to Replacement – Robustness Analysis for Important Predictors*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Outcome: Gain score | Sample 1: Model 2 | Sample 2: Model 2 | Sample 3: Model 2 | Sample 4: Model 2 |
| Deviation from class average pretest. | 19,114 (98%) | 18,285 (96%) | 18,177 (96%) | 15,569 (96%) |
| Percent of free and reduced lunch. | 9,415 (48%) | 8,698 (46%) | 8,142 (43%) | 7,559 (47%) |
| Observations | 19,571 | 19,025 | 18,913 | 16,195 |
| Number of Teacher\_Year | 285 | 283 | 272 | 245 |

***Note.*** The RIR number means how many students need to be replaced with students for whom the effect is zero to invalidate the inference at a 0.05 level (i.e., change the inference from statistically significant to not statistically significant). The percentage in the parenthesis indicates the percentage of RIR out of the total sample.