Methods Supplemental Material: Quantitative Analyses of Students' Learning Gains

This supplemental material explains the data analyses that underlie Figures 1 and 3 in the article: Designing Educational Systems to Support Enactment of the Next Generation Science Standards.

- Figure 1: How we conducted IRT-based analyses to develop estimates of students individual and collective proficiency in terms of logits and learning progression levels.
- Figure 3: how we applied a two-level hierarchical linear model to study learning gains of students participating in *Carbon TIME*.

Figure 1: IRT Analyses

Figure 1 compares IRT-based estimates of student pretest and posttest proficiencies with end of school year baseline levels (students of the same teacher the year before) for the first two years of this study.

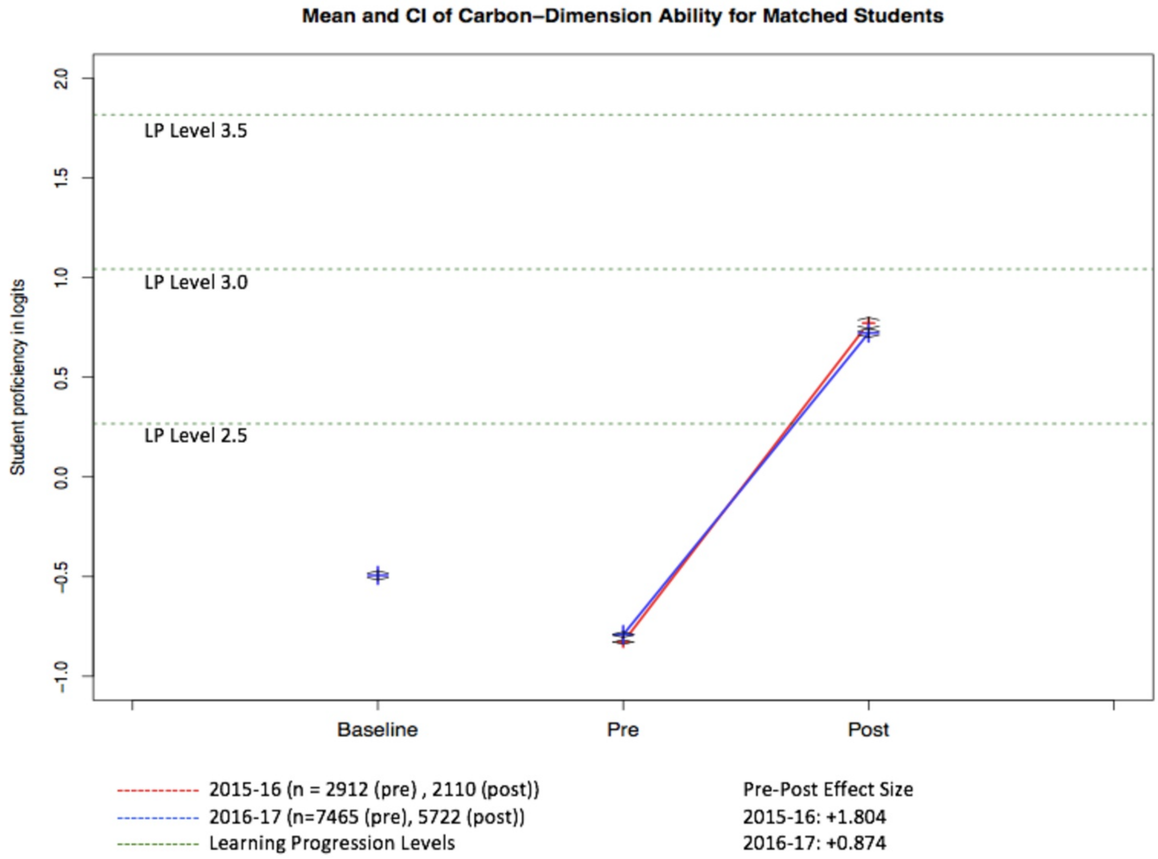


Figure 1: Mean learning progression (LP) levels of students in Carbon TIME and baseline (classes of participating teachers the year before they started using Carbon TIME). Error bars

represent 95% confidence intervals. LP Level 4 is equivalent to full achievement of NGSS performance expectations in this domain.

These results can be interpreted from norm-referenced and criterion-referenced perspectives. Students improved significantly compared to both pretest and baseline performances. In other studies we have shown that high school students participating in *Carbon TIME* show higher proficiency on learning progression-based assessments than college science majors in biology courses (Rice, Doherty, & Anderson, 2014). However, from a criterion-referenced perspective, these results also show that most students still fall short of Level 4 of the learning progression, which is equivalent to the *NGSS* performance expectations in this domain.

This supplemental material explains how we applied a latent regression item response model (Adam, Wilson and Wu, 1997) to study students' learning gains by participating in *Carbon TIME*.

Data:`

We used the individual item responses of students in the full tests to determine learning gains. For each student, we have their responses to each item both at the beginning (Pre Test) and at the end (Post Test) of a school year. For the 2016-17 school year, we have a total 5722 responses for the pre test and 7465 responses for the post-test. For the 2015-16 school year, we have a total of 2912 responses for the pre test and 2110 responses for the post test. All following analyses were conducted with these data.

Method:

We used a one-parameter item response modeling approach, based on the partial credit model, to determine both the difficulties of the various item steps on the pretest/posttest items, and the proficiencies of the students, using the ConQuest software (Wu, Adams, Wilson, and Haldane, 2007). This software is based on the item response model (the Multidimensional Random Coefficients Multinomial model, or MRCML) described in Adams, Wilson, and Wang (1997), and allows estimation of both unidimensional and multidimensional models, as well as latent regression models and item effects.

Using this model allowed us to anchor the difficulties of identical items used in the pretest and the posttest, as well as items that were the same in the 2015-16 and 2016-17 tests, which provides person estimates that are directly comparable across all time periods and test forms.

We used a latent regression model to determine the pretest-posttest differences in each of the two school years. The advantage of latent regression is that it provides means and standard errors which have been adjusted for measurement error, resulting in more accurate confidence intervals and statistical significance tests. The mean achievement scores of the two groups are directly estimated from the item response data without first producing individual student scores. Using a unidimensional latent regression model, the pre test and post test data is analyzed in the ConQuest software (used for item response data).

The mean scores for pre test and post test along with the standard errors are computed. They are represented in the logit scale. Logits are a measure of how likely a student of some proficiency is to get a particular item right or wrong. The zero logit is set to be the student average.

Results:

We use the latent regression output from ConQuest to compute the mean achievement scores for each group. We use the weighted likelihood cases estimates to compute standard deviation of the groups.

Table 1 : Descriptive Analysis for Full Test 2015 - 16

	N	Mean	SD*	Difference in Mean	Pooled SD	Effect Size
Pre Test	2912	-0.828	0.56	1.60	0.89	1.80
Post Test	2110	0.772	1.20			

Table 2 : Descriptive Analysis for Full Test 2016 - 17

	N	Mean	SD*	Difference in Mean	Pooled SD	Effect Size
Pre Test	7465	-0.791	1.98	1.51	1.73	0.87
Post Test	5722	0.722	1.34			

As we can see from Table 1 and Table 2, the difference in overall learning gains between pre test and post test is 1.6 for 2015-16 and 1.513 for 2016-2017. This difference is statistically significant in both the cases. ($p<0.001$). Based on this evidence, we draw the conclusion that the mean achievement scores in the pre-test and post-test significantly differ in both the tests.

We also compute the effect size using the Cohen's d formula. Cohen's d is determined by calculating the mean difference between the two groups, and then dividing the result by pooled standard deviation. We use Cohen's for effect size measure because in our case, the two groups (pre and post) have similar standard deviations and are of similar size.

Based on the above, the effect size for learning gains in 2015-2016 is + 1.80 and for 2016-17 is +0.874. This means that the two groups differ by 1.80 standard deviation in 2015-16 and 0.87 standard deviation in 2016-17. Cohen suggested that an effect size greater than 0.8 is considered to be a 'large' effect size.

We also represent the learning progression levels 2.5, 3 and 3.5 on the graph as the lines that pass through 0.267, 1.043 & 1.818 logits respectively. Essentially these learning progression levels represent the median thresholds or cut points of all explanatory items of carbon dimensions at those particular levels.

Figure 3: HLM Analyses

Figure 3 compares student pre-post learning gains for 58 individual teachers in the 2016–17 school year. (The online supplemental materials include descriptions of hierarchical linear models (HLM)-based analyses that support Figure 3 and the other conclusions below; see Methods Supplement.) The differences among classrooms are both statistically and educationally significant. Students gained an average of 1.41 logits (about 0.91 learning progression levels on a scale from Level 2 to Level 4). In the most successful classrooms learning gains were about twice the average; in the four least successful classrooms students did worse on the posttest than on the pretest. (See the supplemental materials for a detailed discussion of HLM analysis methods and checking data quality.)

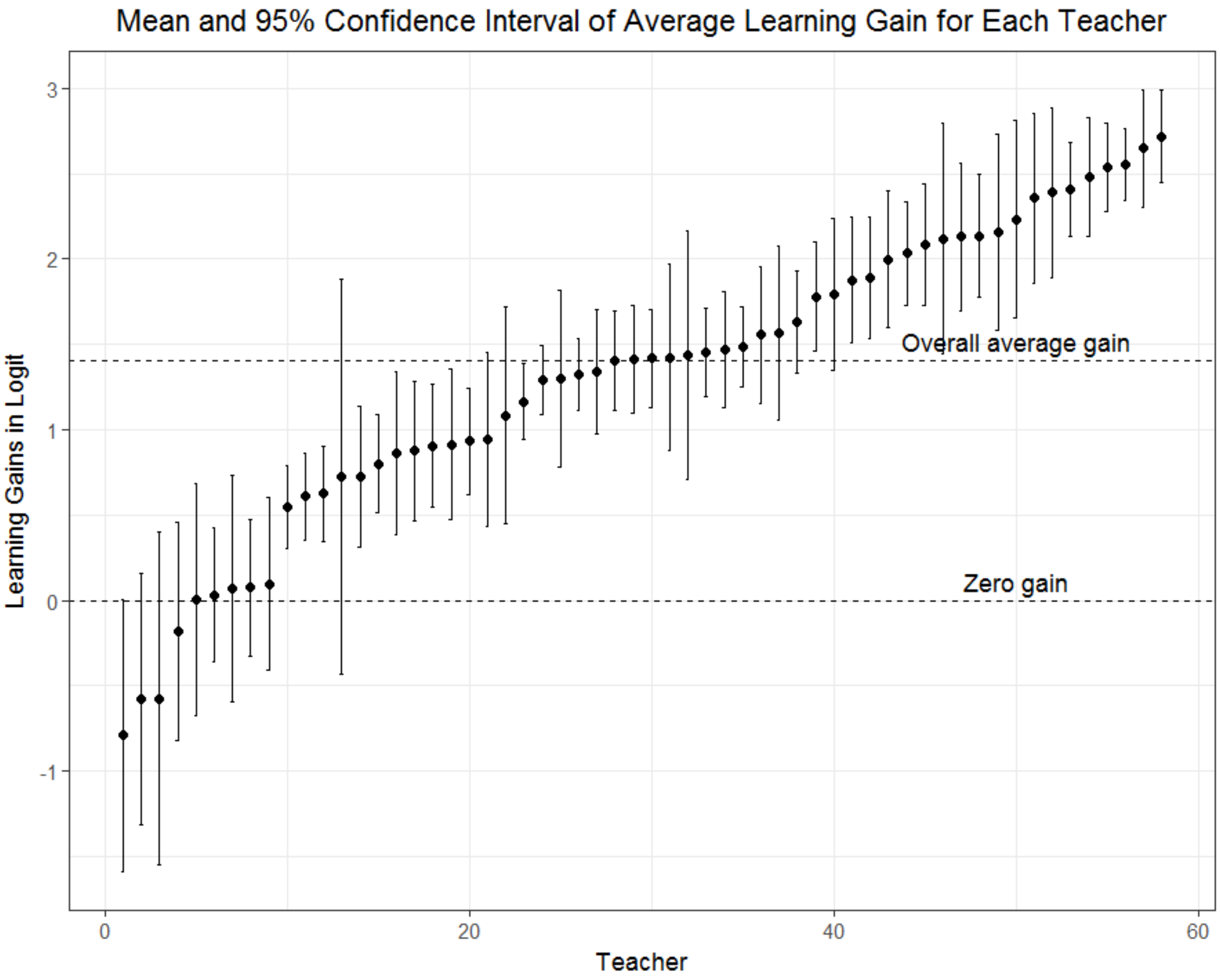


Figure 3: Comparing learning gains for macroscopic explanations (carbon dimension) for individual teachers for the 2016–17 school year. All teachers (N = 58) who had at least 20 students taking both pretest and posttest are included. Zero logits indicates no difference in average pretest and posttest proficiencies. One learning progression level is a gain of about 1.55 logits. The average learning gain for all students of these teachers was 1.41 logits.

Additional HLM analyses investigated associations between student learning gains and other variables associated with diversity in students and schools. We applied a two-level model in which students' learning gains are predicted by (a) how much individual students deviate from their class-average pretest proficiencies, (b) class-average pretest proficiencies and (c) the school's percent of free and reduced lunch. Separate analyses of 2015–16 and 2016–17 data show consistent patterns:

- *Carbon TIME reduced the achievement gap between higher-pretest and lower-pretest students within classrooms.* Within classes, students with lower pretest proficiencies showed significantly higher learning gains.
- *Carbon TIME was less successful in higher-poverty schools with fewer organizational resources.* The school percentage of free and reduced lunch was negatively associated with class-average learning gain. That is to say, classrooms from schools with higher percent of free and reduced lunch benefit less from implementing *Carbon TIME*. We discuss this finding in more detail below; we interpret it as evidence that schools with more organizational resources are more successful in implementing *Carbon TIME*. Previous studies have shown the percent of free and reduced lunch can be a proxy measure for material, social, and human material resources such as students' access to qualified and experienced teachers (Darling-Hammond, 2004; Rice, 2010) and the overall quality of conditions in which teachers work (Johnson, Kraft & Papay, 2012).
- *Other variables were not significantly associated with student learning gains.* We also investigated other variables, including grade band (middle school vs. high school), racial composition of students, and class average pretests. None of these variables added significantly to the predictive power of models that included the three key variables above: individual teachers, student pretest, and school percentage of free and reduced lunch.

Data:

We used the estimated latent proficiency scores from our IRT analysis as students' proficiencies in the Carbon dimension in Full Tests. For each student, we have their proficiency scores both at the beginning (Pretest) and end (Posttest) of a school year. We excluded students who miss either Pretest or Posttest and teachers who have fewer than 20 students. After that, we have 58 teachers and 3527 students for 2016-17 and 25 teachers and 1800 students for 2015-16. All following analyses were conducted with these data.

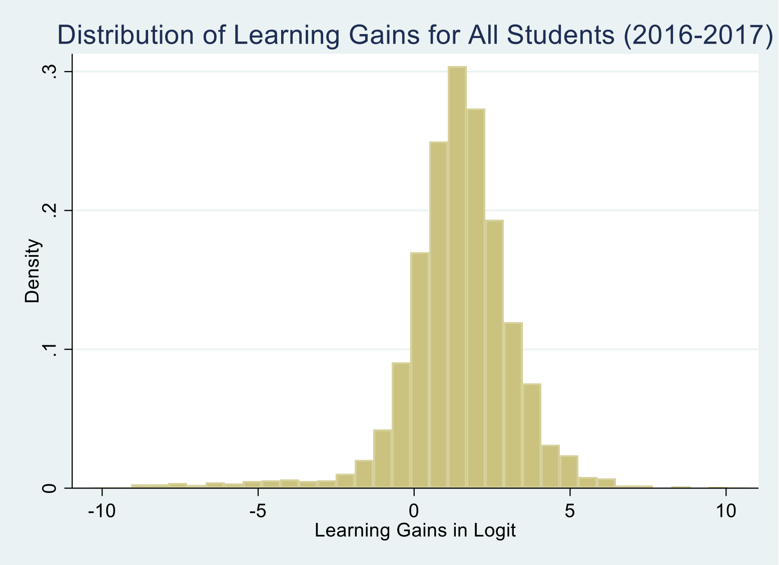
We also collected teacher-level variables including affluence of schools measured by percentage of free and reduced lunch (FRL), grade levels measured as high school or middle school (Gradeband) and percentage of minority students (Minority).

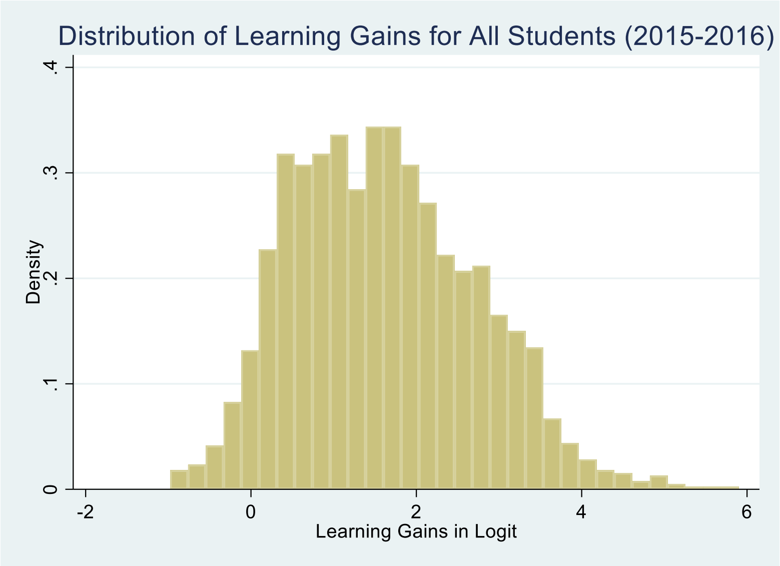
Method:

We applied two-level hierarchical linear models to study what factors may affect how much students learn from participating in *Carbon TIME*. The first level is at student level and the second level is the teacher level. We did not have school as the third level because most of the teachers are from different schools. We used the learning gains (the difference between pretest and posttest) as our outcome variable because we think this provides us a good measure of students' learning through *Carbon TIME*. We acknowledge that measurement errors in the gain score may bias our finding and interpretation. Therefore, we applied the method introduced by Willett (1988) to adjust the results for measurement error in test scores. We started from the unconditional model to decompose the variations of students' learning into within-classroom and between-classroom. Then we fit the most complicated model with all the related variables of interest to study what factors can help explain all these variations. Then we excluded those predictors that did not make significant contributions in explaining the variation in students' learning and interpreted the results in the most parsimonious model.

Results and findings:

- *Students learn by participating Carbon TIME but they vary greatly from each other in terms of how much they learn from Carbon TIME.* Following two figures show the distribution of all students' learning gains for 2015-16 and 2016-17. Most of the students improved from Pretest to Posttest (the Learning Gain is larger than 0). The overall average learning gain is about 1.611 for 2015-2016 and 1.405 for 2016-2017 in terms of the logit ($p < 0.001$ in paired t-test). However, we can tell from these two Figures that students differ greatly from each other in terms of how much learning growth they gained from participating in *Carbon TIME*. Different teachers and many other factors could lead to variation in learning through *Carbon TIME*.





- *Teachers/Classrooms do make a significant role in helping students learn from Carbon TIME.* As the Figure 3 in the paper shows, there is large variation among the average learning gains for each teacher. To illustrate this point in a more precise way, we applied an unconditional model as follows.

  Level 1: $\text{Gain}_{ij} = \beta_{0j} + r_{ij}$

  Level 2: $\beta_{0j} = \gamma_{00} + u_{0j}$

  where Gain_{ij} is the learning gain for student i taught by teacher j. The results show that the intra-class correlation (ICC) in this unconditional model is 19.56% for 2016-2017 and 28.76% for 2015-2016, indicating a considerable amount of between-classroom/teacher variation. This is a quite high ICC among other research (Frank, 1998). Therefore, teachers/classrooms do make a difference in students' learning in *Carbon TIME*.
- *Carbon TIME increased the achievement gap between high-poverty and low-poverty schools. And the percentage of free and reduced lunch is the most important variable that helps explain the between-classroom variations in terms of learning gains from Carbon TIME.* We started with the following two-level hierarchical linear model:

Level 1: $\text{Gain}_{ij} = \beta_{0j} + \beta_{1j} \cdot \left(Pretest_{ij} - \overline{Pretest_j}\right) + r_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01} \cdot \overline{Pretest_j} + \gamma_{02} \cdot FRL_j + \gamma_{03} \cdot GradeBand_j + \gamma_{04} \cdot Minority_j + u_{ij}$

$\beta_{1j} = \gamma_{10}$

where Gain_{ij} is the learning gain for student i taught by teacher j;

$Pretest_{ij}$ is the learning proficiency for student i measured in the Pretest;

$\overline{Pretest_j}$ is the average learning proficiency of teacher j's classroom measured in the Pretest;

FRL_j is the percentage of free and reduced lunch for teacher j's school;

$GradeBand_j$ is whether teacher j is in a high school or middle school;

$Minority_j$ is the percentage of minority students in teacher j's school.

The results indicate that once we controlled the percentage of free and reduced lunch, the other teacher-level variables did not make significant contributions to explaining the variance in gain scores. Therefore, we excluded the GradeBand and percentage of minority students from our model. We fit the following model as our final model for results interpretation.

Level 1: $\text{Gain}_{ij} = \beta_{0j} + \beta_{1j} \cdot \left(Pretest_{ij} - \overline{Pretest}_j\right) + r_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01} \cdot \overline{Pretest}_j + \gamma_{02} \cdot FRL_j + u_{ij}$

$\beta_{1j} = \gamma_{10}$

The following table presents the result of this model for 2016-2017 and 2015-2016.

<table>
<tr><th colspan="6">Parameter Estimates for Two-level Hierarchical Linear Model (2016-2017)</th></tr>
<tr><td>Fixed Effect</td><td>Coefficient</td><td>Standard Error</td><td>T-ratio</td><td>Approx.d.f.</td><td>P-value</td></tr>
<tr><td>For β_{0j}</td><td colspan="5"></td></tr>
<tr><td>γ_{00}</td><td>1.98</td><td>0.23</td><td>8.76</td><td>55</td><td><0.001</td></tr>
<tr><td>$\overline{Pretest_j}, \gamma_{01}$</td><td>-0.27</td><td>0.25</td><td>-1.10</td><td>55</td><td>0.277</td></tr>
<tr><td>FRL_j, γ_{02}</td><td>-0.02</td><td>0.01</td><td>-3.36</td><td>55</td><td>0.001</td></tr>
<tr><td>For β_{1j}</td><td colspan="5"></td></tr>
<tr><td>$(Pretest_{ij} - \overline{Pretest_j}), \gamma_{10}$</td><td>-0.50</td><td>0.03</td><td>-17.52</td><td>3468</td><td><0.001</td></tr>
<tr><th colspan="6">Parameter Estimates for Two-level Hierarchical Linear Model (2015-2016)</th></tr>
<tr><td>Fixed Effect</td><td>Coefficient</td><td>Standard Error</td><td>T-ratio</td><td>Approx.d.f.</td><td>P-value</td></tr>
<tr><td>For β_{0j}</td><td colspan="5"></td></tr>
<tr><td>γ_{00}</td><td>1.85</td><td>0.24</td><td>7.68</td><td>22</td><td><0.001</td></tr>
<tr><td>$\overline{Pretest_j}, \gamma_{01}$</td><td>-0.24</td><td>0.42</td><td>-0.57</td><td>22</td><td>0.576</td></tr>
<tr><td>FRL_j, γ_{02}</td><td>-0.01</td><td>0.01</td><td>-1.82</td><td>22</td><td>0.083</td></tr>
<tr><td>For β_{1j}</td><td colspan="5"></td></tr>
<tr><td>$(Pretest_{ij} - \overline{Pretest_j}), \gamma_{10}$</td><td>-0.38</td><td>0.04</td><td>-9.25</td><td>1774</td><td><0.001</td></tr>
</table>

From the table above, we can see that the percentage of free and reduced lunch has a statistically significant negative effect on students' learning gains. Though the coefficient here is only -0.02 (2016-2017), the standardized coefficient is around -0.2. (After adjusting for the measurement error in the gain score, this partial correlation got to -0.23. See Appendix for this correction procedure.) That is to say, classrooms from schools with higher percent of free and reduced lunch benefit less from implementing *Carbon TIME*. (2015-2016 data also shows a consistent pattern.)

- *Carbon TIME reduced the achievement gap between high-pretest and low-pretest students within classrooms.* The other important finding from the table above is that the coefficient for $(Pretest_{ij} - \overline{Pretest_j})$ is around -0.5 with p value smaller than 0.001 (for 2016-2017). $(Pretest_{ij} - \overline{Pretest_j})$ measures how far the student i deviates from the class-average learning gain. This indicates that within classes students with lower pretest proficiencies gained significantly higher learning growth. The standardized coefficient is around -0.28. This partial correlation is around -0.23 after adjusting for measurement error in pretest and learning gains, which is still a quite strong correlation educationally.

Check students with negative learning gains:

Among the 3527 students in our analysis (2016-2017), there are 498 students whose learning gains are negative. This means that their learning proficiency level is lower in their Posttest compared to their Pretest. We took a closer look at these students' data to see what factors can lead to these students' poor performance in the Posttest. We found that there could be several other reasons other than students have not learned anything. One of these explanations is that students have not tried hard on the Posttest because the Full Posttest is low-stake for them. We presented some pieces of evidence for their lack of effort in Posttests as follows.

- Among all the 996 tests taken by these 498 students, there are 361 tests are incomplete. Only 80 out of these 361 are pretest while the other 281 tests are posttests. This is saying that students are more likely to leave the posttest incomplete.
- We calculated the missing item proportion for each of these 498 students. That is, for each student, we listed how many items they left in blank. A paired t-test shows that the item missing rate is statistically significantly higher in the posttest than in the pretest. On average, one student may leave 6% items in blank for pretest but may leave as many as 32.4% items in blank for posttest.

References:

Adams, R. J., Wilson, M., & Wang, W. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. Applied Psychological Measurement, 21(1), 1–23.

Frank, K. A. (1998). Chapter 5: Quantitative Methods for Studying Social Context in Multilevels and Through Interpersonal Relations. *Review of Research in Education*, *23*(1), 171–216. https://doi.org/10.3102/0091732X023001171.

Willett, J. B. (1988). Chapter 9: Questions and answers in the measurement of change. Review of research in education, 15(1), 345-422.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). ACER ConQuest Version 2.0. Generalized Item Response Modelling Software. Melbourne: Australian Council for Educational Research.

Appendix: Procedures to Adjust Measurement Errors

In our hierarchical linear models, we used the learning gains (the difference between pretest proficiency and posttest proficiency) as our outcome variable. We acknowledge that this could lead to biased estimates due to unreliability in the learning gains. Therefore, we applied the approach presented by Willet (1998) to adjust the coefficients of interest for measurement errors. The steps are summarized as follows. (We report these results based on 2016-17 data.)

- Step 1: calculate the reliability of the difference score/learning gain
  From the IRT analyses, we know that reliability for pretest and posttest are both 0.832. Applying Willet's formula (1998, p. 368), we calculated the reliability of the difference score, which turns out to be about 0.77.
- Step 2: get the standardized coefficients of interest (partial correlation of interest)

The two coefficients that are of most interest to us are those for $\left(Pretest_{ij} - \overline{Pretest_j}\right)$ and FRL_j. The two standardized coefficients are around -0.28 and -0.20. These are the partial correlations. For example, -0.2 is the partial correlation between the school percentage of free and reduced lunch and student learning gains controlling other variables in the model.

- Step 3: adjust the partial correlation for measurement errors
  Based on Willet's formula (1998, p. 374), we adjusted the partial correlations from Step 2 to get the adjusted partial correlations for measurement errors in gain scores and pretest scores. The adjusted partial correlations are both around -0.23.