Supporting Information

Holmes et al. 10.1073/pnas.1505329112

Quantitative Comparison Tools

The first type of comparison encountered in a typical introductory physics laboratory is to compare two independently measured values of the same physical parameter, a task that is known to be challenging for students (3, 5, 10). In many instructional laboratories, students do so by assessing whether the uncertainty ranges defined by the measurements overlap. Scientists, however, generally refer to a continuous scale associated with the measurements' probability distributions (22), such as the number of units of uncertainty by which two measurements differ (so-called $1-\sigma$, $2-\sigma$, or $3-\sigma$ differences in physics, for example). Following the Guide to Uncertainty in Measurement (23), this could be calculated as

$$t' = \frac{A-B}{\sqrt{\delta_A^2+\delta_B^2}}, \qquad \text{[S1]}$$

where A and B are two measured values and δ_A and δ_B are their uncertainties, respectively. As such, a large t' score means that the measurements differ by more than their combined uncertainties, and a small t' score means the measurements are similar within their combined uncertainties. We use the letter t for the index in reference to the structural similarity to the Student's t value, but we do not imply the index applies to the t distribution.

Interpreting the outcome of this comparison provides the necessary structure for deciding how to act on the comparison. For example, because overestimated uncertainties can lead to an artificially small t' score, a low t' score could mean that poor precision has hidden a small disagreement. As such, one could choose to improve the quality of the measurements. Under a model that predicts the two measurements should agree, a large t' score could mean that the model is limited or inappropriate. One could then choose to evaluate, adjust, or discard this model. One could also attempt to identify possible measurement errors that are causing a systematic effect. In all of these cases, the statistic compares the difference between measured quantities within units of variability. Rather than specifically comparing sample means according to the sample SDs, however, the t' score uses any measurement value with its uncertainty. As such, we do not try to compare the t' scores on the t distribution or make inferences about probabilities. Indeed, if the measurements were sample means from populations with the same variance, the t' score would be equivalent to Student's t for comparing independent samples (or, if homogeneity of variance is violated, the t' score would be equivalent to Welch's t).

The χ^2 equation for least-squares fitting lends itself to the same quantitative framework defined by the weighted or reduced χ^2 statistic

$$\chi_w^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - f(x_i)}{\delta y_i}\right)^2, \qquad \text{[S2]}$$

where x_i and y_i are the measured independent and dependent values, δy_i is the uncertainty associated with each y_i, N is the number of data points, and $f(x_i)$ are the model values associated with each x_i. This parameter evaluates the average difference between measured data and a model in units of uncertainty (squared). Values, therefore, are subject to the same interpretation and follow-up measurements as with the t' score (see Table S1).

Students were also taught a number of additional statistical analysis tools. The full set of tools taught to each condition are found in Table S2, which also specifies whether the tool informs a comparison or is primarily procedural.

Comparison Cycles Instruction Across the Year

Students in the experimental group were given explicit instructions to make comparisons between their measurements and/or models and iterate to improve their measurements. These behaviors were also graded and present in a grading rubric. This support was faded across the course. The explicit instructions in the text were the first to be removed, followed by assigned marks, and eventually instructor support was also removed. A map of this fading process across the year is included in Table S3.

Student Experiments Included in the Study

Week 2: Period of a Pendulum as a Function of Amplitude. In this experiment, students were asked to measure the period of a pendulum at two (experimental group, 10° and 20°) or three (control group, 5°, 10°, and 20°) angles of amplitude and compare their measurements. Students were not given a model for the process, but most of the students believed from previous experience (high school or college-level physics class) that the period was independent of angle according to the equation

$$T = 2\pi\sqrt{\frac{L}{g}}, \qquad \text{[S3]}$$

where L is the length of the pendulum, g is the acceleration due to gravity, and T is the period of the pendulum. The derivation of this equation, however, involves an approximation that

$$\sin\theta \approx \theta \qquad \text{[S4]}$$

for small angles, θ. High-precision measurements, therefore, expose this approximation and reveal the difference in the periods at different amplitudes from the second-order correction to this approximation.

Week 16: Resistor–Capacitor Circuit 2. In this experiment, students studied the voltage decay across a resistor in a parallel resistor–capacitor (RC) circuit. This was the second experiment with this equipment and circuit. They measured the time constant (τ) of the voltage decay across the resistor as a function of resistance of the resistor, which is given by the model

$$\tau = RC. \qquad \text{[S5]}$$

In addition to verifying that the relationship between τ and R was in fact linear with an intercept through the origin, students could compare the capacitance of the capacitor with the value of the slope from a graph of τ versus R. Resistance from other parts of the circuit were negligible in this experiment.

Week 17: Inductor–Resistor Circuit. Using a similar measurement procedure to the week 16 experiment, students studied the time constant of the voltage decay (τ) across a resistor in a series inductor–resistor (LR) circuit, which is given by the model

$$\tau = \frac{L}{R}. \quad [S6]$$

For this model, the time constant as a function of resistance, plotted as $\frac{1}{\tau}$ versus resistance, would give a straight line with an intercept through the origin. Resistance in the additional components in the circuit, however, is nonnegligible here, resulting in a nonzero intercept in the plot. Students could choose whether to perform a one-parameter ($y = mx$) or two-parameter ($y = mx + b$) linear fit to their data, which would cause them to confront the issue of the intercept. Students did not know the inductance of the inductor and so could not make a comparison with the value from the fit. Students could check their circuit for a finite (noninfinite) time constant with the resistor set to zero resistance.

Sophomore Laboratory: *LRC* Circuit. In the *LRC* circuit experiment, an inductor (L), resistor (R), and capacitor (C) are connected in series, and the equation governing the voltage decay across the resistor is

$$\frac{V_R}{V_0} = \frac{1}{\sqrt{\left(1 + \left(\left(\omega^2 + \omega_0^2\right)/(\gamma\omega)\right)^2\right)}}, \quad [S7]$$

where V_R is the voltage across the resistor, V_0 is the amplitude of the input AC voltage source, ω is the angular frequency of the voltage source, ω_0 is the resonant frequency, and γ is the bandwidth. Students fit their data of $\frac{V_R}{V_0}$ as a function of frequency, ω, to determine the parameters ω_0 and γ. Additional resistance in the circuit beyond the resistance in the resistor, however, means that the ratio of V_R to V_0 will never be exactly 1, and so it is necessary to add a third scaling factor, A, to the model, such that

$$\frac{V_R}{V_o} = \frac{A}{\sqrt{\left(1 + \left(\left(\omega^2 + \omega_0^2\right)/(\gamma\omega)\right)^2\right)}}. \quad [S8]$$

Students also measured the parameters ω_0 and γ through another experiment and could calculate their values (using measurements of the components R, L, and C) through the definition of these parameters. As such, students had multiple comparisons to make to inform the quality of the fit beyond the analysis of the fit itself.

Interrater Reliability

For all of the data presented, one rater coded all items and another rater coded ~10% of the items. The primary coder was never blind to condition because of the nature of the student products. In the control group, students printed their analysis work from spreadsheets and pasted them into their laboratory notes, whereas the experimental group submitted their spreadsheets electronically. The second rater, however, was given copies that made the rater blind to condition.

Interrater-reliability analysis using Cohen's κ statistic was performed to evaluate consistency between raters. Values greater than 0.6 were considered substantial agreement and so do not suggest a need for blind coding. For the quality of reflective comments, the interrater reliability for the raters was found to be $\kappa = 0.657, P < 0.001$. For identifying whether students proposed or proposed and carried out changes to their methods and measurements, the interrater reliability for the raters was found to be $\kappa = 0.714, P < 0.001$. For identifying whether students identified and/or physically interpreted the disagreements with models, the interrater reliability for the raters was found to be $\kappa = 0.881, P < 0.001$.

Participants

Included in the study were two cohorts (groups) of students enrolled in the same introductory undergraduate physics course at a research-intensive university in Canada. The control group consisted of students enrolled in 2012/2013, whereas the experimental group consisted of students enrolled in 2013/2014. The course, both years, was spread across two semesters of eight or nine 3-h laboratory weekly laboratory sessions. Each laboratory session included no more than 48 students and was facilitated by two graduate student teaching assistants and the course instructor. The number of students included in the analysis is found in Table S4. The variability in the number of students each week is attributable to students not attending all laboratories. In the control group, 109 students conducted all three first-year laboratories, and only 31 students conducted all three first-year laboratories and the sophomore laboratory. In the experimental group, 108 students conducted all three first-year laboratories and only 36 students conducted all three first-year laboratories and the sophomore laboratory. Because the effects of the laboratory occurred throughout more than just the four laboratories evaluated, we include any students who participated each particular week.

On entering the course, the two groups had statistically equivalent pretest scores on the Force Concept Inventory (FCI) (24): control, $M = 77\%, SE = 2\%$; experiment, $M = 76\%, SE = 2\%, t(266) = 0.20, P = 0.839$. By the end of the first term, the groups had statistically equivalent scores on the Mechanics Baseline Test (MBT) (25): control, $M = 72\%, SE = 2\%$; experiment, $M = 68\%, SE = 2\%, t(288) = 1.21, P = 0.227$. By the end of the second term, the groups also had statistically equivalent scores on the Brief Electricity and Magnetism Survey (BEMA) (26): control, $M = 70\%, SE = 2\%$; experiment, $M = 64\%, SE = 2\%, t(177) = 1.96, P = 0.052$. These assessments have been used to evaluate the introductory physics students in the department for over 20 y, and, in the last decade, students' incoming scores have been consistent within a 2% SD.

The critical thinking behaviors assessed in this study relate primarily to evaluating data and physical measurement systems. The questions on the FCI, MBT, and BEMA evaluate students' ability to apply specific physics concepts in idealized situations. There is very little overlap between the knowledge and reasoning required to answer those questions, and the real-world, data-driven critical thinking about data and measurement systems learned in the laboratory course. We also would expect that the lecture and other components of the courses would dominant over a possible effect related to the laboratory. Therefore, it is not surprising that the scores are not correlated.

Students in the course both years were almost all intending to major in a science, technology, engineering, or math field, although they do not declare their majors until their second year. The breakdown of students' intended majors in the experimental group by the end of the course are in Table S5. Unfortunately, these data were unavailable for the control group. We do have data that show that ~15% of students in the control group and 20% of the students in the experimental group chose physics as a major by their second year of study.

Evaluation of the Sophomore Students. We will further evaluate the students who continued into the sophomore laboratory course to explore whether the results seen in the sophomore laboratory are attributable to transfer or selection effects. First, we will do a two-by-two comparison on the end-of-first-year MBT and BEMA scores (Table S6), comparing between students who did and did not take the sophomore laboratory course and between the experiment and control groups in the first-year course.

Overall, the students who went on to take the sophomore physics laboratory course outperformed the students who did not take the sophomore laboratory, as measured on both the MBT

and the BEMA (note that, of the students in the control group, there was no difference between students who did and did not take the sophomore laboratory course on the BEMA). This tells us that the students in the sophomore physics laboratories generally had a stronger conceptual physics background than the students who did not continue in an upper-year physics laboratory course. This is consistent with the expected selection bias of students who choose to pursue more physics courses. Of the students who took the sophomore physics laboratory, however, there is a nonsignificant difference between the experimental and control groups on both the MBT and BEMA. This is consistent with the overall lack of differences on these measures between the full experiment and control conditions in the first-year laboratory course discussed in the previous section.

Next, we compare these two subgroups on their evaluation, iteration, and reflection behaviors throughout the first-year laboratories. The trends in the Fig. S1 *A–C* showing only the sophomore students are very similar to those for the whole course (Figs. 1–3). This suggests that the students who continued into the sophomore course were not exceptional in their behaviors in first-year. This further suggests that the effect seen in the sophomore laboratory experiment are not attributable to selection effects. It remains that the upwards shift in the control group's reflective comments and evaluation of the model are attributable to something inherent in the sophomore laboratory course. Most likely these shifts can be attributed to the prompt in the instructions to explain why there may be extra parameters in the model. This instruction would explain a shift in the model evaluation and reflective comments but not in iteration, as seen in the data.

Reflection Analysis

To analyze students' reflection in the laboratory, we evaluated students' reflective comments associated with their statistical data analysis and conclusions. The reflective comments were coded using a set of four classes based on Bloom's Taxonomy classes (13). Fig. S2 *A* and *B* provide samples of this coding applied to student work. The four comments levels were:

i) Application: a written reflection statement that offers the outcome of the procedural application of data analysis tools (e.g., The χ^2 value is 2.1.) These comments were distinct from procedural statements (e.g., then we calculated the χ^2 value).

ii) Analysis: a written reflection statement that analyzes or interprets their data analysis or results (e.g., our χ^2 value is 0.84, which is close to one, indicating that our model fits the data well).

iii) Synthesis: a written reflection statement that synthesizes multiple ideas, tool analyses, or reflections to propose a new idea. This could include suggesting ways to improve measurements (e.g., we will take more data in this range, because the data are sparse) or models (e.g., our data has an intercept so the model should have an intercept), as well as making comparisons (e.g., the χ^2 value for the $y=mx$ fit was 43.8 but for the $y=mx+b$ fit χ^2 was 4.17, which is much smaller).

iv) Evaluation: a written reflection statement that evaluates, criticizes, or judges the previous ideas presented. Evaluation can look similar to analysis, but the distinction is that evaluation must follow a synthesis comment. For example, after a synthesis that compared two different models and demonstrated that adding an intercept lowered the χ^2 value, an evaluation could follow as, "…the intercept was necessary due, most likely, to the inherent resistance within the circuit (such as in the wires)."

Fig. S2 *A* and *B* demonstrate how the coding scheme is applied to three excerpts from students' books in the *LR* experiment (week 17). Each of the levels build on each other, so a student making a level 4 evaluation statement would also have made lower level statements, although level 1 comments (application) need not be present. Although it is important that students reflect on various parts of the data analysis, only the maximum reflection level a student reached was coded. It should be noted that the comments were not evaluated on correctness.

Analysis

For the first-year experiments, generalized linear mixed-effects models were performed using R (27) and the linear mixed-effects models using Eigen and S4 package (28) to analyze all three outcome measures (proposing and/or carrying out measurement changes, identifying and/or interpreting disagreements with models, and levels of reflection/comments). For measurement changes and evaluating models, logistic regression analysis was performed because of the dichotomous nature of the outcome variables. For the reflection data, Poisson regression was used to account for the bounded nature of the outcome variables. All three analyses used condition, laboratory week, and the interaction between condition and laboratory week as fixed effects and student identifier code (student ID) as a random effects intercept. Type 3 analysis of variance (ANOVA) was performed on the logistic regression models using the R Companion to Applied Regression package (29) to assess the overall impact of the variables. Sophomore laboratory data were analyzed using χ^2 tests for independence of proportions.

Proposing and/or Carrying Out Measurement Changes. A logistic regression was carried out to compare the proportion of students in each group and across each experiment proposing and/or carrying out changes to their measurements (Table S7). Note, for this analysis, proposing versus proposing and carrying out changes were collapsed to a single dichotomous variable of proposing or carrying out changes. The logistic regression model was statistically significant, $\chi^2(5)=470.55, P<0.001$. A type 3 ANOVA of the logistic regression model demonstrated that condition and the interaction between condition and laboratory week were highly significant in the model, but laboratory week alone was not significant.

With significant effects for the interaction, we can compare the groups each week to explore where the significant differences exist. To do this, we use a χ^2 test of proportions comparing groups on the distribution of the number of students who did not propose or change their measurements, who proposed changes to their measurements, and who proposed and made changes to their measurements (returning to the three-level, rather than dichotomous, variable). Taking into account the multiple comparisons across weeks, we use a Bonferroni correct to set the α level at 0.01. This gave statistically significant differences between groups on all four experiments: week 2, $\chi^2(2)=270.38$, $P<0.001$; week 16, $\chi^2(2)=107.51, P<0.001$; week 17, $\chi^2(2)=128.39, P<0.001$; sophomore laboratory, $\chi^2(2)=17.58, P<0.001$. This demonstrates that the experimental group outperformed the control group on this measure on all experiments.

Evaluating Models. A logistic regression was carried out to compare the proportion of students in each group and across each experiment identifying the disagreement with the model and/or physically interpreting the issue (Table S8). Note, for this analysis, identifying versus physically interpreting the disagreement with the model were collapsed to a single dichotomous variable. The logistic regression model was statistically significant, $\chi^2(3)=171.96, P<0.001$. A type 3 ANOVA of the logistic regression model demonstrated that condition and the interaction between condition and laboratory week were highly significant in the model, but laboratory week alone was not significant.

With significant effects for the interaction, we can compare the groups each week to explore where the significant differences

exist. To do this, we use a χ^2 test of proportions comparing groups on the distribution of the number of students who did not identify the disagreement with a model, who did identify the disagreement, and who identified and interpreted the disagreement. Taking into account the multiple comparisons across weeks, we use a Bonferroni correct to set the α level at 0.02. This gave significant differences between groups on all three experiments: week 2, $\chi^2(2) = 8.60, P = 0.014$; week 17, $\chi^2(2) = 99.04, P < 0.001$; sophomore laboratory, $\chi^2(2) = 10.32, P = 0.006$.

Reflection Behaviors. A Poisson regression was carried out to analyze the quality of the reflective comments in each group across each experiment (Table S9). The regression model was statistically significant, $\chi^2(5) = 109.03, P < 0.001$. A type 3 ANOVA of the logistic regression model demonstrated that condition and the interaction between condition and laboratory week were highly significant in the model, but laboratory week alone was not significant.

With a significant interaction, we can compare the groups each week to explore where the significant differences exist. To do this, we use a χ^2 test of proportions comparing the distribution of the numbers of students in each group who reached each maximum comment level. Taking into account the multiple comparisons across weeks, we use a Bonferroni correct to set the α level at 0.01. This gave significant differences between groups on all three first-year experiments, but nonsignificant differences on the sophomore-laboratory: week 2, $\chi^2(3) = 25.44, P < 0.001$; week 16, $\chi^2(3) = 51.86, P < 0.0001$; week 17, $\chi^2(3) = 155.83, P < 0.0001$; sophomore laboratory, $\chi^2(3) = 7.58, P = 0.056$.

Time on Task in the *LR* Experiment

One confounding issue to the week 17 *LR* circuit experiment was that students in the control group worked through a computer-based inquiry activity at the beginning of the experiment session. The activity taught students how to calculate the uncertainty in the slope of a best-fitting line, which they also used to reanalyze the previous week's data. As such, the control group spent approximately 2 h on the *LR* circuit laboratory, whereas the experimental group spent 3 h. Not having enough time to reflect on data and act on that reflection may explain the different outcomes observed in the main text. As a precautionary measure, we observed students in the experimental group 2 h into the laboratory session to evaluate what analysis they had performed by that time. The observer recorded whether the group had by that time produced a one-parameter mx fit or a two-parameter $mx + b$ fit.

The results, shown in Fig. S3, demonstrate that if the students in the experimental group had been given the same amount of time on task as students in the control group, more of them still would have made the modification to the model and included an intercept in their fit. Given additional time, however, even more students were able to think critically about the task and make better sense of their data. From this result, we conclude that the effects seen in this experiment are still primarily attributable to students' overall improved behaviors. Indeed, the effect is much larger because of the additional time, which is an important feature of the intervention itself. It takes time for students to engage deeply in a task, think critically, and solve any problems that arise (30). Comparing between students in the experimental group at the 2-h mark and the final 3-h mark demonstrates the striking effect that an extra hour can make to students' productivity.

The number of single-parameter mx fits decreased slightly from the 2-h observations and the final submitted materials for the experimental group. This could have occurred if students recognized that the mx fit was not helpful in understanding their data, because of the additional intercept required. This is interesting to note in light of the limitations of the analysis methods used in this study. Analyzing laboratory books can only keep track of recorded activity and many behaviors may have occurred without record. The result that some students created additional fits and then did not submit them at the end of the laboratory period demonstrates that students in the experimental group still may have engaged in additional reflective and iterative behaviors beyond what was recorded. Differences between the control and experimental groups, then, are unlikely attributable to students in the experimental group simply recording more while engaging in the same behaviors as students in the control group.

The slope uncertainty activity provided to the students in the control group just before the *LR* circuit laboratory may, however, have narrowed the focus of students' analysis. That is, the activity first introduced students to the uncertainty in the slope of a one-parameter best fitting line (that is, with the intercept fixed at the origin). As such, it could be argued that these students were more likely to fix the intercept at the origin so that they could apply the learned formula. The activity, however, also included a follow-up task that introduced the uncertainty in the slope of a two-parameter best fitting line (intercept not fixed), and so students did have access to both options. Students also could have used their analysis to identify the issue even if they did not change their fit.

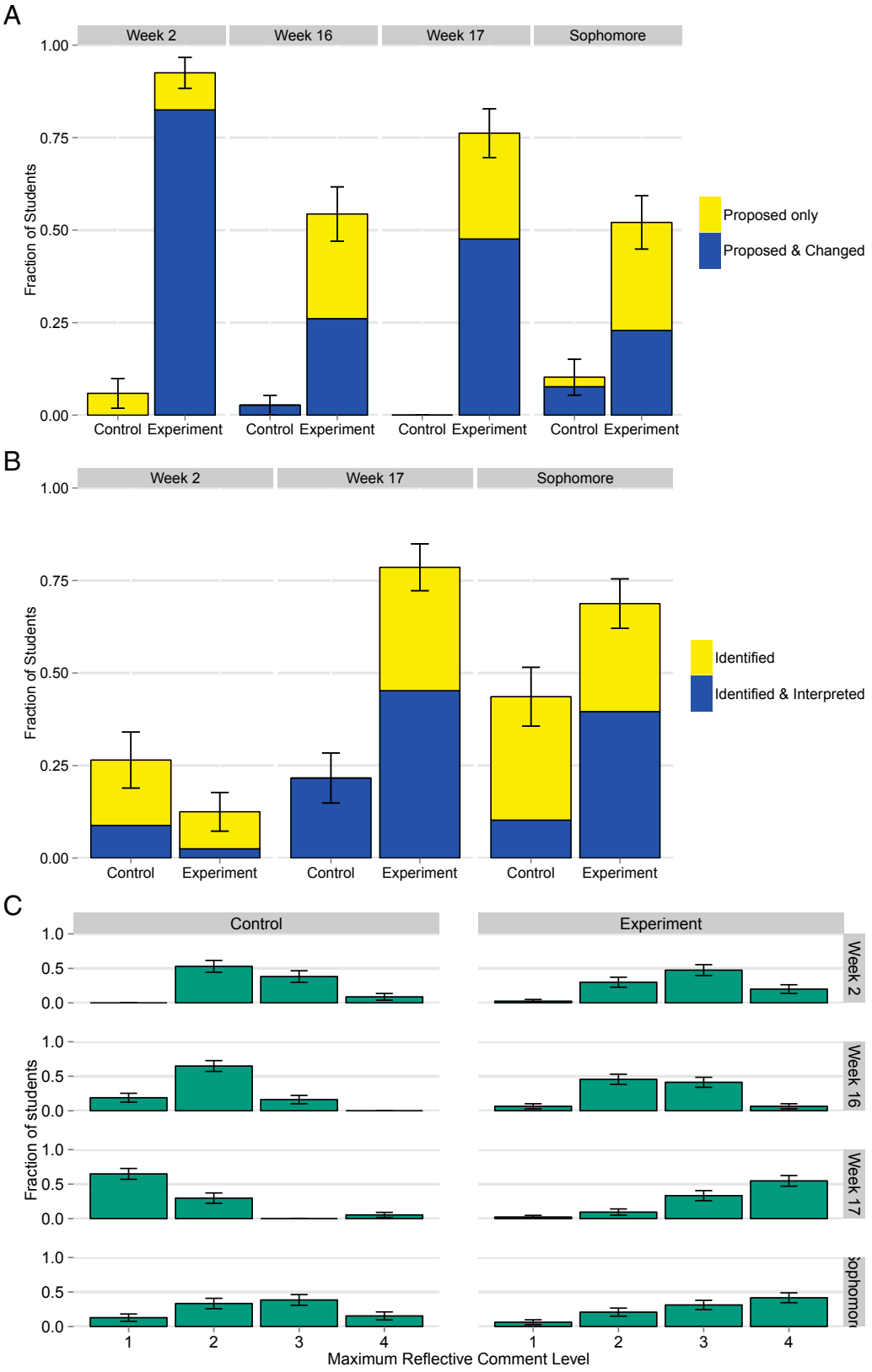

Fig. S1. Evaluating the sophomore laboratory students. *A–C* show the measures in the core of the analysis including only the students who moved into the sophomore-level physics course. The data are very similar to the class as a whole, demonstrating that the students in the sophomore laboratory are representative of the full first-year class on these measures. (*A*) Proposing and/or carrying out changes to their experimental methods. (*B*) Identifying and interpreting model disagreements. (*C*) Maximum reflection comment level reached.

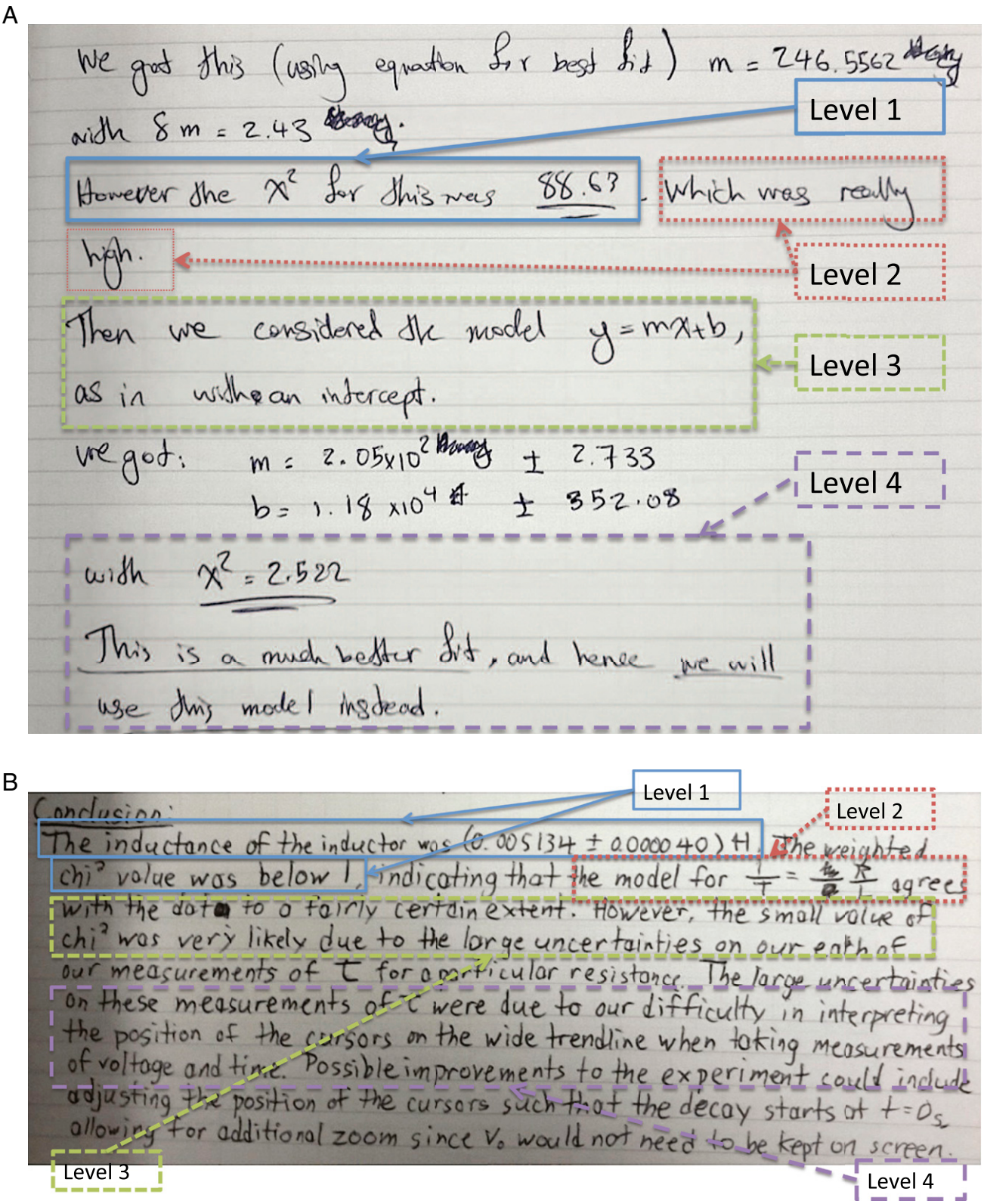

Fig. S2. Two students' reflections during an experiment provide examples of the reflection coding scheme. (*A*) The student makes a level 1 comment about applying χ^2 to the student's experiment and then shows that this value is high (level 2). A level 3 statement describes considering a different model, and then the student finally evaluates the new model by describing the much lower χ^2 value. (*B*) The students starts with a level 1 comment about χ^2 and the inductance and then analyzes the fit line compared with the model (level 2). The student then comments on χ^2 being small, attributing it to large uncertainties (level 3). The student justifies the uncertainty as attributable to limitations of the measurement equipment (level 4). Finally the student provides further suggestions for improvement (additional level 3).

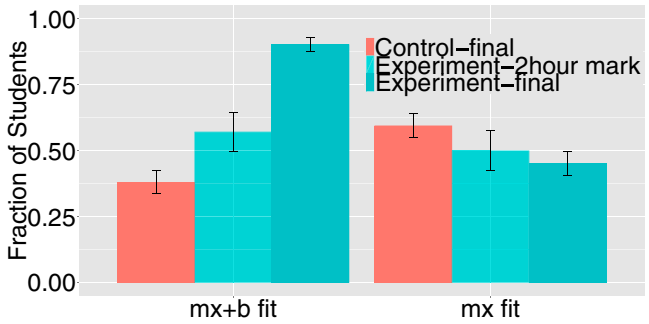

Fig. S3. The distribution of graphical analyses made by students by the end of the *LR* circuits laboratory in the control and experimental groups and within the first 2 h of the laboratory for the experimental group. Uncertainty bars represent 67% confidence intervals on the proportions. The bars are larger for the "Experiment-2hour mark," because only groups, rather than individuals, were assessed. Bars in each group may add to more than 1, because students may have analyzed either or both fits.

Table S1. Interpretations of and follow-up behaviors from comparisons

t' score	Interpretation of measurements	Follow-up investigation	χ^2
$0 < \lvert t' \rvert < 1$	Unlikely different, uncertainty may be overestimated	Improve measurements, reduce uncertainty	$0 < \chi^2 < 1$
$1 < \lvert t' \rvert < 3$	Unclear whether different	Improve measurements, reduce uncertainty	$1 < \chi^2 < 9$
$3 < \lvert t' \rvert$	Likely different	Improve measurements, correct systematic errors, evaluate model limitations or approximations	$9 < \chi^2$

t' score comparisons are between pairs of measurements and χ^2 comparisons are between datasets and models.

Table S2. Statistical tools taught to students in each condition

Comparison tools		Procedural tools
Control and experiment condition	Experiment condition only	Control and experiment condition
Overlapping uncertainty ranges	t' score	Histograms
Unweighted χ^2	Residual plots	Mean
Weighted χ^2		SD
		Standard uncertainty in the mean (SE)
		Semilog and log–log plots
		Weighted average
		Uncertainty in fit parameters of fit lines

The statistical tools taught to students in each condition are specified by whether they are procedural or inform the comparison cycles.

Table S3. Support given to experimental condition to make and act on comparisons

	Week																		
Support	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Compare: instructions		X	X	X	X	X					X	X							
Compare: marking		X	X	X	X	X	X			X	X	X		X		X			
Iterate: instructions		X			X	X													
Iterate: marking		X			X	X	X												

The experimental group received explicit support to make and act on comparisons. The support came in the form of explicit instructions and/or reference in the marking scheme and was faded over time. In the table, an X indicates that the behavior (comparing or iterating) was supported that week.

Table S4. Sample sizes on each measure in the study between groups and experiments

Group	Week 2	Week 16	Week 17	Sophomore laboratory
Control	146	132	131	39
Experiment	159	138	133	48

Table S5. Students in the experimental group who have declared a variety of STEM majors

Intended Major	Experimental group, %
Physics or astronomy	14
Life sciences	13
Engineering physics	7
Non-STEM	2
Computer science	1
Chemistry	1
Other STEM or undecided	62

STEM, science, technology, engineering, and mathematics.

Table S6. Evaluating students who went into the sophomore physics laboratory

	Sophomore Laboratory		Comparisons	
Group	Took laboratory, %	Did not take laboratory, %	Took laboratory vs. did not take laboratory	Experimental vs. control group
MBT				
Control Group	77 (12)	70 (16)	$t(76.6) = 2.46$; $P = 0.016$*	
Experimental Group	75 (17)	66 (16)	$t(80.6) = 2.81$; $P = 0.006$**	
Took laboratory				$t(71.2) = 0.59$; $P = 0.556$
BEMA				
Control Group	74 (9)	65 (20)	$t(34.8) = 1.85$; $P = 0.073$	
Experimental Group	68 (16)	61 (16)	$t(70.8) = 2.06$; $P = 0.04$*	
Took laboratory				$t(44.3) = 1.71$; $P = 0.094$

Students from the first year course (both from the control and experimental conditions) who did and did not take the sophomore laboratory are compared on MBT and BEMA diagnostics. Numbers are mean percentage on the test with SD in parentheses. *$P < 0.05$; **$P < 0.01$.

Table S7. Analysis of students' iteration behaviors

Model coefficients and variables	Estimate	*SE*	Wald *z*	*df*	χ^2	*P*
Model coefficients						
Condition = Experiment	7.97	0.94	8.49			<0.0001***
Week = Week 16	−0.82	0.86	−0.96			0.336
Week = Week 17	−0.41	0.75	−0.55			0.582
[Condition = experiment] × [week = week 16]	−2.64	1.03	−2.56			0.010**
[Condition = experiment] × [week = week 17]	−2.54	0.93	−2.72			0.007**
Model variables						
Condition				1	83.02	<0.001***
Week				2	28.99	<0.001***
Condition × week				2	9.28	0.01*

Analysis used logistic regression to compare the control and experimental groups across four experiments, three in the first-year course and one in the sophomore course. *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

Table S8. Analysis of how students evaluated models

Model coefficients and variables	Estimate	*SE*	Wald *z*	*df*	χ^2	*P*
Model coefficients						
Condition = experiment	−0.83	0.33	−2.55			0.011*
Week = week 17	−0.27	0.30	−0.88			0.379
[Condition = experiment] × [week = week 17]	3.60	0.60	5.97			<0.001***
Model variables						
Condition				1	6.49	0.011*
Week				1	0.77	0.379
Condition × week				1	35.62	<0.001***

Analysis used logistic regression to compare the control and experimental groups across three experiments, two in the first-year course and one in the sophomore course. *$P < 0.05$; ***$P < 0.001$.

Table S9. Analysis of students' reflection behaviors

Model coefficients and variables	Estimate	*SE*	Wald *z*	*df*	χ^2	*P*
Model coefficients						
Condition = experiment	0.13	0.07	1.89			0.059
Week = week 16	−0.29	0.08	−3.48			<0.001***
Week = week 17	−0.40	0.09	−4.59			<0.001***
[Condition = experiment] × [week = week 16]	0.17	0.11	1.52			0.130
[Condition = experiment] × [week = week 17]	0.58	0.11	5.29			<0.001***
Model variables						
Condition				1	3.57	0.059
Week				2	24.48	<0.001***
Condition × week				2	28.55	<0.001***

Analysis used regression to compare the control and experimental groups across four experiments, three in the first-year course and one in the sophomore course. ***$P < 0.001$.