Laboratory Courses with Guided-Inquiry Modules Improve Scientific Reasoning and Experimental Design Skills for the Least-Prepared Undergraduate Students

Lawrence S. Blumer[†‡]* and Christopher W. Beck[†§]

[‡]Department of Biology, Morehouse College, Atlanta, GA 30314; [§]Department of Biology, Emory University, Atlanta, GA 30322

ABSTRACT

Past studies on the differential effects of active learning based on students' prior preparation and knowledge have been mixed. The purpose of the present study was to ask whether students with different levels of prior preparation responded differently to laboratory courses in which a guided-inquiry module was implemented. In the first study, we assessed student scientific reasoning skills, and in the second we assessed student experimental design skills. In each course in which the studies were conducted, student gains were analyzed by pretest quartiles, a measure of their prior preparation. Overall, student scientific reasoning skills and experimental design skills did not improve pretest to posttest. However, when divided into quartiles based on pretest score within each course, students in the lowest quartile experienced significant gains in both studies. Despite the significant gains observed among students in the lowest quartile, significant posttest differences between lowest and highest quartiles were observed in both scientific reasoning skills and experimental design skills. Nonetheless, these findings suggest that courses with guided-inquiry laboratory activities can foster the development of basic scientific reasoning and experimental design skills for students who are least prepared across a range of course levels and institution types.

INTRODUCTION

Efforts to transform undergraduate education, including laboratory pedagogy, have been pursued for more than two decades (National Research Council [NRC], 2003; American Association for the Advancement of Science, 2011; President's Council of Advisors on Science and Technology [PCAST], 2012). Studies on the efficacy of inquiry-based laboratory learning almost always show some positive outcomes (reviewed in Beck *et al.*, 2014). However, the effects of inquiry-based learning in lecture and laboratory courses are not necessarily uniform for all students. The learning gains that can be achieved as a result of inquiry-based teaching may vary due to student gender, race, ethnicity, age, and first-generation at college status. For example, Preszler (2009) found that replacing traditional instruction with peer-led workshops in a biology course benefited all students, but that the benefits were more pronounced for female students and underrepresented minority (URM) students. These benefits included higher scores on examination questions, higher grades compared with preintervention semesters, and increased student retention. Similarly, Eddy and Hogan (2014) found significantly increased effectiveness of an active-learning intervention (use of guided-reading questions, online preclass homework, and in-class problem-solving activities) with the greatest benefits (examination points earned and course pass rates) experienced by URMs and first-generation at college students. They suggest that these benefits resulted from students spending more time before and after class in course material review and students feeling they were part of a learning community.

In addition, Haak *et al.* (2006) found that increasing course structure with required inquiry-based activities (frequent problem solving and data analysis) improved the course grades of all students and had a disproportionate benefit for educationally or economically disadvantaged students in an introductory biology course. Similar findings have been reported in other science, technology, engineering, and mathematics (STEM) disciplines. For instance, in a physics course, active-learning interventions that increased interactions between students improved the performance of all students on a physics concept inventory (Force Concept Inventory) and decreased the gender gap between male and female learning outcomes (Lorenzo *et al.*, 2006). Thus, female students experienced a disproportionate benefit from the active-learning interventions.

In addition to demographic factors, prior skills and knowledge may matter. Differences between students in past experience and preparation predispose them to differences in learning outcomes, even in the absence of any intervention (Theobald and Freeman, 2014). For example, prior grade point average (GPA) in biological sciences courses predicted final grades in an anatomy and physiology course taught in a traditional, non-flipped format (DeRuisseau, 2016). In another study, predicted grades based on Scholastic Aptitude Test (SAT) and prior college GPA had the strongest effect on the number of questions answered correctly in an analysis of two forms of active-learning interventions (Eddy *et al.*, 2013). Similarly, in a study on the efficacy of small-group engagement activities in undergraduate biology courses, prior GPA predicted student improvement, independent of intervention (Marbach-Ad *et al.*, 2016). Likewise, Belzer *et al.* (2003) found that student performance on a pretest high school–level biology content survey was positively correlated with the final grade students earned in an introductory undergraduate general zoology course. However, the pretest score variation within each grade category was very large (Belzer *et al.*, 2003). In another study, prior experience in a prerequisite biology course improved performance on examination questions in a subsequent course, but only for concepts that were deemed very familiar, given the concept had been thoroughly addressed in the prerequisite course (Shaffer *et al.*, 2016). In addition to the independent effects of prior experience on student learning gains, the efficacy of inquiry-based interventions may be influenced by the previous academic experiences of students. For instance, the number of previous laboratory courses had a negative effect on gains in a pretest/posttest assessment of scientific reasoning skills for students in guided-inquiry laboratory courses in which students were presented with a research question and developed an experimental design, with guidance from their instructor, to address that question (Beck and Blumer, 2012).

Because demographic factors and prior experiences can affect learning gains, researchers have traditionally tried to control for differences among students in an effort to look for the main effects of an intervention. A common experimental approach is to compare an intervention class with a closely matched nonintervention class. For example, Brownell *et al.* (2012) sought to match an intervention and nonintervention group for gender distribution, class year, major, previous laboratory research experience, and GPA. Propensity score matching is one statistical approach for creating these closely matched groups (e.g., Rodenbusch *et al.*, 2016; Crimmins and Midkiff, 2017). Another approach is using analysis of covariance (ANCOVA) to control for differences between students' past experience and preparation so that the main effects of an intervention may be observed (Theobald and Freeman, 2014; Gross *et al.*, 2015). When using a pretest/posttest approach, these methods can determine whether an intervention is responsible for the observed gains in student performance while controlling for pretest differences. *However, if outcomes of an intervention depend on the starting point of a student, we might be missing the differential effects of the intervention by controlling for differences among students.* Documenting these differences in the effects of active-learning pedagogies for different student populations is exactly the kind of "second-generation research" that Freeman *et al.* (2014) suggested as the next steps for advancing studies on active-learning pedagogies.

Conducting separate outcome analyses on each (or a specific) quartile of students based on pretest performance or another metric of prior preparation, such as GPA or SAT scores, is a common approach in K–12 pedagogy evaluation research (Marsden and Torgerson, 2012). However, this approach is less common for undergraduate-level discipline-based education research. Past studies have shown mixed results with respect to which students benefit the most from pedagogical interventions. For example, Marbach-Ad *et al.* (2016) found that high-GPA students experienced significant improvement as a result of small-group engagement activities compared with traditional instruction, but low-GPA students did not show significant learning gains resulting from the same active-learning intervention. Similarly, Shapiro *et al.* (2017) categorized students on the basis of self-reported prior knowledge of physics as either high or low prior knowledge. Compared with a simple control, students with low prior knowledge performed worse on conceptual examination questions when they experienced either factual knowledge clicker questions or an enhanced control in which some content was flagged as important. However, these effects were not observed in high prior knowledge students. In contrast, in this same study, the use of conceptual clicker questions resulted in similar performance on both factual and conceptual questions by all students (Shapiro *et al.*, 2017). Consequently, this intervention had a disproportionate benefit for low prior knowledge students. Similarly, Gross *et al.* (2015), in a study evaluating how flipped classrooms improved student outcomes, found the greatest gains in attempting and accurately answering homework questions among students in the lowest two precourse GPA quartiles. Jensen and Lawson (2011) used a scientific reasoning pretest to categorize students into three groups to assess the effects of inquiry instruction (vs. traditional didactic instruction) and the composition of collaborative student groups. Gains from pretest to posttest for high and medium reasoners were not influenced by either the instructional method or group composition. In contrast, low reasoners in inquiry instruction showed greater gains in homogeneous groups, while in didactic instruction they showed greater gains in heterogeneous groups. In addition, Beck and Blumer (2012) found all students gained in confidence and scientific reasoning skills resulting from a guided-inquiry laboratory activity, but the gains were greatest for those students whose pretest scores were in the lowest quartile.

The results of past studies on the differential effects of active learning based on students' prior preparation and knowledge have been mixed. Therefore, the purpose of the present study was to ask whether students with different levels of prior

preparation responded differently to laboratory courses in which a guided-inquiry module using the bean beetle (*Callosobruchus maculatus*) model system was implemented. We used pretest assessment scores as a proxy for student preparation. This study was conducted in diverse biology courses from a broad range of institutions. We predicted that students with low pretest scores would benefit the most from guided inquiry for a couple of reasons. First, in a previous smaller-scale study, the least-prepared students benefited the most from guided inquiry (Beck and Blumer, 2012). Second, least-prepared students would benefit more from the scaffolding provided by guided-inquiry activities, which would lead to greater learning gains (D'Costa and Schlueter, 2013).

METHODS

We conducted a 4-year study (2009–2012) on whether laboratory courses incorporating a guided-inquiry module impacted the development of student scientific reasoning and experimental design skills. In the first 2 years of the study, we assessed student scientific reasoning skills (study 1), and in the second 2 years, we assessed student experimental design skills (study 2). In both studies, faculty teams were trained by the authors in guided-inquiry laboratory pedagogy at 2½-day (~20 hours) faculty professional development workshops. These workshops began with discussions on the range of laboratory teaching methods and the evidence for improved student outcomes with guided inquiry. This was followed by a hands-on laboratory activity in which the faculty participants were the students and the presenters instructed the group using guided-inquiry methods. In this approach, the faculty participants (the students) were presented with a research question, and they developed an experimental design to address that question, with guidance from the authors (the instructors), using the Socratic method. The presenters then deconstructed the steps that they used in conducting guided inquiry and addressed questions about the functions of each step. Based on preworkshop/postworkshop surveys, faculty participants reported both increased confidence in applying guided-inquiry methods and significant changes in their instructional practices toward more guided-inquiry teaching (unpublished data). After the workshops, the teams developed and implemented their own new guided-inquiry laboratory modules using the bean beetle model organism in undergraduate courses at their institutions. These protocols were reviewed by the authors and posted on the bean beetle website (www.beanbeetles.org). In addition, some of these protocols have appeared in peer-reviewed journals (study 1: Butcher and Chirhart, 2013; D'Costa and Schlueter, 2013; Schlueter and D'Costa, 2013; Pearce *et al.*, 2013; study 2: Fermin *et al.*, 2014; Smith and Hicks, 2014). The courses in which these bean beetle guided-inquiry protocols were included were the same courses in which we studied student learning gains in scientific reasoning skills (study 1) and experimental design skills (study 2).

The institutions, courses, and students in the two studies were not the same. In the courses that were part of study 1, the duration of laboratory modules varied between courses. Most often, modules lasted 2–3 weeks (46% of modules). Faculty self-reported on a range of instructional practices, such as science process skills that included scientific reasoning and experimental design, using a four-point Likert scale with values of

2 = seldom and 3 = often (Beck and Blumer, 2016). The median emphasis on science process skills was 2.9 (unpublished data). In the courses that were part of study 2, most modules were of 2- to 3-week duration (40% of modules), and median emphasis on science process skills was 3.3 (unpublished data). The common student assessments used in this study were administered by faculty teams in the context of their own courses. This study was approved by the institutional review boards at Emory University (IRB#00010542), Morehouse College (IRB#025), and participating institutions, when required.

Study 1. Impact of Courses with Guided-Inquiry Laboratory Modules on Scientific Reasoning Skills

Student scientific reasoning skills were assessed at the beginning (pretest) and end (posttest) of the semester. The assessment consisted of four problems, each with two questions, obtained from a test of scientific reasoning developed by Lawson (1978; version 2000). Only a subset of the problems in the instrument were used, as we wished to focus on scientific reasoning related to experimental design (see the Supplemental Material). Course instructors implemented the assessment in class using a Scantron sheet to record answers. Students were instructed that the assessment was part of a research study and no credit or extra credit would be given. Pairs of questions for each problem were contingent on each other, so they were scored as a pair in an all-or-none manner (i.e., either 2 points or 0 points for each pair of questions). We scored the questions in this manner to assess reasoning skill more accurately, as this approach decreases the likelihood that students will answer correctly based on random responses.

A total of 570 students from 51 courses at 12 different institutions completed both the pretest and the posttest. In some courses, only a small number of students ($N < 10$) completed the assessment. We removed these courses from our data set to minimize potential bias associated with low response numbers and to eliminate inadequate sample sizes when students were distributed into quartiles for data analysis. After courses with small numbers of responses were removed, we had data on 479 students from 33 courses at 11 different institutions. Students in this study represented the full range of course levels and institution types, both genders, and individuals self-identifying as URMs and non-URMs (Table 1). Because students of different genders and URM status were not evenly distributed across courses, we did not include student demographics in our analyses. Doing so would confound demographic effects with course effects.

To examine overall changes in scientific reasoning skills, we compared pretest and posttest scores for all students using a paired *t* test. The results were the same when we used a nonparametric Wilcoxon signed-rank test. To determine whether changes in scientific reasoning skills differed depending on course level, we used a linear mixed-effects model with absolute gain (posttest score – pretest score) as the dependent variable, course level as a fixed factor, pretest score as a covariate, and course as a random effect, as students within a course are not independent of one another (Eddy *et al.*, 2014; Theobald and Freeman, 2014; Theobald, 2018). To test for homogeneity of slopes (i.e., that the slope of the relationship between absolute gain and pretest score was the same for each course level), we included the interaction between course level and pretest

TABLE 1. Distribution of students and institutions completing the scientific reasoning assessment

	Liberal arts colleges	Research university	Regional comprehensive university	Community college
Number of institution types	8	1	1	1
	Minority serving	**Majority serving**		
Number of institution types	2	9		
	Non–biology majors	**Introductory biology**	**Upper-level biology**	
Number of each course type	8	22	3	
Students in each course type	103	314	62	
	Males	**Females**	**Not reported**	
Number of students	196	254	29	
	URM[a]	**Non-URM**	**Not reported**	
Number of students	85	338	56	

[a]URM, Self-reported identification as an underrepresented minority.

score in our initial model (Beck and Bliwise, 2014). Because the interaction was not significant ($F(2, 455.7) = 0.38$, $p = 0.68$), we removed the interaction from the final model.

As we were particularly interested in whether the impact of guided-inquiry laboratory modules differed based on student preparation, we divided students in each course into quartiles based on their pretest scores (quartile 1: highest scores; quartile 4: lowest scores), based on the assumption that lower pretest scores indicated less prior preparation. Then, we repeated the analyses described above for each quartile independently. All statistical analyses were conducted using SPSS Statistics version 24 (IBM Corporation, 2016).

Study 2. Impact of Courses with Guided-Inquiry Laboratory Modules on Experimental Design Skills

Student experimental design skills were assessed at the beginning (pretest) and end (posttest) of the semester, using the Experimental Design Ability Test (EDAT) developed by Sirum and Humburg (2011). As in study 1, course instructors imple-

mented the assessment in class. Students were instructed that the assessment was a part of a research study and no credit or extra credit would be given.

Student responses were scored blindly by the authors. Each written student response was assigned a unique random number, and all student responses were mixed before being scored. EDAT scores were matched to the source of the response only after all responses were scored. A standard scoring rubric was used (Sirum and Humburg, 2011; see Table 2). This rubric has 10 items for scoring responses that counted 1 point each for a maximum score of 10 (Sirum and Humburg, 2011). Interrater reliability on a random subset of EDAT responses ($N = 10$) was high (Pearson's correlation $r = 0.96$); therefore, each response was only scored by one of us. In addition, we maintained consistency in ratings by sitting together during scoring of all responses and discussing those for which there were questions about scoring. Following the analysis of Shanks *et al.* (2017), we partitioned the first four items, which were deemed lower-level experimental skills (basic understanding), and items 5, 8, and 10, which were deemed

TABLE 2. Rubric for scoring the EDAT (Sirum and Humburg 2011)[a]

Rubric item	Rubric answer
0	Did not answer
0.0	Answered, but no points
1	Recognition that an experiment can be done to test the claim (vs. simply reading the product label).
2	Identification of what variable is manipulated (independent variable is ginseng vs. something else).
3	Identification of what variable is measured (dependent variable is endurance vs. something else).
4	Description of how dependent variable is measured (e.g., how far subjects run will be measure of endurance).
5	Realization that there is one other variable that must be held constant (vs. no mention).
6	Understanding of the placebo effect (subjects do not know if they were given ginseng or a sugar pill).
7	Realization that there are many variables that must be held constant (vs. only one or no mention).
8	Understanding that the larger the sample size or number of subjects, the better the data.
9	Understanding that the experiment needs to be repeated
10	Awareness that one can never prove a hypothesis; that one can never be 100% sure; that there might be another experiment that could be done that would disprove the hypothesis; that there are possible sources of error; that there are limits to generalizing the conclusions (credit for any of these)

[a]Each item included in a student response to a writing prompt was given 1 point, except for the first two items (rubric items 0 and 0.0). All responses in which students did not answer were removed from analysis. The sum of responses for items 1–10 was the composite score; the sum of responses for items 1–4 was the basic understanding score; and sum of responses for items 5, 8, and 10 was the advanced understanding score, based on Shanks *et al.* (2017). The writing prompt for both pretest and posttest was: "Advertisements for an herbal product, ginseng, claim that it promotes endurance. To determine if the claim is fraudulent and prior to accepting this claim, what type of evidence would you like to see? Provide details of an investigative design."

TABLE 3. Distribution of students and institutions completing the EDAT

	Liberal arts colleges	Research university	Regional comprehensive university	Community college
Number of institution types	1	0	5	0

	Minority serving	Majority serving		
Number of institution types	3	3		

	Non–biology majors	Introductory biology	Upper-level biology	
Number of each course type	0	3	6	
Students in each course type	0	59	86	

	Males	Females	Not reported	
Number of students	37	69	39	

	URM[a]	Non-URM	Not reported	
Number of students	31	77	37	

[a]URM, Self-reported identification as an underrepresented minority.

higher-level experimental skills (advanced understanding). The total score based on all 10 items is referred to as the "composite score." For basic understanding, advanced understanding, and composite score, we calculated absolute gains as posttest score minus pretest score. Students who did not respond on either the pretest or the posttest were excluded from the analysis.

A total of 164 students from 12 courses at nine different institutions completed the pretest and the posttest. After dropping small-enrollment courses ($N < 10$ students) from the data set, we had pretest/posttest data for 145 students from nine courses at six different institutions. Only 74 of 145 students provided complete demographic data, which prevented us from examining the effects of student demographic factors on changes in experimental design skills (Table 3).

We analyzed the data on changes in experimental design skills for overall effects and effects by quartile in the same way as described above for changes in scientific reasoning skills. As in study 1, we assumed that lower pretest scores indicated less prior preparation. In separate analyses, we considered composite score, basic understanding, and advanced understanding as dependent variables. All statistical analyses were conducted using SPSS Statistics version 24 (IBM Corporation, 2016).
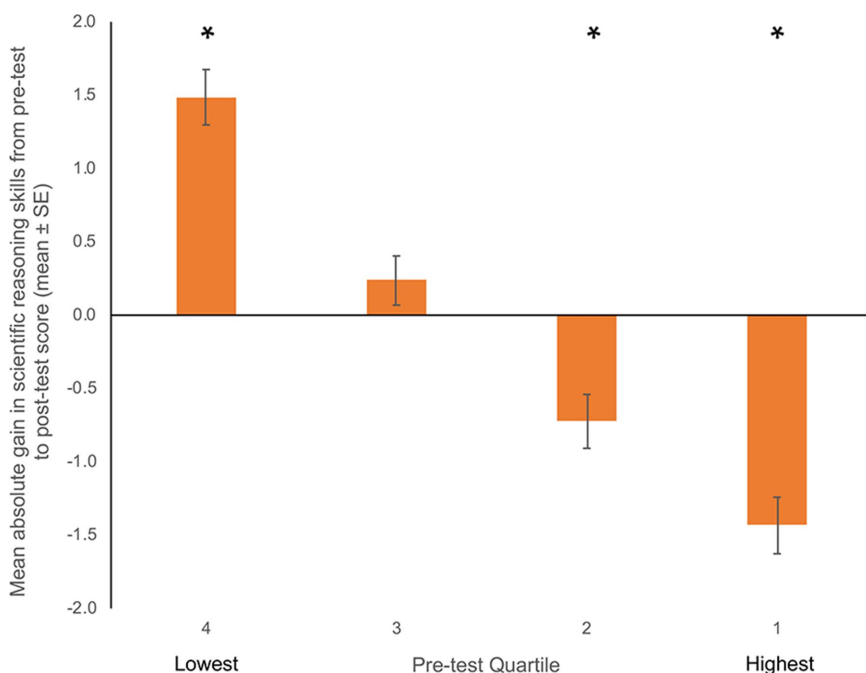

FIGURE 1. Mean absolute gain in scientific reasoning skills from pretest to posttest score (mean ± SE) by pretest quartile in study 1. The problems and questions used to assess these skills were from a test of scientific reasoning developed by Lawson (1978; version 2000). These results are for 479 students from 33 courses at 11 different institutions. Significant changes in absolute scores from pretest to posttest, within a quartile, are indicated with an asterisk (paired t tests, $p < 0.001$).

RESULTS

Study 1. Impact of Courses with Guided-Inquiry Laboratory Modules on Scientific Reasoning Skills

Overall, student scientific reasoning skills did not increase significantly between the pretest at the beginning of the semester (3.32 ± 0.11) and the posttest at the end of the semester (3.24 ± 0.11; $t_{478} = -0.77$, $p = 0.44$). Across all courses, absolute gains in scientific reasoning were significantly negatively related to pretest score (slope $= -0.46$, $t = -11.7$, $p < 0.001$), such that students with higher pretest scores showed lower gains. Gains did not vary significantly among courses (Wald $Z = 0.51$, $p = 0.61$); however, course level had a significant effect on gains ($F(2, 23.9) = 5.32$, $p = 0.012$). Students in upper-level courses showed significantly higher absolute gains (0.65 ± 0.28) than students in nonmajors courses (-0.50 ± 0.21) and students in introductory biology courses (-0.09 ± 0.12). Changes in scientific reasoning pretest to posttest were not significantly different between nonmajors and introductory biology students ($p = 0.10$).

When we divided the students into quartiles based on pretest score within each course, we found significant gains
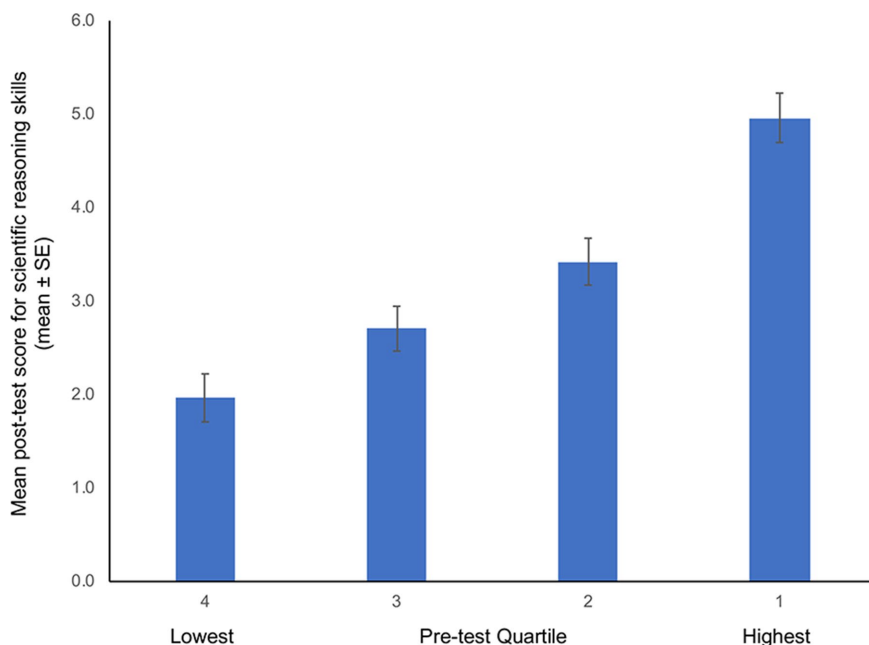
FIGURE 2. Mean posttest score (mean ± SE) in scientific reasoning skills among pretest quartiles in study 1. The problems and questions used to assess these skills were from a test of scientific reasoning developed by Lawson (1978; version 2000). These results are for 479 students from 33 courses at 11 different institutions. There are significant differences between all the quartiles in posttest scores ($F(3,460) = 42.7$, $p < 0.001$).

for the students in the lowest quartile ($t_{108} = 7.88$, $p < 0.001$), but significant decreases pretest to posttest for students in the highest two quartiles (1st quartile: $t_{98} = -7.44$, $p < 0.001$; 2nd quartile: $t_{126} = -3.94$, $p < 0.001$; Figure 1). Students in the third quartile showed no significant gains in scientific reasoning skills ($t_{143} = 1.41$, $p = 0.16$; Figure 1). Despite differences among the quartiles in the change in scientific reasoning skills between the pretest and posttest, all quartiles still showed significant differences in posttest scores ($F(3,460) = 42.7$, $p < 0.001$; Figure 2).

Dividing students into quartile based on pretest score reduces the variation in pretest score within a quartile, so absolute gains

in scientific reasoning were unrelated to pretest score in all but the 2nd quartile (Table 4). Similar to when we considered the entire data set combined, gains did not differ significantly among courses for most quartiles, with the exception of the lowest quartile (Wald $Z = 1.98$, $p = 0.048$). Course level also did not have a significant effect on gains in scientific reasoning when each quartile group was considered independently (Table 4).

Study 2. Impact of Courses with Guided-Inquiry Laboratory Modules on Experimental Design Skills

Across all courses, students' basic understanding, advanced understanding, and composite score on the EDAT did not increase significantly from pretest to posttest (Table 5). Absolute gains in composite score were significantly negatively related to pretest score (slope = -0.66, $t = -7.83$, $p < 0.001$), such that students with higher pretest scores showed lower gains. Gains did not vary significantly among courses (Wald $Z = 1.91$, $p = 0.23$); however, course level had a significant effect on gains ($F(1, 9.4) = 5.30$, $p = 0.046$). Students in upper-level courses showed significantly lower absolute gains (-0.28 ± 0.29) than students in introductory biology courses (0.85 ± 0.39). The effects of course level on absolute gains in composite score appear to be due to increases in the understanding of the placebo effect (rubric item 6; see Table 2) for students in introductory courses (Figure 3). Course level also influenced changes in advanced understanding ($F(1, 9.6) = 5.06$, $p = 0.049$) for which students in upper-level courses showed significantly lower absolute gains (-0.14 ± 0.11) compared with students in introductory biology courses (0.26 ± 0.14). Absolute gains in basic understanding did not differ significantly based on course level ($F(1, 9.0) = 2.75$, $p = 0.13$; lower: 0.35 ± 0.24; upper: -0.16 ± 0.18).

TABLE 4. Effect of course level and pretest score on absolute gain in scientific reasoning by quartile[a]

Quartile	Source	F	Numerator df	Denominator df	p value
1st	Intercept	0.15	1	95	0.70
	Course level	0.038	2	95	0.96
	Pretest score	2.96	1	95	0.089
2nd	Intercept	3.54	1	123	0.062
	Course level	3.40	2	123	0.036
	Pretest score	6.65	1	123	0.011
3rd	Intercept	1.05	1	140	0.31
	Course level	1.39	2	140	0.25
	Pretest score	0.83	1	140	0.36
4th	Intercept	34.76	1	36.0	0.001
	Course level	1.57	2	33.6	0.22
	Pretest score	0.92	1	54.5	0.34

[a]Statistics for 4th quartile calculated using linear mixed-effects model with course as random effect. For other quartiles, course did not have a significant effect on gains in scientific reasoning. As a result, general linear models were used.

TABLE 5. Effect of guide-inquiry laboratory pedagogy on experimental design skills as assessed with the EDAT

Measure	Pretest score	Posttest score	t value	p value
Composite score	3.74 ± 0.16	3.90 ± 0.17	0.810	0.42
Basic understanding	2.66 ± 0.10	2.74 ± 0.10	0.758	0.45
Advanced understanding	0.57 ± 0.05	0.56 ± 0.06	−0.094	0.92

FIGURE 3. Experimental design skills observed on pretest and posttest for students in (A) introductory biology ($N = 59$ students from three courses) and (B) upper-level biology ($N = 86$ students from six courses) courses in study 1. The observed experimental design skills are the percentage of students whose response to the EDAT prompt included a given rubric category. The categories on the x-axis represent the EDAT scoring rubric items (Table 2) listed in the order of experimental design sophistication (Sirum and Humburg, 2011). Categories in bold font are components of the basic understanding score, and those in capital letters are components of the advanced understanding score based on Shanks *et al.* (2017).

When we divided the students into quartiles based on pretest score within each course, we found significant gains for the students in the lowest quartile for basic understanding ($t_{32} = 4.25$, $p < 0.001$), advanced understanding ($t_{32} = 3.14$, $p = 0.004$), and composite score ($t_{32} = 5.67$, $p < 0.001$; Figure 4). In fact, a greater percentage of students showed an understanding of all aspects of experimental design on the posttest compared with the pretest (Figure 5). In contrast, students in the top quartile showed a significant decrease in composite score ($t_{34} = -2.90$, $p = 0.006$), likely due to a decrease in advanced understanding ($t_{34} = -2.42$, $p = 0.02$), as their basic understanding scores did not change pretest to posttest ($t_{34} = -1.39$, $p = 0.17$; Figure 4). In particular, students in the top quartile were less likely to include a description of factors that might need to be controlled in their experimental designs on the posttest as compared with the pretest (Figure 5). That these students forgot about controls is unlikely; rather, motivation to respond with complete answers on the posttest could have been low. Students in the second quartile exhibited a significant decrease in basic understanding pretest to posttest ($t_{38} = -2.54$, $p = 0.02$), but not in the other measures of experimental design skill (Figure 4; composite score: $t_{38} = -1.71$, $p = 0.10$; advanced understanding: $t_{38} = -1.36$, $p = 0.18$). They were less likely to articulate the independent and dependent variables in their experimental designs on the posttest compared with the pretest (Figure 5), again perhaps suggesting a lack of motivation. Students in the third quartile showed no significant change from the beginning of the semester to the end of the semester in experimental design skills (Figures 4 and 5; composite score: $t_{37} = 0.17$, $p = 0.87$; basic understanding: $t_{37} = -0.47$, $p = 0.64$; advanced understanding: $t_{37} = 0.55$, $p = 0.88$). Despite differences in absolute gains by quartile, posttest scores still differed among quartile groups for composite score (composite score: $F(3,135) = 6.74$, $p < 0.001$) and basic understanding ($F(3,135) = 3.96$, $p = 0.01$), especially for the top-quartile students in comparison with students in the other quartiles (Figure 6). Differences among the quartiles in advanced understanding were only marginally significant ($F(3,135) = 2.52$, $p = 0.06$; Figure 6).

To examine whether the differences in gains across quartiles were influenced by course level, we examined the effect of course level on gains for each quartile separately. Similar to when we considered the entire data set combined, absolute gains in composite score did not differ
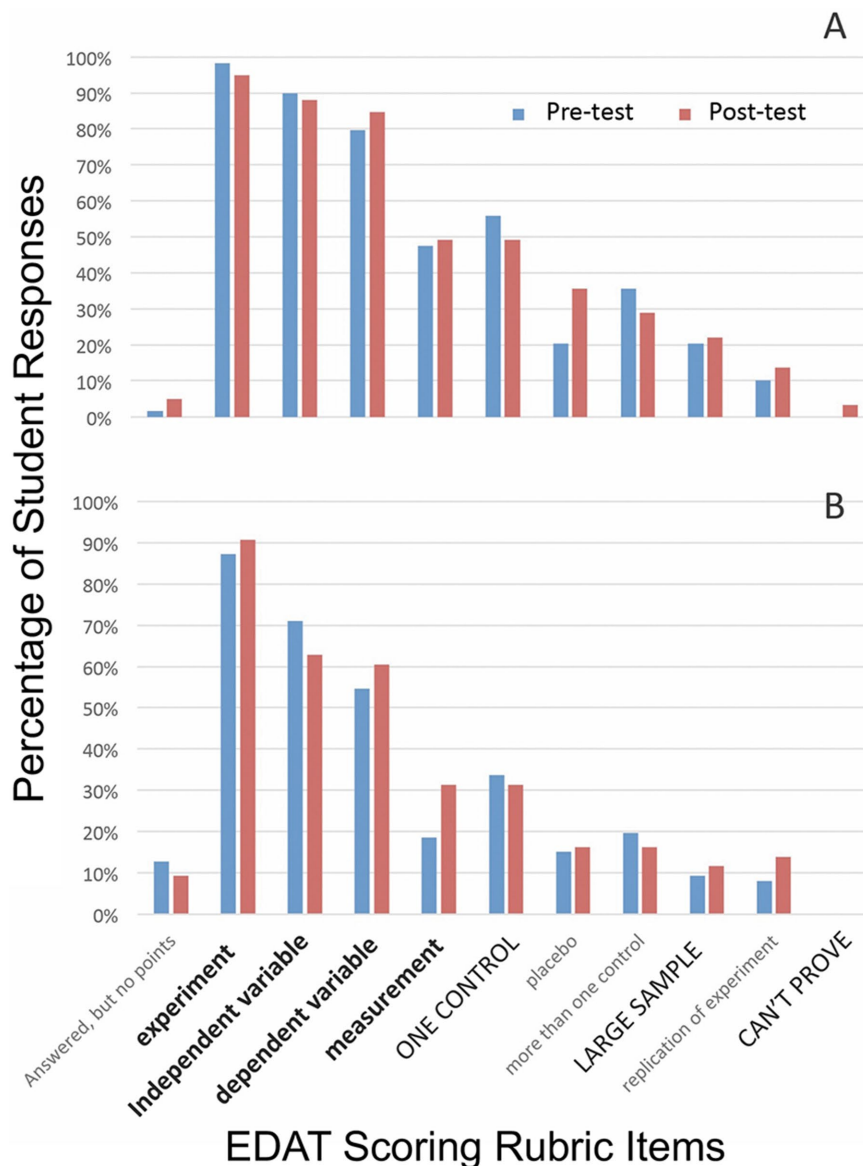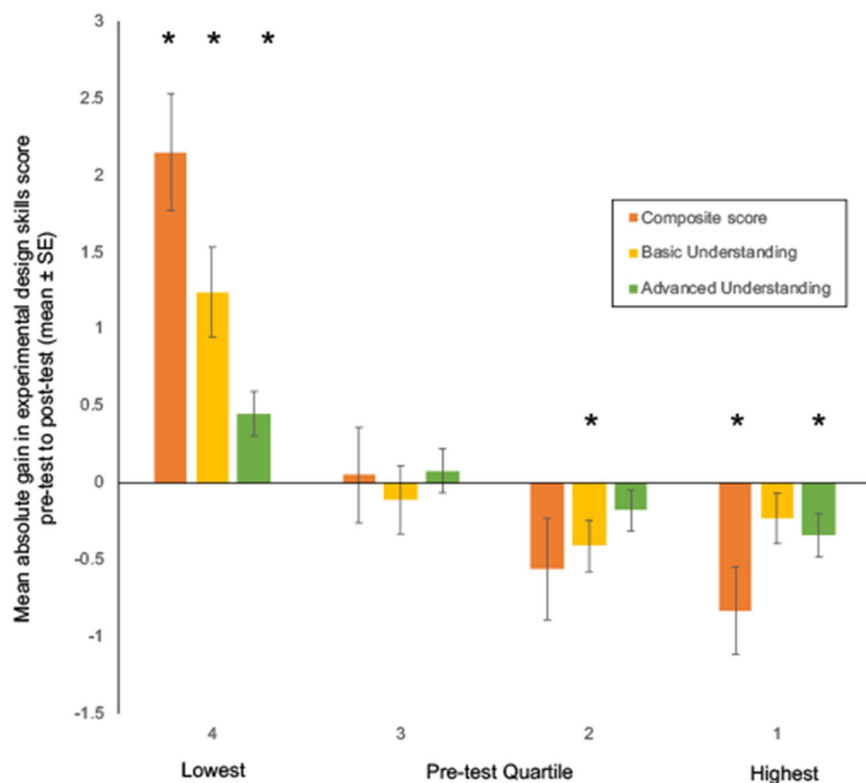
FIGURE 4. Mean absolute gain in experimental design skills from pretest to posttest score (mean ± SE) by pretest quartile in study 2. Experimental design skills were assessed using the EDAT (Sirum and Humburg, 2011). In each quartile, we show the mean absolute gain in composite EDAT score (orange, maximum possible gain = 10), the mean absolute gain in basic understanding score (yellow, maximum possible gain = 4), and the mean absolute gain in advanced understanding score (green, maximum possible gain = 3). These results are for 145 students from nine courses at six different institutions. Significant changes in EDAT scores from pretest to posttest, within quartile, are indicated with an asterisk (paired *t* test, $p \leq 0.02$).

significantly among courses for all quartiles ($p > 0.5$ in all cases). Generally, gains in experimental design skills were negatively related to pretest scores and not significantly affected by course level in any quartile (Table 6). However, for absolute gains in composite score, we found a positive relationship between pretest score and gains for students in introductory courses and a negative relationship between pretest score and gains for students in upper-level courses for students in the top quartile (significant pretest score*course level interaction for 1st quartile; Table 6). As a result, upper-level students showed a greater gain when they had lower pretest scores, but introductory students exhibited a greater gain when they had higher pretest scores. Students in the lowest quartile also showed differences in gains in advanced understanding depending on course level (Table 6). Students in introductory courses had greater gains (0.84 ± 0.16) compared with students in upper-level courses (0.17 ± 0.14), independent of pretest scores.

DISCUSSION

In general, students exhibit significant learning gains when they take inquiry-based laboratory courses (Beck *et al.*, 2014). In contrast, we found no significant gains in students' scientific reasoning and experimental design skills in guided-inquiry labora-

tory courses across a broad range of course levels and institutions, although these skills were emphasized in the guided-inquiry approach presented at our faculty development workshops. However, the lack of an overall effect of guided-inquiry on learning gains appears to be due to significant improvements by the least-prepared students (i.e., students in the lowest quartile on the pretest) being counterbalanced by significant decreases in posttest scores by the students in the highest quartile on the pretest. Previous studies have shown similar disproportionate benefits of pedagogical innovations to students with less preparation (e.g., Beck and Blumer, 2012; Gross *et al.*, 2015; Shapiro *et al.*, 2017). Indeed, our study provides evidence that it is important to include guided-inquiry modules in laboratory courses because of the positive gains in scientific reasoning and experimental design skills for the least-prepared students.

In study 1, we examined the effect of laboratory courses with a guided-inquiry module on changes in students' scientific reasoning skills. The students who were least prepared, based on lowest quartile pretest scientific reasoning skills (0.59 ± 0.11 out of a possible 8 points), showed significant increases in posttest scientific reasoning skills (2.00 ± 0.22). Yet these gains were not sufficient to close the gap between students in the lowest quartile and those in the other quartiles. In study 2, we explored whether students' experimental design skills improved as a result of taking a laboratory course with a guided-inquiry module. Previous studies have used the EDAT assessment tool to examine student outcomes in courses that implemented either authentic inquiry or guided-inquiry. Sirum and Humburg (2011) showed that students in laboratory courses in which students had to design their own experiments demonstrated significant increases in experimental design skills, as measured with the EDAT, but that students in traditional laboratory courses exhibited no change. Similarly, Shanks *et al.* (2017) showed that students in an introductory biology laboratory course that had an authentic research experience integrated into the existing laboratory curriculum exhibited significant gains in experimental design skills (i.e., EDAT score). In addition, students in a guided-inquiry biochemistry laboratory course showed significant increases in EDAT score, whereas those in cookbook sections of the same course did not (Goodey and Talgar, 2016). Using the same assessment, at the beginning of the semester, we found that the majority (64%) of the least-prepared students (i.e., those in the lowest quartile based on pretest scores) articulated that an experiment should be conducted to test a claim; however, few of these students could describe the components of an experimental design for the experiment. By the end of the semester, students in the lowest quartile exhibited significant gains in their experimental
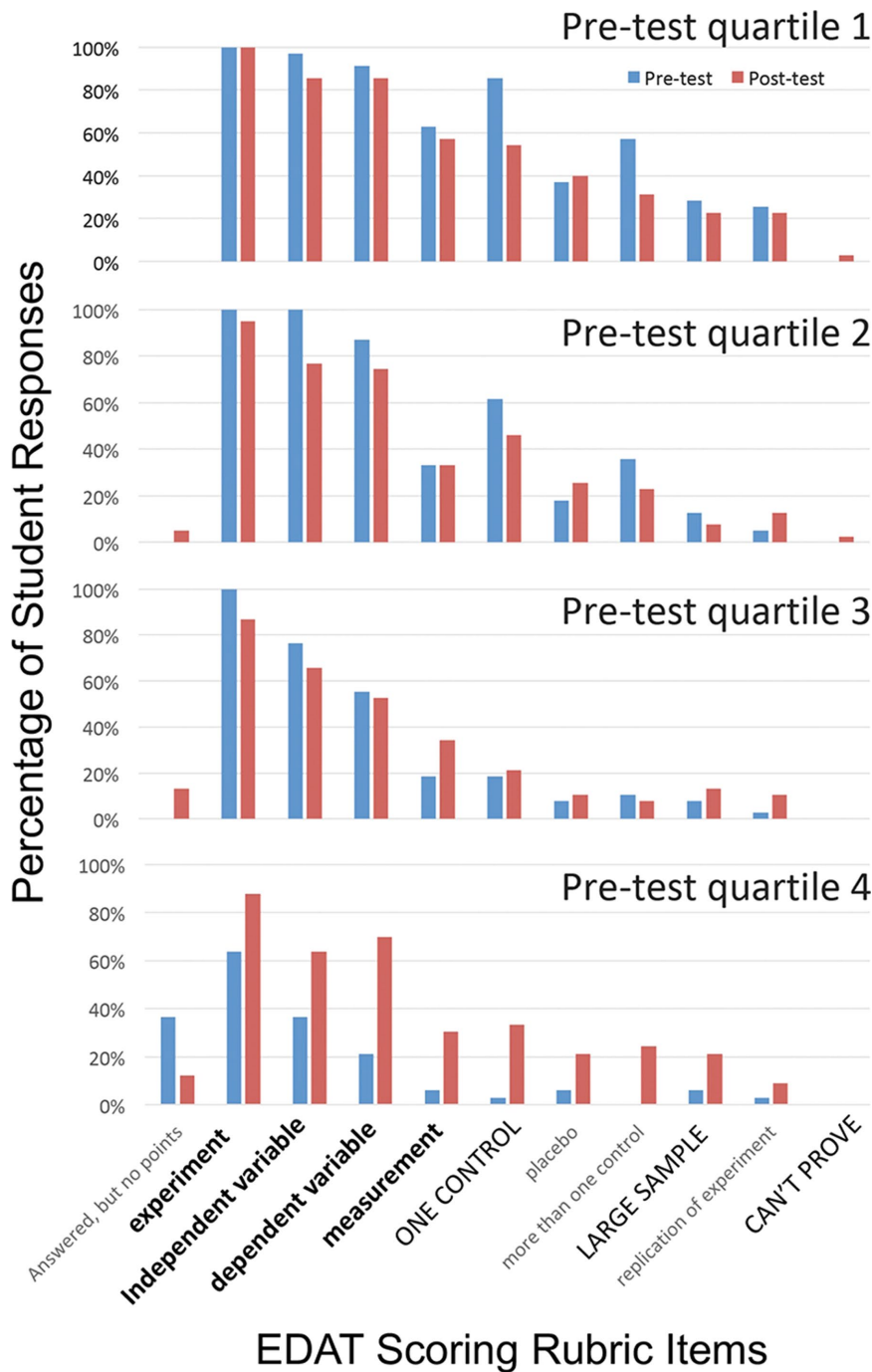
FIGURE 5. Experimental design skills observed on pretest and posttest for students in different quartiles. The observed experimental design skills are the percentage of students whose response to the EDAT prompt included a given rubric category in study 2. The categories on the *x*-axis represent the EDAT scoring rubric items (Table 2) listed in the order of experimental design sophistication (Sirum and Humburg, 2011). Categories in bold font are components of the basic understanding score, and those in capital letters are components of the advanced understanding score based on Shanks *et al.* (2017). The sample size was 145 students from nine courses at six different institutions.

design skills. Importantly, these gains were not only in basic understanding (identification of independent and dependent variables and how to measure the dependent variable), but also in advanced understanding of experimental design (that at least

one other variable must be controlled and that larger sample sizes result in more reliable conclusions). Gains by students in the lowest quartile were sufficient that differences in posttest scores were no longer significant between students in the lowest quartile and those in the middle two quartiles. However, the posttest scores for students in these three quartiles was similar to the pretest scores for students in the cohorts studied by Sirum and Humburg (2011) and Shanks *et al.* (2017). Yet the results from our two experiments taken together suggest that courses that include a guided-inquiry module can result in the development of basic scientific reasoning and experimental design skills for students who are least prepared across a range of course levels and institution types.

In contrast to students in the lowest quartile, based on pretest scores, students in the top quartile showed significant decreases in scientific reasoning skills (study 1) and experimental design skills (study 2) across the course of the semester. While these results were initially disconcerting, multiple possible explanations might account for them. First, students in the top quartile might exhibit a ceiling effect with the assessments we used. For example, for scientific reasoning skills, students in the top quartile scored 6.73 ± 0.15 points out of a possible 8 points on the pretest. As a result, the possibility for increases in scientific reasoning were limited. In contrast, for experimental design skills (study 2), the top-quartile students only scored a 5.86 ± 0.19 points out of a possible 10 points on the pretest. Although the pretest scores on the EDAT were similar to the posttest scores for the cohort in Shanks *et al.* (2017), a ceiling effect for the EDAT seems unlikely. The students in the experimental sections of Sirum and Humburg's (2011) study averaged 6.6 on the EDAT at the end of the semester. In addition, fewer than 50% of the students in the top quartile of our study articulated the importance of large sample sizes, the possibility of a placebo effect, the need to repeat experiments, and uncertainty in science on the pretest.

Second, we avoided the phenomenon of regression to the mean (Nesselroade *et al.*, 1980; Marsden and Torgerson, 2012) by minimizing the impact of random responses to multiple-choice assessment questions (study 1) and by using an open-ended response assessment scored with a rubric (study 2). In an extreme case, regression to the mean could occur if all students were to respond to questions
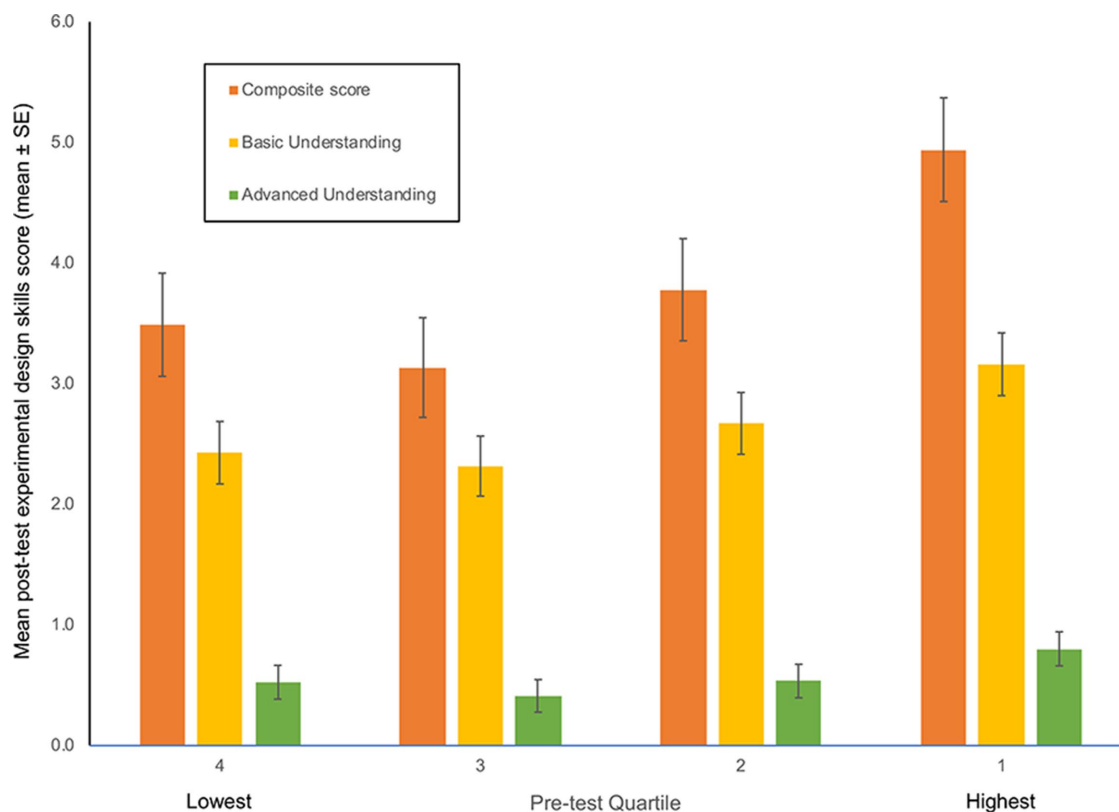
FIGURE 6. Mean posttest experimental design skills score (mean ± SE) among pretest quartiles in study 2. Experimental design skills were assessed using the EDAT (Sirum and Humburg, 2011). In each quartile, we show the mean composite EDAT posttest score (orange, maximum possible score = 10), the mean basic understanding posttest score (yellow, maximum possible score = 4), and the mean advanced understanding score (green, maximum possible score = 3). There are significant differences among quartiles in composite posttest score ($F(3,135) = 6.74$, $p < 0.001$) and basic understanding posttest score ($F(3,135) = 3.96$, $p = 0.01$), and marginally significant differences among quartiles in advanced understanding posttest scores ($F(3,135) = 2.52$, $p = 0.06$). These results are for 145 students from nine courses at six different institutions.

on a multiple-choice pretest assessment at random; some students would fall within the top quartile and others in the lowest quartile. If these same students were to respond at random to the same assessment as a posttest, the mean score for students in the highest quartile would decrease and that for students in the lowest quartile would increase. For any assessment in which random error of measurement of true scores occurs, such regression to the mean is possible. Minimizing the random error of measurement will reduce this effect. In study 1, pairs of multiple-choice questions related to the same scenario were used to assess scientific reasoning skills. As a result, students received credit for those questions only if they answered both questions in a pair correctly, thus reducing the likelihood that students could answer questions correctly by chance alone. Therefore, decreases in scientific reasoning for students in the top quartile and increases in scientific reasoning for the lowest quartile are unlikely to be explained by regression to the mean. In study 2, experimental design skills were assessed with an open-ended assessment that was scored using a rubric. By their nature, open-ended assessments scored using a rubric should have lower random error of measurement than multiple-choice assessments. At a minimum, students cannot guess the correct answer at random, which should make regression to the mean

less likely for open-ended assessments. Furthermore, decreases due to regression to the mean for top-quartile students should be evident across all elements of a rubric for an open-ended assessment. However, students in the top quartile in study 2 did not show decreases in those elements of experimental design related basic understanding, but only those associated with advanced understanding. As a result, decreases in experimental design skills for top-quartile students are unlikely due to regression to the mean.

Finally, students in the top quartile on the pretest might not have been sufficiently motivated to perform well on the posttest, and therefore their scores declined. In both experiments, the assessments were implemented by diverse faculty across a wide range of courses and institutions. As a result, the degree to which faculty motivated their students to perform well on assessments could not be standardized. While faculty were encouraged to emphasize the importance of the study to their students, we had no way of insuring how and the degree to which this was done. Indeed, student motivation can significantly decrease performance on an assessment when it is not graded compared with when it is graded (e.g., Napoli and Raymond, 2004). Documenting a motivation effect on our assessment of scientific reasoning skills is difficult. However,

TABLE 6. Effect of course level and pretest score on absolute gain in experimental design skills by quartile

Quartile	Source	F	Numerator df	Denominator df	p value
Composite score					
1st quartile	Intercept	0.12	1	31	0.74
	Course level	8.06	1	31	0.008
	Pretest score	0.12	1	31	0.73
	Pretest*level	8.87	1	31	0.006
2nd quartile	Intercept	7.12	1	36	0.011
	Course level	2.54	1	36	0.12
	Pretest score	9.48	1	36	0.004
3rd quartile	Intercept	2.14	1	35	0.15
	Course level	0.56	1	35	0.46
	Pretest score	2.26	1	35	0.14
4th quartile	Intercept	20.62	1	30	<0.001
	Course level	2.42	1	30	0.13
	Pretest score	4.57	1	30	0.04
Basic understanding					
1st quartile	Intercept	3.37	1	32	0.08
	Course level	0.86	1	32	0.36
	Pretest score	4.45	1	32	0.04
2nd quartile	Intercept	3.26	1	36	0.08
	Course level	2.30	1	36	0.13
	Pretest score	5.26	1	36	0.03
3rd quartile	Intercept	5.93	1	35	0.08
	Course level	1.43	1	35	0.37
	Pretest score	7.65	1	35	0.04
4th quartile	Intercept	58.95	1	30	<0.001
	Course level	2.81	1	30	0.23
	Pretest score	25.39	1	30	0.001
Advanced understanding					
1st quartile	Intercept	7.32	1	32	0.01
	Course level	1.92	1	32	0.18
	Pretest score	22.06	1	32	<0.001
2nd quartile	Intercept	5.95	1	36	0.02
	Course level	0.08	1	36	0.78
	Pretest score	17.40	1	36	<0.001
3rd quartile	Intercept	20.36	1	35	<0.001
	Course level	3.25	1	35	0.08
	Pretest score	48.55	1	35	<0.001
4th quartile	Intercept	36.53	1	30	<0.001
	Course level	9.65	1	30	0.004
	Pretest score	32.08	1	30	0.001

for experimental design skills, the lack of motivation for top-quartile students on the posttest was more apparent. These students were less likely to include a description of factors that might need to be controlled in their experimental designs on the posttest as compared with the pretest. That they forgot about controls or "unlearned" the importance of controls in experimental design seems unlikely. It is more likely that these students were unmotivated to write longer and more complete answers. In fact, some students were excluded from our final sample because their responses to the posttest prompt were similar to "see my answer for the pretest."

Because the guided-inquiry laboratory modules were implemented and assessed at different course levels, we explored whether changes in scientific reasoning skills and experimental design skills differed based on course level. In study 1, students in upper-level courses showed significantly higher gains than students in nonmajors and introductory biology courses. However, course level did not have a significant effect on gains when we analyzed the data by pretest quartile. In contrast to study 1, in study 2, students in upper-level courses showed significantly lower gains than students in introductory biology courses. We found a similar effect for students in the lowest quartile in terms of advanced understanding of experimental design. The

contrasting results between the two experiments suggest that the impact of laboratory courses with a guided-inquiry module on student learning gains is variable across course levels and might be influenced by individual courses or institutions themselves. In our sample, we only had data on courses at one particular level for most institutions, which led to course-level effects potentially being confounded with institution-level effects. To clarify the effects of course level on the impact of guided-inquiry laboratories, future studies should examine courses at different levels at the same institution.

While inquiry-based laboratory courses generally lead to improved learning gains, the absence of an effect has occasionally been found (Beck *et al.*, 2014). Likely, other instances of negligible impacts of educational interventions on student learning abound, but are not published, due to the "file drawer" effect. However, our results suggest that the absence of overall effects of an intervention might be masking positive effects for the least-prepared students. Therefore, we recommend that future studies examine the impact of interventions for different pretest quartiles separately. When taking this approach, investigators need to be cautious of the possibility that differences in gains among quartiles might simply be indicative of regression to the mean, especially when using multiple-choice assessments (Marsden and Torgerson, 2012). For studies with control and intervention groups, ANCOVA can be used to control for regression to the mean (Barnett *et al.*, 2005), and a significant interaction between pretest score and treatment group on posttest score would indicate differential effects of an intervention based on student preparation (Beck and Bliwise, 2014).

REFERENCES

American Association for the Advancement of Science. (2011). *Vision and change in undergraduate biology education*. Washington, DC.

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, *34*, 215–220.

Beck, C. W., & Bliwise, N. G. (2014). Interactions are critical. *CBE—Life Sciences Education*, *13*, 371–372.

Beck, C. W., & Blumer, L. S. (2012). Inquiry-based ecology laboratory courses improve student confidence and scientific reasoning skills. *Ecosphere*, *3*, art112.

Beck, C. W., & Blumer, L. S. (2016). Alternative realities: Faculty and student perceptions of instructional practices in laboratory courses. *CBE—Life Sciences Education*, *15*, ar52. doi: 10.1187/cbe.16-03-0139

Beck, C. W., Butler, A., & Burke da Silva, K. (2014). Promoting inquiry-based teaching in laboratory courses: Are we meeting the grade? *CBE—Life Sciences Education*, *13*, 444–452.

Belzer, S., Miller, M., & Shoemake, S. (2003). Concepts in biology: A supplemental study skills course designed to improve introductory students' skills for learning biology. *American Biology Teacher*, *65*, 30–40.

Brownell, S. E., Kloser, M. J., Fukami, T., & Shavelson, R. (2012). Undergraduate biology lab courses: Comparing the impact of traditionally based "cookbook" and authentic research-based courses on student lab experiences. *Journal of College Science Teaching*, *41*, 36–45.

Butcher, G. Q., & Chirhart, S. E. (2013). Diet and metabolism in bean beetles. In McMahon, K. (Ed.), *Proceedings of the 34th conference of the Association for Biology Laboratory Education (ABLE)* (Tested studies for laboratory teaching, Vol. 34, pp. 420–426). Chapel Hill, NC: Association for Biology Laboratory Education.

Crimmins, M. T., & Midkiff, B. (2017). High structure active learning pedagogy for the teaching of organic chemistry: Assessing the impact on academic outcomes. *Journal of Chemical Education*, *94*, 429–438.

D'Costa, A. R., & Schlueter, M. A. (2013). Scaffolded instruction improves student understanding of the scientific method and experimental design. *American Biology Teacher*, *75*, 18–28. doi: 10.1525/abt.2013.75.1.6

DeRuisseau, L. R. (2016). The flipped classroom allows for more class time devoted to critical thinking. *Advances in Physiology Education*, *40*, 522–528.

Eddy, S. L., Brownell, S. E., & Wenderoth, M. P. (2014). Gender gaps in achievement and participation in multiple introductory biology classrooms. *CBE—Life Sciences Education*, *13*, 478–492.

Eddy, S. L., Crowe, A. J., Wenderoth, M. P., & Freeman, S. (2013). How should we teach tree-thinking? An experimental test of two hypotheses. *Evolution Education and Outreach*, *6*, 13.

Eddy, S. L., & Hogan, K. A. (2014). Getting under the hood: How and for whom does increasing course structure work? *CBE—Life Sciences Education*, *13*, 453–468.

Fermin, H., Singh, G., Gul, R., Noh, M. G., Ramirez, V. C., & Firooznia, F. (2014). Pests be gone! Or not?! Looking at the effect of an organophosphate insecticide on acetylcholinesterase activity in the bean beetle. In McMahon, K. (Ed.), *Proceedings of the 35th conference of the Association for Biology Laboratory Education (ABLE)* (Tested studies for laboratory teaching, Vol. 35, pp. 436–444). Calgary, AB: Association for Biology Laboratory Education.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences USA*, *111*, 8410–8415.

Goodey, N. M., & Talgar, C. P. (2016). Guided inquiry in a biochemistry laboratory course improves experimental design ability. *Chemistry Education Research and Practice*, *17*, 1127–1144

Gross, D., Pietri, E. S., Anderson, G., Moyano-Camihort, K., & Graham, M. J. (2015). Increased preclass preparation underlies student outcome improvement in the flipped classroom. *CBE—Life Sciences Education*, *14*, ar36. doi: 10.1187/cbe.15-02-0040

Haak, D. C. C., HilleRisLambers, J., Pitre, E., & Freeman, S. (2006). Increased structure and active learning reduce the achievement gap in introductory biology. *Science*, *332*, 1213–1216.

IBM Corporation. (2016). *IBM SPSS Statistics for Windows (Version 24.0)*. Armonk, NY.

Jensen, J. L., & Lawson, A. (2011). Effects of collaborative group composition and inquiry instruction on reasoning gains and achievement in undergraduate biology. *CBE—Life Sciences Education*, *10*, 64–73.

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*, *15*, 11–24.

Lorenzo, M., Crouch, C., & Mazur, E. (2006). Reducing the gender gap in the physics classroom. *American Journal of Physics*, *74*, 118–122.

Marbach-Ad, G., Rietschel, C. H., Saluja, N., Carleton, K. L., & Haag, E. S. (2016). The use of group activities in introductory biology supports learning gains and uniquely benefits high-achieving students. *Journal of Microbiology & Biology Education*, *17*, 360–369.

Marsden, E., & Torgerson, C. J. (2012). Single group, pre- and posttest research designs: Some methodological concerns. *Oxford Review of Education*, *38*, 583–616.

Napoli, A. R., & Raymond, L. A. (2004). How reliable are our assessment data? A comparison of the reliability of data produced in grades and un-graded conditions. *Research in Higher Education*, *45*, 921–929.

National Research Council. (2003). *BIO 2010. Transforming undergraduate education for future research biologists*. Washington, DC: National Academies Press.

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression toward the mean and the study of change. *Psychological Bulletin*, *88*, 622–637.

Pearce, A. R., Sale, A. L., Srivatsan, M., Beck, C. W., Blumer, L. S., & Grippo, A. A. (2013). Inquiry-based investigation in biology laboratories: Does neem provide bioprotection against bean beetles? *Bioscene*, *39*, 11–16.

President's Council of Advisors on Science and Technology. (2012). *Engage to excel: Producing one million additional college graduates with degrees in science*. Washington, DC: U.S. Government Office of Science and Technology.

Preszler, R. W. (2009). Replacing lecture with peer-led workshops improves student learning. *CBE—Life Sciences Education*, *8*, 182–192.

Rodenbusch, S. E., Hernandez, P. R., Simmons, S. L., & Dolan, E. L. (2016). Early engagement in course-based research increases graduation rates and completion of science, engineering, and mathematics degrees. *CBE—Life Sciences Education*, *15*, ar20. doi: 10.1187/cbe.16-03-0117

Schlueter, M. A., & D'Costa, A. R. (2013). Guided-inquiry labs using bean beetles for teaching the scientific method and experimental design. *American Biology Teacher*, *75*, 214–218.

Shaffer, J. F., Dang, J. V., Lee, A. K., Dacanay, S. J., Alam, U., Wong, H. Y., … Sato, B. K. (2016). A familiar(ity) problem: Assessing the impact of prerequisites and content familiarity on student learning. *PLoS One*, *11*, e0148051. doi: 10.1371/journal.pone.0148051

Shanks, R. A., Robertson, C. L., Haygood, C. S., Herdliska, A. M., Herdliska, H. R., & Lloyd, S. A. (2017). Measuring and advancing experimental design ability in an introductory course without altering existing lab curriculum. *Journal of Microbiology & Biology Education*, *18* doi: 10.1128/jmbe.v1118i1121.1194

Shapiro, A. M., Sims-Knight, J., O'Rielly, G. V., Capaldo, P., Pedlow, T., Gordon, L., & Monteiro, K. (2017). Clickers can promote fact retention but impede conceptual understanding: The effect of the interaction between clicker use and pedagogy on learning. *Computers & Education*, *111*, 44–59.

Sirum, K., & Humburg, J. (2011). The Experimental Design Ability Test (EDAT). *Bioscene*, *37*, 8–16.

Smith, J. M., & Hicks, K. A. (2014). Detecting genetic polymorphisms in different populations of bean beetles (*Callosobruchus maculatus*). In McMahon, K. (Ed.), *Proceedings of the 35th conference of the Association for Biology Laboratory Education (ABLE)* (Tested studies for laboratory teaching, Vol. 35, pp. 281–289). Calgary, AB: Association for Biology Laboratory Education.

Theobald, E. (2018). Students are rarely independent: When, why, and how to use random effects in discipline-based education research. *CBE—Life Sciences Education*, *17*, rm2. doi: 10.1187/cbe.17-12-0280

Theobald, R., & Freeman, S. (2014). Is it the intervention or the students? Using linear regression to control for student characteristics in undergraduate STEM education research. *CBE—Life Sciences Education*, *13*, 41–48.