Self-Efficacy and Performance of Research Skills among First-Semester Bioscience Doctoral Students

Katherine Lachance,[†‡] Ronald J. Heustis,[§‖] Joseph J. Loparo,[‖] and Madhvi J. Venkatesh[‖]*

[†]Department of Genetics, Blavatnik Institute, [§]Program in Graduate Education, and [‖]Department of Biological Chemistry and Molecular Pharmacology, Blavatnik Institute, Harvard Medical School, Boston, MA 02115; [‡]Program in Bioinformatics and Integrative Genomics, Harvard University, Cambridge, MA 02138

ABSTRACT

Research skills, especially in experimental design, are essential for success in bioscience doctoral training. While there is a growing body of literature on the development of research skills among science, technology, engineering, and mathematics doctoral students, very little is specific to biosciences. We seek to address this gap by characterizing aptitude and self-perceived facility with research skills among incoming bioscience doctoral students, as well as how and why they change over the first semester of doctoral training. Our results reveal variability in research skills self-efficacy and a wide range in aptitude and self-perceived facility with experimental design at the beginning of the semester, both of which are uncorrelated with the duration of predoctoral research experience. We found that students significantly improved in both experimental design performance and research skills self-efficacy over their first semester; students attributed their experience and comfort with experimental design to a variety of factors, including laboratory research, course work, mentoring, and interaction with colleagues. Notably, we found that the largest research skills self-efficacy gains were aligned with material that was covered in students' first-year course work about experimental design. Together, these results demonstrate the importance of explicit training in experimental design and other research skills early in bioscience doctoral training.

INTRODUCTION

A fundamental goal of doctoral education is training students in research skills, thereby enabling them to transition from consumers to producers of knowledge (Weidman, 2010). Indeed, recent national reports and select training grant guidelines have explicitly emphasized research skills training for doctoral trainees in the science, technology, engineering, and mathematics (STEM) disciplines (National Institute of General Medical Sciences, 2017; National Academies of Sciences, Engineering, and Medicine [NASEM], 2018). However, there is limited scholarly work investigating how students' research skills develop during STEM doctoral training and which experiences are most important in contributing to these changes (Feldon, 2016). Moreover, most of this work is distributed across multiple STEM disciplines and is not specific to the biosciences.

Recent scholarship indicates that the trajectories of research skill development among doctoral trainees begin in undergraduate studies: Students who participate in undergraduate research have persistently higher research skill levels than their peers (Gilmore et al., 2015). Additionally, the duration of an individual's undergraduate research experience has been shown to correlate with both demonstrated research skill levels and self-rating of ability to perform scientific skills (Thiry et al., 2012; Gilmore et al., 2015). Because research skill development at the graduate level builds on pre-existing skills, students with more undergraduate research experience are able

to develop proficiency in higher-level research competencies more quickly than their less experienced counterparts (Timmerman *et al.*, 2013).

The literature also suggests that multiple factors, rather than simply engaging in laboratory research, contribute to the development of proficiency and self-perception of ability (or self-efficacy) in research skills among STEM doctoral students (Feldon, 2016). Faculty mentoring has been cited as playing a central role in the research skill development and research self-efficacy of doctoral trainees (Paglis *et al.*, 2006; Walker *et al.*, 2008; Barnes and Austin, 2009), despite concerns about mentors' limited time and pedagogical training and the variability of mentors' investment (Lovitts, 2001; Bianchini *et al.*, 2002; Anderson *et al.*, 2011). Additionally, doctoral student teaching experiences, which are sometimes considered to interfere with research training, actually lead to greater skill growth in key research skills (Feldon *et al.*, 2011). Other structured settings outside the research laboratory, including interdisciplinary teaching labs for expedited problem solving (Vale *et al.*, 2012) and a course in which students evaluate the validity of research conclusions based on experimental design (Zolman, 1999), are effective mechanisms for developing students' research skills.

While many of these findings can be generalized across doctoral training programs, research skills training must be customized for each field, as methodological research skills have nuances specific to individual disciplines (Gilbert *et al.*, 2004). For doctoral students in cellular and molecular biology specifically, research skills are built on a foundation of critically engaging with the primary literature and using appropriate control conditions to design experiments and interpret results (Feldon *et al.*, 2017b).

Understanding the purpose of experiments and importance of controls represents a subset of research skills within the framework of experimental design (Deane *et al.*, 2014). Experimental design is a core competency in the biosciences (Coil *et al.*, 2010; American Association for the Advancement of Science [AAAS], 2011) and has been taught and assessed in a discipline-specific manner at the undergraduate level. These efforts have included the development of content and activities for teaching the overarching concepts in experimental design (e.g., Hiebert, 2007; Pollack, 2010; D'Costa and Schlueter, 2013; Brownell *et al.*, 2014; Fry, 2014), instructional strategies targeting specific concepts such as proper controls (e.g., Lin and Lehman, 1999; Shi *et al.*, 2011), and the development of assessment rubrics (Sirum and Humburg, 2011; Brownell *et al.*, 2014; Dasgupta *et al.*, 2014; Killpack and Fulmer, 2018) and validated concept inventories (Deane *et al.*, 2014; University of British Columbia, 2014a,b). However, development of neither proficiency nor self-efficacy in experimental design has been explored among bioscience doctoral students.

Conceptual Framework

Self-efficacy and its theoretical underpinnings serve as a useful framework to understand the factors influencing how doctoral students perceive their own ability, experience, and comfort in performing research skills such as experimental design. Self-efficacy is defined as the belief in one's ability to accomplish a task or effect change (Bandura, 1977). Skill-specific self-efficacy is reported to be derived from four sources: 1) performance accomplishments, which involve demonstrating success on tasks requiring a given skill; 2) emotional states, or the feelings that arise when completing such tasks; 3) vicarious experiences derived from comparing oneself to others to determine norms and possible opportunities; and 4) social persuasion, which emanates from the encouragement of peers, instructors, and mentors (Bandura, 1977; Trujillo and Tanner, 2014). The related self-reported measures of experience and comfort with a given skill allow us to further understand the role of the sources of self-efficacy: Experience indicates the amount of practice that one believes one has in performing a given skill and is most closely aligned with performance accomplishments, while comfort captures feedback on performance and many of the sociocultural and emotional factors encapsulated by the other three sources. As doctoral students often encounter research challenges, self-efficacy, experience, and comfort in research skills, particularly in aspects of experimental design, are important in determining training outcomes, including research productivity (Brown *et al.*, 1996; Szymanski *et al.*, 2007; Gökçek *et al.*, 2014; Lambie *et al.*, 2014).

In STEM fields, women often have lower self-efficacy than their counterparts because of decreased opportunities for performance accomplishments, fewer positive role models from whom to draw vicarious experiences, and less social persuasion from colleagues who may have internalized biases (Hackett and Betz, 1981; Kardash, 2000). These can manifest in the STEM doctoral experiences of women through gender-based isolation and marginalization, a limited number of positive role models, disciplinary stereotypes, and inequities in academic recognition (MacLachlan, 2006; Hill *et al.*, 2010; De Welde and Laursen, 2011; Feldon *et al.*, 2017a).

Considering that self-efficacy is influenced by socioemotional factors like identity and is poorly correlated with performance-based evidence (Falchikov and Boud, 1989; Dunning *et al.*, 2003), it is important to study performance and self-efficacy independently. During early graduate training, these independent outcomes are dynamic, because students participate in formative experiences such as completing course work, performing research rotations, and engaging with the scientific community (Golde, 1998; Thakore *et al.*, 2014). Both factors contribute to doctoral training outcomes: Performance relates to students' ability to complete the tasks expected by graduate programs and faculty mentors (Feldon *et al.*, 2010), while self-efficacy affects levels of aspiration, motivation, and persistence (Bandura, 1977; Trujillo and Tanner, 2014).

Research Questions

Considering the importance of both self-efficacy and performance of research skills in early doctoral training, we have embedded experimental design training in one of our core bioscience courses. This study uses assessments to better characterize both the self-efficacy and performance of research skills of students enrolled in this course, with a particular focus on biological experimental design, by focusing on the following research questions:

1. Does the amount of time spent doing predoctoral research predict self-reported experience and comfort with experimental design, research skills self-efficacy, or experimental design aptitude upon entering a bioscience doctoral program?

2. In our student population, how do research skills self-efficacy and experimental design aptitude, as measured using assessments, change during the first semester of doctoral training?
3. Are there differences between the research skills self-efficacy of men and women in our population?
4. What first-semester doctoral training experiences do students report as the most important factor contributing to their levels of experience and comfort with experimental design?

METHODS

Study Population

This study was conducted with first-year doctoral students enrolled in the Principles of Molecular Biology course offered at a private, R1 institution in the northeastern United States. This course is open for enrollment to incoming life sciences doctoral students in several programs across the institution; it is required for the largest bioscience graduate program and highly recommended for many others. In addition to covering concepts in molecular biology, the course embeds experimental design training in lectures, discussion sections, and assessments by emphasizing the selection and justification of appropriate experimental approaches to test given hypotheses, the prediction and interpretation of results, and the identification of appropriate control conditions. Throughout the course, students receive feedback from peers and teaching assistants as they perform these tasks through written assessments and oral presentations. A majority of students enrolled in the course were also concurrently taking another course that emphasized statistics and quantitative skills training.

A total of 45 students in Fall 2017 and 58 students in Fall 2018 consented to participate in this study and successfully completed the pre- and postcourse surveys and Biological Experimental Design Concept Inventory (BEDCI), when applicable. All study participants completed all questions administered in the year during which they were enrolled in the course, so there were no missing data in this study. The assessments of each individual were linked but anonymous, using student-generated alphanumeric identifiers. We excluded data from students who completed the assessments as a part of their enrollment in the course, but did not consent to participate in this study.

This study population consisted of first-year doctoral students. There were 48 men (47%) and 52 women (50%) in the study population, which is representative of the gender balance across the institution's life sciences programs. Also included in the study were three students who declined to provide gender information; these students were excluded from analyses to test whether gender was a predictive factor in research skills self-efficacy, but were included in all other analyses.

All data collected and analyzed for this study were approved for exemption by our institution's institutional review board (IRB) and is covered by protocol IRB17-0668.

Comparison of Fall 2017 and Fall 2018 cohorts

The students who participated in the study in 2017 and 2018 were nearly identical in demographic and scientific background. We used a chi-squared test to test associations between categorical variables, including duration of students' previous lab research experience, students' gender and underrepresented minority status, and students' scientific backgrounds across the 2017 and 2018 student cohorts. We also employed a Wilcoxon signed-rank test to determine whether the two cohorts significantly differed in self-reported experience or comfort with experimental design and total research skills self-efficacy. While there was a significant difference between cohorts in the doctoral programs in which students were enrolled, there was no statistical difference between the cohorts in terms of gender, race, previous years of research lab experience, subjects of prior degrees, experience and comfort with experimental design, and self-reported research skills self-efficacy (Supplemental Table S1). Because the backgrounds of students across demographic, experience, and skill-based metrics were similar, and because this study took place in the first semester of their doctoral programs, we expect the impact of different program enrollment to be negligible. As such, we combined cohorts in analyses to boost the statistical power of these studies. This assumption, however, cannot be tested, as some programs had so few students enrolled in the course that a question about program affiliation on survey instruments would de-identify students. Additionally, there were no changes in the admissions standards or processes between 2016 and 2017 to create or explain major differences between the cohorts.

Pre- and Postcourse Surveys

During the first and last weeks of the semester, students were asked to complete online Qualtrics-based surveys for a small number of participation points. On both the pre- and postcourse surveys, students were asked identical questions about their levels of experience and comfort with experimental design. Students answered the question "How would you rate your experience practicing experimental design?" using a five-point Likert scale ranging from "no experience" to "extensive experience" and the question "How uncomfortable or comfortable are you with designing experiments?" using a seven-point Likert scale ranging from "extremely uncomfortable" to "extremely comfortable."

In addition, both the pre- and postcourse surveys included the research skills survey instrument developed by Kardash (2000). For each of the 14 items (Figure 3A, discussed later in the paper), students were asked "To what extent do you feel you can do each of the following?" Responses were collected on a five-point Likert scale with prompts ranging from "not at all" to "a great deal." The reliability of this research skills self-efficacy scale in our population was established by calculating Cronbach's alpha for both the pretest ($\alpha = 0.90$) and posttest ($\alpha = 0.93$). We recognize that research skills encompass both practical and theoretical aspects of conducting research in the laboratory. Even within the biosciences, practical research skills vary widely by subdiscipline, while theoretical skills are more foundational and broadly applicable (AAAS, 2011); therefore, this instrument focuses more on theoretical research skills.

The precourse survey also asked students "How many years have you worked in a research lab?," with answer choices for "no previous experience," "0.1–1.9 years," "2–2.9 years," "3–3.9 years," "4–4.9 years," "5–5.9 years," "6–6.9 years," and "7 or more years." When answering this question, students were instructed to consider themselves working in a lab for a given week if they worked more than 5 hours, on average, for that week. The final question on the postcourse survey asked

students "With which gender do you identify?" Though multiple options (including "not listed" and "prefer not to answer") were listed, all but three students selected either "male" or "female."

In 2018, as part of the postcourse assessment, students were asked, "In the past semester, which of the following experiences have contributed to your current level of experience practicing experimental design?" and "In the past semester, which of the following experiences have contributed to your current level of comfort in designing experiments?" For each question, students could select all that applied from a list of the following options: "attending research seminars," "completing course work," "discussing scientific topics with colleagues," "giving scientific presentations," "participating in laboratory research," "reading scientific literature," "receiving advice from mentors," and "writing a project proposal." Students were also given a choice for "other" with space to specify the experience, though no students used this option. From these options, students were also asked to select "Which experience was the most important in changing your level of experience practicing experimental design?" and "Which experience was the most important in changing your comfort in designing experiments?"

Pre- and Posttests of the BEDCI

The BEDCI consists of 14 multiple-choice questions based on three scenarios and is designed to test eight central concepts in biological experimental design at the undergraduate level (Deane *et al.*, 2014). Because the majority of our first-semester doctoral students have little to no additional experience beyond their undergraduate training, we believe that this instrument can provide valid results for first-semester doctoral students. Additionally, Deane and colleagues had graduate students with teaching experience act as experts when validating the BEDCI, further supporting the assertion that doctoral students understand and correctly interpret the questions on the instrument. While the BEDCI does not directly test students' procedural experimental design skills in the laboratory setting, it does act as a proxy in assessing these skills.

The BEDCI pre- and posttests were administered on the fourth and last days of class in Fall 2017, respectively, as per the instructions given by the developers for using the validated concept inventory. During each administration, the scenarios and questions were shown to the students on PowerPoint slides for the allocated lengths of time while students recorded their responses on answer sheets. For a small subset of students who were unable to attend the last day of class in person, the BEDCI was administered the day before or the following week using the same format and timing. Students were awarded a small number of participation points for completing both BEDCI tests. Due to the amount of in-class time required for this assessment, the BEDCI was not administered during Fall 2018.

Data Analysis

Data analyses are reported in the following five sections. A significance level of $p < 0.05$ was used in all analyses. All analyses and visualization were performed using R 3.5.1. Code for all analyses is available at https://github.com/harvard-cfp/research-skills.

Analyses of Incoming Students' Research, Self-Efficacy, and Performance. We initially sought to characterize the extent of previous research experience of our incoming doctoral student population and whether this was related to how capable students felt about performing essential research skills. On the precourse survey, we asked students to report the number of years that they had worked in a research lab and to rate the extent to which they felt that they could accomplish each of 14 research tasks (Kardash, 2000). To see whether total research skills self-efficacy varied with the length of time spent working in a lab, we separated student responses by the ranges of time that students had used to report their previous research experience.

Considering the scope of most undergraduate and postbaccalaureate research projects, many students at the beginning of their doctoral training may not have had experience with some research skills such as writing a research paper or relating results to larger concepts within the field. Therefore, we postulated that the amount of time spent working in a lab may be more predictive of the subset of skills required for experimental design than total research skills self-efficacy. Specifically, we hypothesized that time spent doing predoctoral research was predictive of self-reported experience and comfort with experimental design, as reported on the precourse survey, and experimental design aptitude, as measured by total score on the BEDCI at the beginning of the semester. To test our hypotheses, we used linear regression to calculate the adjusted R^2 values between each of these measures and the number of years students spent working in the laboratory environment before entering doctoral training. Weak predictors were considered to be those with an adjusted $R^2 \leq 0.7$.

Changes in Experimental Design Aptitude. Our next set of analyses examined changes in experimental design aptitude that occur over the duration of the first semester of doctoral training. To accomplish this, we administered the BEDCI to the same cohort of students at the end of the semester and linked each participant's pre- and postcourse responses using student-generated alphanumeric identifiers. We tested our hypothesis that experimental design aptitude changed over the course of the semester using a Wilcoxon signed-rank test, which determines whether the pre- and postcourse aptitude scores were likely drawn from the same distribution. As this test does not assume normality, we applied it to determine whether performance on the BEDCI substantially differed at the beginning and the end of the semester, leveraging the paired nature of the data by matching each student's responses at the beginning and end of the semester.

We also sought to assess student improvement on individual BEDCI questions to identify whether changes in overall performance were related to changes in student understanding of specific concepts within the umbrella of experimental design. Due to the paired nature of pre- and postassessments, we used McNemar's test to assess student improvement on individual BEDCI questions over the semester. We performed a multiple hypothesis testing correction with Benjamini-Hochberg type I error adjustment and report adjusted p values (Supplemental Table S3).

Changes in Research Skills Self-Efficacy. We also examined changes in research skills self-efficacy during the first semester of doctoral training, using student responses to the Kardash (2000) survey instrument included in the surveys at the beginning and end of the semester. Because we used the same

instrument at both time points, we were able to identify areas in which students showed the largest growth during their first semester of doctoral training. In analyzing the research skills self-efficacy data, the responses were not treated as linearly related, because the Likert-scale ratings were presented as categorical rather than numerical to the students. Therefore, we tested our hypothesis that there were changes in the distribution across the five Likert-scale categories over the semester using a Wilcoxon signed-rank test. We controlled for employing this test across the 14 questions using the Benjamini-Hochberg type I error adjustment for multiple hypotheses and report adjusted p values (Supplemental Table S2). We also leveraged the paired nature of the data in this analysis, as each student's responses from the beginning and end of the semester could be matched.

We examined self-efficacy changes at the individual level by equating the Likert-scale categories from "not at all" through "a great deal" with numerical scores of 1 through 5, respectively. We calculated changes by subtracting scores from the beginning of the semester from those at the end of the semester for individual self-efficacy items for each student (Supplemental Figure S3). Additionally, we calculated changes in net self-efficacy by summing changes of each student on all 14 of the self-efficacy items and determining how this quantity differed from the beginning to the end of the semester (Supplemental Figure S4).

To facilitate discussion of our research skills self-efficacy results, we used pairwise Spearman correlations to group research skills self-efficacy items that covaried with one another. Students' responses for each self-efficacy question were correlated with their responses for every other question in a pairwise manner. These correlation coefficients were used to group items that consistently had high pairwise correlations for each student's responses on the pre- and posttest, as well as net change in student self-efficacy (Supplemental Figure S2).

Comparison of Men's and Women's Research Skills Self-Efficacy. The next set of analyses compared the data for men and women in our cohort to look for potential gender-based differences. To test our hypothesis that gender was a strong factor in student self-efficacy in research skills, we used ordinal logistic regression, as this analysis takes advantage of the ordered nature of Likert-scale ratings. We regressed on each of the 14 items using the pre- and posttest ratings as the response variables, with gender as the predictive factor. To test whether there was a gender difference for total research skills self-efficacy, we used the Wilcoxon signed-rank test with a false discovery correction for multiple hypotheses to compare the distributions of men's and women's responses.

To check for confounding variables, we assessed differences between women and men in our study in terms of experimental design performance (Supplemental Figure S6) as well as background and prior experience (Supplemental Table S4). Specifically, we tested whether there were gender differences in experimental design performance by comparing the overall BEDCI scores among men and women using Student's t test. We also tested differences in proportions of race, previous degree subject, and program with a chi-squared statistical test. Differences in distributions of self-reported experience and comfort with experimental design were assessed using the Wilcoxon signed-rank test. To determine whether there was a significant difference in the number of years worked in a laboratory environment (as a categorical variable) between men and women, we performed ordinal logistic regression, using years of research experience as the response variable and gender as the predictive factor.

Factors Contributing to Experience and Comfort with Experimental Design. The final portion of this study looked at the aspects of training that students selected as the most important contributors to their experience and comfort with experimental design. To address this, we included questions on the 2018 postcourse survey that asked students to identify the factor that was most important in changing their level of experience or comfort with experimental design during the past semester. Considering the central importance that performing laboratory research is given in doctoral training in the biological sciences, we were specifically interested in whether students differentially attributed this activity to influencing their experimental design experience relative to their comfort with designing experiments. To test our hypothesis that there was a difference in the proportion of students who selected "participating in laboratory research" when asked about the main factor that contributed to changes in their experimental design experience versus comfort, we employed the chi-squared test.

RESULTS

Length of Prior Research Experience Is Not Predictive of Self-Efficacy in Research-Related Tasks

All students in our cohort reported that they had spent some time working at least 5 hours per week in a research lab before beginning their doctoral training. However, the students varied in the duration of their experience: 40% of the students had spent fewer than 3 years, 47% of the students had spent between 3 and 5 years, and 13% had spent more than 5 years (Figure 1A). The majority of students acquired this research experience while they were undergraduates, though approximately 25% of the students also performed research as postbaccalaureates, interns, or master's students.

In the survey administered at the beginning of the semester, students most frequently reported "a moderate amount" of self-efficacy for each of the 14 Likert-scale items of the research skills self-efficacy instrument (Figure 1B). However, the distribution of responses was skewed toward positive responses, with a greater number of responses of "a lot" or "a great deal" compared with those for "a little" or "not at all." When we separated student responses by the length of time that they had spent working in a lab, we found that students who had done more research before enrolling in a doctoral program did not rate their own abilities to perform research skills more highly than their peers (Figure 1B). This result suggests that our students do not equate the amount of time spent doing predoctoral lab work with their capacity to do research.

In the precourse survey, we also asked students to rate their levels of experience and comfort in a narrower area of research skills, namely experimental design. While there was variability in how the students rated their levels of experience and comfort with designing experiments, a majority of students reported medium to high levels of experience and at least some level of comfort in practicing experimental design (Figure 1C and D, respectively). We used linear regression analyses to determine whether there was a relationship between the amount of time spent doing research and either self-reported experience or
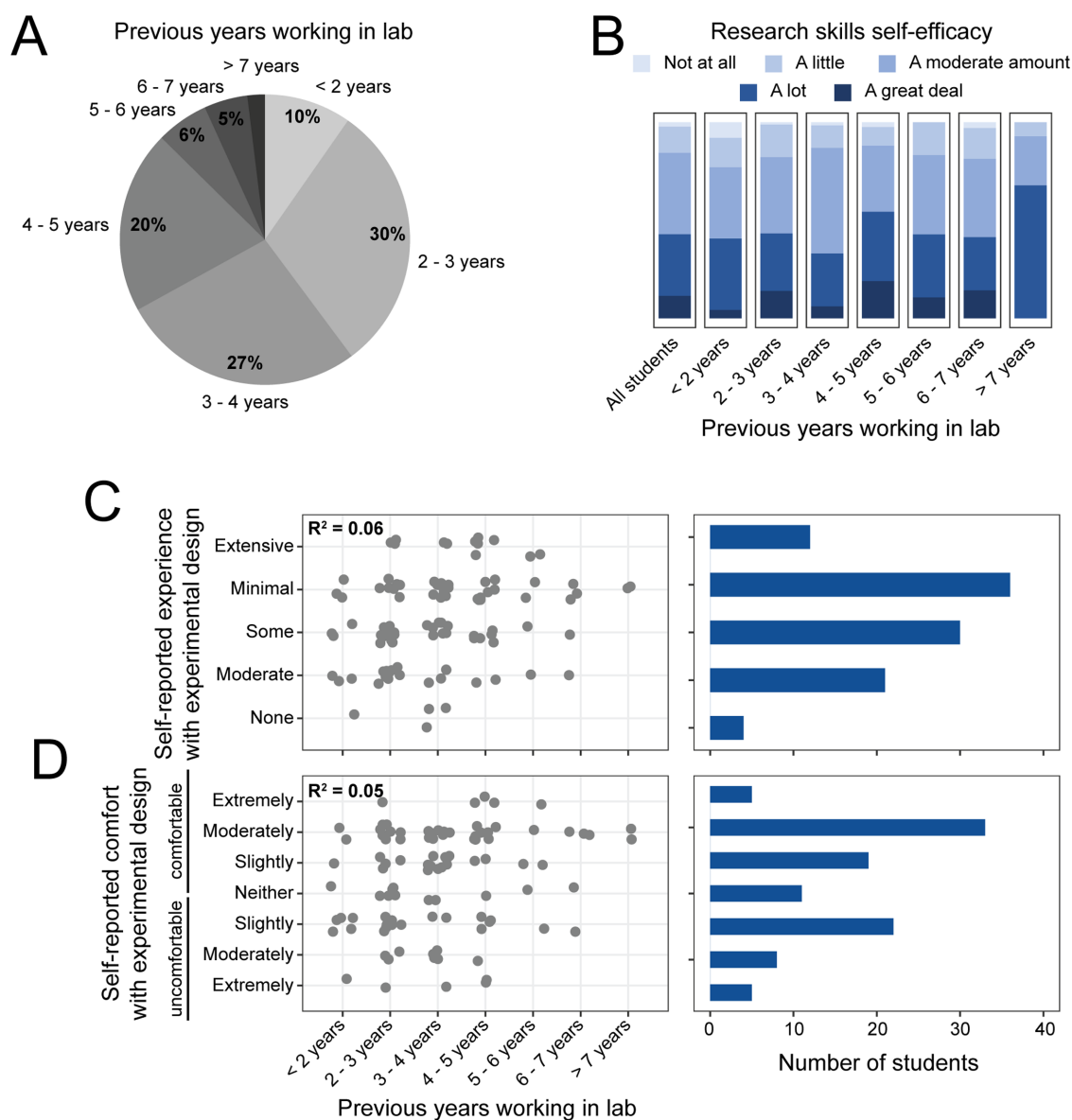
FIGURE 1. Length of time spent doing predoctoral research is not predictive of student research skills self-efficacy or self-reported experience and comfort with experimental design (*n* = 103 students). (A) Pie chart showing the length of time that students had spent working an average of more than 5 hours per week in a lab. All students reported having some previous research experience. (B) Aggregate distributions of student self-rating on a 14-item self-efficacy instrument included in a precourse survey. Responses divided by students' prior amount of time spent doing research show that this parameter is not predictive of how students rate themselves. (C, D) Histograms of the student responses to questions about self-reported experience (C) and comfort (D) with experimental design show that the student population varies, but there is a greater proportion of students reporting positive levels for both measures. Scatter plots show that the number of previous years spent working in a lab was not predictive of self-reported experience (C) and comfort (D) with experimental design (adjusted R^2 = 0.06 and 0.05, respectively). All responses were collected during surveys of first-year students during the first week of the Fall 2017 or Fall 2018 semester and thus are representative of the incoming levels of training, self-efficacy, experience, and comfort of our study population.

comfort with experimental design. Similar to what we observed with research skills self-efficacy, we found that both measures were not predicted by the number of years spent working in a lab, as indicated by adjusted R^2 values of 0.06 for experience and 0.05 for comfort (Figure 1C and D, respectively). Thus, our students do not feel that they are getting practice with designing experiments during the time that they spend doing predoctoral research.

Length of Prior Research Experience Is Not Predictive of Student Performance on an Experimental Design Concept Inventory

We assessed our incoming students' experimental design aptitude by administering the BEDCI at the beginning of the Fall 2017 semester. Though this assessment was designed for undergraduate students, we found that our students received a mean score of only 63.8% (Figure 2B). Fewer than 50% of students
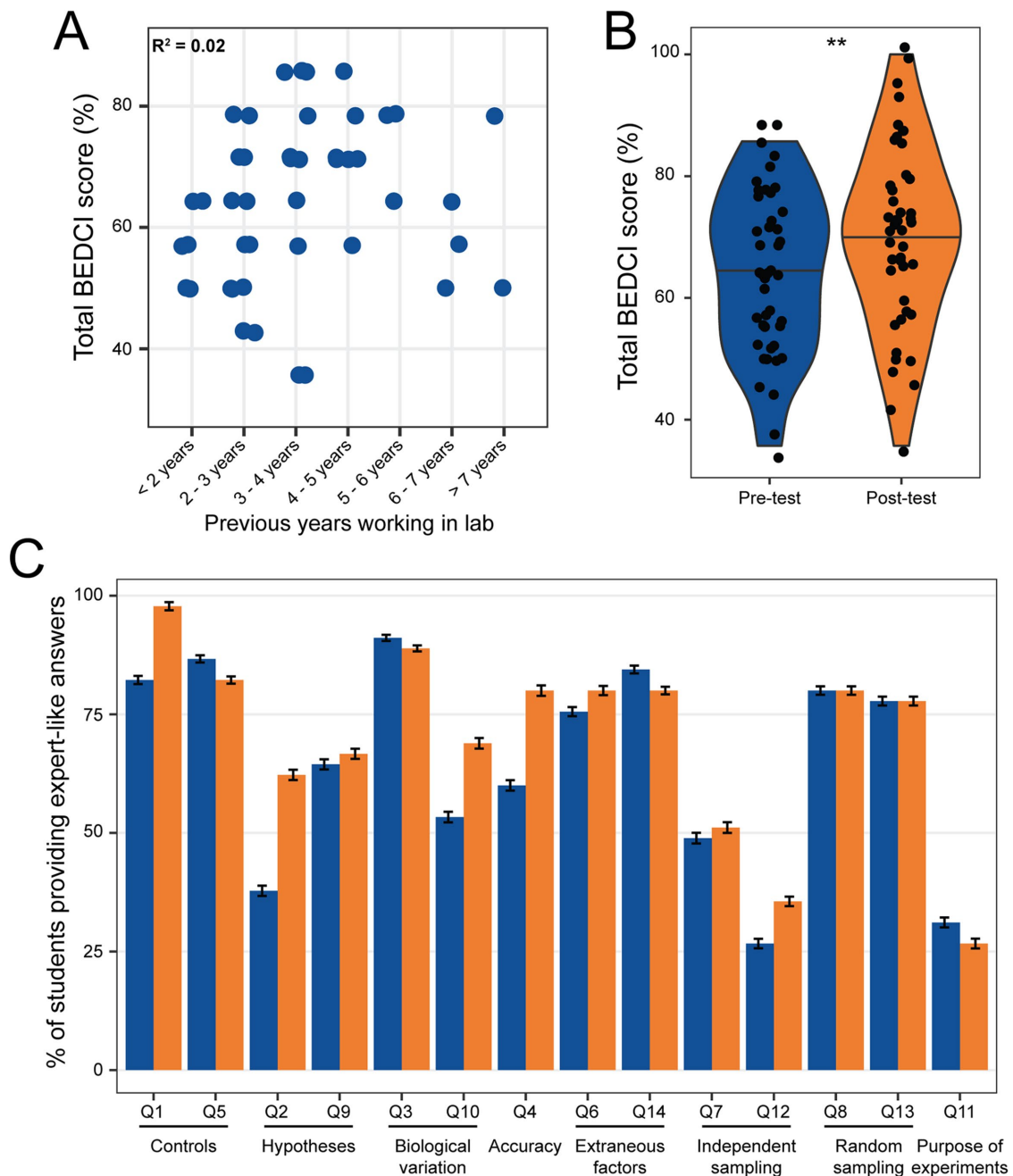
FIGURE 2. Students improved their performance of experimental design over their first semester of doctoral training ($n = 45$ students). (A) Scatter plot of total BEDCI score vs. length of time spent working in a lab shows that duration of previous lab experience is not predictive of BEDCI score (adjusted $R^2 = 0.02$). (B) Violin plots of student scores on the BEDCI pre- and posttests show a wide distribution of student scores on both administrations of the BEDCI. A 5 percentage point change in mean total score (central horizontal line on each violin) was observed on the posttest relative to the pretest. Significance was determined using a Wilcoxon signed-rank test for differences in students' pre- to postcourse scores (*$p < 0.05$). (C) Comparisons of the percentage of students providing expert-like answers on each of the BEDCI questions during the pre- and posttests. The questions are grouped into the eight concepts tested by the BEDCI. The student population trended toward improvement in the posttest in questions related to controls, hypotheses, biological variation, and accuracy. Significance was determined by McNemar's chi-squared test for differences in students' pre- to postcourse scores, with a false discovery correction for multiple hypotheses (1 df).

correctly answered one or more of the questions related to independent sampling, the purpose of experiments, and hypotheses (Figure 2C). However, for other concepts such as controls, extraneous factors, and random sampling, 75% or more of the students correctly answered both questions related to the concept. When considering the relationship between total BEDCI score and predoctoral research experience, linear regression analysis showed that the length of time that students had spent

working in a lab was not predictive of their performance on the BEDCI (adjusted $R^2 = 0.02$; Figure 2A). This reinforces our finding that performing a greater amount of predoctoral research does not necessarily give a student more training with essential research skills, such as experimental design.

Students' Experimental Design Aptitude Increases Modestly during the First Semester of Doctoral Training

Because the average student performance on the BEDCI at the beginning of the semester was relatively low, we sought to determine whether our students would improve on the assessment after one semester of doctoral training. When we administered the BEDCI at the end of the semester, we saw a modest, but statistically significant, increase in the mean from 63.8% to 68.9% (Figure 2B). Individual students varied in how their performance changed from the beginning to the end of the semester, with the net change in scores (posttest minus pretest) ranging from −29% to 29%. Within this range, 65% of students had a net positive change, 13% had no net change, and 22% had a net negative change (Supplemental Figure S1A). The subset of students who had a net negative change was heterogeneous and did not disproportionately represent women or students who had spent less time doing predoctoral research (Supplemental Figure S1B).

Analysis of the pre- and posttest responses to individual questions using McNemar's chi-squared test revealed that our students trended toward improvement on questions about controls, hypotheses, biological variation, and accuracy (Figure 2C). For the first three of these concept areas, we saw trends toward improvement in student performance in *only one of the two BEDCI questions about each concept* (Figure 2C). The question pairs differed in the experimental scenarios upon which they were based, the wording of the questions, and the misconceptions being tested. For instance, while questions 1 and 5 were both about controls, question 1 explicitly asked for the appropriate control treatment, while question 5 asked which addition to the experimental design would be most valuable in supporting a given conclusion. Thus, question 1 tested whether students could identify the control, while question 5 asked students to weigh the importance of adding a control treatment relative to other changes to the experimental design, such as increasing the sample size. Therefore, the differences in the level of improvement between pairs of questions on hypotheses, controls, and biological variation likely reflect changes in student understanding on some, but not all aspects, of these concepts.

Students' Research Skills Self-Efficacy Increases Significantly during the First Semester of Doctoral Training

To examine changes in students' self-efficacy in designing experiments and performing other aspects of research, we administered the research skills self-efficacy instrument in surveys at both the beginning and end of the semester. Initially, our students reported the lowest levels of self-efficacy for statistically analyzing data and writing a research paper and the highest levels for understanding the importance of controls and collecting data. Wilcoxon signed-rank tests comparing the pre- and posttest results revealed statistically significant increases for all of these items, except collecting data, with one of the largest increases for statistically analyzing data. While there was an

increase in the proportion of students who reported higher levels of self-efficacy on all of the remaining 10 items included in the instrument, Wilcoxon signed-rank tests revealed statistically significant increases for eight of the 10 items. When combining the way that students rated themselves across all 14 items, our student cohort showed a significant overall increase in their self-efficacy from the beginning to the end of the semester (Figure 3). At the individual level, 71% of the students showed a net increase, 5% showed no net change, and 24% showed a net decrease in self-efficacy from the beginning to the end of the semester when summing across all 14 items (Supplemental Figure S4A). Students who had net negative changes in self-efficacy did not significantly differ from the overall cohort in the duration of predoctoral research experience that they reported (Supplemental Figure S4B), indicating that early doctoral training does not disproportionately benefit those who enter graduate school with less research experience.

To determine whether student self-efficacy on different items was related, we performed pairwise Spearman correlations between the 14 items and grouped skills that covaried with one another. This analysis gave us four groups that we labeled field knowledge, experimental design, interpretation & iteration, and science communication. Each of these groups included two to four items for which individual students consistently reported similar self-efficacy values. The analysis also revealed that students' self-efficacy in the three items about controls, experimentation, and statistics did not consistently correlate with self-efficacy in any of the other skills addressed in the instrument (Supplemental Figure S2). Thus, our data show that students do not correlate their ability to understand the importance of controls or statistically analyze data with their ability to design an experiment or theoretical test of a hypothesis.

Grouping the items based on covariance showed that the two areas in which students showed the largest self-efficacy gains were experimental design and interpretation & iteration (Figure 3). Within these categories, the most significant growth was seen in items about the use of hypotheses in research, including how they are formulated, experimentally tested, and employed in data interpretation. Out of the three items that varied independently, students showed significant increases in self-efficacy in the items on controls and statistics. Other areas of significant growth in self-efficacy included conceptual knowledge of the field, oral communication, and thinking independently.

Men and Women Show No Significant Differences in Total Research Skills Self-Efficacy

We found that women were slightly overrepresented in the subpopulation of students that decreased in research skills self-efficacy (Supplemental Figure S4B). To further elucidate any gender-based differences in research skills self-efficacy, we performed ordinal logistic regression on each of the 14 self-efficacy items using the pre- and posttest ratings as the response variables and gender as the predictive factor. While women rated themselves significantly lower than men on a single item in the pretest in 2017 (Supplemental Figure S5), Wilcoxon signed-rank tests showed that there was no significant difference between the total research self-efficacy of women and men on either the pre- or posttest ($p = 0.3$ and 0.8 for the pre- and posttest, respectively; Figure 3C). Thus, the minor

A

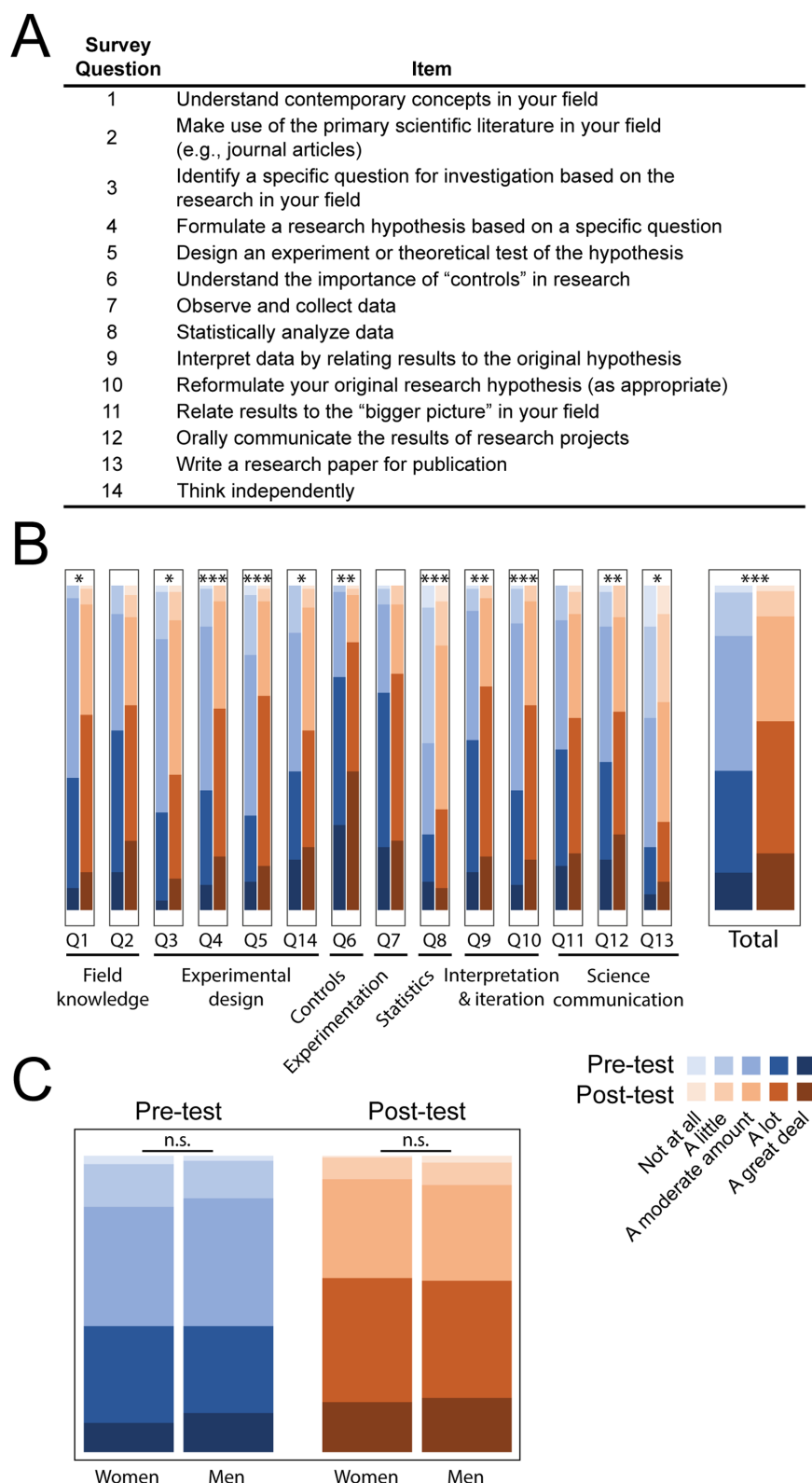| Survey Question	Item
1	Understand contemporary concepts in your field
2	Make use of the primary scientific literature in your field (e.g., journal articles)
3	Identify a specific question for investigation based on the research in your field
4	Formulate a research hypothesis based on a specific question
5	Design an experiment or theoretical test of the hypothesis
6	Understand the importance of "controls" in research
7	Observe and collect data
8	Statistically analyze data
9	Interpret data by relating results to the original hypothesis
10	Reformulate your original research hypothesis (as appropriate)
11	Relate results to the "bigger picture" in your field
12	Orally communicate the results of research projects
13	Write a research paper for publication
14	Think independently

FIGURE 3. Students significantly improved in many aspects of research skills self-efficacy during their first semester of doctoral training. (A) The 14 research skills self-efficacy items included on our pre- and postcourse surveys. For each item, students were asked to respond to the question "To what extent do you feel you can do each of the following?," with prompts ranging from "not at all" to "a great deal." (B) Histogram of student responses to the 14 self-efficacy items comparing the pre-and posttest levels ($n = 103$ students). Students show significant improvements in self-efficacy in 11 of the 14 items. The items have been grouped into categories that covaried with one another or varied independently, as measured by Spearman correlations between individual items in the pretest, posttest, and changes between them. The clusters with multiple items that varied together were designated field knowledge, experimental design, interpretation & iteration, and science communication. Significance was determined using a Wilcoxon signed-rank test for differences in students' pre- to post-course scores, with a false discovery correction for multiple hypotheses (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$, two-tailed). (C) Total pre- and posttest self-efficacy distributions separated by gender show that there is no significant difference between men and women in either the pre- or posttest ($p = 0.3$ and 0.8, respectively). Lack of significance was determined using a Wilcoxon signed-rank test for differences in the distribution of men's and women's responses, with a false discovery correction for multiple hypotheses.

self-efficacy differences that we observed between men and women are not present when comparing the total research self-efficacy of the two subpopulations.

We did not find any other major differences between men and women in our population that would contribute to differences in self-efficacy between the two groups. Specifically, we found that both groups demonstrated similar levels of experimental design performance, as measured by total BEDCI scores (Supplemental Figure S6). The only difference that we found in the background and academic preparation of the two groups was that women had spent significantly more time performing research than men (Supplemental Table S4); however, as we have shown above, time spent doing predoctoral research is not predictive of self-efficacy in research-related tasks.

Factors Contributing to Self-Reported Experience and Comfort with Designing Experiments

The survey at the end of the Fall 2018 semester showed that students vary from one another in the modes of training and feedback that they find most important in changing their experience and comfort with experimental design (Figure 4).
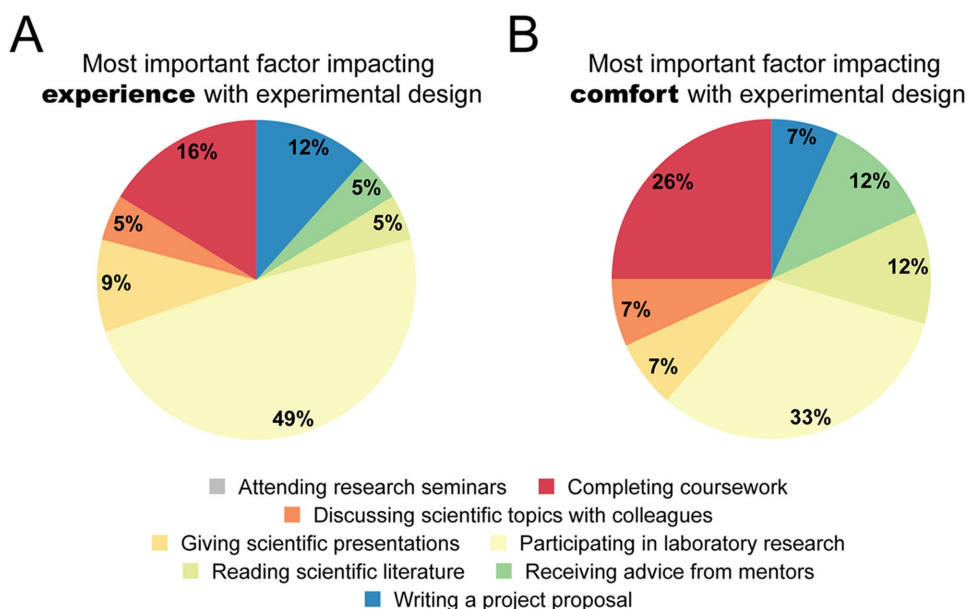
FIGURE 4. Students with a net increase in research skills self-efficacy select different factors as the most important for contributing to their experience or comfort with experimental design (*n* = 43 students). Pie charts showing what students cited as the primary factor contributing to their experience (A) and comfort (B) with experimental design. There were significant differences (*p* = 0.03) by chi-squared test between the proportion of students who cited participating in laboratory research for the two questions (49% for experience and 33% for comfort). Other notable changes include the increase in students citing completing course work and receiving advice from mentors for comfort (26% and 12%, respectively) in comparison to experience (16% and 5%, respectively).

When we compared individual student responses to the questions about experience and comfort, we found that 64% of the students cited different factors for each of the two questions. Notably, there was a significant difference (*p* = 0.03) in the proportion of students who selected "participating in laboratory research" as the primary contributor for each question, with 49% of students selecting it for experience and only 33% citing it for comfort. Additionally, many aspects of doctoral training were more frequently cited for comfort relative to experience; these included completing course work, receiving advice from mentors, discussing scientific topics with colleagues, and reading scientific literature. This indicates that these doctoral training activities have value in facilitating growth in designing experiments by making students feel more comfortable in performing this skill.

DISCUSSION

Multiple reports and publications have highlighted the importance of training in research skills, and experimental design in particular, throughout science education (National Research Council, 2009; Coil *et al.*, 2010; AAAS, 2011). The literature also indicates that increasing objective measures of student capability must be coupled with bolstering levels of students' self-efficacy for students to persist through challenges in applying these skills (Trujillo and Tanner, 2014). This study addresses both objective assessment and self-perception of research skills ability, with a particular focus on experimental design, during the critical period early in doctoral training. Our results show that the majority, but not all, of our students improve in their experimental design aptitude and research skills self-efficacy, with some of the most significant increases in self-efficacy relating to experimental design skills. Notably, our study also

provides insight into factors that students identify as important in contributing to self-efficacy in experimental design, which can help inform resource allocation to different training experiences for first year bioscience doctoral students. The combination of assessments of both performance and self-efficacy while identifying underlying factors provides a novel contribution to the limited body of literature on research skills training and assessment for doctoral-level bioscience students.

It is important to note unavoidable limitations in the study design that influence the interpretation of these results. The surveys and assessments for our study were administered in the Principles of Molecular Biology course, which is required for the institution's largest bioscience doctoral program and highly recommended for several other life science programs. As a result, the program composition of the students in our study population is not representative of the full body of bioscience doctoral students at the institution, and our sample size is limited based on enrollment in the course. Additionally, student responses on surveys and assessments may have been inadvertently influenced by its administration in the course, even though the questions contained no specific references to the course.

Furthermore, the incorporation of experimental design–related instruction and assignments in the course potentially influences the interpretation of our findings. Specifically, increases in both performance and self-efficacy related to controls and hypotheses may be influenced, at least in part, by the training provided on these topics during the course. However, we cannot test this in the absence of a matched control group of students enrolled in the same doctoral programs, but not in the course. Additionally, the proportion of students citing completing course work as the primary contributor to aspects of experimental design self-efficacy may have been inflated by the course's

explicit training in this skill. Nevertheless, this result indicates that many doctoral students value the integration of experimental design training in a content-based bioscience course and that course work can serve as an important avenue for delivering this type of skills training (Gutlerner and Van Vactor, 2013; Glass, 2014; Heustis *et al.*, 2019).

To our knowledge, this study is the first use of the BEDCI to determine learning gains in experimental design within a doctoral student population. Our students did not perform as well as we had expected on this assessment, despite the fact that it was designed for undergraduates and our participants were doctoral students who reported having predoctoral research experience. This indicates that many of our students did not learn certain critical aspects of experimental design through their predoctoral training or those that learned these concepts are unable to apply them outside the specific context in which they learned them (i.e., a narrow field of research in which they had worked). Despite performing more poorly than expected, our students scored higher, on average, on the BEDCI than the first- and third-year undergraduate students included in the foundational studies validating the assessment (Deane *et al.*, 2014). Changes in performance between the pre- and posttest administrations of the BEDCI allowed us to identify experimental design concepts on which our students trended toward improvement, but we feel that more sensitive instruments are necessary to assess the development of experimental design understanding among our doctoral students. To further pursue this and circumvent concerns that concept inventories only reveal limited aspects of students' conceptual understanding (Smith and Tanner, 2010), we are currently in the process of developing and validating tools to score student responses to open-ended experimental design questions based on our course content. Future studies may also gain more nuanced views of student growth in experimental design aptitude by using rubrics developed to score graduate students' written research proposals (e.g., Timmerman *et al.*, 2011) or oral presentations of experimental designs (e.g., Heustis *et al.*, 2019).

Similar to what we saw with experimental design performance, our incoming doctoral students showed room for improvement in research skills self-efficacy and their ratings did not correspond to the length of time that they had spent doing predoctoral research. These results suggest that time spent doing practical laboratory tasks may not make students feel more capable of performing cognitive aspects of research such as designing experiments. The lack of a relationship between duration of predoctoral research and self-efficacy was particularly striking, as previous work has shown that involvement in research activities is a significant predictor of research self-efficacy (Bieschke *et al.*, 1996) and undergraduate students with multiple years of research experience rate their ability to perform scientific skills more highly than less experienced counterparts (Thiry *et al.*, 2012). However, we are aware that undergraduate research experiences vary greatly in the expectations of the students involved and the quality of mentoring provided, so it is difficult to capture the impact of these experiences solely by their duration (Linn *et al.*, 2015). Thus, research skills self-efficacy, as well as experience and comfort with experimental design, may be more strongly correlated with other aspects of predoctoral research experiences such as supportive mentoring, intellectual agency, and professional development resources,

which the literature has shown to be important for student skill development (Hunter *et al.*, 2007; Johnson *et al.*, 2015; NASEM, 2017).

Our data show that total research skills self-efficacy increased in our student population during the first semester of doctoral training, with improvements in 11 of the 14 items in the self-efficacy instrument. Importantly, we did not observe significant differences in total research skills self-efficacy among men and women in our study population at either the beginning or end of the first semester of doctoral training. This suggests that factors that adversely impact women during doctoral training in STEM fields such as gender-based marginalization, few positive role models, disciplinary stereotypes, and inequities in academic recognition (MacLachlan, 2006; Hill *et al.*, 2010; De Welde and Laursen, 2011; Feldon *et al.*, 2017a) do not have measurable impacts on the research skills self-efficacy of women relative to men at early stages of doctoral training. However, further research is necessary to determine whether women and men maintain comparable levels of research skills self-efficacy more longitudinally throughout their doctoral training.

The areas of formulating, testing, and reformulating hypotheses were the areas of largest growth in self-efficacy for our doctoral students, while previous research has shown that these were among the skills for which undergraduates reported the lowest levels of self-efficacy (Kardash, 2000). This indicates that early graduate training can assist students in acquiring self-efficacy in these higher-order research skills that they may not have employed during their undergraduate research experiences. The observed self-efficacy changes in skills related to hypotheses and the importance of controls were particularly notable in that they aligned with tasks that our students are asked to perform on assessments throughout our course, such as justifying their selection of experimental approaches to test given hypotheses, articulating why specific controls are necessary, and interpreting the anticipated results in relation to the hypotheses. Additionally, both hypotheses and controls were concepts for which students trended toward improvement in their performance on the BEDCI, suggesting that the improvements in self-efficacy on the uses of hypotheses and controls in research are reflected in changes in student performance on questions relating to these concepts. Other areas of observed growth in self-efficacy, such as conceptual knowledge of the field and oral communication, reflect other goals and teaching strategies of the course, such as increasing knowledge of the field of molecular biology and requiring students to give oral presentations. We also saw a large self-efficacy increase for the item on statistical analysis of data. This may be a reflection of the fact that the majority of students enrolled in our course were required to take a concurrent course that explicitly emphasized quantitative skills training.

Approximately 25% of our students had net negative changes in their research skills self-efficacy over the course of their first semester of doctoral training. One potential explanation is that a heightened understanding of the complexity of research skills led some students to rate their abilities lower at the end of the semester. After entering a doctoral program, some students may have had their research skills more closely scrutinized and held up to a higher level of rigor than they had in their previous experiences. For instance, more detailed feedback may have led students to recognize caveats and potential

pitfalls in their proposed experimental designs, which in turn may have decreased their self-efficacy in designing an experiment to test a hypothesis. This explanation is consistent with literature demonstrating that undergoing training leads novices to rate their performance lower due to heightened metacognitive awareness (Kruger and Dunning, 1999). Furthermore, it posits that a decrease in self-efficacy may be an indication of a more accurately calibrated understanding of one's abilities relative to objective measures of performance.

Alternatively, reductions in self-efficacy may be a by-product of the presence of high-achieving classmates or challenges during the first semester of graduate training; both of these possibilities fit within the framework of the imposter phenomenon (Parkman, 2016). Such reductions in self-efficacy can be mitigated by teaching and mentoring practices such as providing opportunities for active, hands-on learning; giving prompt and encouraging feedback; establishing explicit guidelines of mutual respect and support; and encouraging students to value one another's contributions (Colbeck *et al.*, 2001; Trujillo and Tanner, 2014; Dewsbury and Brame, 2019). These practices, which promote multiple sources of self-efficacy, can benefit all students regardless of whether they display measurable decreases in self-efficacy. In our sample, women were slightly overrepresented among the students whose net research skills self-efficacy decreased, indicating that they may disproportionately benefit from the implementation of the aforementioned practices during the early stages of doctoral training.

Our students varied in the factors that they cited as the most important in contributing to their experience or comfort with experimental design, which underscores the diverse experiences that contribute to self-efficacy in this composite research skill. Participating in laboratory research was prominently represented in these answers despite our finding that predoctoral research does not contribute to experimental design training. We attribute this apparent contradiction to differences between predoctoral and doctoral research experiences (Delamont and Atkinson, 2001): Predoctoral students are often limited to executing experiments that have already been designed, so actively participating in the design of well-controlled experiments is unique to doctoral-level research for many students (Feldon *et al.*, 2017b). However, our results indicate that slightly more than half of our students cited factors other than participating in laboratory research as the primary contributor to their experience in experimental design, and an even greater fraction cited factors other than research as the primary contributor to their comfort with this skill. This underscores the need to supplement mentor-dependent training in laboratories for the development of research skills self-efficacy among graduate trainees. Laboratory training may not be the most effective for all students and can vary significantly between labs, so graduate programs should invest in providing centralized skill development opportunities that can benefit all students. These include embedding skills training in required course work and providing opportunities to engage in structured learning experiences such as reading scientific literature, giving presentations, or writing proposals.

The variety of factors that students cite as contributing to their experience and comfort with experimental design are consistent with the four sources of self-efficacy described by Bandura's framework (Bandura, 1977). Bandura theorized that personal mastery experiences were the most important; we also find that these experiences, including the participation in research and the completion of research-related tasks such as doing course assignments, writing proposals, and giving scientific presentations, are the strongest contributors to improving experience and comfort in experimental design. Additionally, our results emphasize the impact of social persuasion, vicarious experiences, and emotional states as contributors, as these are the means by which mentoring, reading primary literature, and participating in scientific discussion likely influence student experience and comfort with experimental design. The aspects of feedback and peer interaction embedded in the practice of experimental design in our course may also contribute to these sources of self-efficacy. The importance of a variety of factors indicates that no single modality meets the needs of all students and that educators should use course work, mentoring, peer interactions, and scientific engagement, in addition to laboratory experiences, to positively influence the research skills self-efficacy of doctoral students.

REFERENCES

American Association for the Advancement of Science. (2011). *Vision and change in undergraduate biology education: A call to action* (Final report). Washington, DC. Retrieved May 10, 2019, from http://visionandchange.org/finalreport

Anderson, W. A., Banerjee, U., Drennan, C. L., Elgin, S. C. R., Epstein, I. R., Handelsman, J., ... & Warner, I. M. (2011). Changing the culture of science education at research universities. *Science, 331*(6014), 152–153. https://doi.org/10.1126/science.1198280

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Barnes, B. J., & Austin, A. E. (2009). The role of doctoral advisors: A look at advising from the advisor's perspective. *Innovative Higher Education, 33*(5), 297–315. https://doi.org/10.1007/s10755-008-9084-x

Bianchini, J. A., Whitney, D. J., Breton, T. D., & Hilton-Brown, B. A. (2002). Toward inclusive science education: University scientists' views of students, instructional practices, and the nature of science. *Science Education, 86*(1), 42–78. https://doi.org/10.1002/sce.1043

Bieschke, K. J., Bishop, R. M., & Garcia, V. L. (1996). The utility of the research self-efficacy scale. *Journal of Career Assessment, 4*(1), 59–75. https://doi.org/10.1177/106907279600400104

Brown, S. D., Lent, R. W., Ryan, N. E., & McPartland, E. B. (1996). Self-efficacy as an intervening mechanism between research training environments and scholarly productivity: A theoretical and methodological extension. *Counseling Psychologist, 24*(3), 535–544. https://doi.org/10.1177/0011000096243012

Brownell, S. E., Wenderoth, M. P., Theobald, R., Okoroafor, N., Koval, M., Freeman, S., ... & Crowe, A. J. (2014). How students think about experimental design: Novel conceptions revealed by in-class activities. *BioScience, 64*(2), 125–137. https://doi.org/10.1093/biosci/bit016

Coil, D., Wenderoth, M. P., Cunningham, M., & Dirks, C. (2010). Teaching the process of science: Faculty perceptions and an effective methodology. *CBE—Life Sciences Education*, *9*(4), 524–535. https://doi.org/10.1187/cbe.10-01-0005

Colbeck, C. L., Cabrera, A. F., & Terenzini, P. T. (2001). Learning professional confidence: Linking teaching practices, students' self-perceptions, and gender. *Review of Higher Education*, *24*(2), 173–191. https://doi.org/10.1353/rhe.2000.0028

D'Costa, A. R., & Schlueter, M. A. (2013). Scaffolded instruction improves student understanding of the scientific method & experimental design. *American Biology Teacher*, *75*(1), 18–28. https://doi.org/10.1525/abt.2013.75.1.6

Dasgupta, A. P., Anderson, T. R., & Pelaez, N. (2014). Development and validation of a rubric for diagnosing students' experimental design knowledge and difficulties. *CBE—Life Sciences Education*, *13*(2), 265–284. https://doi.org/10.1187/cbe.13-09-0192

Deane, T., Nomme, K., Jeffery, E., Pollock, C., & Birol, G. (2014). Development of the Biological Experimental Design Concept Inventory (BEDCI). *CBE—Life Sciences Education*, *13*(3), 540–551. https://doi.org/10.1187/cbe.13-11-0218

Delamont, S., & Atkinson, P. (2001). Doctoring uncertainty: Mastering craft knowledge. *Social Studies of Science*, *31*(1), 87–107. https://doi.org/10.1177/030631201031001005

De Welde, K., & Laursen, S. (2011). The glass obstacle course: Informal and formal barriers for women Ph.D. students in STEM fields. *International Journal of Gender, Science, and Technology*, *3*(3), 571–595. Retrieved December 15, 2019, from http://genderandset.open.ac.uk/index.php/genderandset/article/view/205

Dewsbury, B., & Brame, C. J. (2019). Inclusive teaching. *CBE—Life Sciences Education*, *18*(2), fe2. https://doi.org/10.1187/cbe.19-01-0021

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*(3), 83–87. https://doi.org/10.1111/1467-8721.01235

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research*, *59*(4), 395–430. https://doi.org/10.3102/00346543059004395

Feldon, D. F. (2016). The development of expertise in scientific research. In Scott, R. A., Buchmann, M. C., & Kosslyn, S. M. (Eds.), *Emerging trends in the social and behavioral sciences* (pp. 1–14). New York: Wiley. https://doi.org/10.1002/9781118900772.etrds0411

Feldon, D. F., Maher, M. A., & Timmerman, B. E. (2010). Performance-based data in the study of STEM Ph.D. education. *Science*, *329*(5989), 282–283. https://doi.org/10.1126/science.1191269

Feldon, D. F., Peugh, J., Maher, M. A., Roksa, J., & Tofel-Grehl, C. (2017a). Time-to-credit gender inequities of first-year PhD students in the biological sciences. *CBE—Life Sciences Education*, *16*(1), ar4. https://doi.org/10.1187/cbe.16-08-0237

Feldon, D. F., Peugh, J., Timmerman, B. E., Maher, M. A., Hurst, M., Strickland, D., … & Stiegelmeyer, C. (2011). Graduate students' teaching experiences improve their methodological research skills. *Science*, *333*(6045), 1037–1039. https://doi.org/10.1126/science.1204109

Feldon, D. F., Rates, C., & Sun, C. (2017b). Doctoral conceptual thresholds in cellular and molecular biology. *International Journal of Science Education*, *39*(18), 2574–2593. https://doi.org/10.1080/09500693.2017.1395493

Fry, D. J. (2014). Teaching experimental design. *Institute for Laboratory Animal Research Journal*, *55*(3), 457–471. https://doi.org/10.1093/ilar/ilu031

Gilbert, R., Balatti, J., Turner, P., & Whitehouse, H. (2004). The generic skills debate in research higher degrees. *Higher Education Research & Development*, *23*(3), 375–388. https://doi.org/10.1080/0729436042000235454

Gilmore, J., Vieyra, M., Timmerman, B., Feldon, D., & Maher, M. (2015). The relationship between undergraduate participation and subsequent research performance of early career STEM graduate students. *Journal of Higher Education*, *86*(6), 834–863. https://doi.org/10.1080/00221546.2015.11777386

Glass, D. J. (2014). *Experimental design for biologists* (2nd ed.). New York, NY: Cold Spring Harbor Laboratory Press.

Gökçek, T., Taşkin, D., & Yildiz, C. (2014). Investigation of graduate students' academic self-efficacy beliefs. *Procedia—Social and Behavioral Sciences*, *141*, 1134–1139. https://doi.org/10.1016/j.sbspro.2014.05.191

Golde, C. M. (1998). Beginning graduate school: Explaining first-year doctoral attrition. *New Directions for Higher Education*, *1998*(101), 55–64. https://doi.org/10.1002/he.10105

Gutlerner, J. L., & Van Vactor, D. (2013). Catalyzing curriculum evolution in graduate science education. *Cell*, *153*(4), 731–736. https://doi.org/10.1016/j.cell.2013.04.027

Hackett, G., & Betz, N. E. (1981). A self-efficacy approach to the career development of women. *Journal of Vocational Behavior*, *18*(3), 326–339. https://doi.org/10.1016/0001-8791(81)90019-1

Heustis, R. J., Venkatesh, M. J., Gutlerner, J. L., & Loparo, J. J. (2019). Embedding academic and professional skills training with experimental-design chalk talks. *Nature Biotechnology*, *37*, 1523–1527. https://doi.org/10.1038/s41587-019-0338-1

Hiebert, S. M. (2007). Teaching simple experimental design to undergraduates: Do your students understand the basics? *Advances in Physiology Education*, *31*(1), 82–92. https://doi.org/10.1152/advan.00033.2006

Hill, C., Corbett, C., & St. Rose, A. (2010). *Why so few? Women in science, technology, engineering, and mathematics*. Washington, DC: American Association of University Women. Retrieved December 15, 2019, from www.aauw.org/files/2013/02/Why-So-Few-Women-in-Science-Technology-Engineering-and-Mathematics.pdf

Hunter, A-B., Laursen, S. L., & Seymour, E. (2007). Becoming a scientist: The role of undergraduate research in students' cognitive, personal, and professional development. *Science Education*, *91*, 36–74. https://doi.org/10.1002/sce.20173

Johnson, W. B., Behling, L. L., Miller, P., & Vandermaas-Peeler, M. (2015). Undergraduate research mentoring: Obstacles and opportunities. *Mentoring & Tutoring: Partnership in Learning*, *23*(5), 441–453. https://doi.org/10.1080/13611267.2015.1126167

Kardash, C. M. (2000). Evaluation of undergraduate research experience: Perceptions of undergraduate interns and their faculty mentors. *Journal of Educational Psychology*, *92*(1), 191–201. https://doi.org/10.1037/0022-0663.92.1.191

Killpack, T. L., & Fulmer, S. M. (2018). Development of a tool to assess interrelated experimental design in introductory biology. *Journal of Microbiology & Biology Education*, *19*(3). 19.3.98. https://doi.org/10.1128%2Fjmbe.v19i3.1627

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Lambie, G. W., Hayes, B. G., Griffith, C., Limberg, D., & Mullen, P. R. (2014). An exploratory investigation of the research self-efficacy, interest in research, and research knowledge of Ph.D. in education students. *Innovative Higher Education*, *39*(2), 139–153. https://doi.org/10.1007/s10755-013-9264-1

Lin, X., & Lehman, J. D. (1999). Supporting learning of variable control in a computer-based biology environment: Effects of prompting college students to reflect on their own thinking. *Journal of Research in Science Teaching*, *36*(7), 837–858. https://doi.org/10.1002/(SICI)1098-2736(199909)36:7%3C837::AID-TEA6%3E3.0.CO;2-U

Linn, M. C., Palmer, E., Baranger, A., Gerard, E., & Stone, E. (2015). Undergraduate research experiences: Impacts and opportunities. *Science*, *347*(6222), 1261757. https://doi.org/10.1126/science.1261757

Lovitts, B. (2001). *Leaving the ivory tower: The causes and consequences of departure from doctoral study*. Lanham, MD: Rowman & Littlefield.

MacLachlan, A. J. (2006). The graduate experience of women in STEM and how it could be improved. In Bystydzienski, J. M., & Bird, S. R. (Eds.), *Removing barriers: Women in academic science, technology, engineering, and mathematics* (pp. 237–254). Bloomington: Indiana University Press.

National Academies of Sciences, Engineering, and Medicine (NASEM). (2017). *Undergraduate research experiences for STEM students: Successes, challenges, and opportunities*. Washington, DC: National Academies Press. https://doi.org/10.17226/24622

NASEM. (2018). *Graduate STEM education for the 21st century*. Washington, DC: National Academies Press. https://doi.org/10.17226/25038

National Institute of General Medical Sciences. (2017, updated 2019). *National Institute of General Medical Sciences Ruth L. Kirschstein*

National Research Service Award (NRSA) predoctoral institutional research training grant (T32) guidelines. Retrieved May 10, 2019, from https://grants.nih.gov/grants/guide/pa-files/par-17-341.html

National Research Council. (2009). *A new biology for the 21st century*. Washington, DC: National Academies Press. https://doi.org/10.17226/12764

Paglis, L. L., Green, S. G., & Bauer, T. N. (2006). Does advisor mentoring add value? A longitudinal study of mentoring and doctoral outcomes. *Research in Higher Education*, *47*(4), 451–476. https://doi.org/10.1007/s11162-005-9003-2

Parkman, A. (2016). The imposter phenomenon in higher education: Incidence and impact. *Journal of Higher Education Theory and Practice*, *16*(1), 51–60. Retrieved May 11, 2019, from www.na-businesspress.com/JHETP/ParkmanA_Web16_1_pdf

Pollack, A. E. (2010). Exploring the complexities of experimental design: Using an on-line reaction time program as a teaching tool for diverse student populations. *Journal of Undergraduate Neuroscience Education*, *9*(1), A47–A50. Retrieved May 11, 2019, from www.ncbi.nlm.nih.gov/pmc/articles/PMC3597424

Shi, J., Power, J. M., & Klymkowsky, M. W. (2011). Revealing student thinking about experimental design and the roles of control experiments. *International Journal for the Scholarship of Teaching and Learning*, *5*(2), ar8. https://doi.org/10.20429/ijsotl.2011.050208

Sirum, K., & Humburg, J. (2011). The Experimental Design Ability Test (EDAT). *Bioscene: Journal of College Biology Teaching*, *37*(1), 8–16. Retrieved May 11, 2019, from https://files.eric.ed.gov/fulltext/EJ943887.pdf

Smith, J. I., & Tanner, K. (2010). The problem revealing how students think: Concept inventories and beyond. *CBE—Life Sciences Education*, *9*(1), 1–5. https://doi.org/10.1187/cbe.09-12-0094

Szymanski, D. M., Ozegovic, J. J., Phillips, J. C., & Briggs-Phillips, M. (2007). Fostering scholarly productivity through academic and internship research training environments. *Training and Education in Professional Psychology*, *1*(2), 135–146. https://doi.org/10.1037/1931-3918.1.2.135

Thakore, B. K., Naffziger-Hirsch, M. E., Richardson, J. L., Williams, S. N., & McGee, R. Jr. (2014). The academy for future science faculty: Randomized controlled trial of theory-driven coaching to shape development and diversity of early-career scientists. *BMC Medical Education*, *14*, 160. https://doi.org/10.1186/1472-6920-14-160

Thiry, H., Weston, T. J., Laursen, S. L., & Hunter, A. B. (2012). The benefits of multi-year research experiences: Differences in novice and experienced students' reported gains from undergraduate research. *CBE—Life Sciences Education*, *11*(3), 260–272. https://doi.org/10.1187/cbe.11-11-0098

Timmerman, B. C., Feldon, D., Maher, M., Strickland, D., & Gilmore, J. (2013). Performance-based assessment of graduate student research skills: Timing, trajectory, and potential thresholds. *Studies in Higher Education*, *38*(5), 693–710. https://doi.org/10.1080/03075079.2011.590971

Timmerman, B. E. C., Strickland, D. C., Johnson, R. L., & Payne, J. R. (2011). Development of a "universal" rubric for assessing undergraduates' scientific reasoning skills using scientific writing. *Assessment & Evaluation in HigherEducation*,*36*(5),509–547.https://doi.org/10.1080/02602930903540991

Trujillo, G., & Tanner, K. D. (2014). Considering the role of affect in learning: Monitoring students' self-efficacy, sense of belonging, and science identity. *CBE—Life Sciences Education*, *13*(1), 6–15. https://doi.org/10.1187/cbe.13-12-0241

University of British Columbia. (2014a, September 24). Experimental design (third/fourth year undergraduate level). In *Q4B Concept Inventories*. Vancouver, BC, Canada. Retrieved May 10, 2019, from http://q4b.biology.ubc.ca/concept-inventories/experimental-design-thirdfourth-year-undergraduate-level

University of British Columbia. (2014b, December 12). Experimental design (first year undergraduate level). In *Q4B Concept Inventories*. Vancouver, BC, Canada. Retrieved May 10, 2019, from http://q4b.biology.ubc.ca/concept-inventories/experimental-design-first-year-undergraduate-level

Vale, R. D., DeRisi, J., Phillips, R., Mullins, R. D., Waterman, C., & Mitchison, T. J. (2012). Interdisciplinary graduate training in teaching labs. *Science*, *338*(6114), 1542–1543. https://doi.org/10.1126/science.1216570

Walker, G. E., Golde, C. M., Jones, L., Bueschel, A. C., & Hutchings, P. (2008). *The formation of scholars: Rethinking doctoral education for the twenty-first century*. San Francisco, CA: Jossey-Bass.

Weidman, J. C. (2010). Doctoral student socialization for research. In Gardner, S. K., & Mendoza, P. (Eds.), *On becoming a scholar: Socialization and development in doctoral education* (pp. 45–55). Sterling, VA: Stylus.

Zolman, J. F. (1999). Teaching experimental design to biologists. *Advances in Physiology Education*, *227*(6), S111–S118. https://doi.org/10.1152/advances.1999.277.6.S111