Null effects of boot camps and short-format training for PhD students in life sciences

David F. Feldon[a,1], Soojeong Jeong[a], James Peugh[b], Josipa Roksa[c,d], Cathy Maahs-Fladung[a], Alok Shenoy[a], and Michael Oliva[a]

[a]Department of Instructional Technology & Learning Sciences, Utah State University, Logan, UT 84322-2830; [b]Department of Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229-3026; [c]Department of Sociology, University of Virginia, Charlottesville, VA 22904; and [d]Curry School of Education, University of Virginia, Charlottesville, VA 22904

Edited by Dale Purves, Duke University, Durham, NC, and approved July 28, 2017 (received for review April 6, 2017)

Many PhD programs incorporate boot camps and summer bridge programs to accelerate the development of doctoral students' research skills and acculturation into their respective disciplines. These brief, high-intensity experiences span no more than several weeks and are typically designed to expose graduate students to data analysis techniques, to develop scientific writing skills, and to better embed incoming students into the scholarly community. However, there is no previous study that directly measures the outcomes of PhD students who participate in such programs and compares them to the outcomes of students who did not participate. Likewise, no previous study has used a longitudinal design to assess these outcomes over time. Here we show that participation in such programs is not associated with detectable benefits related to skill development, socialization into the academic community, or scholarly productivity for students in our sample. Analyzing data from 294 PhD students in the life sciences from 53 US institutions, we found no statistically significant differences in outcomes between participants and nonparticipants across 115 variables. These results stand in contrast to prior studies presenting boot camps as effective interventions based on participant satisfaction and perceived value. Many universities and government agencies (e.g., National Institutes of Health and National Science Foundation) invest substantial resources in boot camp and summer bridge activities in the hopes of better supporting scientific workforce development. Our findings do not reveal any measurable benefits to students, indicating that an allocation of limited resources to alternative strategies with stronger empirical foundations warrants consideration.

graduate training | boot camp | research skills | doctoral education

To increase the efficiency and effectiveness of training for future scientists, many PhD programs incorporate boot camps, summer bridge programs, and other short-format instructional interventions to accelerate the development of doctoral students' research skills and acculturation into their respective disciplines. Boot camp programs (2 d to 2 wk) are often designed to expose graduate students to research design (e.g., refs. 1 and 2), mathematical analysis methods and statistical techniques (e.g., refs. 1–3), or scientific writing skills (e.g., ref. 4). Similarly, bridge program activities (4 wk to 8 wk) include an emphasis on research skills training, as well as socialization activities intended to better embed incoming students into the scholarly community (e.g., ref. 2).

Although such interventions are increasingly popular and often supported by federal funding agencies (a search of the National Science Foundation and National Institutes of Health award databases on December 9, 2016 indicated $27.8 million in active funding supporting boot camps and bridge programs), there are few empirical studies of the effectiveness of these strategies. Relevant studies to date rely primarily on participants' reports of their satisfaction and perceived value of their experiences (e.g., refs. 3 and 5). Such studies are of limited validity, because the accuracy of individuals' judgments about their abilities and what they may have learned from a given experience is notoriously inaccurate (6–10).

Extensive evidence suggests that effective instruction or practice should be spaced out over an extended period to support meaningful learning and long-term retention (11, 12). Consequently, the condensed nature of boot camp or nanocourse training may not be as helpful as students perceive. For example, Budé et al. (13) compared the understanding of statistical concepts (*t* tests, analysis of variance, linear regression analysis, etc.) of first-year college students who studied in a 6-mo statistics course (i.e., distributed practice) to those of students who were in a course that covered the same content and provided the same materials and activities in a period of only 8 wk (i.e., massed practice). They found that students using distributed practice performed significantly better than the students using massed practice on the tests administered both during and right after the course. Further, research on metacognition and students' judgments of their own learning suggest that learners often fail to be aware of the impact of spaced instruction. Rather, they tend to experience massed instruction as more effective for their learning, in contrast to the empirical assessments of their performance (14–17), which could account for the positive qualitative reports obtained from boot camp and bridge program participants without necessarily yielding demonstrable effects.

In this study, we compared the skill development, scholarly productivity, and socialization of a national cohort of 294 PhD students in the life sciences (i.e., microbiology, cellular biology, molecular biology, developmental biology, genetics) who did or

Significance

To increase the effectiveness of graduate research training, many universities have introduced boot camps and bridge programs lasting several days to several weeks. National Science Foundation and National Institutes of Health currently support such interventions with nearly $28 million in active awards. Previous evidence for the efficacy of this format exists primarily in the form of anecdotes and end-of-course surveys. Here we show that participation in such short-format interventions is not associated with observable benefits related to skill development, scholarly productivity, or socialization into the academic community. Analyzing data from 294 PhD students in life sciences from 53 US institutions, we found no evidence of effectiveness across 115 variables. We conclude that boot camps and other short formats may not durably impact student outcomes.

Author contributions: D.F.F. and J.R. designed research; D.F.F. and A.S. performed research; J.P., C.M.-F., and M.O. analyzed data; and D.F.F., S.J., J.P., and J.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: david.feldon@usu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1705783114/-/DCSupplemental.

did not participate in boot camp or summer bridge programs immediately before or following the first year of their doctoral programs. Participants were drawn from 53 institutions in the United States and provided data in the form of annual surveys and sole-authored samples of their scholarly writing over 2 y. Of the 294, 48 (16.3%) reported boot camp or summer bridge program participation. Given the prior research on effective learning and metacognition, we hypothesized that boot camp and summer bridge participants would not differ significantly from nonparticipants on any measure.

Research skill development was measured using both a self-report instrument of confidence in ability to perform specific research tasks and the independent scoring of sole-authored research reports or proposals provided by study participants. The Research Experience Self-Rating scale (18) assessed PhD students' beliefs about their abilities to understand contemporary concepts in the field, make use of primary scientific research literature in the field, identify a productive research question, formulate a research hypothesis, design an experiment or theoretical test of a hypothesis, understand the importance of controls in research, observe and collect data, statistically analyze data, interpret results, and reformulate hypotheses as appropriate.

Writing samples were either research proposals or reports of empirical findings and were collected at three time points: before entry into the PhD program, at the end of the first academic year, and at the end of the second academic year. Each writing sample was blinded and scored by two expert raters using a previously validated rubric to assess research skills represented in the submitted manuscripts (19). Mean scores of the two ratings for each measured skill were used in analyses. All raters possessed a PhD in life sciences and had attained robust interrater reliability on all rated skills when scoring participants' writing samples (intraclass correlations ≥ 0.75). The research skills measured were as follows: setting context for a study, framing testable hypotheses, attention to validity and reliability of methods, experimental design, appropriate selection of data for analysis, presentation of data, data analysis, basing conclusions on data, identifying limitations, and effective use of primary literature.

Scholarly productivity was measured using annual self-reported counts of peer-reviewed journal articles, conference papers, and published abstracts that were independently confirmed through researcher verification of participant-provided citation information.

Socialization is defined as "a process of internalizing the expectations, standards, and norms of a given society, which includes learning the relevant skills, knowledge, habits, attitudes, and values of the group that one is joining" (ref. 20, p. 400). The society that doctoral students aspire to join is that of the scholars conducting and publishing research within their chosen discipline. Graduate students are expected to learn their new roles and the skills, values, attitudes, and expectations attached to those roles (21) within the nested contexts of the discipline, institution, department or program, and supervisor's laboratory. Acculturating into these contexts is typically a slow process, and students who do not do so successfully are less likely to complete their programs of study (22, 23). We measured socialization using the following instruments: the Campus Climate and Commitment Survey (perceptions of academic and intellectual development, PhD goal commitment, and institutional commitment; ref. 24), the Perceived Cohesion Scale (sense of belonging to the research community; ref. 25), Weidman and Stein's (26) instrument eliciting perceptions of department collegiality, the Graduate Advising Survey for Doctoral Students (function of advisor and time to degree; ref. 27), and the Research Infrastructure subscale of the Student Research Experience Questionnaire (28).

Scores on survey subscales, publication counts, and research skills were compared between boot camp/summer bridge participants and nonparticipants at 1 y and 2 y after program matriculation. Gains between time points for each measure were similarly compared. All analyses statistically controlled for gender by including it as a covariate. Replicate analyses included the additional covariates of underrepresented racial/ethnic minority status, international student status, and quantity of undergraduate research experience, to rule out the possibility that boot camp and bridge programs could have targeted students for participation who were deemed to be at greater risk of program attrition based on demographics or limited experience with research. All analyses were conducted controlling for nesting within institutions to allow the ignoring of nesting without producing biased parameter estimates. Comparisons used the multiple-group analysis function in Mplus (Version 7.4) to ensure that the assumption of homogeneity of covariate regression slopes was met through parameter estimate constraints while appropriately handling missing data.

Results

Across 115 separate comparisons with each set of covariates, only two and four results, respectively, yielded outcomes at a $P < 0.05$ level, which all favored participants who did not report a boot camp or summer bridge program experience. However, after adjusting for family-wise Type-1 error using the False Discovery Rate (FDR) method (29), which is more liberal than a traditional Bonferroni correction, no comparisons with either set of covariates resulted in P values below the critical threshold values. Based on the results from first-year and second-year cross-sections, as well as gains over the course of the first and second years, we conclude that, despite prior studies reporting high levels of student satisfaction with boot camps and other short-format training (3, 5), participation in these activities by individuals in our sample is not associated with any quantifiable advantages in research skill development, scholarly productivity, or socialization in comparison to students who did not participate. We find it especially noteworthy that the skills most often targeted in participants' boot camp experiences—data analysis ($n = 26$), computer programming ($n = 23$), experimental design ($n = 22$), and academic writing ($n = 21$)—yielded nonsignificant differences on measures of those skills with P values of at least 0.2, and most in the $0.7 \leq P \leq 0.9$ range.

Table S1 presents results from all pairwise comparisons controlling only for gender. Table S2 presents results from all pairwise comparisons when controlling for additional covariates, including gender, duration of undergraduate research experience, underrepresented racial/ethnic minority status, and international student status. All obtained Cohen's d effect sizes (computed only for comparisons in which $P \leq 0.1$) were small (30). Similarly, Monte Carlo analyses failed to reject the null hypothesis in greater than 73% of cases, indicating that there is a low likelihood of attained results attributable to chance (31, 32). Further, inverse sampling weights were included in follow-up analyses to ensure that differential participation rates across institutions did not influence the results (33, 34). Consistent with the two prior analyses, their inclusion did not result in any significant differences after controlling for FDR. These outcomes support the conclusion that there are no significant differences associated with participation in boot camps and summer bridge programs in our sample.

Discussion

How can students' high levels of satisfaction and perceived value reported elsewhere (e.g., refs. 3 and 5) be reconciled with null findings reported here? Research on metacognition and students' judgments of their own learning suggests that, despite the well-established advantages of instruction or practice spaced out over an extended period, learners often fail to recognize the positive impacts of such instruction. Rather, they tend to experience fewer, longer (i.e., massed) blocks of instruction as more effective for their learning, in contrast to the empirical assessments of their performance (15, 17). In short, they conflate the intensity of the experience with its effectiveness.

Convergent findings are evident in studies of summer bridge programs intended to facilitate the transition from high school into undergraduate study. Similar to the short-format instructional strategies discussed by Gutlerner and Van Vactor (4), much of the available evidence regarding their efficacy relies on self-reported perceptions of value and lacks performance-based or longitudinal assessment (35–37). However, the few studies that have assessed longer-term outcomes and/or used more rigorous designs find limited, if any, benefits (38, 39). Most retention and degree completion results yield null (e.g., refs. 40 and 41) or small-magnitude effects of limited duration (39, 42, 43). For instance, Barnett et al. (38) reported that the eight programs they studied had small, positive effects on passing math and writing courses in the first semester compared to a control group. However, these differences were no longer statistically significant after 2 y, and no effect on persistence was found. Similarly, Cabrera et al. (35) reported that bridge program participants did not differ from nonparticipants on GPA or persistence after controlling for traditional forms of training in the first year of undergraduate study.

Another possible explanation for the lack of impact on skill development is the specific set of skills often emphasized in boot camp training, including the curriculum described by Gutlerner and Van Vactor (4) (i.e., experimental design and data analysis strategies). Emerging evidence suggests that graduate students' research skill development follows a specific progression, in which some skills must meet certain thresholds before others can be developed (44, 45). In these studies, skills related to experimental design and data analysis do not demonstrate substantial improvement until students are able to both effectively use primary literature in the framing of their research and generate appropriate and testable hypotheses. Thus, providing experimental design or data analysis training for students who have not yet acquired skills that develop earlier in a learning progression may not be an effective strategy.

Previous research also reports positive faculty views of these interventions (5). However, the published articulation of faculty enthusiasm did not identify student outcomes as the basis for endorsement. Instead, faculty contributing to the delivery of short-format training observed that major benefits of the approach were reduced teaching demands on their time and opportunities to interact with other faculty during delivery of the training (5). Given the pressures on faculty to allocate time to nonteaching activities, it may be that assessments of value are due to self-serving bias, in which faculty judgments of value are skewed by the extent to which it helps them further other goals that are not teaching-related (46).

Conclusions

The consistent pattern of nonsignificant differences in outcomes between short-format training participants and nonparticipants in our sample has direct implications for ongoing efforts to improve doctoral training in life sciences. Currently, many universities and government agencies are investing substantial resources in boot camp and summer bridge activities in the hopes of supporting a better-qualified and more effectively retained scientific workforce (47). The proliferation of these specific strategies is based on preliminary evidence reflecting highly enthusiastic self-reports of participants. However, the current findings suggest that a more critical and methodologically diverse approach should be taken to determine the extent to which boot camps and other short-format instructional activities can contribute to vital training goals. While the generalizability of the current study is limited by its descriptive observational design, it does provide a robust warrant for further investigation. If future studies do not demonstrate measurable benefits to students' research skills, scholarly productivity, or socialization processes compared with students who do not participate in boot camps and other short-format interventions, limited resources available may be better allocated to alternative strategies with stronger empirical foundations.

Materials and Methods

Participant Recruitment. Participants were recruited in two ways. First, program directors and department chairs of the 100 largest biological sciences doctoral programs in the United States were contacted by email to describe the study and request that they inform incoming PhD students about the research project. Following, to diversify the prospective pool of participants, all public flagship universities (research intensive), historically black colleges and universities, and Hispanic-serving institutions offering PhD programs in appropriate biology subfields were contacted. Collectively, emails were sent to administrators at 203 postsecondary institutions. Those who agreed forwarded recruitment information on behalf of the study to students entering PhD programs in Fall 2014 or provided students' email addresses for recruitment materials to be disseminated by project personnel. Interested students then contacted the research team, expressing a willingness to participate. In instances where incoming cohorts were six students or more, campus visits were arranged for a member of the research team to present information to eligible students and answer questions during program orientation or an introductory seminar meeting. Second, emails describing the study and eligibility criteria were forwarded to several listservs, including those of the American Society for Cell Biology and the Center for the Integration of Research, Teaching, and Learning Network for broader dissemination. All students who responded to these emails already attended programs contacted in the first phase of recruitment, suggesting that recruitment efforts approached saturation at the institutional level.

Those individuals who responded to the recruitment emails or presentations were screened to ensure that they met the criteria for participation (i.e., beginning the first year of a PhD program in microbiology, cellular biology, molecular biology, or genetics in Fall 2014) and fully understood the expected scope of participation over the course of the funded project (4 y with possible renewal). It was further explained that all data collected would remain confidential, that all writing samples were scored blindly, and that no information disseminated regarding the study would individually identify them in any way. Participants signed consent forms, and the data collection and analyses were conducted per the requirements specified by the institutional review board (IRB) for human subjects research at Utah State University (protocol 5888). Participants who remained active in the study received a $400 annual incentive, paid in semiannual increments.

Participants were informed that, if they failed to provide two or more consecutive annual data items (i.e., annual surveys) or more than 50% of the biweekly surveys in a single academic year, they would be withdrawn from the study. In addition, any participants who took a leave of absence from their academic program greater than one semester would be withdrawn. All data points were checked and followed up by research assistants for timely completion and appropriate responding. Of $n = 336$ participants from $C = 53$ institutions in the United States, 13 participants were withdrawn during the time these data were collected (nine due to low response rate and four due to taking leave from the degree program in excess of one semester). Twenty-three participants left the study when they withdrew from their academic programs. Two participants chose to end their participation in the study while persisting in their PhD programs, and one participant is deceased. An additional three participants did not provide data regarding their participation in boot camp or bridge programs and were excluded from the current analyses. Deducting these 42 individuals from the sample yielded a sample size for the current study of $n = 294$.

Data regarding the demographic distribution of participants, including gender, across institutions are presented in Tables S3 and S4. Participant age ranged from 23 y to 55 y. Based on Carnegie classification, 42 institutions are R1 (highest research activity), seven institutions are R2 (higher research activity), and the remaining four institutions fall in other Carnegie categories.

Data Collection. For the analysis reported here, relevant data were obtained through web-based surveys and the collection of annual sole-authored writing samples. Surveys were completed during the academic year for the first 2 y of participants' PhD programs. Writing samples were collected at three time points: one sample written by students within 1 y before the start of their PhD programs (i.e., before boot camp/bridge program participation, PhD coursework, and supervised research associated with participants' PhD programs), one sample written during the spring or summer of their first year, and one sample written during the spring or summer of their second year. Details on specific survey instruments and scoring of writing samples are provided in *Annual Survey Battery* and *Measurement of Research Skills*.

Annual Survey Battery.

Background variables. At the outset of the study, participants completed a background survey that elicited their self-identified gender and race/ethnicity, as well as the extent of prior research experience differentiated by setting (high school, undergraduate, graduate, industrial), international student status, and current doctoral program (program, department, institution).

Self-efficacy. Self-efficacy for specific research skills was assessed using the Research Experience Self-Rating Scale (18), which presents individual research competencies and asks respondents to evaluate "To what extent do you feel you can…." on a Likert scale of 1 to 5 ("not at all," "less capable," "capable," "more capable," "a great deal"). Following are the individual task items: Understand contemporary concepts in your field, Make use of the primary science literature in your field (e.g., journal articles), Identify a specific question for investigation based on the research in your field, Formulate a research hypothesis based on a specific question, Design an experiment or theoretical test of the hypothesis, Understand the importance of "controls" in research, Observe and collect data, Statistically analyze data, Interpret data by relating results to the original hypothesis, and Reformulate your original research hypothesis (as appropriate). For the current sample, the scale yielded a reliability of $\alpha = 0.903$.

Goal commitment and institutional commitment. To assess value for and commitment to degree attainment, participants also completed the Degree Commitment and Institutional Commitment subscales (24). The Degree Commitment subscale includes three items, which require respondents to rate the importance of earning a doctoral degree (e.g., "It is important for me to get a PhD") and completing the program of studies (e.g., "It is important for me to finish my program of studies") on a Likert scale of 1 to 3 ("disagree," "neither agree nor disagree," "agree"). For the current sample, this subscale yielded a reliability of $\alpha = 0.998$.

The Institutional Commitment subscale includes three items, which require respondents to rate the certainty of their choice of an institution (e.g., "I am confident I made the right decision in choosing this institution") and the sense of belonging to the institution (e.g., "I feel I belong at this institution") on a Likert scale of 1 to 3 ("disagree," "neither agree nor disagree," "agree"). For the current sample, this subscale yielded a reliability of $\alpha = 0.968$.

Scholarly socialization. Four subscales assessing participants' socialization experiences were also used to assess participants' scholarly engagement and social interactions with faculty and peers (26). The Participation in Scholarly Activities subscale included a checklist of 11 items describing scholarly and research activities, such as "Asked a fellow student to critique your work," "Presented a paper at a conference or convention," or "Held membership in a professional organization." Participants were asked to check the activities on the list in which they were involved during doctoral training, and the total number of checks of the 11 items were used as scores for this scale. For the current sample, this subscale yielded a reliability of $\alpha = 0.930$.

The Student–Faculty and Student–Peer Interactions subscale asked respondents to indicate "yes" or "no" to the follow-up four items, with the stem question, "Is there any professor (or student) in your department with whom you…" Four individual endings followed: "Sometimes engage in social conversation," "Often discuss topics in his/her field," "Often discuss other topics of intellectual interest," and "Ever talk about personal matters." For the current sample, this subscale yielded a reliability of $\alpha = 0.966$.

The Department Collegiality subscale included three items to ask respondents to evaluate the extent to which they perceive the department as a collaborative community of scholars where respect and collaboration are internalized (e.g., "I am treated as a colleague by the faculty," "The faculty sees me as a serious scholar") on a Likert scale of 1 to 5 ("strongly disagree," "disagree," "neither agree nor disagree," "agree," "strongly agree"). For the current sample, this subscale yielded a reliability of $\alpha = 0.883$.

The Student Scholarly Encouragement subscale included four items to ask respondents to evaluate the extent to which the departmental climate encourages the scholarly activities and aspirations of students (e.g., "An environment that promotes scholarly interchange between students and faculty," "An educational climate that encourages the scholarly aspirations of all students") on a Likert scale of 1 to 3 ("not at all true," "somewhat true," "completely true"). For the current sample, this subscale yielded a reliability of $\alpha = 0.960$.

Mentorship. The characteristics and qualities of participants' relationships with their advisors were assessed using two relevant subscales from the Graduate Advising Survey for Doctoral Students (27). The two subscales are Function of Advisor, with 16 items (e.g., "My primary advisor is readily available to talk with me when needed," "My primary advisor gives me constructive feedback on my progress toward degree completion"), and Time to Degree, with four items (e.g., "My academic program has structure in place to help graduate students make timely progress toward their degree," "How helpful has your primary advisor been to you in terms of progressing toward the completion of your degree?"). All subscale items used a three-point Likert scale (e.g., "disagree," "neither agree nor disagree," "agree"). For the current sample, the Function of Advisor subscale had an attained reliability of $\alpha = 0.973$, and Time to Degree had an attained reliability of $\alpha = 0.870$.

Academic and social climate. To examine participants' perceptions of the social and academic climate within their assigned research laboratories, programs, departments, and institutions, the Perceived Cohesion/Sense of Belonging scale (25) and the Academic & Intellectual Development subscale (24) were used. The Perceived Cohesion/Sense of Belonging scale included three items (e.g., "I feel a sense of belonging to my lab/research group," "I see myself as part of the lab/research group community") that were accompanied by a Likert scale, ranging from 1 (strongly disagree) to 10 (strongly agree). This scale yielded a reliability of $\alpha = 1.000$ for the current sample. The Academic & Intellectual Development subscale included three items (e.g., "I am satisfied with the extent of my intellectual development since attending this institution," "I am satisfied with my academic experience at this institution") that were accompanied by a three-point Likert scale ("disagree," "neither agree nor disagree," "agree"). This scale yielded a reliability of $\alpha = 0.976$ for the current sample.

Access to research infrastructure. In the process of engaging in research opportunities and developing research skills, access to the necessary resources and equipment may be an important factor. To assess this, participants completed the Research Infrastructure subscale of the Student Research Experience Questionnaire (24). Seven items were included in this subscale (e.g., "I have access to a suitable working space", "I am able to organize good access to necessary equipment"), and each item was rated on a three-point Likert scale ("not at all true," "somewhat true," "completely true"). For the current sample, the scale yielded a reliability of $\alpha = 0.960$.

Publications survey. At the conclusion of the Spring semester, participants received another survey that asked them to identify any journal articles, conference papers, or published abstracts for which they had received authorship credit during the academic year.

Measurement of Research Skills. To examine participants' research skill development, their sole-authored writing samples, reports of empirical findings, or research proposals were collected at three time points: before entry into the doctoral program, at the end of the first academic year, and at the end of the second academic year. The writing samples were received from participants electronically, checked for plagiarism using TurnItIn (48), and assigned to raters based on subject matter.

Two expert raters, with PhDs in relevant subfields of biology, blindly and independently scored each writing sample using the rubric to measure discrete research skills. This rubric was an integrated version of two that have each been previously validated (19, 43) and yielded intraclass correlations (ICC; two-way, random effects) for individual planks between 0.782 and 0.944. The rubric measured the following research skills: setting context for a study (ICC = 0.803), generating testable hypotheses (ICC = 0.862), establishing appropriate controls (ICC = 0.845), research/experimental design (ICC = 0.917), appropriate selection of data for analysis (ICC = 0.834), presentation of data (ICC = 0.905), data analysis (ICC = 0.789), drawing conclusions based on data (ICC = 0.782), exploring alternative interpretations of data (ICC = 0.815), identifying research design limitations (ICC = 0.877), generating implications for findings (ICC = 0.845), effective use of primary literature (ICC = 0.944), and overall writing quality (ICC = 0.832). For the rubric criterion of each research skill, the raters scored a participant's writing sample on the following levels: not addressed (0 points), novice (one points), intermediate (two points), or proficient (three points). Raters could augment scores by adding or subtracting 0.25 from the criterion-anchored integer scores to reflect stronger or weaker cases of performance that met the criteria for the designated level. Mean scores of the two ratings for each research skill were used for all statistical analyses. Full criteria for all planks are provided in *Supporting Information*.

Statistical Analyses. Based on their survey responses indicating whether or not they had participated in a boot camp or a bridge program in the summer immediately before or following their first academic year in their PhD program, participants were dummy coded as 1 = participant ($n = 48$) or 2 = nonparticipant ($n = 246$) as the independent variable. Analyses of covariance (ANCOVA) were then computed comparing T1, T2, and T2 controlling for T1 (i.e., gain) as dependent variables for each of the survey measures identified in *Annual Survey Battery*. Analyses of research skills assessed through writing samples compared T1, T2, and T2 controlling for T1 (i.e., gain) under conditions of controlling for T0 (i.e., gain from before beginning the PhD program) and not controlling for T0 as dependent variables.

Participant gender (dummy coded) was used as a covariate for all analyses, based on substantial influences of gender observed previously on multiple variables of interest with this data set (49).

All analyses were conducted controlling for nesting within institution using specific commands ("Type = Complex") in Mplus (Version 7.4) that allow the ignoring of nesting without producing biased parameter estimates. Comparisons used the multiple-group analysis function in Mplus to ensure that the ANCOVA assumption of homogeneity of covariate regression slopes is met through parameter estimate constraints while appropriately handling missing data. In addition to the above, analyses were repeated using additional covariates: duration of undergraduate research experience, underrepresented racial/ethnic minority status, and international student status. These were selected to rule out effects stemming from the possibility that boot camp and bridge programs could have targeted students for participation who were deemed to be at greater risk of program attrition based on demographics or limited experience with research.

While our sample size ($n = 294$) is admirable given the natures of the data collected and the population studied, it cannot be considered optimal for statistical analyses. Bootstrap resampling, effect size estimate computations, and Monte Carlo simulation testing represent methods that can serve as a check of the accuracy of population inferences made based on the results of a sample of size $n = 294$. For all results with $P \leq 0.1$, two additional analyses were undertaken. First, Cohen's d effect size estimates were generated. Second, Monte Carlo analyses of 5,000 generated datasets of size $n = 294$ enabled the determination of the number of times in 5,000 samples the null hypothesis (H_0:) of a zero mean difference for all dependent variables was rejected. Further, to ensure that the variable number of respondents from each university did not bias the outcomes of the statistical analyses, inverse sampling weights were computed and included in a second series of replication analyses (33, 34). However, their inclusion did not yield any significant differences between groups after applying FDR Type-1 error correction.

ACKNOWLEDGMENTS. The authors gratefully acknowledge the support of the National Science Foundation. This material is based upon work supported under Awards 1431234 and 1431290. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

1. Vale RD, et al. (2012) Graduate education. Interdisciplinary graduate training in teaching labs. *Science* 338:1542–1543.
2. Brandon DH, Collins-McNeil J, Onsomu EO, Powell DL (2014) Winston-Salem State University and Duke University's bridge to the doctorate program. *N C Med J* 75: 68–70.
3. Stefan MI, Gutlerner JL, Born RT, Springer M (2015) The quantitative methods boot camp: Teaching quantitative thinking and computing skills to graduate students in the life sciences. *PLOS Comput Biol* 11:e1004208.
4. Gutlerner JL, Van Vactor D (2013) Catalyzing curriculum evolution in graduate science education. *Cell* 153:731–736.
5. Bentley AM, Artavanis-Tsakonas S, Stanford JS (2008) Nanocourses: A short course format as an educational tool in a biological sciences graduate curriculum. *CBE Life Sci Educ* 7:175–183.
6. Bowman NA (2010) Can 1st-year college students accurately report their learning and development? *Am Educ Res J* 47:466–496.
7. Dunning D, Johnson K, Ehrlinger J, Kruger J (2003) Why people fail to recognize their own incompetence. *Curr Dir Psychol Sci* 12:83–87.
8. Feldon DF, Maher MA, Timmerman BE (2010) Performance-based data in the study of STEM graduate education. *Science* 329:282–283.
9. Feldon DF, Maher MA, Hurst M, Timmerman B (2015) Faculty mentors', graduate students', and performance-based assessments of students' research skill development. *Am Educ Res J* 52:334–370.
10. Stajkovic AD, Luthans F (1998) Self-efficacy and work-related performance: A meta-analysis. *Psychol Bull* 124:240–261.
11. Carpenter SK, Cepeda NJ, Rohrer D, Kang SH, Pashler H (2012) Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educ Psychol Rev* 24:369–378.
12. Rohrer D (2015) Student instruction should be distributed over long time periods. *Educ Psychol Rev* 27:635–643.
13. Budé L, Imbos T, van de Wiel MW, Berger MP (2011) The effect of distributed practice on students' conceptual understanding of statistics. *High Educ* 62:69–79.
14. Dunlosky J, Nelson TO (1992) Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Mem Cognit* 20:374–380.
15. Logan JM, Castel AD, Haber S, Viehman EJ (2012) Metacognition and the spacing effect: The role of repetition, feedback, and instruction on judgments of learning for massed and spaced rehearsal. *Metacogn Learn* 7:175–195.
16. Son LK, Simon DA (2012) Distributed learning: Data, metacognition, and educational implications. *Educ Psychol Rev* 24:379–399.
17. Toppino TC, Cohen MS (2010) Metacognitive control and spaced practice: Clarifying what people do and why. *J Exp Psychol Learn Mem Cogn* 36:1480–1491.
18. Kardash C (2000) Evaluation of undergraduate research experience: Perceptions of undergraduate interns and the faculty mentors. *J Educ Psychol* 92:191–201.
19. Feldon DF, et al. (2011) Graduate students' teaching experiences improve their methodological research skills. *Science* 333:1037–1039.
20. Austin AE, McDaniels M (2006) Preparing the professoriate of the future: Graduate student socialization for faculty roles. *Higher Education: Handbook of Theory and Research*, ed Smart JC (Springer, Dordrecht, The Netherlands), pp 397–456.
21. Weidman JC (2010) Doctoral student socialization for research. *On Becoming a Scholar: Socialization and Development in Doctoral Education*, eds Gardner SK, Mendoza P (Stylus, Sterling, TX), pp 45–56.
22. Golde CM (2005) The role of department and discipline in doctoral student attrition: Lessons from four departments. *J High Educ* 76:669–700.
23. Lovitts BE (2001) *Leaving the Ivory Tower: The Causes and Consequences of Departure from Doctoral Study* (Rowman and Littlefield, Lanham, MD).
24. Nora A, Cabrera AF (1996) The role of perceptions of prejudice and discrimination on the adjustment of minority students to college. *J High Educ* 67:119–148.
25. Bollen KA, Hoyle RH (1990) Perceived cohesion: A conceptual and empirical examination. *Soc Forces* 69:479–504.
26. Weidman JC, Stein EL (2003) Socialization of doctoral students to academic norms. *Res Higher Educ* 44:641–656.
27. Barnes BJ, Chard LA, Wolfe EW, Stassen ML, Williams EA (2011) An evaluation of the psychometric properties of the Graduate Advising Survey for Doctoral Students. *Int J Doctr Stud* 6:1–17.
28. Ginns P, Marsh HW, Behnia M, Cheng JH, Scalas LF (2009) Using postgraduate students' evaluations of research experience to benchmark departments and faculties: Issues and challenges. *Br J Educ Psychol* 79:577–598.
29. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc* 57:289–300.
30. Cohen J (1988) *Statistical Power Analysis for the Behavioral Sciences* (Erlbaum, Hillsdale, NJ), 2nd Ed.
31. Cohen J (1992) A power primer. *Psychol Bull* 112:155–159.
32. Nunnally JC, Bernstein IH (1994) *Psychometric Theory* (McGraw-Hill, New York), 3rd Ed.
33. Stapleton LM (2002) The incorporation of sample weights into multilevel structural equation models. *Struct Equ Modeling* 9:475–502.
34. Stapleton LM (2008) Variance estimation using replication methods in structural equation modeling with complex sample data. *Struct Equ Modeling* 15:183–210.
35. Cabrera NL, Miner DD, Milem JF (2013) Can a summer bridge program impact first-year persistence and performance?: A case study of the New Start Summer Program. *Res Higher Educ* 54:481–498.
36. Garcia LD, Paz CC (2009) Bottom line: Evaluation of summer bridge programs. *About Campus* 14:30–32.
37. Kezar A (2000) Summer BRIDGE PROGRAMS: Supporting all students. ERIC Digest: ED442421.
38. Barnett EA, et al. (2012) *Bridging the Gap: An Impact Study of Eight Developmental Summer Bridge Programs in Texas* (Natl Cent Postsecondary Res, New York).
39. Murphy TE, Gaughan M, Hume R, Moore SG, Jr (2010) College graduation rates for minority students in a selective technical university: Will participation in a summer bridge program contribute to success? *Educ Eval Policy Anal* 32:70–83.
40. DeRoma VM, Bell NL, Zaremba BA, Albee JC (2005) Evaluation of a college transition program for students at-risk for academic failure. *Res Teach Dev Educ* 21:20–33.
41. Walpole M, et al. (2008) Bridge to success: Insight into summer bridge program students' college transition. *J First Year Exper Stud Transit* 20:11–30.
42. Gleason J, et al. (2010) Integrated engineering math-based summer bridge program for student retention. *Adv Eng Educ* 2:1–17.
43. Wathington H, et al. (2011) *Getting Ready for College: An Implementation and Early Impacts Study of Eight Texas Developmental Summer Bridge Programs* (Natl Cent Postsecondary Res, New York).
44. Kiley M, Wisker G (2009) Threshold concepts in research education and evidence of threshold crossing. *High Educ Res Dev* 28:431–441.
45. Timmerman BC, Feldon D, Maher M, Strickland D, Gilmore J (2013) Performance-based assessment of graduate student research skills: Timing, trajectory, and potential thresholds. *Stud High Educ* 38:693–710.
46. Ditto PH, Lopez DF (1992) Motivated skepticism: Use of differential decision criteria for preferred and non-preferred conclusions. *J Pers Soc Psychol* 63:568–584.
47. McGee R, Jr, Saran S, Krulwich TA (2012) Diversity in the biomedical research workforce: Developing talent. *Mt Sinai J Med* 79:397–411.
48. Gilmore J, Strickland D, Timmerman B, Maher M, Feldon DF (2010) Weeds in the flower garden: An exploration of plagiarism in graduate students' research proposals and its connection to enculturation, ESL, and contextual factors. *Int J Educ Integr* 6: 13–28.
49. Feldon DF, Peugh J, Maher MA, Roksa J, Tofel-Grehl C (2017) Time-to-credit gender inequities of first-year PhD students in the biological sciences. *CBE Life Sci Educ* 16:ar4.