EDUCATION RESEARCH

10.1126/sciadv.1501422

Assessing faculty professional development in STEM higher education: Sustainability of outcomes

Terry L. Derting,[1]* Diane Ebert-May,[2] Timothy P. Henkel,[3] Jessica Middlemis Maher,[4] Bryan Arnold,[5] Heather A. Passmore[1]

We tested the effectiveness of Faculty Institutes for Reforming Science Teaching IV (FIRST), a professional development program for postdoctoral scholars, by conducting a study of program alumni. Faculty professional development programs are critical components of efforts to improve teaching and learning in the STEM (Science, Technology, Engineering, and Mathematics) disciplines, but reliable evidence of the sustained impacts of these programs is lacking. We used a paired design in which we matched a FIRST alumnus employed in a tenure-track position with a non-FIRST faculty member at the same institution. The members of a pair taught courses that were of similar size and level. To determine whether teaching practices of FIRST participants were more learner-centered than those of non-FIRST faculty, we compared faculty perceptions of their teaching strategies, perceptions of environmental factors that influence teaching, and actual teaching practice. Non-FIRST and FIRST faculty reported similar perceptions of their teaching strategies and teaching environment. FIRST faculty reported using active learning and interactive engagement in lecture sessions more frequently compared with non-FIRST faculty. Ratings from external reviewers also documented that FIRST faculty taught class sessions that were learner-centered, contrasting with the teacher-centered class sessions of most non-FIRST faculty. Despite marked differences in teaching practice, FIRST and non-FIRST participants used assessments that targeted lower-level cognitive skills. Our study demonstrated the effectiveness of the FIRST program and the empirical utility of comparison groups, where groups are well matched and controlled for contextual variables (for example, departments), for evaluating the effectiveness of professional development for subsequent teaching practices.

INTRODUCTION

Despite the continued availability of, and interest in, teaching development opportunities for STEM (Science, Technology, Engineering, and Mathematics) faculty, there is little empirical evidence to support the relative impact of professional development programs on teaching practice (*1*–*4*). Studies that generate reliable evidence of transfer of training following professional development in teaching are needed (*3*, *5*). Research on transfer (that is, Do people apply to their job what they learned in training?) (*6*) merits empirical evidence of whether, and how, professional development affects subsequent teaching practices (*7*–*10*). After all, the goal of most educational professional development programs is to produce long-lasting changes through the implementation of new approaches to future teaching endeavors (*11*). Furthermore, assessment of the impact of professional development programs is necessary to identify the types of professional development activities that are most closely associated with changes in participants' teaching. Knowledge of effective activities and approaches will allow us to improve the efficiency and impact of professional development programs.

Our knowledge of transfer of training following professional development is bounded by the type of research designs used (that is, qualitative, quantitative, and quasi-experimental), experimental rigor, and type of data collected. Many studies of the impacts of faculty professional development use self-reported or interview data to determine how the participants enacted what they learned during professional training [for example, (*1*)]. However, it is well documented that disjunction between instructors' conceptions of their teaching and their claimed and actual educational practices exists (*7*, *12*–*15*). Numerous qualitative and quantitative studies of the impacts of professional development have been conducted, but few are designed to provide strong evidence of outcomes. Stes *et al.* (*16*) reviewed the effects of instructional development on teachers' learning of skills, and they found that, across 108 studies, a comparison/control group was used in only 14% of the quantitative or mixed-methods studies and none was used in the qualitative studies. Furthermore, implementation by faculty does not occur in a vacuum; therefore, it is challenging to assess program success across different institutions [for example, (*15*)]. We examined the impacts of a national faculty professional development program, Faculty Institutes for Reforming Science Teaching IV (FIRST), on the teaching practices of faculty. Results from our study provided strong evidence of the impacts of FIRST on learner-centered teaching at multiple institutions.

The FIRST professional development program targeted 201 biology postdoctoral scholars (postdocs) at universities nationwide. Postdocs were recruited through announcement of the FIRST program on the electronic mailing lists of professional societies. We used learning theory and evidence-based instructional strategies in a mentored, team-based approach to transforming teaching practice throughout FIRST, which was designed around the principles of scientific teaching (*17*). Over a 2-year period, the postdocs engaged in an iterative process of curriculum development and teaching practicum, with the goal of developing and applying learner-centered teaching practices in undergraduate biology courses (*18*). At the beginning of each year of participation, postdocs completed a 4-day summer workshop. The broad objectives of the workshops are described by Ebert-May *et al.* (*18*). During the workshop, the postdocs learned to actively engage students in both large and small

[1]Department of Biological Sciences, Murray State University, Murray, KY 42071, USA. [2]Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA. [3]Department of Biology, Valdosta State University, Valdosta, GA 31698, USA. [4]Delta Program, University of Wisconsin–Madison, Madison, WI 53706, USA. [5]Department of Biology, Illinois College, Jacksonville, IL 62650, USA.
*Corresponding author. E-mail: tderting@murraystate.edu

enrollment courses, use individual and group learning strategies, use backward design, and write assessments that were aligned with learning objectives and instructions. Key to the first workshop was establishing teams of postdocs who designed an entire learner-centered lower-level biology course based on the principles of scientific teaching (*17*), that is, teaching using scientific practices (for example, creating models and arguments, working with data) to learn core concepts. Following the first workshop, the postdocs completed a teaching experience in their department or in another unit on campus. The length of the experience varied from one class session to an entire course, and the regional team leaders in the project served as long-distance teaching mentors for the groups.

Year 2 was the critical follow-up summer workshop. During this time (3 or 4 days), the postdocs reflected on and discussed the challenges they encountered during their teaching experience. They reviewed videos from their courses and, with assistance from their teaching mentors, identified strengths and elements that needed work. During the year 2 workshop, postdocs analyzed their course design, which was now informed with some student assessment data from their teaching during the previous academic year, and continued to revise their entire course. During year 2, the postdocs completed a second teaching experience (a full or partial course), and some taught the course that they had developed in the context of a new faculty position. Again, long-distance teaching mentors and a closed FIRST LISTSERV were valuable tools for maintaining contact and providing immediate support to these early-career teachers.

Evaluation of the postdocs' teaching during their participation in FIRST demonstrated the program's success in developing learner-centered instructors (*18*). The question of whether participants continued using learner-centered practices in their subsequent roles as faculty remained. In the present study, we followed a subset of FIRST postdocs into their initial faculty position and collected evidence of their teaching, along with identical data from a paired colleague. Our research is both compelling and timely because it provides a new line of investigation into whether faculty who completed FIRST transferred skills and beliefs about teaching that they addressed during the program and practiced more learner-centered teaching compared with other early-career faculty in the same departments. Our research also demonstrates the utility of a paired research design for identifying areas of significant difference following professional development.

RESULTS

We first verified that FIRST and non-FIRST faculty had similar backgrounds in relation to teaching. Participation in prior professional development activities and confidence in their current level of preparation as a teacher did not differ significantly within pairs (n = 18 pairs; Student's t test, t = 0.66, P = 0.52; Wilcoxon signed-rank test, S = 14.5, P = 0.17). Faculty pairs were recruited from a broad range of institution types (seven research institutions, seven comprehensive institutions, six liberal arts institutions, and one community college), and the percentage of appointment (as per contract) that was devoted to teaching averaged 60% for both groups. The pairs also reported similar first-hand knowledge of and experience with active learning (Student's t test, t = −0.03, P = 0.97 and t = 0.06, P = 0.95, respectively) in the context of science education reform, assessment, theories of learning, and teaching practices (for example, case studies, cooperative/collaborative learning, and problem-based learning). Faculty pairs also reported similar levels of departmental commitment to undergraduate education (Student's t test, t = −1.29, P = 0.21; Table 1). However, non-FIRST faculty perceived greater challenges to implementing an active-learning course compared with their FIRST faculty pair, specifically in terms of cooperation of faculty in other departments (but not in their own department) and of Teaching Assistants or other instructional staff (Student's t test, t = −2.85, P < 0.01 and t = −2.11, P = 0.04, respectively). Perceptions of other potential barriers to active-learning course implementation (for example, time to develop materials, grade, training of Teaching Assistants, support of campus administrators, tenure-related issues, and student attitudes and evaluations) did not differ significantly between the two groups. Background data were incomplete for two pairs. Box plots of data from other assessments used in our study indicated similarity between these individuals and the other 18 faculty members in their respective group (FIRST or non-FIRST). Thus, these two pairs were included in subsequent analyses.

Perceptions of teaching practice

Non-FIRST and FIRST faculty reported similar perceptions of their teaching strategies at the end of the semester on the Approaches to Teaching Inventory (ATI) (*19*, *20*). Course enrollments ranged from 10 to 212 students, with the average course size being 53 and 59 students for FIRST and non-FIRST faculty, respectively. On average, all faculty reported a high level of self-efficacy; however, non-FIRST faculty reported a slightly higher level of self-efficacy in their teaching ability compared with FIRST faculty at the end of the semester [n = 20 pairs; mean ± SE: non-FIRST, 4.6 ± 0.1; FIRST, 4.3 ± 0.1 (using a five-point Likert scale where 5 = *strongly agree*); Student's t test, t = −2.30, P = 0.03; Cohen's d = 0.24]. No significant differences in the pairs' support for the use of conceptual-change/student-focused (CCSF) or information-transmission/teacher-focused (ITTF) teaching strategies in their course (n = 20 pairs; Student's t test, t = 1.15, P = 0.26 and Student's t test, t = −0.68, P = 0.50, respectively) were noted. Both groups had a higher mean score for the

Table 1. Characteristics of matched pairs of a FIRST faculty participant and a non-FIRST faculty colleague. Knowledge of and experience with active learning, perception of departmental commitment, and challenges to implementing active learning were calculated as summed Likert responses to the Teaching Background Survey (Appendix S1). Large course sizes are those with >75 students per course. Data are presented as mean ± SE.

Group	Female participants (%)	Teaching experience (years)	Active-learning knowledge	Active-learning experience	Departmental commitment	Active-learning challenge	Large course size (%)
FIRST	56	2.1 ± 2.3	35.2 ± 2.3	34.8 ± 1.8	21.3 ± 4.8	51.5 ± 1.2	18
Non-FIRST	50	4.1 ± 4.3	35.3 ± 2.7	34.7 ± 2.5	23.2 ± 3.2	48.7 ± 2.4	18

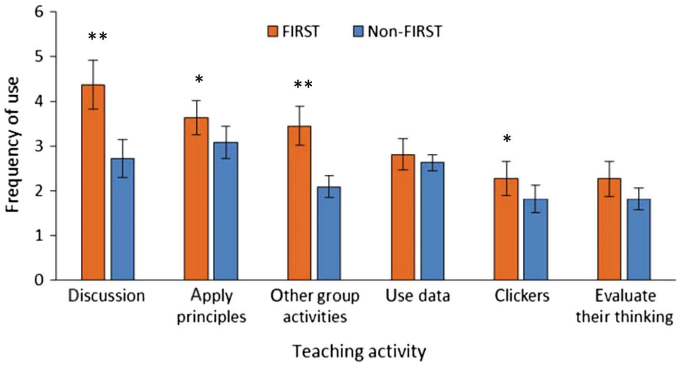

Fig. 1. Teaching practice at the end of a semester. The frequency (mean ± SE) with which FIRST and non-FIRST faculty (n = 11 pairs) used various types of interactive activities during a semester (paired t test or Wilcoxon signed-rank test, *P < 0.05, **P < 0.01), as reported on the Teaching Practice Survey. "Discussion"—student discussions in pairs or small groups to answer a question; "Apply principles"—classroom interactions that required students to apply principles presented in class to a novel question; "Use data"—activities in which students use data to answer questions while working in small groups; "Clickers"—clicker questions that test conceptual understanding; "Evaluate their thinking"—individual writing activities that require students to evaluate their own thinking (1 = never; 2 = several times per semester; 3 = once per week; 4 = once per class; 5 = more than once per class).

CCSF subscale (FIRST, 3.89 ± 0.12; non-FIRST, 3.63 ± 0.19) than for the ITTF subscale (FIRST, 3.33 ± 0.12; non-FIRST, 3.48 ± 0.12), with the difference between the subscale scores being similar for the two groups (Student's t test, t = 1.11, P = 0.28).

Faculty pairs (n = 20 pairs) also had similar perceptions of their teaching environment, specifically in terms of the potential impacts of a large class size (Student's t test, t = −0.97, P = 0.35), their department's commitment to student learning (Student's t test, t = −0.34, P = 0.74), impacts of their workload (Student's t test, t = 0.55, P = 0.59), and potential impacts of a course composed of students with diverse abilities (Student's t test, t = −0.58, P = 0.57). On average, the members of pairs differed in their perceived level of control over course content and how their course was taught, with FIRST faculty reporting greater control compared with non-FIRST faculty (Student's t test, t = 2.77, P = 0.01).

Self-reported information from faculty about their own teaching at the end of a semester (Teaching Practice Survey) indicated significant differences within the pairs. FIRST faculty reported using active learning, interactive engagement, and other interactive activities within the lecture portion of their courses more than once per class. In contrast, non-FIRST faculty used interactive activities significantly less, usually averaging once per week and sometimes once per class session (n = 11 pairs; Wilcoxon signed-rank test, S = 27.5, P = 0.002). When specific types of class activities were examined, FIRST faculty more frequently engaged students in group discussions to answer a question (Student's t test, t = 2.96, P = 0.01) and in application of principles to address a novel question (Student's t test, t = 2.47, P = 0.03), and used other small group activities (Student's t test, t = 3.13, P = 0.01) compared with non-FIRST faculty (Fig. 1). No differences in how frequently they presented clicker questions that tested conceptual understanding (Wilcoxon signed-rank test, S = 27.5, P = 0.002) or in how frequently they used individual writing activities that required students to evaluate their own thinking (Student's t test, t = 0.20, P = 0.85) were noted among faculty pairs.

Observed teaching practice

Participation in FIRST had a significant positive impact on the extent to which class sessions were learner-centered when controlling for potential effects of instructor gender, class size, and perceptions of challenges to implementing active-learning classes (linear regression, $F_{4,\ 23}$ = 10.7, P < 0.0001, R^2 = 0.65; Table 2). On average, FIRST faculty taught class sessions that were learner-centered, contrasting with the teacher-centered class sessions of most non-FIRST faculty (Fig. 2). The mean Reformed Teaching Observation Protocol (RTOP) (*21*) score for FIRST faculty (51.8 ± 2.3) was within RTOP category III, which is characterized by significant student engagement with some minds-on and hands-on involvement of students (*22*, *23*). The mean score for non-FIRST faculty was markedly lower (37.8 ± 1.9) within RTOP category II, which is characterized as primarily lecture with minor student participation (*22*, *23*). The effect size was very large, as measured using Cohen's d (d = 1.64) (*24*). As indicated by the parameter estimate for the treatment effect, all else being equal, FIRST faculty had an expected RTOP score that was 16 points—or an entire RTOP category—higher than that of non-FIRST faculty (Table 2).

We tested for consistency between external reviews of faculty teaching practice and self-reported faculty perceptions of their own teaching strategies. For FIRST and non-FIRST faculty, RTOP score was significantly and positively correlated with support for the use of CCSF approaches when teaching (n = 19, r^2 = 0.44, P = 0.002 and n = 17, r^2 = 0.34, P = 0.015, respectively). For FIRST (but not non-FIRST) faculty, lack of support for the use of ITTF approaches was significantly and negatively correlated with RTOP score (r^2 = 0.31, P = 0.02).

Students' perceptions of the classroom, as determined from responses to the Experiences of Teaching and Learning Questionnaire (ETLQ) (*25*), differed between courses taught by FIRST faculty and courses taught by non-FIRST faculty. These differences were not likely attributable to differences among the students because students in both types of courses had similar responses on each of the three sections of the Learning and Studying Questionnaire (LSQ) (*25*) at the beginning of the semester (n = 13 pairs; learning orientations: Student's t test, t = −0.25, P = 0.81; reasons for taking the particular course: Student's t test, t = −1,45, P = 0.17; student approaches to learning and studying: Student's t test, t = −0.23, P = 0.82). At the end of a course, students differed significantly in their perceived approaches to learning and studying (n = 11 pairs; Student's t test, t = −2.37, P = 0.04) primarily because of significantly less agreement with a surface approach by students in courses taught by FIRST faculty compared with courses taught by non-FIRST faculty.

Students also differed significantly in their perceptions of the teaching-learning environment (n = 11 pairs; Student's t test, t = 2.52, P = 0.03), mainly in two areas. First, students in courses taught by FIRST faculty reported greater instructor enthusiasm and support (for example, the instructor was patient in explaining topics, helped them see how to think and reach conclusions, and valued student views more than in courses taught by non-FIRST faculty), and, second, they reported greater support from other students (for example, talking with other students helped their understanding, students supported their peers and tried to help when it was needed). The two groups of students did not differ in their perceptions of the demands made of them during the course (n = 11 pairs; Student's t test, t = 1.31, P = 0.22) or of learning achieved (t = 2.14, P = 0.058).

Assessment of student learning

In contrast with teaching practice, there was no significant difference in the levels of cognition targeted by the course goals stated in the syllabi of

Table 2. Regression analysis results for the effects of faculty participation in the FIRST project (treatment) on RTOP scores from expert reviews of faculty teaching. Model $r^2 = 0.65$. Challenges to active-learning implementation were calculated as summed Likert responses to the Teaching Background Survey (Appendix S1). *P* values in the right-hand column are from two-tailed *t* tests.

Coefficient	Parameter estimate	SE	*P*	Standardized estimate
Intercept	61.19	14.92	0.0004	0
Treatment	15.95	2.74	<0.0001	0.73
Gender	2.06	3.16	0.52	0.09
Class enrollment	−0.05	0.03	0.13	−0.22
Challenges to active learning	−0.43	0.30	0.16	−0.18

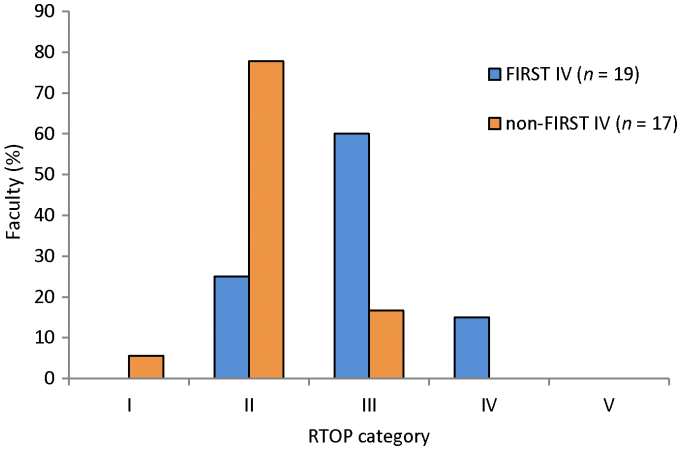

Fig. 2. Comparison of RTOP scores per category. Distribution of RTOP scores for teaching videos of FIRST and non-FIRST faculty.

FIRST faculty ($n = 14$) and non-FIRST faculty ($n = 12$; logistic regression, likelihood ratio $\chi^2 = 1.59$, $P = 0.90$; Fig. 3). Non-FIRST faculty tended to state more goals that focused on content knowledge and fewer goals that focused on application of knowledge compared with FIRST faculty, but the variability among faculty was high and the patterns were not statistically significant. Likewise, there was no statistically significant difference in the levels of cognition assessed through high-stakes assessments ($n = 17$ FIRST faculty and $n = 15$ non-FIRST faculty, logistic regression, likelihood ratio $\chi^2 = 3.60$, $P = 0.61$). Most of the assessment questions by FIRST and non-FIRST faculty targeted student knowledge and comprehension of concepts, or Bloom cognitive skill levels I and II.

DISCUSSION

As postdocs, FIRST participants engaged in a 2-year professional development program designed to develop learner-centered teaching practices. We provide evidence that FIRST faculty continued to use these practices early in their initial faculty positions, demonstrating successful transfer of professional development training from one career stage to another. Furthermore, FIRST faculty taught in a more learner-centered way compared with early-career colleagues, indicating that the FIRST program may lead to differential outcomes for faculty and students.

On the basis of our results, we rejected our null hypothesis that there were no differences in teaching practice between FIRST and non-FIRST faculty. At the same time, we accepted the null hypothesis that there were no differences in the perceptions of teaching strategies and the levels of cognition targeted by the two groups of faculty. Our results demonstrate the utility of a paired approach as a means of determining the ways by which faculty professional development programs succeed in creating changes in teaching practice and the ways by which they do not.

Perceptions of teaching strategies

We found no measurable effect of the FIRST program of professional development on participants' perceptions of the teaching strategies that they used in the courses taught during our research. Moreover, although the self-efficacy score of non-FIRST faculty was statistically greater than that of FIRST faculty, the effect size was small and the difference was unlikely to be of practical consequence. These results suggest that change in the perceptions of one's teaching that are influenced by underlying beliefs (*26*) occurs slowly. Frequently, teachers with more instructional training score significantly higher on the CCSF subscale of the ATI and have higher self-efficacy compared with teachers with less training [for example, (*27*)]. Furthermore, the impact of instructional training is often greater on the CCSF subscale compared with the ITTF subscale [for example, (*26*, *28*)]. For example, Gibbs and Coffey (*2*) found that, 1 year after instructional training, treatment teachers had a significantly higher CCSF score compared with control teachers, but that there was no difference in ITTF scores. The absence of significant differences in CCSF and ITTF scores between FIRST and non-FIRST faculty in our study may be a function of the duration of training and the current pedagogical/teaching activities of participants. Studies of the impacts of instructional training on instructional strategies (*27*, *28*) showed that changes in CCSF scores did not increase linearly with pedagogical training and only increased to a level that differed from untrained faculty after a year of coursework on learning and instruction in higher education. Thus, the training of FIRST faculty may not have been sufficient to increase their support for CCSF teaching strategies to a level above that of their non-FIRST colleagues. Alternatively, support for learner-centered strategies may decline once participants are actually "in the classroom" following training [for example, (*28*, *29*)].

To better contextualize the perceptions of FIRST faculty about their teaching, we examined data from the literature. Addy and Blanchard

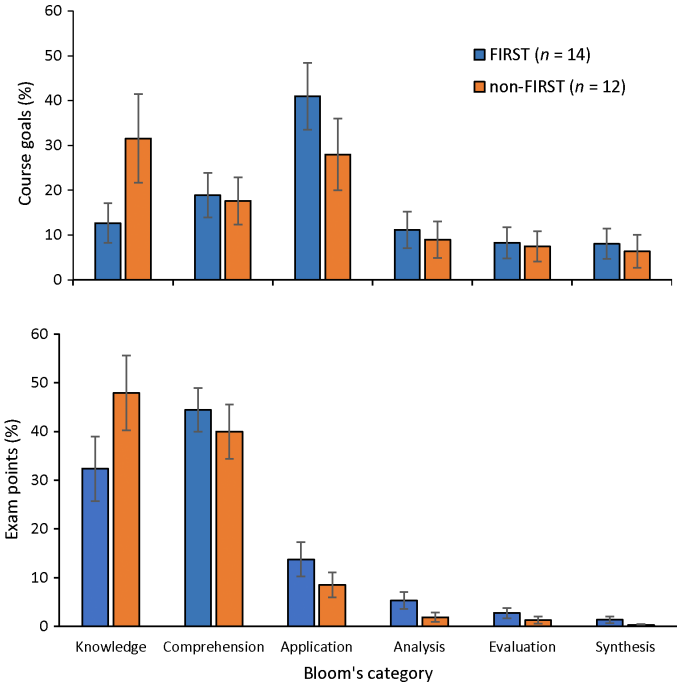

Fig. 3. Distribution of Bloom category levels for goals and assessments. Percentage (mean ± SE) of course goals as stated in the course syllabi (**top**) and assessment points per course (**bottom**) that were categorized into each Bloom category.

(*30*) reported ATI scores for science faculty with educational specialties. They calculated average raw scores for the CCSF and ITTF subscales of the ATI, which we express here as the percentage of total possible points for each subscale (that is, raw score/highest possible score). For the CCSF subscale, the scores for FIRST faculty, non-FIRST faculty, and faculty with educational specialties were similar, ranging from 70 to 78% of the highest possible score. However, for the ITTF subscale, the scores for our participants (FIRST faculty, 73%; non-FIRST faculty, 69%) were markedly higher than the scores for science faculty with education specialties (mean score, 50%). Our results were consistent with those of previous studies, which showed that belief in a teacher-centered approach to learning is stable and more difficult to change compared with a learner-centered approach (*2*, *26*, *28*, *29*).

Teaching practice

We found that the FIRST and non-FIRST faculty taught differently despite similarities in teaching background and perceptions of teaching practices. We measured several facets of faculty teaching and found differences particularly among faculty pairs' self-reported use of teaching methods (Fig. 1) and ratings from independent observers of actual classroom practice (Fig. 2), as well as student reports of the teaching practices used. FIRST faculty were more likely to use a learner-centered approach in the classroom and reported greater use of interactive activities to support learning, with the latter further confirmed by student reports. In particular, students in courses taught by FIRST faculty reported more elements of a learner-centered learning environment and a richer learning community. Although we do not have direct evidence of students' accomplishments, empirical evidence shows that students' positive perceptions of the learning environment have a positive effect on learning [for example, (*31*)]. Furthermore, high-impact practices, such as those used by FIRST faculty, were demonstrated to more effectively engage students in the learning process and to result in more positive outcomes for students (*32*).

The most striking difference between faculty pairs came from direct observations of teaching practice (Fig. 2): participation in FIRST was a strong predictor of RTOP score (Table 2). These results demonstrated that the FIRST program of professional development resulted in the successful transfer of faculty teaching practices that differed significantly from those of well-matched colleagues. Indeed, the range and mean scores of the FIRST faculty were virtually identical with those of science faculty trained in educational specialties (*30*). The RTOP scores of the non-FIRST faculty were very similar to those of faculty from two other professional development programs (*18*), suggesting that even faculty who are not involved in a formal instructional training program incorporate some level of student engagement in their teaching. The differences in the degree to which FIRST and non-FIRST faculty implemented learner-centered teaching may have been due, in part, to differences in their perceptions of challenges to such implementation. The pairs only differed in their perception of cooperation from faculty outside their department and support staff as a challenge. Nevertheless, because of their training, FIRST faculty may have perceived themselves as being less in need of external sources of support, and thus, more readily used learner-centered approaches.

Most of the faculty pairs recruited for this study were appointed at teaching-intensive institutions. As a result, we might predict that teaching excellence is an important evaluation criterion for both members of a faculty pair at these institutions, meaning that our comparison group may already be highly motivated to use effective pedagogy. The differences we detected then are even more striking. Moreover, course enrollment had no significant predictive value for the implementation of learner-centered teaching.

Assessment of student learning

Despite marked differences in the teaching practices of FIRST and non-FIRST faculty, the training that FIRST participants received on alignment of assessment with teaching practices was not implemented. That result was consistent with outcomes from the FIRST participants during training (*18*) and from faculty in a previous FIRST program (*33*). One explanation for the lack of difference in assessment practices is that beliefs about student learning comprise core beliefs that are resistant to change [for example, (*29*)]. In addition, large course size and time required for grading assessments might lead faculty to use assessments that focus largely on lower-level cognition, such as multiple-choice tests. That possibility was supported by a significant, but weak, negative correlation between course size and the average Bloom score for assessments used in courses ($n = 26$; Pearson $r = -0.428$, $P = 0.033$). Regardless of the explanation, collectively, our results indicated that alignment of assessments with the types of learning that students practice in a course was not readily accomplished by FIRST participants.

Implications

Our study demonstrated the empirical utility of comparison groups, where groups are well matched and control for contextual variables (for example, departments) when evaluating the effectiveness of professional development for subsequent beliefs and behaviors. The success of a program is not measured along a single axis; as our results demonstrate, some aspects of teaching are more difficult to change than others.

Results from FIRST can be used formatively to inform the design of future professional development in STEM disciplines, particularly with regard to addressing instructors' beliefs about how students learn and perceptions of instructional practices and designing assessments. In part, our results suggest that developing alignment between learning objectives and assessments is challenging for many instructors, and that professional development programs that place greater emphasis on assessment design may be more effective in achieving this goal.

How to scale projects nationally and how to sustain them over time are key questions about professional development programs that investigators must address (*8*, *9*). By assessing the transfer of professional development in the long run, we can learn more about its underlying dynamics and the factors that impact the long-term use of trained knowledge and skills (*6*). Although the transfer process and behavioral change continue to unfold years after training and development activities, researchers tend to investigate transfer soon after training is completed (*34*). We understand little about the longer-term transfer process, particularly the persistence of training after the training experience (*6*, *35*). Specifically, we need longitudinal studies to provide evidence that the impact of professional development is maintained, or even compounded, over time (*10*). Our study demonstrated that teaching practices developed during FIRST transferred into the initial period of a faculty appointment, but the data only represent a snapshot in time. We know little about how those behaviors may change—positively or negatively—as early-career STEM faculty continue to establish themselves as teachers and scholars. Organizational systems and contextual factors may be important to the transfer process: institutional and departmental cultures, pretenure expectations, incentives, and faculty professional identity are variables that might influence the long-term transfer and longitudinal impacts of teaching professional development (*6*, *8*, *36*–*38*).

Limitations and next steps

This study is limited in several ways. First, the study population was not a random sample—rather, individuals represented FIRST participants who both had begun a faculty position during the term of the study and could find a suitably matched, and willing, non-FIRST faculty member in their departments. Although the study pairs represented four types of institutions (research, comprehensive, liberal arts, and community college), it is important to avoid overgeneralization given the relatively small sample size (≤20 pairs). Data from students were incomplete in several instances, thereby limiting our ability to analyze students' perceptions and measures of their performance. Furthermore, this study was completed with a group of faculty immediately following their participation in professional development; therefore, we cannot make claims about how FIRST faculty continue to teach beyond the first few years of a faculty position. Next steps must include broadening the scope to address longer-term transfer of training and the contextual factors at the department and institution levels that may catalyze or inhibit successful outcomes for faculty professional development programs. Objective measures of student learning and skills are also needed, as these are ultimately the outcomes that we strive to improve through faculty professional development.

MATERIALS AND METHODS

Study design

We determined whether the perceptions of teaching and teaching practices of faculty who participated in FIRST differed from those of faculty in the same department who did not participate. We used a paired quasi-experimental design and comprehensive assessment plan to address our research question. We designed our research to test the null hypothesis that there were no differences in the perceptions of teaching, teaching practices used, and levels of cognition targeted by course assessments between FIRST participants and non-FIRST faculty.

We recruited 20 faculty members based on the following: (i) employment in a full-time, tenure-track position soon after completing the FIRST program (that is, 2011 for cohort 1 and 2013 for cohort 2); (ii) willingness to complete assessments required for the study; and (iii) availability of a colleague who was willing to participate and to complete the required assessments. We evaluated the equivalency of the 20 recruited faculty and the nonrecruited FIRST participants by comparing key outcomes that were measured through an end-of-project survey completed at the end of participation in FIRST. These comparisons indicated that faculty in our paired study were representative of other FIRST participants (Table 3). Both groups reported similar knowledge of and experience with active-learning pedagogy and practices and level of confidence in teaching (Wilcoxon rank sum test, $P > 0.05$ for all variables), and demonstrated similar levels of student engagement in the classroom (that is, RTOP score; t test, $t = -1.32$, $P = 0.19$) upon completion of FIRST. Both groups averaged less than a year of teaching experience upon entry to the FIRST program. The percentage of female participants was slightly lower in our paired study, but gender was not a significant factor in teaching practice (that is, RTOP score).

Faculty pairs were established by matching a FIRST participant who was employed in a tenure-track position with a non-FIRST tenure-track faculty member at the same institution. Power analyses were conducted using G*Power 3.1.9.2, with a calculated effect size of ~1 and an α of 0.05. The results indicated that we needed 9 to 11 faculty pairs to determine statistical differences in outcomes from our measures of teaching practice (that is, the RTOP) and surveys completed by the participants. We established 20 faculty pairs (six in 2010–2011 and the remainder during the 2012–2013 academic year) to ensure that we obtained complete data sets from at least 11 faculty pairs. Each pair was employed at a different institution. Lack of baseline data on the equivalence of comparison groups is a common pitfall in quasi-experimental designs in education (*39*). Therefore, we used multiple variables to ensure that our two faculty groups were comparable. To determine similarities and differences between FIRST and non-FIRST faculty at the beginning of their participation in our study, we asked each participant to complete a Teaching Background Survey [Appendix S1; available as supplemental information in Ebert-May *et al.* (*18*)] that we developed. The Teaching Background Survey provides information about the participants' prior teaching experience, prior professional development activities, number of courses taught, perception of departmental support, and theoretical knowledge of, and experience with, active-learning pedagogy and teaching strategies. All but two of the non-FIRST faculty had five or fewer years of teaching. Members of each pair taught courses that were as similar as possible in topic, enrollment, and level. Most (78%) taught an introductory-level course. One-third of the pairs taught the same course; the remainder taught matched courses.

To determine whether the teaching practices of FIRST participants were more learner-centered than those of non-FIRST faculty, we compared faculty perceptions of teaching strategies, perceptions of environmental factors that influence teaching, and actual teaching practices at the end of a semester. Information about the instruments used for these and other assessment purposes is presented in Table 4. Faculty

Table 3. Comparison of FIRST participants who were selected for this paired study with nonselected FIRST participants upon completion of the FIRST program. Sample size is given in parentheses. Teaching experience is expressed as years before participation in FIRST. Knowledge of and experience with active learning and teaching confidence were calculated as summed Likert responses to survey questions identical to those in the Teaching Background Survey (Appendix S1). The RTOP score refers to the average RTOP score for videos of classroom teaching. Data are presented as mean ± SE.

Group	Female participants (%)	Teaching experience (years)	Active-learning knowledge	Active-learning experience	Teaching confidence	RTOP score
Paired study	55 (n = 20)	0.44 ± 0.1 (17)	42.2 ± 2.9 (18)	40.6 ± 2.4 (18)	3.0 ± 0.2 (18)	49.6 ± 1.4 (20)
Nonparticipants	66 (n = 180)	0.73 ± 0.1 (141)	43.4 ± 0.9 (108)	37.5 ± 0.8 (108)	3.1 ± 0.1 (108)	46.0 ± 0.7 (150)

perceptions were characterized using the ATI (*19*, *20*); Experience of Teaching Questionnaire (ETQ) (*40*), a four-item Self-Efficacy Survey from Lindblom-Ylänne *et al.* (*26*); and the Teaching Practice Survey (Appendix S2), which we designed. The ATI measures qualitative variation in two key dimensions of teaching, specifically CCSF and ITTF. Instructors who use an ITTF approach see their roles as mainly to transmit information to students and to focus on the development of skills that improve competency in information transfer. Instructors who use a CCSF approach aim to change students' thinking about the material studied, with a focus on ways to challenge students' current ideas so that students construct their own knowledge. The two dimensions are independent, rather than being ends of a continuum (*40*). Use of the ATI is context-specific; thus, faculty participants completed the ATI at the end of the course that they taught. A score for each subscale (CCSF and ITTF) was calculated for each participant.

Faculty perceptions of their teaching environment were measured using the ETQ, which consists of five subscales characterizing environmental factors that are likely to influence teaching practice: course size, heterogeneity of students (that is, ability, preparation, language, and other skills), degree to which the faculty members controlled what they taught and how it was taught, department's commitment to student learning, and workload. We calculated subscale scores for each participant. For faculty teaching in 2012, we also determined the frequency with which they reported using active learner-centered practices in their courses using the Teaching Practice Survey.

Classroom teaching practice was characterized using data from two sources: teaching observations completed by experts in biology education and students' perceptions of teaching in courses taught by the faculty participants. Participants submitted a video recording of at least one complete class session of the course that they taught during the study; 86% of participants submitted videos for two class sessions. Videos were recorded as described in Ebert-May *et al.* (*12*). Evaluation of the videos was conducted using the same assessment procedure that was used in the FIRST project (*18*). Expert reviewers rated each video recording using the RTOP. Sawada (*21*) designed the RTOP to measure the extent of "reformed teaching" used in the classroom. The RTOP is a validated observational instrument that focuses on the nature of student learning and is aligned with constructivist theories about teaching and learning (*22*, *23*, *41*, *42*). Across institutions and users, the RTOP is a highly consistent instrument for both item and interrater reliability (*42*, *43*), and RTOP scores have positive correlations with student learning gains (*44*–*46*). Details of RTOP subcategories and score interpretations are explained fully in Budd *et al.* (*47*). We trained and calibrated biology education experts in the use of the RTOP. During the initial calibration, all potential reviewers (n = 18) watched a set of 8 to 14 videos, followed by a discussion of their RTOP scores. Upon completion of the initial calibration, reviewers who had an intraclass correlation coefficient (ICC) (*48*) of at least 0.7 and the time to actually review videos were selected as the final pool of reviewers (n = 13). Videos from the study reported here were reviewed within a much larger pool of videos from the FIRST project. A randomly selected video from the larger pool was assigned to all reviewers each month, without their knowledge of its purpose, to monitor their calibration. Interrater reliability of the reviewers was determined each month using the cumulative monthly ICC. The average ICC for the total review period was 0.71 (range, 0.46 to 0.85).

Each video from our participants was reviewed by two experts who did not know the instructor in the video and did not know that the video resulted from our paired study. If the scores from the two reviewers were not in agreement, then further expert reviews were completed until similar scores were achieved. Final expert scores were averaged to obtain a final total RTOP score for each video. When two videos were submitted for a course, the final scores were averaged and their mean was used for the analyses. The average difference in RTOP score between videos for a single course was very small (0.86 ± 1.6). Further details of the review process are provided in Ebert-May *et al.* (*18*).

Students' perceptions of teaching by FIRST and non-FIRST faculty were examined using the ETLQ (*25*), which consists of four sections: (i) student approaches to learning and studying, (ii) students' perceptions of the teaching-learning environment in a course, (iii) students' perceptions of the demands made on them by the course, and (iv) students' perceptions of the learning that they achieved in the course. Students completed the ETLQ at the end of a course. To determine whether students in FIRST and non-FIRST courses were similar, we asked the students to complete the LSQ (*25*) at the beginning of the course. The LSQ consists of three sections that focus on (i) reasons for taking the degree program, (ii) reasons for taking the particular course, and (iii) student approaches to learning and studying.

During professional development, FIRST participants learned to engage students in scientific practices that required higher cognitive thinking (for example, creating and testing models, constructing arguments, and working with data) to learn concepts. Therefore, we predicted that, as faculty, the FIRST participants would use assessments that incorporated questions requiring higher-order thinking. We compared the assessments of faculty pairs by determining the level of cognitive skills that were targeted in their high-stakes assessments (that is, exams and

Table 4. Characteristics of instruments used in the paired study. Sample size refers to the number of faculty participants who submitted complete data for an instrument.

Instrument	Acronym	Purpose	When used	Items and subscales	Response type	Sample size
Teaching Background Survey (Appendix S1)	BK	Participants' confidence, knowledge of and experience with teaching and pedagogy	Beginning of the study	79 items	Ordinal, written	18 pairs
Approaches to Teaching Inventory (*19*, *20*)	ATI	Participants' perceptions of teaching strategies used in a course	Beginning of the course	22 items, 2 subscales	Likert	20 pairs
Self-Efficacy Survey (*26*)	SE	Participants' confidence in their teaching ability	Beginning of the course	4 items	Likert	20 pairs
Teaching Practice Survey (Appendix S2)	TPS	Participants' perceived use of different classroom teaching practices and approaches to assessment	End of the course	30 items	Likert	11 pairs
Experience of Teaching Questionnaire (*40*)	ETQ	Participants' perceptions of environmental factors that are likely to influence teaching practices	End of the course	32 items, 5 subscales	Likert	20 pairs
Learning and Studying Questionnaire (*25*)	LSQ	Students' perceptions of their reasons for taking the course and approaches to learning and studying	Beginning of the course	56 items, 3 sections, 5 subscales	Likert	13 pairs
Experiences of Teaching and Learning Questionnaire (*25*)	ETLQ	Students' perceptions of course demands and learning achieved, the teaching-learning environment, and their approaches to learning and studying	End of the course	77 items, 4 sections	Likert	11 paired courses
Reformed Teaching Observation Protocol (*21*)	RTOP	Expert ratings of the extent to which learner-centered teaching practices are used in a class	End of the study	25 items, 5 subscales	Ordinal	19 FIRST faculty, 17 non-FIRST faculty

quizzes). We classified the cognitive skills assessed by each quiz/exam question using Bloom's taxonomy (*49*). Bloom's taxonomy is composed of six cognitive skill levels that represent a continuum from simple to complex cognitive tasks: (i) knowledge, (ii) comprehension, (iii) application, (iv) analysis, (v) synthesis, and (vi) evaluation. The first two categories typically describe lower-order cognitive skills, and the latter four categories describe higher-order cognitive skills (*50*). Each assessment item on the quizzes and tests was assigned a cognitive skill level by two independent raters who had achieved a Cohen's κ of 0.87 ($n = 109$ assessments). We calculated the percentage of points on each quiz or exam that was assigned in each Bloom category [for example, (25 points Bloom category 1/80 points total) × 100] and averaged the values within each Bloom category for all assessments used in a course.

Statistical analysis

In analyses of faculty, faculty participants were used as the experimental unit. Because not all faculty completed all assessments, the sample size varied among the analyses used. In particular, for paired analyses, only pairs for which data were available from both members could be included. In paired statistical tests, for each pair, the response or score of the non-FIRST faculty was subtracted from that of the FIRST faculty. The differences were tested for normality. We then tested whether the calculated difference between the pairs of observations was significantly different from zero, using paired-sample Wilcoxon signed-rank test or Student's *t* test, as determined by the outcome of tests of normality. The effect of participation in FIRST on the RTOP score was analyzed using linear regression analysis. We examined the data using outlier and leverage diagnostics, as well as fit diagnostics, to ensure alignment with the assumptions for linear regression. Coded variables were used for treatment (0 = non-FIRST; 1 = FIRST faculty). Instructor gender (coded variable; 0 = male, 1 = female) and class size were included as covariates because of their potential effect on RTOP score (*47*).

The cognitive levels of course goals that were stated on faculty syllabi, as well as the assessment questions (that is, quizzes and tests) used by faculty, were determined by categorizing goals and assessment questions using Bloom's taxonomy (*49*). We tested for differences in the number of course goals in each Bloom category between FIRST and non-FIRST faculty using logistic regression. Likewise, we tested for differences in the percentage of assessment points allocated to each Bloom category using logistic regression.

Analyses of student responses to surveys were conducted with course as the sample unit. Accordingly, student responses on the sections of the LSQ and the ETLQ were averaged for each course. We tested whether the calculated differences in mean scores for each faculty pair differed significantly from zero, using Student's *t* test or paired-sample Wilcoxon signed-rank test, after testing for normality. Two-sided testing was used throughout. All statistical analyses were conducted using SAS version 9.4 (SAS, Cary, NC). Results of statistical analyses were considered to be significant at $P < 0.05$. All protocols used in the FIRST project were approved by the Michigan State University Institutional Review Board (X08-550 exempt, category 2).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/2/3/e1501422/DC1

Appendix S1. Paired Teaching Background Survey.

Appendix S2. Teaching Practice Survey.

REFERENCES AND NOTES

1. M. S. Garet, A. C. Porter, L. Desimone, B. F. Birman, K. S. Yoon, What makes professional development effective? Results from a national sample of teachers. *Am. Educ. Res. J.* **38**, 915–945 (2001).
2. G. Gibbs, M. Coffey, The impact of training of university teachers on their teaching skills, their approach to teaching and the approach to learning of their students. *Active Learn. Higher Educ.* **5**, 87–100 (2004).
3. C. Henderson, A. Beach, N. Finkelstein, Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature. *J. Res. Sci. Teach.* **48**, 952–984 (2011).
4. C. Henderson, M. Dancy, M. Niewiadomska-Bugaj, Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process? *Phys. Rev. Phys. Educ. Res.* **8**, 020104 (2012).
5. C. Amundsen, M. Wilson, Are we asking the right questions? A conceptual review of the educational development literature in higher education. *Rev. Educ. Res.* **82**, 90–126 (2012).
6. S. Yelon, J. K. Ford, S. Bhatia, How trainees transfer what they have learned: Toward a taxonomy of use. *Perform. Improv. Q* **27**, 27–52 (2014).
7. L. M. Desimone, Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educ. Res.* **38**, 181–199 (2009).
8. E. Seymour, K. DeWelde, C. Fry, in *A White Paper Commissioned for the Forum "Characterizing the Impact and Diffusion of Engineering Education Innovations"* (National Academy of Engineering, Washington, DC, 2011), 30 pp.
9. National Research Council, in *Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education* (National Academies Press, Washington, DC, 2012), 327 pp.
10. S. H. Shaha, K. F. Glasset, H. Ellsworth, Long-term impact of on-demand professional development on student performance: A longitudinal multi-state study. *J. Int. Educ. Res.* **11**, 29–34 (2015).
11. K. A. Ericsson, Deliberate practice and acquisition of expert performance: A general overview. *Acad. Emerg. Med.* **15**, 988–994 (2008).
12. D. Ebert-May, T. L. Derting, J. Hodder, J. L. Momsen, T. M. Long, S. E. Jardeleza, What we say is not what we do: Effective evaluation of faculty professional development programs. *Bioscience* **61**, 550–558 (2011).
13. L. Fung, L. P. Y. Chow, Congruence of student teachers' pedagogical images and actual classroom practices. *Educ. Res.* **44**, 313–321 (2002).
14. R. Kane, S. Sandretto, C. Heath, Telling half the story: A critical review of research on the teaching beliefs and practices of university academics. *Rev. Educ. Res.* **72**, 177–228 (2002).
15. K. Murray, R. Macdonald, The disjunction between lecturers' conceptions of teaching and their claimed educational practice. *Higher Educ.* **33**, 331–349 (1997).
16. A. Stes, M. Min-Leliveld, D. Gijbels, P. Van Petegem, The impact of instructional development in higher education: The state-of-the-art of the research. *Educ. Res. Rev.* **5**, 25–49 (2010).
17. J. Handelsman, S. Miller, C. Pfund, *Scientific Teaching* (W. H. Freeman, New York, NY, 2006).
18. D. Ebert-May, T. L. Derting, T. P. Henkel, J. M. Maher, J. L. Momsen, B. Arnold, H. A. Passmore, Breaking the cycle: Future faculty begin teaching with learner-centered strategies after professional development. *CBE Life Sci. Educ.* **14**, ar22 (2015).
19. K. Trigwell, M. Prosser, Development and use of the Approaches to Teaching Inventory. *Educ. Psychol. Rev.* **16**, 409–424 (2004).
20. K. Trigwell, M. Prosser, P. Ginns, Phenomenographic pedagogy and a revised *Approaches to teaching inventory*. *Higher Educ. Res. Dev.* **24**, 349–360 (2005).
21. D. Sawada, *Reformed Teacher Education in Science and Mathematics: An Evaluation of the Arizona Collaborative for Excellence in the Preparation of Teachers* (Arizona State University Document Production Services, Tempe, AZ, 2003).
22. M. D. Piburn, D. Sawada, K. Falconer, J. Turley, R. Benford, I. Bloom, "Reformed Teaching Observation Protocol (RTOP). ACEPT IN-003. The RTOP rubric form, training manual and reference manual containing statistical analyses" (2000); available at http://PhysicsEd.BuffaloState.Edu/AZTEC/rtop/RTOP_full/PDF/.
23. D. Sawada, M. D. Piburn, E. Judson, J. Turley, K. Falconer, R. Benford, I. Bloom, Measuring reform practices in science and mathematics classrooms: The Reformed Teaching Observation Protocol. *School Sci. Math.* **102**, 245–253 (2002).
24. J. M. Maher, J. C. Markey, D. Ebert-May, The other half of the story: Effect size analysis in quantitative research. *CBE Life Sci. Educ.* **12**, 345–351 (2013).
25. N. Entwistle, V. McCune, J. Hounsell, *Approaches to Studying and Perceptions of University Teaching-Learning Environments: Concepts, Measures and Preliminary Findings. Enhancing Teaching-Learning Environments in Undergraduate Courses Project* (University of Edinburgh, Edinburgh, UK, 2002).
26. S. Lindblom-Ylänne, K. Trigwell, A. Nevgi, P. Ashwin, How approaches to teaching are affected by discipline and teaching context. *Stud. Higher Educ.* **31**, 285–298 (2006).
27. L. Postareff, S. Lindblom-Ylänne, A. Nevgi, The effect of pedagogical training on teaching in higher education. *Teach. Teacher Educ.* **23**, 557–571 (2007).
28. L. Postareff, S. Lindblom-Ylänne, A. Nevgi, A follow-up study of the effect of pedagogical training on teaching in higher education. *Higher Educ.* **56**, 29–43 (2008).
29. S. S. Fletcher, J. A. Luft, Early career secondary science teachers: A longitudinal study of beliefs in relation to field experiences. *Sci. Educ.* **95**, 1124–1146 (2011).
30. T. M. Addy, M. R. Blanchard, The problem with reform from the bottom up: Instructional practices and teacher beliefs of graduate teaching assistants following a reform-minded university teacher certificate programme. *Int. J. Sci. Educ.* **32**, 1045–1071 (2010).
31. A. Lizzio, K. Wilson, R. Simons, University students' perceptions of the learning environment and academic outcomes: Implications for theory and practice. *Stud. Higher Educ.* **27**, 27–52 (2002).
32. S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, M. P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8410–8415 (2014).
33. J. L. Momsen, T. M. Long, S. A. Wyse, D. Ebert-May, Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills. *CBE Life Sci. Educ.* **9**, 435–440 (2010).
34. B. D. Blume, J. K. Ford, T. T. Baldwin, J. L. Huang, Transfer of training: A meta-analytic review. *J. Manage.* **36**, 1065–1105 (2010).
35. S. W. Stirman, J. Kimberly, N. Cook, A. Calloway, F. Castro, M. Charns, The sustainability of new programs and innovations: A review of the empirical literature and recommendations for future research. *Implement. Sci.* **7**, 17 (2012).
36. J. J. Walczyk, L. L. Ramsey, P. Zha, Obstacles to instructional innovation according to college science and mathematics faculty. *J. Res. Sci. Teach.* **44**, 85–106 (2007).
37. A. E. Austin, *Promoting Evidence-Based Change in Undergraduate Science Education* (a paper commissioned by the National Academies National Research Council Board on Science Education, Washington, D.C., 2011).
38. S. E. Brownell, K. D. Tanner, Barriers to faculty pedagogical change: Lack of training, time, incentives, and… tensions with professional identity? *CBE Life Sci. Educ.* **11**, 339–346 (2012).
39. K. S. Yoon, T. Duncan, S. W.-Y. Lee, B. Scarloss, K. Shapley, *Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement, Issues & Answers Report, REL 2007-No. 033* (U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest, Washington, DC, 2007); http://ies.ed.gov/ncee/edlabs.
40. M. Prosser, K. Trigwell, Relations between perceptions of the teaching environment and approaches to teaching. *Br. J. Educ. Psychol.* **67**, 25–35 (1997).
41. D. MacIsaac, K. Falconer, Reforming physics instruction via RTOP. *Phys. Teach* **40**, 479–485 (2002).
42. J. C. Marshall, J. Smart, C. Lotter, C. Sirbu, Comparative analysis of two inquiry observational protocols: Striving to better understand the quality of teacher-facilitated inquiry-based instruction. *School Sci. Math.* **111**, 306–315 (2011).
43. A. Amrein-Beardsley, S. E. Osborn Popp, Peer observations among faculty in a college of education: Investigating the summative and formative uses of the Reformed Teaching Observation Protocol (RTOP). *Educ. Assess. Eval. Account.* **24**, 5–24 (2012).
44. K. Falconer, S. Wyckoff, M. Joshua, D. Sawada, *Annual Conference of the American Educational Research Association, Technical Report No. C01-4* (American Educational Research Association, Seattle, WA, 2001).
45. A. Lawson, R. Benford, I. Bloom, M. Carlson, K. Falconer, D. Hestenes, E. Judson, M. Piburn, D. Sawada, J. Turley, S. Wyckoff, Evaluating college science and mathematics instruction: A reform effort that improves teaching skills. *J. Coll. Sci. Teach.* **31**, 388–393 (2002).
46. B. V. Bowling, C. A. Huether, L. Wang, M. F. Myers, G. C. Markle, G. E. Dean, E. E. Acra, F. P. Wray, G. A. Jacob, Genetic literacy of undergraduate non-science majors and the impact of introductory biology and genetics courses. *Bioscience* **58**, 654–660 (2008).
47. D. A. Budd, K. J. van der Hoeven Kraft, D. A. McConnell, T. Vislova, Characterizing teaching in introductory geology courses: Measuring classroom practices. *J. Geosci. Educ.* **61**, 461–475 (2013).
48. K. L. Gwet, *How to Compute Intraclass Correlation With MS EXCEL: A Practical Guide to Inter-Rater Reliability Assessment for Quantitative Data* (Advanced Analytics LLC, Gaithersburg, MD, 2010).
49. B. S. Bloom, Ed., *Taxonomy of Educational Objectives: The Classification of Educational Goals* (D. McKay, New York, NY, 1956).
50. L. W. Anderson, D. R. Krathwohl, Eds., *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives (Complete Edition)* (Longman, New York, NY, 2001).

Acknowledgments: We are appreciative of the faculty who participated in the study. We thank the experts who reviewed the video recordings. We also thank C. J. Mecklin for his statistical advice. We appreciate the comments of two anonymous reviewers, which resulted in significant improvements to our manuscript. **Funding:** The research was funded by the NSF under the Division of Undergraduate Education Award 08172224 (to D.E.-M. and T.L.D.).

Author contributions: T.L.D. and D.E.-M. conceived and designed this work. T.L.D., D.E.-M., T.P.H., J.M.M., B.A., and H.A.P. contributed to the acquisition, analysis, and/or interpretation of data and to the critical review of writing for important intellectual content. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors. Data reported in this paper are archived at the Dryad Digital Repository at http://dx.doi.org/10.5061/dryad.186m0.

Submitted 10 October 2015
Accepted 30 January 2016
Published 18 March 2016
10.1126/sciadv.1501422

Citation: T. L. Derting, D. Ebert-May, T. P. Henkel, J. M. Maher, B. Arnold, H. A. Passmore, Assessing faculty professional development in STEM higher education: Sustainability of outcomes. *Sci. Adv.* **2**, e1501422 (2016).

Assessing faculty professional development in STEM higher education: Sustainability of outcomes

Terry L. Derting, Diane Ebert-May, Timothy P. Henkel, Jessica Middlemis Maher, Bryan Arnold and Heather A. Passmore

Sci Adv **2** (3), e1501422.
DOI: 10.1126/sciadv.1501422