International Journal for Academic Development

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rija20

Understanding the impact of educational development interventions on classroom instruction and student success

Lindsay B. Wheeler & Dorothe Bach

To cite this article: Lindsay B. Wheeler & Dorothe Bach (2020): Understanding the impact of educational development interventions on classroom instruction and student success, International Journal for Academic Development, DOI: 10.1080/1360144X.2020.1777555

To link to this article: https://doi.org/10.1080/1360144X.2020.1777555

INTERNATIONAL JOURNAL FOR ACADEMIC DEVELOPMENT
https://doi.org/10.1080/1360144X.2020.1777555

Understanding the impact of educational development interventions on classroom instruction and student success

Lindsay B. Wheeler and Dorothe Bach

Center for Teaching Excellence, University of Virginia, Charlottesville, VA, USA

ABSTRACT
This study explored three US educational development (ED) programs: a weeklong course design institute, a new faculty learning community (NFLC), and a STEM learning community (STEM-LC). We compared observed instruction and student achievement for 239 STEM undergraduate courses taught by instructors who had or had not engaged in ED. Courses taught by NFLC and STEM-LC instructors had significantly more learning-focused syllabi and active learning than courses taught by non-engaged instructors, controlling for class size and type. We conclude that instructors need support in implementing active learning to ensure all students benefit. Additional research is needed to explore ED and active learning.

ARTICLE HISTORY
Received 26 April 2019
Accepted 9 March 2020

KEYWORDS
Program assessment; student achievement; teaching development; undergraduate STEM

Introduction

Tracing the impact of educational development (ED) from intervention to student learning has historically been difficult (Beach et al., 2016; Chism et al., 2012). A growing number of studies attempt to move beyond program participation and satisfaction (Condon et al., 2016; Meizlish et al., 2018; Tomkin et al., 2019). However, these studies are limited to demonstrating changes at the student level (for example, they often measure only practice, or use self-reported data). The present study uses a novel approach to directly measure the impact of ED programs. Using classroom observation and student success data collected at a US institution, it compares courses taught by Science, Technology, Engineering and Mathematics (STEM) instructors who have participated in ED programs with courses taught by instructors who have not participated. 'Instructors' refers to any individual who serves in the formal role of instructor, whether graduate student or staff/faculty member.

Undergraduate instruction

Both in the US and internationally, research studies continue to demonstrate the importance of instructor-student and student-student engagement for student success (Freeman et al., 2014; Kuh et al., 2008). In a review of meta-analyses, Schneider and

CONTACT Lindsay B. Wheeler Lsb4u@virginia.edu
Supplemental data for this article can be accessed at https://doi.org/10.1080/1360144X.2020.1777555.

Preckel (2017) found that student achievement is highest in courses that utilize small group work, a finding also found in some STEM education literature (Freeman et al., 2014). However, a recent study confirmed that STEM undergraduate courses continue to rely heavily on lecture as primary mode of instruction (Stains et al., 2018).

In the US, which served as the context for the present study, STEM fields are focusing their attention on the academic success of marginalized students (President's Council of Advisors on Science and Technology [PCAST], 2012) to address gaps in achievement and retention for students from these under-represented minority groups (URM students) compared to non-URM students (Daempfle, 2003; Gasiewski et al., 2012; National Science Foundation, National Center for Science and Engineering Statistics, 2017). Because active and collaborative learning has the potential to reduce the achievement gap (Haak et al., 2011), US ED efforts with STEM faculty are largely aimed at supporting instructors in learning about and implementing such practices in their undergraduate classrooms.

Educational development interventions

This study explores the impact of three educational development programs at a US research university, including: a week-long course design institute (CDI), a new faculty learning community (NFLC), and a STEM learning community (STEM-LC). A 'faculty learning community' (FLC) is a type of community of practice (CoP), but while FLCs are typically intentionally formed within an institution and have a defined duration (Kezar & Gehrke, 2017), CoPs – particularly in higher education STEM education – arise from a community around a shared idea or interest, and are sustained over time (for example, discipline-based education communities). CDI is the initial module that instructors complete before participating in NFLC and STEM-LC. We briefly describe the format, characteristics, and goals of each of these.

Course design institute (CDI)

Course design institutes are widely adopted in educational development. Our university's CDI is a week-long intensive workshop (35 contact hours) that, like many others, draws on ideas from backward and integrated course design (Fink, 2013), educative assessment (Wiggins, 1998), active learning (Bonwell & Eison, 1991), student motivation (Schunk et al., 2007), and transparent assignments (Winkelmes et al., 2016) to support instructors in designing learning-focused courses. At the time of the present study, 508 individuals had participated in CDI, 144 of whom were STEM instructors.

During CDI, instructors produce promising course syllabi (Bain, 2004) that compellingly communicate to students why they should care about material (relevance), what they will learn (goals and objectives), what students will do to achieve the objectives (transparent, equitable assessments aligned with those objectives), and how they will be supported to succeed in the course (environment, inclusivity). Throughout the week, instructors engage in small groups and are supported by an educational developer. Instructors also have opportunities to meet one-on-one with other instructors, educational developers, and trained undergraduate students to receive feedback on their course design and syllabus. Prior research on our CDI demonstrates its positive impact on instructors' self-reported understandings of and confidence with learning-focused teaching practices as well as shifts in course syllabi away from an exclusive focus on content (Palmer et al., 2016). The present study aimed to extend this research by comparing CDI participants to those who have not engaged in ED interventions.

New faculty learning community (NFLC)

Programs focused on new faculty are among the services most frequently offered in educational development (Beach et al., 2016) and research suggests that efforts focused on future and new faculty have the potential of being more impactful than those aiming to change the practices of more established colleagues (Ebert-May et al., 2011). As at many other institutions, the goal of this program was to help a new generation of faculty adopt more learner-focused teaching practices and become more reflective practitioners. Faculty enrolled in NFLC first complete the CDI described above (35 hours). They then participate in a semester- or year-long learning community consisting of a half-day retreat, seven 90-minute meetings led by ED staff, and an individual teaching consultation (16 hours total).

Unlike other early-career faculty programs that focus on instructional development though introductory workshops on a range of different and often disconnected teaching topics, the NFLC is tightly focused on supporting faculty in implementing the learning-focused courses they designed during CDI. NFLC engages participants in a cycle of deliberate classroom experimentation, analysis of, and reflection on their experience, and gives them the opportunity to provide and receive peer- and expert feedback before implementing a new teaching strategy. Meizlish et al. (2018) speculate that programs with more extended emphasis on course design, might be anticipated to have more impact. The present study focused on a subset of NFLC faculty from the 2015–2017 cohorts who teach in STEM disciplines to provide evidence to support, or refute, the speculations of Meizlish et al. (2018).

STEM learning community (STEM-LC)

Calls for reform in undergraduate STEM education (for example, PCAST, 2012) demonstrate the need for ED programs focused on STEM instructors, particularly for instructors teaching large-enrollment courses (Seymour & Hewitt, 1997). The goal of the STEM-LC was to support STEM instructors teaching large-enrolment courses. Engagement in the STEM-LC included CDI (35 hours) and a year-long learning community (14 hours), which involved seven 60-minute meetings run by ED staff, peer-observations of other STEM-LC participants, and an individual consultation with ED staff. During the meetings, members could learn from and encourage one another in the redesign and administration of their new courses. Topics and literature concerning learner-focused pedagogy were often discussed, and members of the cohort were able to discuss their own questions and experiences as part of the process.

While other ED programs have engaged instructors from a single department (Ebert-May et al., 2011), we embedded a cohort of six instructors from the same department within the 2015–2016 STEM-LC community (comprised of a total of 10 faculty across four STEM departments). This allowed for development of their department-level community of practice but also provided opportunities to engage with and develop cross-disciplinary relationships, which research suggests is one effective change strategy (Henderson et al., 2011). These six instructors from Department A were the focus for assessing the STEM-LC in the present study.

Educational development intervention assessments

Historically, educational developers have had limited capacity to study the effectiveness of their interventions (Kolomitro & Anstey, 2017), and educational researchers have traditionally shown limited interest in researching this work (Chism et al., 2012). More recently, a growing number of studies have systematically traced the impact of instructor participation in ED programs, which can be organized according to Kreber & Brooks' six-step impact model (Kreber & Brook, 2001). For the sake of brevity, we have collapsed these into three areas: instructor characteristics (for example, perceptions, beliefs), instructor performance (course design, instructional practice), and student outcomes (perceptions and learning). First, some studies have explored the impact of ED programs in improving instructors' knowledge of teaching practices, teaching beliefs, and teaching self-efficacy (Palmer et al., 2016). Second, positive changes in instructor performance following ED interventions have been documented through analyzing course syllabi, self-reported instructional practices, and classroom observations (Lauridsen & Lauridsen, 2018). Third, studies use student evaluations of teaching and self-reported learning gains to demonstrate the efficacy of ED programs (Condon et al., 2016; Meizlish et al., 2018).

Purpose

Despite the potential of using evidence-based practices in STEM undergraduate instruction, there is little research exploring the impact of ED programs on these courses. When exploring the ED literature, no study, to our knowledge, has sought to use direct measures of instructional practice and student outcomes to assess the impact of ED programs. Further, very few studies of ED programs compare outcomes of instructors who have and have not engaged in these programs. Thus the research questions for the study were:

What differences, if any, exist in the instructional practices and student achievement for STEM courses taught by instructors who have or have not:

(1) participated in ED interventions? (RQ1)
(2) participated in NFLC? (RQ2)
(3) participated as a departmental cohort in STEM-LC? (RQ3)

Methods

This quantitative study was conducted at a US mid-Atlantic, research-intensive university during the fall 2016 and spring 2017 semesters. Data included a 25% representative sample of all undergraduate STEM lecture and seminar courses, which included any credit-bearing course that was not considered a laboratory, studio, or independent study during the data collection period (see Supplemental Methods for further details of the participants, data collection, and analysis).

Participants and recruitment

All instructors-of-record who taught the 1006 undergraduate STEM course during the fall 2016 and spring 2017 semesters were emailed to voluntarily participate in the Institutional Review Board (IRB)-approved study. We aimed to obtain a representative instructor and course sample across all STEM disciplines. Participants included 150 of the 447 instructors (33.6%) (Table 1). We use the term ‘instructor’ to refer to all participants, ‘faculty’ for all non-graduate student instructors, and ‘professor’ for tenure-track, tenured, and full professors. Courses included 239 of the 984 courses (24.3%), taught across all four years across all STEM departments, a fairly representative sample of the population.

To assess the NFLC program specifically, we compared new faculty (those hired within three years of data collection) who had engaged in the program with those who had not. To assess the STEM-LC program, we focused on Department A, where a subset of six instructors participated in the program. These instructors comprised 37.5% of instructors (n = 16) teaching in Department A during the time when observations were conducted. A comparable ‘control’ department, Department B, included observed instructors (n = 16 of 51 total Department B instructors, or 31.4% study participation) who had not engaged in any ED interventions. Departments A and B were similar in disciplinary content, and student racial demographics; however, they differed in course level and instructor status. Department A instructors were mostly non-tenure track faculty teaching across all course levels, while Department B instructors were graduate students teaching in lower-level courses and tenure track/tenured teaching in upper-level courses.

Data collection and coding

Data were collected during the fall 2016 and spring 2017 semesters. The unit of analysis for the study was at the course level. Data sources included classroom observations,

Table 1. Instructor Demographics for Sample of STEM Undergraduate Courses.

		n (%)
Intervention	None	140 (58.6)
	CDI only	28 (11.7)
	NFLC	35 (14.6)
	STEM-LC	36 (15.1)
Instructor Type	Graduate student	12 (5.0)
	Non-tenure track faculty/staff[+]	78 (32.6)
	Tenure track Assistant Professor	39 (16.3)
	Tenured Associate Professor	54 (22.6)
	Full Professor	56 (23.4)

[+] Non-tenure track faculty/staff may be called adjunct/instructor/general faculty/wage employees at other institutions. The unit of analysis is the course, so if an instructor taught multiple courses they are counted multiple times.

course syllabi, student grade data, and instructor information. We coded and scored each data source individually to create four data points for each individual course.

Observations

We collected observational data using the Classroom Observation Protocol for Undergraduate STEM (COPUS) (Smith et al., 2013), which captures the presence or absence of 26 different behaviors in 2-minute time increments. Trained undergraduate students conducted, on average, two observations of each course using COPUS (observers were reliable following training, all κs >.8). The COPUS data were converted into COPUS profiles (copusprofiles.org) (Stains et al., 2018), ranging from didactic to student-centered, or learning-focused, instruction (Table 2).

Course syllabi

We collected a total of 196 syllabi for the 239 courses (82%). A previously developed and validated syllabus rubric (Palmer et al., 2014) was used by trained graduate researchers to score each syllabus on a scale from learning-focused (a score of 46) to content-focused (a score of 0) (percent agreement = 85% for a subset of syllabi coded). From the syllabus scores, each course then received a score indicating whether the syllabus was content-focused (< 17 = 1), transitional (17–30 = 2), or learning-focused (> 30 = 3).

Student grade data. We obtained student grade data from for all observed courses. These data were used to calculate the percentage of A (A = excellent grade) and DFW rates (D = substandard passing grade, F = failing grade, and W = student withdrew) for each course overall (number DFWs/total number of students x 100). We also calculated these rates for underrepresented minority (URM) students, defined in this study as students who self-identify as African American, Hispanic, or Native American (Chang et al., 2008). We further obtained student grade data for all courses in Department A (STEM-LC cohort department) for the year before (2014–2015) and after (2016–2017) participants engaged in the program.

Data analysis

The independent variable was ED intervention status, and the outcome measures included COPUS profile scores, syllabus scores, and percentage grade rates for students and subgroups of students. Of the 239 courses observed, 24 (10%) did not have any URM

Table 2. Description of COPUS Profiles Used to Characterize Classroom Instruction.

Grouping	Profile #	Description
Didactic	1	Greater than 80% lecture, few student questions
	2	Greater than 80% lecture, few clicker questions/group work
Interactive	3	Lecture with group activities
	4	Lecture with clicker question group work
Student-centered	5	Group work activities consistently used
	6	Process Oriented Guided Inquiry Learning (POGIL)-like instruction with group worksheets and one-on-one instructor-student interactions
	7	Varied group work activities

students. When DFW rates were included in the analyses, the courses with no URM students were excluded from the data set so as to not skew the data.

RQ1 analyses (overall ED intervention impact)

After testing assumptions, we ran an ANCOVA to explore the impact of the three ED interventions on COPUS profile scores while controlling for number of enrolled students and level of course. Assumptions were not met for syllabus scores, so an ANOVA was run instead. Our a priori alpha level for these tests was $p < .05$. We aimed to use hierarchical linear regression to identify variables that predicted grades for white students and URM students; however, the assumptions (homoscedadiscity, collinearity, and multicollinearity) were not met. Therefore we ran an ANOVA with post-hoc testing using a Bonferroni adjustment (adjusted alpha, $p < .05/4 = .0125$) to explore differences in grade data within each intervention group (such as difference in percentage of DFW). To explore differences in grade data for subgroups of students within each intervention group (such as difference in percentage of DFW between white, URM, and Asian students), we split the file by intervention group and ran three t-tests (URM-white, URM-Asian, white-Asian). We used a conservative alpha ($p < .05/3 = .0167$) to account for multiple comparisons.

We also examined differences in outcome measures for smaller courses (<60 students) separately to understand differences in ED interventions specific to these class sizes. Given the small size of each intervention group, we explored the data to ensure normality for each outcome measure. In cases where the homogeneity of variance assumption was violated (Levene's test <.05), a non-parametric Mann-Whitney U test was run.

RQ2 analyses (new faculty learning community, NFLC)

We compared differences in outcome measures for courses taught by new faculty that participated in NFLC ($n = 35$) and those that did not ($n = 29$). We used non-parametric Mann-Whitney U tests to identify differences in outcome measures and calculated frequencies to descriptively characterize differences between the two groups.

RQ3 analyses (STEM-learning community, STEM-LC, cohort)

Similar to the NFLC data, the small number of courses observed in Department A ($n = 17$) and Department B ($n = 29$) dictated the use of non-parametric Mann-Whitney U tests. To triangulate our data set, we also calculated descriptives for overall departmental grade data for Department A in the year prior to and following participation in STEM-LC.

Results

Below we present the findings for each research question. Because sample sizes were very small for some subgroups, we acknowledge that these smaller groups may not represent populations of instructors who have participated in our various interventions. We therefore explicitly differentiate descriptive findings from inferential findings.

Table 3. Differences in instructional practice and student success between ED intervention groups.

ED intervention group+	Instructional practices, Mean (SD) COPUS Group	Instructional practices, Mean (SD) Syllabus Score	Student success, Mean (SD) %DFW	Student success, Mean (SD) %A
None (n = 140)	1.44 (.68)	7.74 (7.42)	5.50 (5.93)	51.53 (21.14)
CDI (n = 28)	1.68 (.77)	16.41 (10.11)**	3.19 (3.50)	53.05 (22.58)
STEM-LC (n = 35)	1.75 (.87)*	17.65 (9.95)**	4.58 (4.82)	43.12 (21.84)
NFLC (n = 36)	1.97 (.75)**	19.89 (8.92)**	5.16 (5.76)	58.38 (19.80)

+ n-value for groups are lower for syllabus scores and student success due to missing data. * significant from none, p <.05, ** significant from none, p <.01. COPUS group ranged from didactic (1), interactive lecture (2) and student centered (3). Syllabus score ranged from content-focused (0) to learning-focused (46).

RQ1: overall ED intervention impact

Overall there were significant differences between ED intervention groups in measures of instructional practices and student success measures (Table 3). We found there were significant differences between ED intervention groups' observed instructional practice, even when controlling for class size and type, $F(3, 233) = 5.27$, $p < .05$; however, ED intervention only accounted for 9.6% of the variance in instructional practice. Courses taught by both NFLC and STEM-LC instructors had significantly more observed active learning than courses taught by participants who had not engaged in ED interventions. We also found syllabi scores were significantly different between the groups, $F(3, 190) = 15.80$, $p < .05$, with 28.5% of variance explained by ED intervention type. Post-hoc comparisons identified that courses taught by instructors who had engaged in any one of our three interventions had significantly more learning-focused syllabi than courses taught by participants who had not. Exploring overall student grades, there were significant differences in percentage of student A grades between the ED interventions, but not between DFW rates. Post-hoc Bonferroni comparisons with an adjusted p-value ($p < .0125$) meant that there were no significant differences found between individual ED intervention groups.

When disaggregating success measures by race and gender, we identified additional differences between ED intervention groups (Table 4). First, female DFW rates were significantly higher than for males in courses where the instructor had not engaged in any ED interventions ('None' group), and finding not present for any of the ED intervention groups. Second, the DFW rates for URM students was significantly higher than for White students for all groups *except* NFLC. These findings were similar when comparing URM

Table 4. Grade outcomes for each ED intervention group disaggregated by race and gender.

Student Success		ED Interventions, Mean (SD) grades: None (n = 125)	CDI (n = 27)	STEM-LC (n = 33)	NFLC (n = 31)
Race	URM %DFW	10.06 (16.04)	7.33 (10.65)	8.45 (10.91)	7.62 (11.33)
	White %DFW	4.93 (6.75)**	2.41 (3.84)**	3.98 (4.76)**	3.99 (4.29)
	Asian %DFW	5.47 (14.32)**	4.69 (11.18)	2.70 (4.85)**	2.14 (3.68)**
	URM %A	33.44 (31.30)	39.80 (34.10)	29.81 (29.81)	38.34 (28.07)
	White %A	51.05 (24.30)**	54.58 (23.84)**	42.32 (21.53)**	57.14 (23.81)**
	Asian %A	45.46 (30.92)**	46.27 (32.81)	42.17 (23.33)**	59.99 (25.06)**
Gender	Female %DFW	4.23 (6.04)	3.02 (4.62)	4.54 (5.92)	3.44 (5.56)
	Male %DFW	6.89 (9.47)*	3.35 (3.76)	4.51 (4.52)	5.84 (6.89)
	Female %A	50.88 (28.30)	54.67 (25.48)	42.13 (23.80)	56.51 (29.98)
	Male %A	51.18 (24.45)	52.95 (25.14)	44.80 (22.39)	57.80 (19.93)

* significant from male, p <.05. ** significant from URM, p <.017 (adjusted p-value .05/3).

and Asian student DFW rates between the groups. Alarmingly, White and Asian students' A rates were significantly higher than URM A rates nearly across the board and do not appear to be impacted by any of the interventions.

With the challenges in engaging and retaining students in large, introductory STEM courses (Seymour & Hewitt, 1997), we focused on DFW rates between URM and White students for smaller courses (n < 60) as a subset of our data (Figure 1). We chose this value as these courses would be considered small for our university context. For NFLC, the larger courses were taught by only two instructors and we did not include these data in the analysis. Descriptively, the DFW rates for White and URM students were largest for small courses taught by instructors who had not engaged in ED interventions (M = 10.54) and smallest for courses taught by NFLC instructors (M = 1.56). Further, it appeared the DFW rate for URM students decreased the more contact hours instructors had with our educational developers. Interestingly, White students' DFW rates for courses taught by NFLC instructors were higher than URM students. This finding is addressed below as it relates to instructional practices.

We further separated out the ED intervention groups by the observed instructional practices (didactic, interactive lecture, and student-centered instruction) (Figure 2). When comparing the URM and White student DFW rate, the gap in URM and White student rates was consistent for instructors across both didactic (~4–7% difference) and interactive lecture courses (~3–6% difference), regardless of intervention status. Further, in courses where didactic instruction was observed, URM student DFW rates were consistently higher than those of White students. In courses characterized by interactive lecture, the DFW rates for both White and URM students were lower than in courses with didactic instruction. The DFW rate for White students was also lowest in interactive lecture courses, which was not the case for URM students.

For courses where student-centered instruction was observed, the DFW rates were descriptively more variable than those in didactic or interactive lecture. For student-

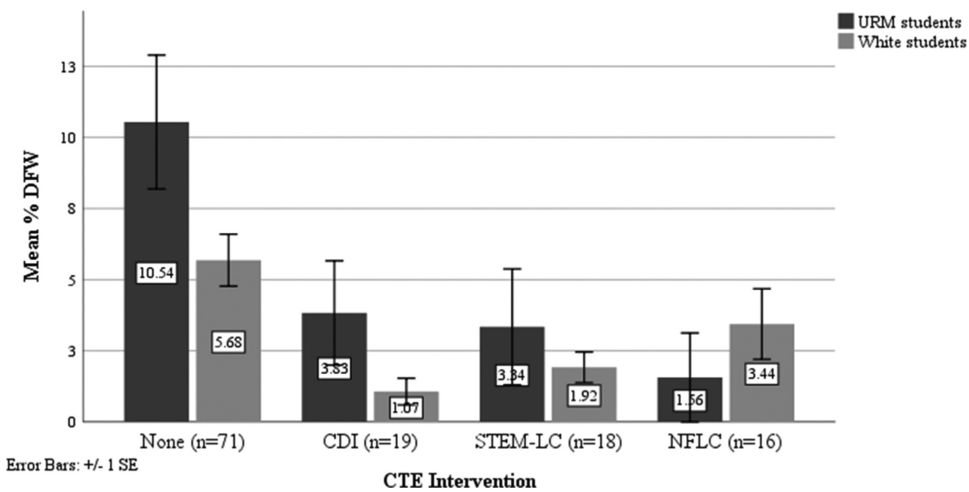

Figure 1. Differences in student failure rates for small course (n < 60) by faculty engagement in different CTE interventions.

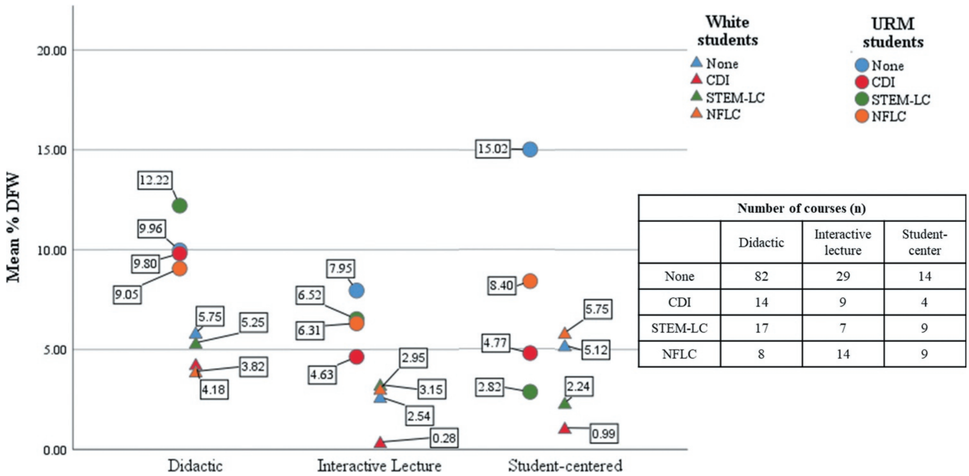

Figure 2. Descriptive Differences in URM failure in courses with different observed instructional practies.

centered instruction in courses taught by non-engaged instructors, the DFW rate for URM students (15.02%) is almost quadruple that of White student DFW rates (5.12%). For student-centered courses taught by CDI and NFLC instructors, the DFW rate gap between URM and White students paralleled interactive lecture courses. For student-centered courses taught by STEM-LC instructors, the DFW rate gap between URM and White students disappeared *and* was low (~2.5%) for both groups.

RQ2: impact of NFLC program on new faculty

For courses taught by new faculty, there were significant differences in instructional practices but not in student success (Table 5). In particular, courses taught by NFLC participants were observed to have significantly more active learning ($Z = -3.05$, $p < .05$) and more learning-focused syllabi ($Z = -3.60$, $p < .05$) when compared with courses taught by new faculty who had not engaged in any ED interventions. There were no significant differences between grade data for courses taught by NFLC faculty or new faculty who had not engaged in ED interventions. There were also no differences between DFW rates for URM and White students for either group; however, the significantly lower A grade rate for URM students compared to White students observed in the overall comparison above persisted when comparing only new faculty.

Categorizing instructional practices further demonstrated the descriptive differences between these two groups. Of the 29 courses taught by new faculty who had not engaged in ED interventions, nearly three quarters (72.41%) utilized didactic instruction, with only 17.4% utilizing student-centered instruction. Conversely, nearly three quarters of the 35 courses taught by NFLC faculty utilized some sort of interactive component to their instruction (45.71% interactive lecture, 25.71% student-centered) and only 28.57% of the courses were categorized as didactic instruction. Similarly, of the 20 syllabi analyzed for courses taught by new faculty who had not engaged in any ED interventions,

Table 5. Differences in New Faculty's Instructional Practices and Student Success.

	Instructional practices, Mean (SD)		Student success, Mean (SD)									
			Race						Gender			
	COPUS Group	Syllabus Score	URM % DFW	White % DFW	Asian % DFW	URM % A	White % A	Asian % A	Female % DFW	Male% DFW	Female % A	Male % A
NFLC (n = 35)	1.97	19. 89	7.62	4.11	2.14	38.34	56.51	59.99*	3.45	5.84	56.51	57.80
	(.75)**	(8.92)**	(11.33)	(16.74)	(3.68)	(28.07)	(23.83)	(25.06)	(5.56)	(6.89)	(29.98)	(19.93)
None+ (n = 29)	1.44	9.60	9.06	6.82	6.99	34.51	50.75	48.03	4.61	8.84	49.27	50.93
	(.78)	(8.76)	(13.24)	(9.98)	(11.88)	(33.07)	(25.01)	(28.33)	(6.11)	(10.44)	(20.05)	(23.58)

+ only includes faculty who have been hired in the last three years. * significant, $p < .05$. ** significant, $p < .01$. COPUS group ranges from didactic (1), interactive lecture (2) and student centered (3). Syllabus score ranges from content-focused (0) to learning-focused (46).

75.00% were content-focused and only 5.00% learning-focused. Two-thirds of the 33 syllabi analyzed for courses taught by NFLC faculty had learning-focused elements (57.59% transitional, 9.09% learning-focused), with only 33.33% identified content-focused.

When exploring DFW for small courses (<60 students) taught by new faculty, the gap between URM and White students was negligible, regardless of whether they had participated in NFLC. For example, DFW rates for URM and White students in small courses taught by non-engaged new faculty (n=16) were 7.50% and 6.81% respectively, and these rates were 1.56% and 3.44% for NFCL faculty teaching small courses (n=16). However, DFW rates for small courses taught by non-engaged faculty were five times higher for URM students (7.50%) than in courses taught by NFLC faculty (1.56%).

RQ3: impact of a cohort model within the STEM-LC program

When comparing Department A, containing a STEM-LC cohort, to Department B, the control department, there were significant differences in instructional practices and student success (Table 6). Courses taught in Department A had significantly more observed active learning ($Z = -2.85$, $p < .05$), more learning-focused syllabi ($Z = -3.05$, $p < .05$), lower URM student DFW rates ($Z = -2.85$, $p < .05$), lower White student DFW rates (-3.19, $p < .05$), higher URM student A rates ($Z = -2.11$, $p < .05$), and higher White student A rates (-2.54, $p < .05$).

Descriptively, there were differences in the DFW rates between URM and White students for small courses in each department. The DFW gaps for Department A were virtually non-existent between URM and White students (1.43 vs. 1.90%), while URM students in Department B have a DFW rate twice that of White students (22.46% vs. 11.88%). The stark differences between Department A and B may be related to the impact of a cohort of instructors engaged in the STEM-LC; however, because there are other confounding factors (for example, student population, instructor type), this is not a definitive conclusion.

We therefore sought to describe the overall grade rates for Department A before (n = 65 courses) and after (n = 63 courses) the cohort of instructors participated in STEM-LC (37.5% of the departmental teaching faculty) to better understand the impact of the program. While the departmental DFW rate for White students did not change pre- to post-intervention (5.91% and 5.10%, respectively), the URM student DFW rate for the department decreased from 15.89% to 9.59%. This suggests that the cohort of STEM-LC instructors may have contributed to decreasing the department's overall URM

Table 6. Differences in Instructional Practices and Student Success for a Cohort of Instructors.

	Instructional practices, Mean (SD)		Student success, Mean (SD)			
	COPUS Group	Syllabus Score	URM %DFW	White % DFW	URM % A	White % A
Dept A (cohort)	2.53 (.87)**	17.40 (10.03)**	2.81 (7.44)**	2.87 (4.18)**	36.02 (31.80)*	48.91 (18.91)*
Dept B (control)	1.48 (.83)	7.76 (8.05)	21.63 (25.37)	10.83 (8.82)	18.54 (25.16)	34.21 (17.76)

* significant, p <.05. ** significant, p <.01. COPUS group ranges from didactic (1), interactive lecture (2) and student centered (3). Syllabus score ranges from content-focused (0) to learning-focused (46).

student failure rate. However, without pre-intervention observation data, we are unable to draw definitive conclusions about the reasons for changes in student failure rates.

Limitations

While we attempted to obtain a large, representative sample of courses, there are limitations in the study design. First, while we aimed to conduct more sophisticated statistics to identify factors that predicted student success (DFW rates, A rates), the data did not meet the assumptions needed to run these tests. Thus, the conclusions we can draw about the relationship between CTE interventions, instructional practices, and student success are constrained by the limitations of the data. Second, the small sub-group sample sizes limit our ability to draw definitive conclusions about the impact of our NFLC and STEM-LC programs when compared to 'comparison' courses. There may be alternate explanations for the differences between groups, particularly around departmental culture, which prior research demonstrates is important to what and how instructors teach (Lund & Stains, 2015). For example, differences observed between course syllabi could be explained by departmental constraints to syllabus structure and differences in instructional practice could be related to the emphasis in the department on teaching. Third, the self-selection of instructors for participation in ED programming may explain differences in the groups. To address these limitations, we compared pre- and post- grade data for Department A in an attempt to eliminate self-selection as a possible explanation for the reported findings. To further address limitations in the present study, our other research explores changes in faculty beliefs, self-efficacy, instructional practices, and student outcomes following participation in NFLC and STEM-LC and suggest positive impacts (Favre et al., In review).

Discussion & implications

In the present study we explored the impact of different ED interventions on instructional practices and student performance in STEM undergraduate courses. The results of our study contribute to the scholarly literature in four ways. First, of the 147 STEM instructors who have participated in CDI since 2008, 99 participated in the present study (67.3%) as either CDI, NFLC, or STEM-LC participants. These three intervention groups had significantly more learning-focused syllabi than the non-engaged CTL group, which suggests that these faculty continued to utilize and implement learning-focused syllabi beyond the course they designed in CDI. These findings add to the work on the impact of CDI on faculty syllabi (Palmer et al., 2014) and suggest that CDI has a lasting impact on the way instructors design courses. Future work could examine how faculty translate their course design to different courses and over time.

Second, the large DFW gap between URM students and White students in student-centered courses taught by non-engaged instructors may be the result of poor implementation of student-centered instruction. It also appears that student-centered instruction has the potential to close the DFW gap between White and URM students when instructors engage in intensive ED interventions. These results add to the recent literature on undergraduate STEM courses that suggests implementation of active learning may differentially impact students (Cooper et al., 2018; Snyder et al., 2016). While there are benefits in using COPUS as an observation tool, it did not allow for differentiation in the quality of student-centered instruction. Further research is needed not only to explore whether active learning is used, but *how* it is implemented. Our study may suggest that instructors indeed need support to ensure active learning is implemented skillfully and in a way that supports all students.

Third, the particularly low DFW rates and closing of the failure gap between URM and White students observed for student-centered courses taught by STEM-LC faculty could potentially be explained by a cohort effect. When participants engage in ED interventions as a cohort, they may be interacting and discussing their course implementation more than instructors who participate solo. This cohort effect aligns with models that link policy, teachers, and students in K-12 professional development (Luft & Hewson, 2014) and provides additional support for the efficacy of local communities of practice (Lakshmanan et al., 2011). Social network analysis and qualitative exploration of these interactions beyond the formal STEM-LC meetings may help us better understand how members of a cohort interact and support each other.

Finally, educational developers are increasingly asked to demonstrate the efficacy of their work and to provide evidence that the investment in educational development is worthwhile. This study contributes to research focused on establishing the link between participation in ED programs, changes in teaching practices, and improved student learning (Condon et al., 2016). Specifically, drawing on research on intensive course design workshops (Palmer et al., 2016) and faculty learning communities (Gehrke & Kezar, 2017), it examines the impact of different ED intervention types and durations. Unique in its design, this study uses a multi-indicator strategy and direct measures such as syllabi and classroom observations to assess changes in teaching as well as institutional grade data to approximate student success. It builds on and moves beyond educational development research that relies on indirect measures such as student evaluations of teaching and self-reported learning gains (Condon et al., 2016; Meizlish et al., 2018). Future work may seek to gather these data in order to be able to run more sophisticated statistical tests exploring relationships between ED programs, instructional practices, and student outcomes.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Lindsay B. Wheeler is assistant director for STEM education initiatives and assistant professor, general faculty, at the University of Virginia's Center for Teaching Excellence. She is also the associate co-editor of *To Improve the Academy*. Her current research interests include how center interventions translate to instructional practice and impact student outcomes. She presents at national and international research conferences and currently serves as chair of the POD Network's Scholarship Committee.

Dorothe Bach is associate director and professor, general faculty, at the University of Virginia's Center for Teaching Excellence. Her current research interests include new faculty development as well as embodied, reflective, and critical pedagogies. She presents workshops nationally and internationally and currently serves on the board of the POD Network.

ORCID

Lindsay B. Wheeler http://orcid.org/0000-0003-2794-0345

References

Bain, K. (2004). What makes great teachers great. *Chronicle of Higher Education*, *50*(31), B7–B9.
Beach, A. L., Sorcinelli, M. D., Austin, A. E., & Rivard, J. K. (2016). *Faculty development in the age of evidence: Current practices, future imperatives*. Stylus.
Bonwell, C. C., & Eison, J. A. (1991). *Active learning: Creating excitement in the classroom. 1991 ASHE-ERIC higher education reports*. ERIC Clearinghouse on Higher Education.
Chang, M. J., Cerna, O., Han, J., & Sàenz, V. (2008). The contradictory roles of institutional status in retaining underrepresented minorities in biomedical and behavioral science majors. *The Review of Higher Education*, *31*(4), 433–464. https://doi.org/10.1353/rhe.0.0011
Chism, N. V. N., Holley, M., & Harris, C. J. (2012). Researching the impact of educational development: Basis for informed practice. *To Improve the Academy*, *31*(1), 129–145. https://doi.org/10.1002/j.2334-4822.2012.tb00678.x
Condon, W., Iverson, E. R., Manduca, C. A., Rutz, C., & Willett, G. (2016). *Faculty development and student learning: Assessing the connections*. Indiana University Press.
Cooper, K. M., Downing, V. R., & Brownell, S. E. (2018). The influence of active learning practices on student anxiety in large-enrollment college science classrooms. *International Journal of STEM Education*, *5*(1), 23. https://doi.org/10.1186/s40594-018-0123-6
Daempfle, P. A. (2003). An analysis of the high attrition rates among first year college science, math, and engineering majors. *Journal of College Student Retention: Research, Theory & Practice*, *5*(1), 37–52. https://doi.org/10.2190/DWQT-TYA4-T20W-RCWH
Ebert-May, D., Derting, T. L., Hodder, J., Momsen, J. L., Long, T. M., & Jardeleza, S. E. (2011). What we say is not what we do: Effective evaluation of faculty professional development programs. *BioScience*, *61*(7), 550–558. https://doi.org/10.1525/bio.2011.61.7.9
Favre, D., Bach, D., & Wheeler, L. B. (In review). *Measuring institutional transformation: A multi-faceted assessment of a new faculty development program*.
Fink, L. D. (2013). *Creating significant learning experiences, revised and updated. An integrated approach to designing college courses*. Jossey-Bass.
Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111
Gasiewski, J. A., Eagan, M. K., Garcia, G. A., Hurtado, S., & Chang, M. J. (2012). From gatekeeping to engagement: A multicontextual, mixed method study of student academic engagement in introductory STEM courses. *Research in Higher Education*, *53*(2), 229–261. https://doi.org/10.1007/s11162-011-9247-y
Gehrke, S., & Kezar, A. (2017). The roles of STEM faculty communities of practice in institutional and departmental reform in higher education. *American Educational Research Journal*, *54*(5), 803–833. https://doi.org/10.3102/0002831217706736
Haak, D. C., HilleRisLambers, J., Pitre, E., & Freeman, S. (2011). Increased structure and active learning reduce the achievement gap in introductory biology. *Science*, *332*(6034), 1213–1216. https://doi.org/10.1126/science.1204820
Henderson, C., Beach, A., & Finkelstein, N. (2011). Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature. *Journal of Research in Science Teaching*, *48*(8), 952–984. https://doi.org/10.1002/tea.20439
Kezar, A., & Gehrke, S. (2017). Sustaining communities of practice focused on STEM reform. *The Journal of Higher Education*, *88*(3), 323–349. https://doi.org/10.1080/00221546.2016.1271694

Kolomitro, K., & Anstey, L. M. (2017). A survey on evaluation practices in teaching and learning centres. *International Journal for Academic Development*, *22*(3), 186–198. https://doi.org/10.1080/1360144X.2017.1313162
Kreber, C., & Brook, P. (2001). Impact evaluation of educational development programmes. *International Journal for Academic Development*, *6*(2), 96–108. https://doi.org/10.1080/13601440110090749
Kuh, G. D., Cruce, T. M., Shoup, R., Kinzie, J., & Gonyea, R. M. (2008). Unmasking the effects of student engagement on first-year college grades and persistence. *The Journal of Higher Education*, *79*(5), 540–563. https://doi.org/10.1353/jhe.0.0019
Lakshmanan, A., Heath, B. P., Perlmutter, A., & Elder, M. (2011). The impact of science content and professional learning communities on science teaching efficacy and standards-based instruction. *Journal of Research in Science Teaching*, *48*(5), 534–551. https://doi.org/10.1002/tea.20404
Lauridsen, K. M., & Lauridsen, O. (2018). Teacher capabilities in a multicultural educational environment: An analysis of the impact of a professional development project. *International Journal for Academic Development*, *23*(2), 98–109. https://doi.org/10.1080/1360144X.2017.1357557
Luft, J. A., & Hewson, P. W. (2014). Research on teacher professional development programs in science. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 899–910). Routledge.
Lund, T. J., & Stains, M. (2015). The importance of context: An exploration of factors influencing the adoption of student-centered teaching among chemistry, biology, and physics faculty. *International Journal of STEM Education*, *2*(1). https://doi.org/10.1186/s40594-015-0026-8
Meizlish, D. S., Wright, M. C., Howard, J., & Kaplan, M. L. (2018). Measuring the impact of a new faculty program using institutional data. *International Journal for Academic Development*, *23*(2), 72–85. https://doi.org/10.1080/1360144X.2017.1364644
National Science Foundation, National Center for Science and Engineering Statistics. (2017). *Women, minorities, and persons with disabilities in science and engineering: 2017, special report NSF 17-310*. www.nsf.gov/statistics/wmpd/
Palmer, M. S., Bach, D. J., & Streifer, A. C. (2014). Measuring the promise: A learning-focused syllabus rubric. *To Improve the Academy: A Journal of Educational Development*, *33*(1), 14–36. https://doi.org/10.1002/tia2.20004
Palmer, M. S., Streifer, A. C., & Williams-Duncan, S. (2016). Systematic assessment of a high impact course design institute. *To Improve the Academy: A Journal of Educational Development*, *35*(2), 339–361. https://doi.org/10.1002/tia2.20041
President's Council of Advisors on Science and Technology (PCAST). (2012). *Engage to Excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics*. Executive Office of the President.
Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin*, *143*(6), 565–600. https://doi.org/10.1037/bul0000098
Schunk, D. H., Pintrich, P. R., & Meece, J. R. (2007). *Motivation in education: Theory, research, and applications* (3rd ed.). Prentice Hall.
Seymour, E., & Hewitt, N. (1997). *Talking about leaving: Why undergraduates leave the sciences*. Westview Press.
Smith, M. K., Jones, F. H. M., Gilbert, S. L., & Wieman, C. E. (2013). The classroom observation protocol for undergraduate stem (COPUS): A new instrument to characterize university STEM classroom practices. *CBE Life Sciences Education*, *12*(4), 618–627. https://doi.org/10.1187/cbe.13-08-0154
Snyder, J. J., Sloane, J. D., Dunk, R. D. P., & Wiles, J. R. (2016). Peer-Led team learning helps minority students succeed. *PLoS Biology*, *14*(3), 1–7. https://doi.org/10.1371/journal.pbio.1002398
Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., Eagan, M. K., Esson, J. M., Knight, J. K., Laski, F. A., Levis-Fitzgerald, M., Lee, C. J., Lo, S. M.,

McDonnell, L. M., McKay, T. A., Michelotti, N., Musgrove, A., Palmer, M. S., Plank, K. M., Sanders, E. R., … Young, A. M. (2018). Anatomy of STEM teaching in North American universities. *Science*, *359*(6383), 1468–1470. https://doi.org/10.1126/science.aap8892

Tomkin, J. H., Beilstein, S. O., Morphew, J. W., & Herman, G. L. (2019). Evidence that communities of practice are associated with active learning in large STEM lectures. *International Journal of STEM Education*, *6*(1), 1–15. https://doi.org/10.1186/s40594-018-0154-z

Wiggins, G. (1998). *Educative assessment. Designing assessments to inform and improve student performance*. Jossey-Bass Publishers.

Winkelmes, M. A., Bernacki, M., Butler, J., Zochowski, M., Golanics, J., & Weavil, K. H. (2016). A teaching intervention that increases underserved college students' success. *Peer Review*, *18*(1/2), 31.