



Cite this: *Chem. Educ. Res. Pract.*,  
2018, 19, 846

## A new approach to supplementary instruction narrows achievement and affect gaps for underrepresented minorities, first-generation students, and women

Cynthia A. Stanich,  \* Michael A. Pelch,  Elli J. Theobald  and Scott Freeman

To help students who traditionally underperform in general chemistry, we created a supplementary instruction (SI) course and called it the STEM-Dawgs Workshops. These workshops are an extension of the Peer-led Team Learning (PLTL) SI. In addition to peer-facilitated problem-solving, we incorporated two components inspired by learning sciences: (1) training in research-based study skills, and (2) evidence-based interventions targeting psychological and emotional support. Here we use an explanatory mixed methods approach to measure the impact of the STEM-Dawgs Workshops, with a focus on four sub-populations that are historically underrepresented in Chemistry: underrepresented minorities, females, low-income students, and first-generation students. Specifically, we compared three groups of students in the same General Chemistry course: students in general chemistry and not the workshops ("Gen Chem students"), students in the workshops ("STEM-Dawgs"), and students who volunteered for the workshops but did not get in ("Volunteers"). We tested hypotheses with regression models and conducted a series of focus group interviews with STEM-Dawgs. Compared to the Gen Chem population, the STEM-Dawg and Volunteer populations were enriched with students in all four under-represented sub-populations. Compared to Volunteers, STEM-Dawgs had increased exam scores, sense of belonging, perception of relevance, self-efficacy, and emotional satisfaction about chemistry. URM STEM-Dawgs had lower failure rates, and exam score achievement gaps that impacted first-generation and female Gen Chem students were eliminated in the STEM-Dawg population. Finally, female STEM-Dawgs had an increased sense of belonging and higher emotional satisfaction about chemistry than women Volunteers. Focus groups suggested that successes came in part from the supportive peer-learning environment and the relationships with peer facilitators. Together, our results indicate that this supplementary instruction model can raise achievement and improve affect for students who are underrepresented in chemistry.

Received 13th February 2018,  
Accepted 22nd May 2018

DOI: 10.1039/c8rp00044a

[rsc.li/cerp](http://rsc.li/cerp)

## Introduction

A recent report projects that to meet demand, the United States needs a 34% increase in the production of bachelor's degrees in science, technology, engineering, and mathematics (STEM) disciplines (Holdren and Lander, 2010). Many or most of these graduates will need to come from four populations that are traditionally underrepresented in STEM (National Science Foundation, 2017): underrepresented minorities (URMs), women, low-income individuals, and first-generation students (first-in-family with a four-year degree). Although individuals from these subpopulations enter college with the same level of interest in STEM careers as their peers, they drop out of STEM at much higher rates (Ishitani, 2006; Chen and Soldner, 2013). For example, women have similar enrollments as men in

introductory STEM courses but underperform in lecture based courses with high-stakes exams (see Appendix 1, Gen Chem; Matz *et al.*, 2017).

Students from underrepresented populations face an array of challenges due to the circumstances of their birth. Many are underprepared due to educational disadvantage or—in the case of women—stereotypes about their ability in quantitative fields. In addition, URM, low-income, and first-generation students are more likely to experience economic disadvantage, requiring them to work while in school and take on debt to pay for tuition and living expenses (Stephens *et al.*, 2012). First-generation students are also entering a world that is unfamiliar to their family members, creating a cultural mismatch that can place them at risk (Orbe, 2008).

Researchers have developed interventions designed to address both the academic and the emotional and psychological obstacles that underrepresented students face. For example,

Department of Chemistry, University of Washington, USA. E-mail: [stanich@uw.edu](mailto:stanich@uw.edu)

courses with high structure—meaning that they require pre-class preparation, intensive active learning, and regular exam practice—can reduce achievement gaps for low-income and URM students because they increase deliberate practice and sense of community (Haak *et al.*, 2011; Eddy and Hogan, 2014). Programs that offer comprehensive support in the form of financial aid, access to research opportunities, supplementary instruction (SI) in key courses, and mentoring have also shown dramatic benefits (*e.g.* Barlow and Villarejo, 2004; Shields *et al.*, 2012; Slovacek *et al.*, 2012; Hall *et al.*, 2014). Finally, less-intensive interventions, often in the form of brief writing exercises or structured group discussions, have been beneficial in reducing achievement gaps caused by stereotype threat (Miyake *et al.*, 2010; Jordt *et al.*, 2017) or for promoting a more positive sense of belonging (Walton and Cohen, 2011).

Research has shown that bad experiences in general chemistry can be especially discouraging for URM and other at-risk students (Barr *et al.*, 2008; De La Franier *et al.*, 2016). As the first step in a comprehensive effort to redesign general chemistry at the University of Washington and improve the performance of underrepresented students, we designed and evaluated an SI workshop as a voluntary, 2-credit companion course for the initial offering in the 3-quarter, year-long general chemistry sequence. SI has a long track record of improving the performance of all students in an array of STEM courses (Arendale, 1994), and in some cases has produced disproportionate gains by underrepresented students (*e.g.* Fullilove and Treisman, 1990). In chemistry, several SI designs have increased student grades (Rath *et al.*, 2012; Batz *et al.*, 2015), with the Peer-led Team Learning (PLTL) model having a particularly large-scale impact (Gosser, 2011).

In almost every case, the core of successful SI in STEM courses has been cooperative group learning—often focused on working exam-like problems. Well-managed cooperative group learning can improve student performance in chemistry (Warfa, 2016) because it encourages students to be metacognitive about their problem solving—meaning that they become more aware of, and better-monitor, their own learning (Schraw *et al.*, 2006; Sandi-Urena *et al.*, 2012; Snyder and Wiles, 2015). Although some SI models employ graduate teaching assistants or professional tutors, peer-facilitated group learning—the PLTL model—has been particularly effective in chemistry (Becvar *et al.*, 2008; Hockings *et al.*, 2008; Snyder *et al.*, 2016).

In addition to building on the strengths of peer-facilitated group practice, we wanted to design and test an SI model that would leverage recent experimental work on effective study strategies. In the absence of explicit training in study skills, many students respond to poor exam scores by doing more of the approaches that produced the initial, disappointing outcome (Stanton *et al.*, 2015). Increased time on task can be helpful, but there is no guarantee that students are spending the additional study time wisely (Chan and Bauer, 2014; Ye *et al.*, 2016). In contrast, making the benefit and “how-to” of evidence-based study skills explicit can promote changes to study behavior (Cook *et al.*, 2013). Explicit practice of evidence-based study habits, in turn, can improve exam performance

(Zhao *et al.*, 2014); effective learning strategies can also improve affect through the positive emotions they engender (Pekrun *et al.*, 2002, 2007).

In addition to PLTL and effective study practices, we also incorporated a third element in the SI model analyzed here: evidence-based interventions that provide psychological and emotional support. The purpose of these exercises was to reduce the impact of stereotype threat, imposter syndrome, test anxiety, and other issues that have a disproportionate impact on underrepresented students. Each of the interventions we chose has been shown to increase some aspect of student performance for disadvantaged students in STEM when used in isolation (citations for each are listed below). Many of these interventions show improvement for underrepresented students in STEM (Cohen *et al.*, 2006; Walton and Cohen, 2007; Miyake *et al.*, 2010; Harackiewicz *et al.*, 2014; Yeager *et al.*, 2014; Paunesku *et al.*, 2015; Yeager *et al.*, 2016; Jordt *et al.*, 2017). The workshop that we designed is the first time that a large suite of psychological and academic interventions has been used in conjunction with the same student population. Additionally, we wanted to brand the course in a positive way, so we called it the “STEM-Dawgs workshop” and developed a logo that was printed on course materials and on a mobile phone carrying case that was given to participants.

We hypothesized that the novel combination of deliberate practice, study skills training, and psychological-emotional interventions, implemented in a weekly workshop that was run by peer facilitators, would provide cognitive and affective benefits to student populations that have historically underperformed in introductory chemistry courses. Our specific research questions were:

(1) What impact does the STEM-Dawgs workshop have on course performance in general chemistry as indexed by exam scores and failure rates? Does it disproportionately benefit students who are traditionally underrepresented in STEM?

(2) What impact does the workshop have on aspects of student affect that may be particularly important for URM, female, low-income, and first-generation students to be successful in general chemistry?

## Methods

We first tested our hypotheses using quantitative data we collected through student performance measures and student responses to surveys. We followed-up these quantitative analyses by implementing an explanatory mixed methods design in an attempt to provide enhanced meaning for our quantitative data (Creswell and Clark, 2011). Specifically, we employed a thematic content analysis of student focus group data to augment results of the quantitative analyses.

### Setting

The general chemistry sequence (3 consecutive quarters) at the University of Washington is predominantly taken by freshmen who intend to pursue a STEM degree. The first quarter of the sequence has a total enrollment of about 3000 students per year.

The demographics of students who participated in this study, based on information provided by the University Registrar's office, generally reflects the population of the university: 54% women, 16% URM, 36% first generation, and 24% under-advantaged (more details are provided in Appendix 1). At the University of Washington, we define "under-advantaged" students as students who qualify for the Educational Opportunity Program (EOP); the EOP supports students from educationally or economically disadvantaged backgrounds and is based on family income and other variables.

This work was conducted under the University of Washington Human Subjects Division application 52402.

### Course context and course design

The STEM-Dawgs workshop was designed as a weekly, two-hour, two-credit companion course for Chemistry 142, which is the first in the 3-quarter (year-long) general chemistry sequence for majors. As currently taught, Chemistry 142 is a 5-credit course consisting of three 50 minute lectures per week taught in a traditional style (no active learning activities), a 50 minute quiz section led by graduate teaching assistants (TAs) who facilitate completion of worksheets or answer questions, and a three-hour lab that meets six times during the 10 week quarter. Lecture sections enroll 275–625 students; quiz sections and lab sections enroll 24 students each. Between one and four professors teach the lectures per quarter, depending on the quarter. Despite the diversity of instructors, all sections are taught in a similar style and are subject to the same criterion for determining final grades—a common median of  $2.6 \pm 0.2$ . Most of the course grade is based on performance on two one-hour midterms and a two-hour final exam.

The STEM-Dawgs workshop was a credit/no credit course, with grades based on completion of assignments and points earned on quizzes. Students were expected to earn at least 70% of the points in the course for credit. Nine STEM-Dawgs sections consisted of a maximum of 24 students each, led by two peer facilitators. Each week, workshop participants were assigned pre-class homework questions and a writing exercise. Prior to most workshop sessions and at other specified times, students did an evidence-based writing intervention designed to help them cope with psychological or emotional difficulties they might face in making the transition from high school to college. We chose these interventions because they have been shown to improve student performance, and especially outcomes for underrepresented students, in courses similar to general chemistry, as we explain in detail below. The writing interventions addressed the following barriers:

(1) **Stereotype threat:** students who are subject to stereotypes about their academic ability may underperform due to the cognitive load of coping with the stereotype, or may take criticism as an inaccurate perception of performance or evidence of not belonging. A values affirmation writing intervention that can mitigate stereotype threat in academic contexts (Cohen *et al.*, 2006; Cohen *et al.*, 2009; Miyake *et al.*, 2010; Jordt *et al.*, 2017) was assigned twice per quarter: at the start and prior to the second midterm. This assignment, called Values Affirmation, prompts

students to select the most important traits about themselves from a list and to write a short essay explaining why they selected those particular traits. In addition, peer facilitators communicated that any critical feedback they gave was motivated by high standards, not racial bias (Cohen *et al.*, 1999; Cook *et al.*, 2013).

(2) **Mindset:** students who view intelligence as a fixed attribute can struggle in response to academic challenges (Dweck and Leggett, 1988; Blackwell *et al.*, 2007). To reduce the impact of this issue, STEM-Dawgs participants did an online assignment that supports better performance by teaching students that the brain is like a muscle and requires exercise to get stronger (Blackwell *et al.*, 2007; Paunesku *et al.*, 2015; Yeager *et al.*, 2016).

(3) **Self-regulation:** self-regulation of behavior, emotions, and impulse is important for academic success (Duckworth, 2011). Mental Contrasting with Implementation Intentions (MCII) is a fantasy-realization intervention designed to convert feasible goals into concrete actions. Early in the term, STEM-Dawgs participants implemented this approach by visualizing a wish, outcome, obstacle and plan (WOOP). The goal was to increase commitment to and confidence in future achievement (Duckworth *et al.*, 2013).

(4) **Test anxiety:** anxiety over assessment can cause decreased levels of performance. A short expressive writing exercise done right before a test improves performance when taking a high-stakes exam (Ramirez and Beilock, 2011). This assignment prompts students to express their pre-exam emotions in writing. STEM-Dawgs participants were encouraged to do the assignment online, as close to the start of each exam as possible.

(5) **Belonging:** feeling accepted in a group has positive effects on health and achievement (Magnuson and Waldfogel, 2008; Meng *et al.*, 2015). To support feelings of belonging in college, students in the workshop did a reading and writing exercise modeled on an intervention about the high school to college transition that has been shown to improve academic outcomes (Walton and Cohen, 2011; Yeager *et al.*, 2016). In this intervention, students read messages from previous students about the challenges inherent in the transition from high school to college. STEM-Dawgs then summarized what they read and told their own transition story. After this writing assignment, STEM-Dawgs were shown a student's video describing her transition story and then discussed it with the peer-mentors.

(6) **Value of learning:** a self-transcendent reason for learning positively correlates with academic outcomes, and can be supported with an intervention, also assigned to the STEM-Dawgs, that prompts students to write about a social injustice and self-transcendent motivations for pursuing higher education (Yeager *et al.*, 2014).

(7) **Expectancy value:** we assigned an intervention that can improve exam performance by increasing the perceived value of a course. The task consists of a short essay on how course content can be relevant to one's own life (Hulleman *et al.*, 2010).

(8) **Post-exam metacognition:** after each exam, STEM-Dawgs completed a writing assignment directing them to examine their study skills. The students answered several open-ended questions regarding study methods and time-on-task, then read

a list of suggested study techniques and designed a new study plan for the next exam (Stanton *et al.*, 2015). This online writing assignment was due 10 days after each of the two midterms.

The schedule for these psychological interventions exercises is provided in Appendix 2.

Workshop sessions started with each student individually taking a 10-minute content quiz, working in a group of three to four to discuss the answers, and then potentially being chosen at random to explain their group's answer to the entire class. Next, the peer facilitators guided students in a discussion about the pre-class writing exercise—the psychological or emotional intervention—and the highlighted study skill for that week. After these discussions, students again worked in small groups to answer chemistry questions that required higher-order cognitive skills and that were relevant to the week's lecture material. During exam weeks, students individually worked on a practice exam, followed by small group work to confirm answers.

The full list of study skills and citations introduced in the course is provided in Appendix 2. During the discussion portion of each week's workshop, a study skill was presented through a worksheet that summarized the methods and results sections of a paper from the peer-reviewed literature. Students then participated in group discussions about why the study skill appeared to be effective and how they could apply the result in their own classes. Annotations about the study skills appeared in the student manual for the workshops and next to assigned problems where the skill was relevant, and peer facilitators were encouraged to point out specific skills as they were being practiced during in-class activities.

### Recruiting workshop participants

We used two methods for recruiting STEM-Dawgs participants: (1) advisors with the Office of Minority Affairs and Diversity introduced the workshop to students who were enrolling in general chemistry and registered individuals who were interested; and (2) students who earned a score in the bottom quartile of the Department's chemistry placement exam were sent recruitment emails describing the workshop as a companion course to Chem 142 where they would learn study skills, self-evaluation, and extra problem-solving practice. They were then asked to indicate their intention to participate *via* an online survey. Students who responded were added to a waiting list. The wait list was then randomized, and students were placed in the open sections of the workshop until they filled.

### Comparison populations

Because many more students volunteered to take the workshops than there were seats available, we were able to analyze data from students who were matched to workshop participants in terms of motivation and demographic characteristics. We call these students the Volunteers, to distinguish them from the STEM-Dawgs. In addition, we also analyzed data from all other students—a population we refer to as Gen Chem Students.

It is important to realize two things about these populations: (1) all students (Volunteers, STEM-Dawgs, and Gen Chem students) were enrolled in general chemistry, and (2) the STEM-Dawgs and

Volunteers represent demographically distinct populations from the Gen Chem Students. As Appendix 1 shows, the STEM-Dawgs and Volunteers populations had a much higher representation of URM, EOP, first-generation, and female students, ranging from a 19–100% increase relative to the Gen Chem population.

### Recruiting and training peer facilitators

Peer facilitators were recruited through notices sent to programs that already used student mentors and tutors, as well as flyers posted on walls around the campus. Peer facilitator selection was done through an in-person interview. Interviews were conducted by the first author. During these interviews, prospective peer facilitators were asked to reflect on their first year in college and how it changed them as a student. Interviewees would often describe their study habits and how they had improved their self-regulatory skills and strategies over the course of their education. Finally, each prospective peer facilitator participated in a mock teaching session where they helped the recruiter through a difficult task. The intention was to observe how prospective peer facilitators might interact with students in the classroom, as well as how they responded to challenges. Originally, 21 students interviewed for the peer facilitator positions, but only eight were picked. Later quarters included the original eight peer facilitators and five additional individuals, drawn from the original 21.

Peer facilitator training began with a two-hour, pre-quarter overview session where the STEM-Dawg mission and techniques were made clear. New facilitators learned about (1) the mechanics of the weekly schedule, and (2) techniques that help facilitate positive group interactions during class. After the workshop's inaugural implementation, we included experienced peer facilitators in the training process, as many of the initial facilitators were eager to stay involved in the course. This development allowed us to remove the pre-quarter training session, because new mentors learned this background information through an apprenticeship with an experienced mentor in the classroom. All peer facilitators attended a weekly meeting to discuss the previous week's session and address tasks for the upcoming week, including a “just-in-time” review of relevant chemistry topics. Finally, peer facilitators would discuss the highlighted study skill for the upcoming class as well as the writing assignment, with the goal of developing strategies to facilitate effective class discussions.

### Quantitative data collection

We collected data from all sections of Chemistry 142, the first quarter in the general chemistry sequence, in fall, winter, and spring quarters of the 2016–2017 academic year. Fall quarter had four sections with three different instructors ( $n = 1438$ ), winter had four sections each with a different instructor ( $n = 833$ ), and Spring quarter had one section with one instructor ( $n = 345$ ); in total, 2616 students took this course in the three consecutive quarters. All quarters were combined for analysis. Student data were collected from (1) a password-protected online learning management system, (2) the instructors' grading records, and (3) the University Registrar's database. The data comprised



responses to pre-post surveys assigned for participation points (at the start and end of the quarter), exam and grade performance, and demographic and academic history characteristics. Demographic data included race, ethnicity, Educational Opportunity Program (EOP) status, first-generation status, gender, and ACT and/or SAT scores. We converted ACT to SAT scores using the published Higher Education College Board's concordance tables (College Board, 2016).

The pre-post survey included a series of Likert-scale questions designed to measure four constructs; mindset, belonging, relevance, and self-efficacy. Mindset is measured through three questions such as, "You can learn new things, but you can't really change your basic intelligence" (Blackwell *et al.*, 2007); belonging through four questions such as, "I feel like I belong in this class" (Walton and Cohen, 2007); relevance *via* four questions such as "My class gives me useful preparation for what I plan to do in life" (Hulleman *et al.*, 2010); and self-efficacy with four questions such as, "I can earn an A in this class" (Harackiewicz *et al.*, 2014). The instrument is made available online by the Project for Education That Scales (PERTS, 2018). In this pre-post survey assignment, students also completed the 8-item Attitude Toward the Subject of Chemistry Inventory V2 (ASCI-V2; Bauer, 2008; Xu and Lewis, 2011) and the Chemical Concept Inventory (CCI; Mulford and Robinson, 2002). The CCI is a 22-item inventory of chemical knowledge and misconceptions, which we administered on a password-protected learning management system called Canvas. Students were given 30 minutes to attempt the questions, which were shown one at a time. We used this inventory as a control for chemical knowledge preparedness in the models described below. We used the PERTS and ASCI-V2 surveys in combination because they are intended to capture related but distinct constructs—how students feel about learning and the course (PERTS), and how they feel about chemistry (ASCI-V2). We specifically chose short surveys to measure the constructs of interest to minimize potential survey fatigue for our students. Factor analysis (described and shown below) confirms that the questions we chose successfully target the affect we were interested in.

### Quantitative data analysis

**Model selection.** We employed a backwards-selection model selection strategy for all of our analyses (Burnham and Anderson, 2002). Following best practice (Zuur *et al.*, 2009), we had two steps in this process: first, we determined the random effect structure, and second, we determined the fixed effects that optimized model fit. We fit multilevel models (with random and fixed effects) because our student observations are nested in sections and terms. Students are grouped into 9 sections and those sections are grouped into three terms (Autumn, Winter, and Spring). As a result, the students in each section and in each term are not truly independent from each other. Fitting multilevel models with term and/or section as random effects with fixed slopes and random intercepts allows for each section or term to have a unique intercept (*i.e.* mean exam score) but fixed slope (*i.e.* relationship between exam score and participation in the STEM-Dawgs workshops). In this way, we account for the non-independence between students within a section or term (Theobald, 2018).

We tested two possible random effects (section and term) by fitting six complex models, each of which included all possible fixed effects but a different random effects structure. Model 1 had term and section as random effects, Model 2 had just term as a random effect, Model 3 had just section as a random effect, Model 4 had section as a random effect and term as a fixed effect, Model 5 had term as a random effect and section as a fixed effect, and Model 6 had no random effects. We compared models using AICc, Akaike's information criterion with a correction for small sample size (Burnham and Anderson, 2002). AICc (and AIC) is a measure of how well a model fits, that relies both on the model maximum likelihood and the number of parameters in the model. Specifically, AICc penalizes model fit as additional parameters are added to the model. In this way, AICc is a relative measure that is particularly useful in comparing nested models (Burnham and Anderson, 2002). Final, best-fitting models had the lowest AICc values; if models differed in AICc by less than 2 units, the model fit was considered identical and the model with the fewest number of parameters was selected as better fitting (Burnham and Anderson, 2002). Starting models are described in association with each hypothesis, and final models are described in the Results.

The second step in our model selection process involved determining the best fitting model by selecting fixed effects. Again, we used backwards selection and assessed model fit by comparing AICc. Using the random effect structure that fit best (from above), we fit complex models that modeled the outcome of interest as a function of the pre-score aligned with the outcome of interest (*e.g.* a PERTS survey construct) and confounding variables that we predicted may affect the outcome. We used a stepwise approach to remove variables, starting with variables that explained the least amount of variation in the outcome. We removed variables until all the variables remained were significant or until the balance between AICc and the number of parameters was minimized.

To investigate the differential effects for students of interest (URM, female, first-generation, and low-income), we fit separate models that included an interaction between student of interest and Group (Gen Chem, STEM-Dawgs, or Volunteer). If in the full starting model the interaction did not help significantly predict the outcome, the interaction parameter was removed first.

As is typical in model selection procedures such as these, each model can be considered a distinct hypothesis being tested (Burnham and Anderson, 2002). For this reason, the *p*-values of best-fitting models are redundant and have not been specifically indicated. Instead, all parameters that were retained in the final model are interpreted with a magnitude and direction. All models were fit in R version 3.4.0 (R Core Team, 2017).

### What impacts does the STEM-Dawgs workshop have on course performance?

**Exam scores.** To test our hypothesis that the intensive workshop improved students' test scores in the first course of introductory chemistry (Chemistry 142) we collected scores from two midterms and a final exam from each section in each of the three terms. Because the number of total exam points and exam difficulty varied among sections, we calculated the percentage of total exam

points earned and converted the percentages to z-scores for each section. This z-score was the outcome in linear models. Because data came from three iterations of the STEM-Dawgs workshop and because student characteristics varied from quarter-to-quarter and section-to-section, we included predictor variables to control for confounding differences in course performance or affect driven by variation in student characteristics. Specifically, we used SAT-total score to control for variation in general academic preparation or ability, and the CCI pre-score to control for variation in incoming chemistry knowledge. Only the students with complete data could be used in the analysis. This criterion resulted in 2045 Gen Chem Student, 197 STEM-Dawgs, and 73 Volunteers being included, which was 88% of the total population. We fit linear models using the `lm` function in R (R Core Team, 2017) with SAT, CCI pre score, term or general chemistry section, comparison group, and demographics as fixed effects. Chemistry lecture section and term were tested as random effects but were never retained in the final model.

**DFW.** To test our hypothesis that the workshop improved (reduced) failure rates, we identified all students who earned a D, F or who withdrew from the course (the DFW rate) for all three quarters of Chem 142. All students who earned a D, F, or W were coded as a 1 and all other students were coded at 0. Using this information, we fit the DFW rate using logistic regression models in R with the same fixed effects as the exam scores. Only students with complete data were used in the analysis. This criterion resulted in 2051 Gen Chem Students, 197 STEM-Dawgs, and 73 Volunteers being included, which was 89% of the total population. We fit logistic regression models using the `glm` function in R (R Core Team, 2017) with SAT, CCI pre score, term or general chemistry section, comparison group, and demographics as fixed effects. Chemistry lecture section and term were tested as random effects but were never retained in the final model.

#### What impact does the workshop have on student affect?

**PERTS and ASCI-V2.** To test our hypothesis that the workshop improved affect as measured by PERTS and ASCI-V2, we fit the post-scores using linear regression in R. In addition to the fixed effects described above, our models included the pre-score on all affect measurements when the outcome variable was the corresponding post-score. In general, our goal was to have model output reflect how the Workshop impacted student performance and affect, all-else-equal. The PERTS instrument looked at four constructs separately; mindset had three questions and the rest had four questions, and all were on a Likert scale ranging from one to six. The ASCI-V2 had two constructs with four questions each, with a Likert scale of one to seven. In each case we converted the Likert-scale choices to a numerical scale and summed the total from each student's responses for each construct. Data for the pre and post scores were collected during each of the three quarters of Chem 142 and combined into one dataset. Students retained in the analysis must have completed both the pre and post surveys. This resulted in 1489 Gen Chem Students, 146 STEM-Dawgs, and 55 Volunteers for PERTS and 1486 Gen Chem Students, 145 STEM-Dawgs, and 55 Volunteers for ASCI-V2, which was 65% of the total population.

The survey data are highly bounded, like most survey data, and have a strong ceiling effect (see Appendix 3). A ceiling effect in survey data commonly occurs when some respondents who gave the highest response (6 in the case of PERTS and 7 in the case of ASCI-V2) would have responded even higher if they had been able to do so. This kind of bounding can be problematic when fitting a linear model because the model may predict values that exceed the highest possible value of the scale. In these cases, a value at or above the threshold (ceiling) will all be forced at the value of the threshold (ceiling). To be sure that this ceiling effect did not qualitatively impact our estimates, we fit censored regression models (also called Tobit models) using the `censReg` package in R (Henningsen, 2017).

Fitting censored regression models with `censReg` allows for specification of the upper and lower bound of the data (the ceiling and the floor respectively). In our case, our data did not approach the floor. Censored regression models account for ceiling (and floor) effects by modeling an uncensored latent outcome in place of the censored observed outcome (Henningsen, 2010).

Nonetheless, model selection with models that were fit using censored regression resulted in identical final, best fitting models. Comparing models fit with the censored regression to models fit with `glm`, the direction of the relationship was identical, but the strength changed (as is common in censored regression). For simplicity, output from the `glm` is reported here but the more numerically precise censored regression coefficients can be found in Appendix 4.

**Instrument population validation.** The ASCI-V2 has been shown to consist of two factors: (1) emotional satisfaction, and (2) intellectual accessibility, as described by Xu and Lewis (2011). Based on this result, we analyzed the two constructs separately. Similarly, we modeled results from the PERTS questions as a four-factor survey, consistent with its design and previous implementation (PERTS, 2018). The four PERTS factors represent mindset, belonging, relevance, and self-efficacy.

To provide a greater degree of confidence in the reliability of these factors, we calculated the Cronbach's  $\alpha$  for both instruments (see Appendix 5). A value 0.7 or greater is generally considered sufficient internal consistency (Murphy and Davidshofer, 2005). To measure the validity based on internal structure, we measured the comparative fit index (CFI) and the standardized root mean square residuals (SRMR) for each instrument (see Appendix 4). The PERTS results were modeled as a four-factor survey and the factors were allowed to correlate. The ASCI-V2 were modeled as a two-factor survey and the factors were allowed to correlate. The CFI and SRMR were determined using the `cfa` function in the `lavaan` package in R (Rosseel, 2012) using the maximum likelihood estimator. Good fit values are CFI > 0.95 and SRMR < 0.08 (Hu and Bentler, 1999).

The PERTS, modeled as a four-factor survey tool, was found to have a good fit as assessed by both CFI and SRMR. The ASCI-V2, modeled as a two-factor survey tool, had a low CFI value, but the SRMR was within good-fit parameters. That result, along with the good Cronbach's  $\alpha$  values and previous use of the ASCI-V2 as a two factor instrument (Brandriet *et al.*, 2011; Xu

and Lewis, 2011) encouraged us to interpret the ASCI-V2 as a two factor survey tool.

### Qualitative data collection

Qualitative data were collected using focus group interviews. All of the students participating in the companion course were sent an email soliciting their participation in the focus groups. Students who volunteered to participate were randomly placed into one of the two focus groups for that quarter. Of the students who did volunteer over the two quarters, four students had irreconcilable scheduling conflicts and were excluded from participating in the focus groups. A total of 16 students were interviewed in four focus groups conducted over two terms (two focus groups per term). Each focus group was held the week before final exams. The interview format was semi-structured, with questions focusing on participants' thoughts and opinions about the workshop, peer facilitators, impressions on the value of the either the chemistry discussion section or workshop, and the impact of the workshop on their overall chemistry grade. Each participant received a twenty-dollar gift card for volunteering. Focus groups were recorded using a digital audio recording device and then sent to a third-party transcription service. Transcripts were checked by one co-author (MP) for accuracy against the original audio recordings.

### Qualitative data analysis

Interview data were analyzed using a thematic content analysis—a qualitative research method that places emphasis on context and interpretation of text data (Vaismoradi *et al.*, 2013). A thematic content analysis is defined as a descriptive approach for identifying, analyzing patterns and reporting themes in qualitative data (Braun and Clarke, 2006). Thematic content analyses are differentiated from standard content analyses in that they typically involve a higher level of interpretation and data transformation than the numerical counting of written occurrences more typical of a “standard” content analyses (Vaismoradi *et al.*, 2013).

Coding was done by one of the co-authors (MP) and began with thorough readings of the transcripts and identification of segments of text in the Atlas.ti software package. Next, each identified segment of text was evaluated independently and characterized with a note. These notations allowed the researcher (MP) to characterize the explicit or implicit meaning of each quotation, which facilitated placing quotations into codes. This preliminary phase of coding led to several iterative and recursive rounds of combining and dividing codes based on the researcher's (MP) evolving interpretations of the data which allowed the analysis to determine the central components, or themes, of the student focus groups. Finally, the researcher used constant comparison to evaluate each quotation and determine if it was reasonable to assign the statement to multiple codes. While codes represent distinct concepts, statements from students are often complex and layered representations of their impressions of the companion course and general chemistry. To account for this complexity, students' statements were often assigned multiple codes where appropriate. Consequently, this allowed the researcher to calculate a code co-occurrence matrix using the Atlas.ti software.

Thematic content analysis employing code co-occurrences used in this study is constructivist in its orientation and assumes a level of subjectivity in the statements made by focus group participants. In order to maintain consistency and depth in the analyses, it was important to constantly reflect upon and revise the qualitative analyses as the quantitative analyses and results evolved over the course of the study. We chose to address this by having a single researcher manage the analysis of all the focus group data instead of relying on the agreement of multiple raters. We acknowledge that interrater reliability can enhance the trustworthiness of some qualitative interpretations and is well-suited to other forms of content analysis. However, our design choice allowed a researcher (MP) experienced with the iterative and recursive nature of thematic analyses to ensure that the data were viewed from a consistent perspective throughout the study. This helped support the trustworthiness and consistency of the qualitative analysis and our choice to omit interrater agreement is consistent with other explanatory study designs in chemistry education (refer to Vishnumolakala *et al.*, 2017).

## Results

### Quantitative results

#### What impact does the STEM-Dawgs Workshop have on course performance?

*Exam scores.* Controlling for SAT-total score, CCI pre-score, and term, the STEM-Dawgs—a population highly enriched in students who historically struggle in general chemistry—achieved exam scores that were indistinguishable from those of the Gen Chem Students. The Volunteers, in contrast, earned 0.26 standard deviations lower than the STEM-Dawgs (Fig. 1).

There was a trend for first-generation students to underperform relative to non-first-generation students in General Chemistry. This gap was slightly reduced for STEM-Dawgs and the trend reversed in the Volunteer population—meaning that first-generation Volunteers did exceptionally well. All of these trends were small, however, even though the interaction term was retained in the final model. First-generation Gen Chem students show a statistically significant achievement gap of 0.08 standard deviations lower for first-generation relative to non-first generation students. Opposite of our expectations, first-generation Volunteers scored 0.49 standard deviations higher as Volunteers (see Appendix 6). No achievement gap was found between first-generation and non-first-generation STEM-Dawgs. Women, on the other hand, scored 0.15 standard deviations lower than men in General Chemistry, in contrast to finding no gender achievement gap in STEM-Dawgs or Volunteers (controlling for SAT, CCI, and Term, see Appendix 6).

*DFW rates.* There were no overall differences in DFW rate among STEM-Dawgs, Gen Chem students, and Volunteers. However, URM students who participated in the workshop had the same chance of failing the course (3%) as URM students in the Gen Chem group (5%), but much lower DFW rates than URM students in the Volunteer Control group (11%). Further,

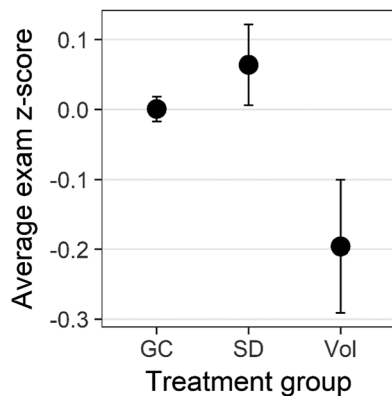


Fig. 1 Fitted standardized exam scores show that Volunteers (Vol) average scores that are 0.20 standard deviations lower than Gen Chem Students (GC) and 0.26 standard deviations lower than STEM-Dawgs (SD) when accounting for CCI scores, SAT-total scores, and Term. Vertical bars indicate standard errors.

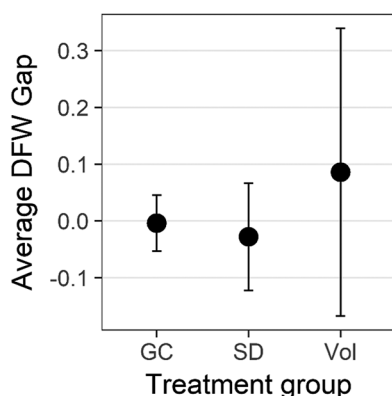


Fig. 2 Modeled DFW gaps in probability between URM and non-URM students show a significant difference between Gen Chem Students (GC) and Volunteers (Vol) when controlling for SAT, CCI, and chemistry lecture section. A gap greater than zero indicates a larger DFW rate for URM students relative to non-URM students. A negative gap means that the DFW rate for URM students is lower than that of non-URM students. Vertical bars are standard errors.

there was no URM *versus* non-URM achievement gap in the STEM-Dawgs and Gen Chem populations, although there was a significant achievement gap in the Volunteer group (Fig. 2). We found similar trends for students who are first-generation or EOP, though they were not statistically significant (see Appendix 4).

**What impact does the workshop have on student affect?** In comparing pre- and post-scores on each of the four measures of affect captured by the PERTS survey, many students show a decrease (see Appendix 7), which is common when measuring affect for lecture courses (Berg, 2005; Adams *et al.*, 2006; Barbera *et al.*, 2008; Semsar *et al.*, 2011; Ding and Mollohan, 2015). There were, however, significant differences among the Gen Chem Students, STEM-Dawgs, and Volunteer populations in the extent of the increase or decline in scores, for all

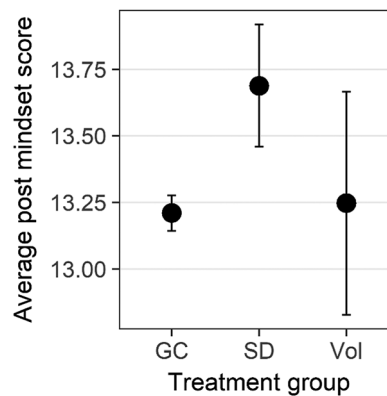


Fig. 3 When accounting for CCI and chemistry lecture section, STEM-Dawgs (SD) show a higher score on the post-mindset survey, on average, than the Gen Chem Students (GC).

measures except Relevance. All affect results below are fitted values of post-scores when taking into account the pre-score.

**PERTS: mindset.** Scoring higher on the Mindset portion of PERTS means that students have more of a growth mindset rather than a fixed mindset. STEM-Dawgs had a small but significantly higher score on mindset than the Gen Chem Students, meaning they had a mindset that was more growth-oriented. STEM-Dawgs scored 0.5 points higher, on average, out of a total 18 points possible (Fig. 3).

**PERTS: sense of belonging.** The STEM-Dawgs felt a higher sense of belonging in the general chemistry course compared to the Gen Chem Students and Volunteers. STEM-Dawgs scored 1.1 points higher than the Volunteers, on average, out of a total of 24 points possible (Fig. 4A). Additionally, the STEM-Dawgs also trend toward having a greater sense of belonging than Gen Chem students, scoring on average 0.5 points higher.

Furthermore, we found that women Volunteers scored 0.9 points less than STEM-Dawgs and 1.1 points lower than the Gen Chem Students, on average. In contrast to the Gen Chem Students, there were gender gaps in belonging for the other two study populations, though the male–female gap was much less in the STEM-Dawgs population (1.4 points) than in the Volunteer population (1.8 points; Fig. 4B).

**PERTS: feelings of relevance.** We found no significant difference among the treatment groups for feelings of relevance, however STEM-Dawgs trended toward showing 0.5 points higher than Gen Chem Students. There was also a trend for URM students in the workshops to score higher than URMs in the Gen Chem populations, non-URM students who were STEM-Dawgs, and all Volunteers (see Appendix 4). Women show a statistically significant difference of 0.4 points, on average, higher than men in the workshop (see Appendix 4).

**PERTS: self-efficacy.** STEM-Dawgs had higher self-efficacy scores than the Volunteers: 1.6 points out of 24 possible—and matched the scores of the Gen Chem Students (Fig. 5). There



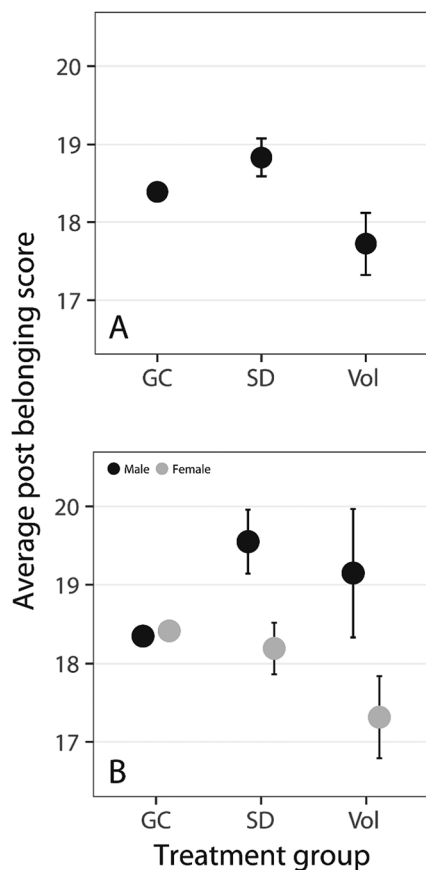


Fig. 4 (A) STEM-Dawgs (SD) show a higher sense of belonging on the post survey compared to Gen Chem Students (GC) and Volunteers (Vol) when controlling for SAT, CCI and general chemistry lecture section. (B) Women Volunteers show a lower sense of belonging than volunteer men. Men and women in the volunteer group are more different than male and female STEM-Dawgs. Vertical bars indicate standard errors and are present in the GC population in (A) and (B) but subsumed by the point.

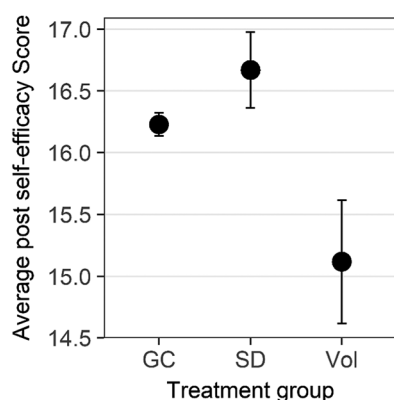


Fig. 5 Post self-efficacy scores show that STEM-Dawgs (SD) had a higher sense of self-efficacy than the Volunteers (Vol) when controlling for SAT, CCI and chemistry lecture section. Error bars are standard error.

were no differences among Gen Chem Students, STEM-Dawgs, and Volunteers in any of the sub-populations of interest.

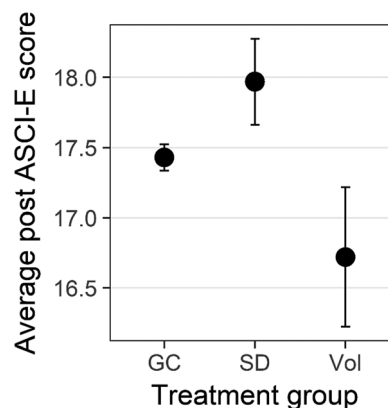


Fig. 6 Average ASCI-V2 emotional satisfaction (ASCI-E) scores are 1.2 points higher for STEM-Dawgs (SD) than the Volunteers (Vol) when controlling for SAT, CCI, and Chem 142 section.

*ASCI-V2.* STEM-Dawgs as a whole ended the term with higher emotional satisfaction with studying chemistry (measured by the ASCI-V2) than the Volunteers by 1.2 points, on average, out of 28 (Fig. 6). They trended toward higher scores than the Gen Chem population by 0.5 points. No sub-groups benefited from the workshop in terms of intellectual accessibility with studying chemistry. An opposite trend was found in EOP *versus* non-EOP, where EOP students in the workshops showed a negative shift in intellectual satisfaction toward chemistry that also occurred in the Gen Chem and Volunteer populations (see Appendix 4).

### Qualitative results

The themes that emerged from the analysis are encompassed in the following six codes:

- *Course Structure*, which encompassed statements about the overall design of the course;
- *Classroom Activities*, which detailed events that took place during workshop sessions;
- *Classroom Culture*, which included interaction(s) between peer facilitators and students, as well as comments about the social aspects of the course;
- *Affect*, which comprised statements dealing with students' attitudes or emotions about science, chemistry, or the workshop;
- *Peer Facilitator*, which consisted of all statements that directly or indirectly related to the undergraduate course leaders;
- *Impact*, which contained statements that dealt with the workshop's influence on students' overall chemistry course experience or academic behavior.

Table 1 summarizes the number of quotations from focus group participants assigned to each of the six codes, and an example quotation from the focus groups that is characteristic of each code. Due to students' statements often being qualified by multiple codes, the counts shown in Table 1 do not represent discrete bins of quotations but rather the number of times that a particular code was applied to statements from the focus group data.

**Table 1** Codes used in the directed content analysis, the number of quotations in each code and an example quotation intended to summarize each code's meaning

Code	Number of quotations	Example quotation
Course structure	17	"It was basically an extra help class. I saw it as a support class to Chemistry 142."
Classroom activities	21	"The peer facilitators would split us into different chem[sic] groups based on our professor, so we would work together. ..."
Culture	33	"...if you got it, then you could explain things to other people and if you didn't get it, then people could explain things to you..."
Affect	24	"Empathy made the class setting more comfortable."
Peer facilitators	67	"They worked together, I wasn't sure if they were friends beforehand, but they seemed like friends."
Impact	30	"It made my learning easier, which made chemistry easier."

Peer facilitators were the most frequently discussed topic during focus groups. Statements from students about peer facilitators were diverse and ranged from positive statements like "it was still fresh in their minds [referring to their content knowledge]" to more critical statements such as "mentors [peer facilitators] were a little too friendly and lost authority." Statements qualified by the Culture code resulted in the second highest numbers of coded statements. Student statements about culture were often reflective of the companion course being a much more friendly, close and group-oriented environment than their chemistry lab or discussion sections. For example, a student stated that "we would all just get the right answer anyway because there were so many people working on it together, we didn't have a problem about finding an answer." Statements from the Impact code had the third highest occurrence among students' talk. Students' perceptions on the impact of the companion varied from explicitly positive statements to statements where they suggested no impact at all. However, some of the comments explicitly stating that it did not help would be associated with segments of talk that alluded to positive impacts of the companion course. For example, a student stated that "I don't think it impacted my grade at all. It impacted how I felt about the course, more positively." In that statement it is clear the student did not think that the companion course improved their grades but they do seem to state that it improved their experience with chemistry. The Affect and Classroom Activities codes have roughly the same number of occurrences. Many of the statements coded as Affect, such as "I felt more comfortable asking questions" and "You could ask anything, whereas if you're in one of those 800-people classes with a professor, it's more intimidating to ask a question you may think is stupid. ...", alluded to the students' perceiving a greater degree of comfort with the companion course. Statements in the Classroom Activities code captured student talk about what actually happened during the companion course classes. Students would often talk about working in groups as well as working through the psychological interventions and study skills activities. The Course Structure code applied to the smallest number of statements made by students, but still provided insight into students' perceptions of the companion course. Statements found within this code encompassed students' frustration with their perception of how much work they felt they had to do in the workshop

**Table 2** Code co-occurrences matrix for all six codes. A dash represents a code co-occurrence of zero

	1	2	3	4	5	6
(1) Course structure	N/A					
(2) Classroom activities	0.03	N/A				
(3) Culture	—	0.15	N/A			
(4) Affect	—	—	0.16	N/A		
(5) Peer facilitators	—	0.02	0.12	0.06	N/A	
(6) Impact	—	—	0.07	0.08	—	N/A

conflicting with the workshop's use of low-stakes assessments. For example, "I feel like I didn't try as hard on homework assignments because they weren't graded. Basically, if you show up, you get points."

Code co-occurrences represent the level of association between codes (Table 2). For example, a student made the following comment about their peer facilitators: "they [peer facilitators] worked together. I wasn't sure if they were friends beforehand, but they seemed like friends. They got along really well." This quotation was qualified by the Peer Facilitators code because the facilitators are the focus of the comment. Additionally, the quotation was also qualified by the culture code due to the student commenting how they perceived that the peer facilitators got along and how they worked together. Because this statement received a Peer Facilitator and Culture code this would mean that those codes are co-occurring. A value of 1 means that two codes always co-occur together while a value of zero means that two codes never co-occur on the same quotation. We interpreted that the higher the code co-occurrence value the greater the conceptual relationship between two codes. The three highest co-occurrences are the Affect, Culture, and Peer Facilitators codes. The highest code co-occurrence was between the Affect and Culture codes while the least co-occurring codes was between the Peer Facilitators code and Classroom Activities. While the peer facilitators in the workshop were in charge of facilitating classroom activities, it is important to note that the focus groups revealed a higher association between Peer Facilitator code and the Culture code, suggesting that peer facilitators had a stronger influence on the classroom culture of the companion course than did the classroom activities. The Impact code also more strongly associated with the Affect and Culture codes (Table 2), suggesting that the impacts of the companion course were less related to direct

performance benefits and more associated with emotional aspects of chemistry.

## Discussion

The STEM-Dawgs workshops improved important aspects of the course performance and affect of participants (Table 3). In some cases, the workshops were also able to close achievement gaps for sub-populations that traditionally underperform in general chemistry. In many cases, the performance and affect of a population that was highly enriched in at-risk students (*e.g.*, STEM-Dawgs) was indistinguishable from a population with a much smaller percentage of at-risk individuals (*e.g.*, Gen Chem), meaning that the workshops leveled the playing field in general chemistry.

Improved exam scores and lower failure rates may be particularly important for underrepresented students who traditionally have been discouraged from pursuing medical school by poor performance in general chemistry (Barr *et al.*, 2008), which is the goal of a large portion of our undergraduate STEM majors. It is notable that unlike Gen Chem Students, there was no achievement gap in exam scores for women and first-generation students in the workshop, and no achievement gap in failure rates for URM individuals who participated.

Data from the focus group interviews supports the hypothesis that increased course performance was due, at least in part, to the deliberate practice that occurred in the workshops (Haak *et al.*, 2011). Many of the student comments coded as Classroom Activities focused on the active learning that occurred in the peer-facilitated sessions, and especially the group work. For example, students made comments such as "...since you had to discuss it in groups you were forced not to lie that you didn't get it..." and "they [peer facilitators] would prompt students to help each other in that way. They weren't necessarily teaching – students were teaching each other and helping each other, which I thought was super useful." The focus on group work may also have facilitated a transfer of learning strategies between students with variable levels of academic skill and

experience. For example, one student summarized the workshop's impact by saying "... (I) definitely think it helped, not because of the class itself, but just the people I met there. I met some really hardcore study geniuses, they dragged me along with them. That pulled me up better than I would've done alone."

### Changes in affect

The declines that occurred from the start to the end of the course in all aspects of affect are consistent with the findings of other studies on science courses (Berg, 2005; Adams *et al.*, 2006; Barbera *et al.*, 2008; Semsar *et al.*, 2011; Ding and Molloyhan, 2015). Given that pattern, changes in most aspects of student affect also indicated that the workshop was beneficial—meaning that workshop participants declined less. One exception was Relevance. Even though the course incorporated content using real world examples, we observed no difference in content relevance among the three treatment groups.

Although the workshops included explicit training in the challenge mindset required to overcome academic obstacles, we observed only a 0.5 point difference in Mindset among the three treatment groups. As Yeager *et al.*, (2016) suggest, the benefits of a challenge mindset may now be so ingrained in the public consciousness that the intervention's efficacy is diluted. The other aspects of affect captured in the PERTS survey questions—belonging and self-efficacy—indicated that STEM-Dawgs had much better attitudes about the general chemistry course than Volunteers, and in the case of belonging, better than Gen Chem Students. Male workshop participants reported a much higher sense of belonging.

Focus group data suggest that these results could be due to the workshop establishing a more-positive classroom culture around the topic of chemistry than the traditional lecture or laboratory. Code co-occurrences suggest that the classroom culture was created by both the classroom activities and peer facilitators (Table 2); an insight summarized in the following quote: "I think the way it was led was more of a really organized study group of friends." This statement suggests that students perceived the workshop to be a more comfortable learning

**Table 3** Summary of the impacts of the STEM-Dawgs workshops through model selection. A complete list of final models and their retained effects can be found in Appendix 4. Each model either tested the overall population (*i.e.*, without student group) or individually tested an interaction between student group (*i.e.*, URM, EOP, First Generation (FG), Gender (women) and course intervention (with Chem 142 and Volunteers compared to STEM Dawgs)). If the interaction was retained in the model, the direction and magnitude of the effect is indicated with letters A/a–f. Magnitude of the effect is approximated with capital letters, indicating that the relationship was significant ( $p < 0.05$ ) and lower case letters indicating that the relationship was retained but not significant to  $p < 0.05$ . A/a indicates that STEM-Dawgs outperform Volunteers. B/b indicates that STEM-Dawgs outperform Gen Chem. Lower case c indicates that there is no difference between STEM-Dawgs and Volunteers, d indicates that there is no difference between STEM-Dawgs and Gen Chem, e indicates that STEM-Dawgs underperformed relative to Gen Chem, and f that STEM-Dawgs underperformed relative to Volunteers. Blank cells indicate that the interaction was not retained in the final model. Coefficients for each model can be found in Appendix 4

	Exams	DFW	Mindset	Belonging	Relevance	Self-efficacy	Intellectual accessibility	Emotional satisfaction
Overall	Ad		Ba	Ab	ab	Ad		Ab
URM		Ad	ce		ab			
EOP		ab						
FG	ef	ab						
Women				ae				

A/a: STEM-Dawgs > Volunteers; B/b: STEM-Dawgs > GenChem; c: STEM-Dawgs = Volunteers; d: STEM-Dawgs = GenChem; e: STEM-Dawgs < GenChem; f: STEM-Dawgs < Volunteers.

environment than their laboratory or lecture sections. The workshop's culture was also perceived as a more-accepting environment, where students felt surrounded by individuals like themselves. "I think it made me more comfortable with chemistry, just being in the class and seeing that other people were struggling and knowing that when you come in with questions, you're not the only one that's struggling." Gaining an "I'm not the only one" perspective can lead to a greater sense of belonging; students who feel like they belong are more likely to persist and perform well in STEM courses (Trujillo and Tanner, 2014; Hanauer *et al.*, 2016).

The more-positive sense of belonging that we observed for female STEM-Dawgs *versus* Volunteers may be particularly important. Research in other fields suggests that women are much more "grade-sensitive" than men—meaning that they are less confident about being successful and less likely to persist than men who are performing at the same level (Rask and Tiefenthaler, 2008; Ellis *et al.*, 2016). Further research should explore whether the higher sense of belonging declared by women in the workshops helps to mitigate the achievement gaps we observed in exam scores and helps them persist in STEM.

Many of the quotes in the Impact code also refer to the workshop's influence on self-efficacy, which has both emotional and academic benefits (Lewis *et al.*, 2009; Sawtelle *et al.*, 2012; Bjørnebekk *et al.*, 2013; Chan and Bauer, 2014; Larson *et al.*, 2015). For example, students made comments such as and "I'd go to lecture and be like [*sic*], 'Oh, I learned that in [the workshop] yesterday.' I could sit there and be like, 'Okay,' and take notes and understand. I learned it previously. . . That was nice," and, "I felt more empowered in the class just because I was capable of showing I could do the problem."

Similar to relevance, the workshop had no impact on the ASCI-V2 intellectual accessibility construct. This scale measures what students believe and know about how to study chemistry and did not change even though STEM-Dawgs had explicit training in research-based study skills. Indeed, gaps in the intellectual accessibility scale occurred for EOP *versus* non-EOP workshop participants—opposite the direction intended by the workshop's design. Did the initial exposure to research-based study skills negatively impact students who realized that their existing study skills are contraindicated by evidence? Would repeated exposure to study skills training produce a better outcome? With respect to the ASCI-V2 emotional scale, however, it was clear that STEM-Dawgs scored higher than both Gen Chem Students and the Volunteers. The workshop helped students feel better about studying chemistry—even though they may have been unsure of how to go about it.

### Comparisons to other SI and PLTL models

A great deal of research into PLTL models in STEM has shown that PLTL is a powerful model for academic improvement (reviewed in Wilson and Varma-Nelson, 2016). Our research design adds to this literature because we were able to compare outcomes with a volunteer-control group. Many SI and PLTL

programs are opt-in programs, making results difficult to interpret due to sampling bias.

The only other SI study in General Chemistry with a volunteer-control, by Chan and Bauer (2014), also used peer facilitators for collaborative group work in 50 or 80 minute supplementary meetings to work on material from that week's lectures. Similar to our study, they show slight decreases in affect over the quarter for the emotional satisfaction factor for the ASCI (as in Appendix 7). However, that study showed no difference in exam scores between men and women or between PLTL and non-PLTL students. The contrast between our results and those of Chan and Bauer (2014) suggest that adding study skill education and psychological interventions during SI may improve students' grades.

PLTL studies that show impacts similar to ours suffer from concerns over student equivalence. For example, a workshop program for student volunteers used peer facilitators for 2 hours a week and showed a lowered rate of students earning D or F in chemistry if participating, as well as a larger increase in grades for minority volunteers compared to non-minority volunteers (Drane *et al.*, 2005).

A more recent study of PLTL avoided this self-selection bias by comparing student outcomes in a traditional course design to a course that dropped one lecture session per week in general chemistry and substituted a required 50 minute PLTL session per week. This model moved PLTL out of the supplemental instruction format and into an element of the course required of all students. Parallel to our study, this data set showed improved retention of all students and a disproportionate improvement for underrepresented minority students (Lewis, 2011).

### The role of peer facilitators

Peer facilitators are a central component in the design of the STEM-Dawgs Workshop; our focus group data suggests that they were also a key element in its success. Statements about peer facilitators often co-occurred with statements about classroom culture. Many of these statements were positive and reflected an open and accepting environment that was guided by interactions with peer facilitators. In particular, focus group participants emphasized the importance of peer facilitators' empathy with respect to students' problems with chemistry. When the interviewer asked whether empathy was important, students would often state something similar to "It [empathy] made the class setting more comfortable," and that "being comfortable" helped them learn. Additionally, the interviews suggested that the peer facilitators' positive attitudes influenced the workshop participants. For example, one student stated "my teachers [peer facilitators] were super positive and they were a bunch of people who wanted to be there. But by the end of the class period, it was like, alright, we learned."

In addition to being empathetic and positive, peer facilitators were perceived as more relatable and approachable than graduate teaching assistants or professors. One student stated that "Having them be [*sic*] peers was really helpful in a lot of cases because they were like, 'I remember when I took this.



I didn't understand it either.' So, you were like, 'okay, I'm not alone in this'." This result echoes a recent study showing that near-peer teachers in a medical course were perceived as more approachable than traditional teachers, and that the near-peers appeared to be more aware of learning outcomes and more invested in exam success (Tayler *et al.*, 2015).

### Limitations of this study

The STEM-Dawgs workshop design had a series of elements that included cooperative group learning, psychological interventions, study skills training, two-hour weekly meetings in small sections, and a prominent role for peer facilitators. We cannot, however, parse whether certain components in the workshops are more important than others. Would the course have worked as well without the study skills training? Were the interventions that were designed to ease the transition to college effective because they were done in combination? We hypothesize that the answers to these questions are no and yes—that the elements included in the STEM-Dawgs design are synergistic.

We also do not know how much of the program's efficacy was due to increased time-on-task, and whether the workshop will "travel." Unlike widely implemented SI models like PLTL in general chemistry or the calculus (Triesman) workshops, the STEM-Dawgs model has only been tested at a single institution and in a single course. Collecting data on total time-on-task and testing this workshop model with other student populations are important topics for future research.

Our explanatory mixed methods design used focus group data to interpret results from the quantitative analyses. Consequently, we did not use interrater agreement to build trustworthiness, rather, we relied on a single researcher to ensure that data were consistently interpreted. We acknowledge that this can lead to a certain level of subjectivity in the interpretation of focus group data. Interpretations from focus groups are also limited by the sample size and non-purposeful sampling of interview participants. Specifically, they are subject to volunteer

bias. In future studies, it would be helpful to explore key hypotheses in a prospective framework: that feelings of belonging and self-efficacy were particularly important aspects of the experience, and that the presence of empathetic and positive peer facilitators is critical to classroom culture.

Finally, although it appears that peer facilitators were a key design element in our workshop, we have yet to study the peer facilitators themselves. Previous work in chemistry has shown that students who mentor other students in a chemistry course show increased grades in a subsequent chemistry course (Amaral and Vala, 2009). Does working as a peer facilitator in the STEM-Dawgs model, and leading discussions about study skills and psychological issues, provide any additional benefit?

## Conclusion

The STEM-Dawgs workshop is, to the best of our knowledge, the first SI design to combine (1) intensive group practice with exam-like problems, (2) training in research-based study skills, and (3) evidence-based interventions that function as psychological and emotional supports. The workshop improved course performance and affect in general chemistry in a population enriched in students who are underrepresented in STEM fields. The workshop also showed benefits in latent traits that promote STEM success and retention, such as a sense of belonging and self-efficacy. Designing SI that specifically targets group practice, study skills, and emotional support holds promise for supporting equity and inclusion in the general chemistry classroom.

## Conflicts of interest

There are no conflicts to declare.

## Appendix 1: student demographics

Table 4.

**Table 4** Demographics of STEM-Dawgs participants, Volunteer controls, and the remaining Chemistry 142 students for all quarters and then combined into the last column on the right

		Fall	Winter	Spring	Total
STEM-Dawgs	Total #	87	82	43	212
	% URM	20%	39%	35%	30%
	% EOP	36%	51%	40%	42%
	% FG	43%	61%	49%	51%
	% Women	67%	59%	60%	62%
Volunteers	Total #	33	47	0	80
	% URM	12%	28%	0	21%
	% EOP	24%	47%	0	38%
	% FG	36%	55%	0	48%
	% Women	73%	79%	0	76%
Gen Chem	Total #	1318	704	302	2324
	% URM	10%	21%	20%	15%
	% EOP	16%	28%	31%	22%
	% FG	30%	40%	41%	34%
	% Women	52%	56%	52%	53%

## Appendix 2: STEM-Dawgs weekly schedule

Table 5.

Table 5 Weekly schedule of the STEM-Dawgs workshops

	Activity to be done before class	In-class activities	Study skill
Week 1	NO CLASS		
Week 2 (week of October 3rd)		<ul style="list-style-type: none"> <li>• Course information</li> <li>• Class introductions</li> <li>• Rubber Band Blast</li> </ul>	Suggestions for doing well on exams as told by previous students. Information sheet for use in study activities. Study Skills: (Dunlosky <i>et al.</i> , 2013)
Week 3	Writing exercise 1: low-stakes practice writing (online)	<ul style="list-style-type: none"> <li>• Quiz 1 (Chapter 2 and 12)</li> <li>• Group work</li> </ul>	Test enhanced learning, (Roediger <i>et al.</i> , 2011)
Week 4	Writing exercise 2: WOOP (online) Online Mock Exam 1	<ul style="list-style-type: none"> <li>• WOOP discussion</li> <li>• Collaborative mock exam key</li> </ul>	Productive failure, (Kapur and Bielaczyc, 2012)
Week 5 CHEM 142 EXAM 1	Writing exercise 3: expressive writing to be done as close to your CHEM 142 exam as possible (online)	<ul style="list-style-type: none"> <li>• Quiz 2 (Chapter 3)</li> <li>• Group work</li> </ul>	Elaborative interrogation, (Woloshyn, Pressley and Schneider, 1992)
Week 6	Writing exercise 4: learning sciences (online) Post exam 1 metacognition	<ul style="list-style-type: none"> <li>• Quiz 3 (Chapter 3)</li> <li>• Group work</li> </ul>	Interleaved practice, (Rohrer and Taylor, 2007)
Week 7	Writing exercise 5: first year transition (online)	<ul style="list-style-type: none"> <li>• Videos from previous students &amp; discussion</li> <li>• Quiz 4 (Chapter 4)</li> <li>• Group work</li> </ul>	Distributed practice, (Richter and Gast, 2017)
Week 8	Writing exercise 6: the value of learning (online) Online Mock Exam 2	<ul style="list-style-type: none"> <li>• Discussion of reasons for learning.</li> <li>• Collaborative mock exam key</li> </ul>	Self-explanation, (Berry, 1983)
Week 9 CHEM 142 EXAM 2	Writing exercise 7: low-stakes writing II Writing exercise 8: expressive writing 2 to be done as close to your CHEM 142 exam as possible (online)	NO CLASS	
Week 10	Writing exercise 9: everyday chemistry (online) Post exam 2 metacognition	<ul style="list-style-type: none"> <li>• Sharing resources used for writing exercise 8</li> <li>• Quiz 5 (Chapter 15)</li> <li>• Group work</li> </ul>	Re-reading, (Rothkopf, 1968) Or sleep deprivation, (Pilcher and Huffcutt, 1996)
Week 11	Online practice final exam	<ul style="list-style-type: none"> <li>• Collaborative mock final exam key</li> </ul>	Concept mapping Suggestions for doing well on exams as told by previous students.
Final exam week	Writing exercise 10: expressive writing 3 to be done as close to your CHEM 142 exam as possible (online)	<ul style="list-style-type: none"> <li>• Group picture, swag</li> </ul>	

## Appendix 3: bounded results

Fig. 7.

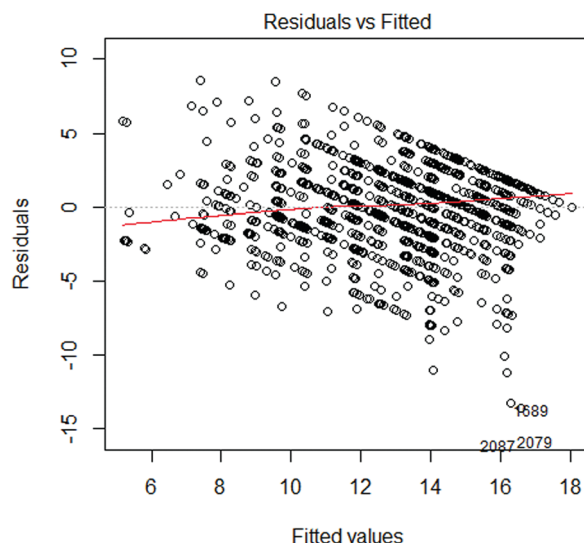


Fig. 7 An example of the residuals as a function of the fitted values in a linear regression of survey data. The bounded nature of the fit results in the ceiling effect seen at the upper right in the figure. This specific example is from the overall Mindset post scores as a function of pre-scores, SAT, CCI, and Gen Chem section. Censored regression is suggested with bounded data such as this. See Appendix 4 for comparisons of linear and censored regression on survey data.

## Appendix 4: beta values from model fits

Tables 6 and 7.

**Table 6** Beta values from models describing the variables retained in the models for Exam performance and DFW rates. All models were fit with STEM-Dawgs as the comparison. A value is given for any variable retained in the model. Coefficients with a  $p \leq 0.05$  are in bold. Note that for DFW, the log(odds) are higher for student groups who perform worse

Construct	Group of interest	Intercept	CHEM 142	Vol	SOI	SOI* CHEM 142	SOI*Vol
Exam	Overall <sup>a</sup>	−0.049	−0.063	<b>−0.260</b>	—	—	—
	URM <sup>b</sup>	−0.057	−0.071	<b>−0.254</b>	−0.088	—	—
	EOP <sup>b</sup>	−0.050	−0.072	<b>−0.251</b>	−0.081	—	—
	First gen. <sup>a</sup>	−0.002	−0.082	<b>−0.514</b>	−0.107	0.020	<b>0.536</b>
	Women <sup>b</sup>	0.078	−0.071	<b>−0.235</b>	<b>−0.146</b>	—	—
DFW <sup>c</sup>	Overall <sup>b</sup>	−3.335	—	—	—	—	—
	URM <sup>b</sup>	<b>−3.205</b>	−0.123	−1.040	−0.627	0.541	<b>2.290</b>
	EOP <sup>b</sup>	<b>−3.005</b>	−0.356	−0.778	−0.861	0.903	1.367
	First gen. <sup>b</sup>	<b>−3.163</b>	−0.380	−0.760	−0.367	0.841	1.117
	Women <sup>b</sup>	<b>−3.612</b>	0.110	−0.093	<b>0.326</b>	—	—

<sup>a</sup> Controlling for SAT, CCI, and Term. <sup>b</sup> Controlling for SAT, CCI, and CHEM 142 section. <sup>c</sup> Betas are log(odds) from a logistic regression.

**Table 7** Beta values from models describing the variables retained in the affect models. Bold values have  $p < 0.05$ . Affect models were fitted using OLR and censored regression. Coefficients of both are shown next to each other for comparison

Construct	Group of interest	Regression	Intercept	Pre score	CHEM 142	Vol	SOI	SOI: CHEM 142	SOI: Vol
Mindset post	Overall	Linear <sup>a</sup>	3.604	0.729	−0.479	−0.441	—	—	—
		Censored <sup>a</sup>	2.617	0.822	−0.585	−0.547	—	—	—
	URM	Linear <sup>a</sup>	3.751	0.729	−0.662	−0.562	−0.628	0.879	0.511
		Censored <sup>a</sup>	2.816	0.822	−0.836	−0.752	−0.846	1.226	0.875
	EOP	Linear <sup>a</sup>	3.604	0.729	−0.479	−0.441	—	—	—
		Censored <sup>a</sup>	2.572	0.818	−0.545	−0.551	0.327	—	—
	First gen	Linear <sup>a</sup>	3.604	0.729	−0.479	−0.441	—	—	—
		Censored <sup>a</sup>	2.617	0.822	−0.585	0.547	—	—	—
	Women	Linear <sup>a</sup>	3.604	0.729	−0.479	−0.441	—	—	—
		Censored <sup>a</sup>	2.617	0.822	−0.585	−0.547	—	—	—
Belonging post	Overall	Linear <sup>b</sup>	8.540	0.557	−0.443	−1.110	—	—	—
		Censored <sup>b</sup>	7.960	0.596	−0.502	−1.186	—	—	—
	URM	Linear <sup>b</sup>	8.540	0.557	−0.443	−1.110	—	—	—
		Censored <sup>b</sup>	7.960	0.596	−0.502	−1.186	—	—	—
	EOP	Linear <sup>b</sup>	8.540	0.557	−0.443	−1.110	—	—	—
		Censored <sup>b</sup>	7.960	0.596	−0.502	−1.186	—	—	—
	First gen	Linear <sup>b</sup>	8.540	0.557	−0.443	−1.110	—	—	—
		Censored <sup>b</sup>	7.960	0.596	−0.502	−1.186	—	—	—
	Women	Linear <sup>b</sup>	9.220	0.559	−1.199	−0.398	−1.357	1.424	−0.478
		Censored <sup>b</sup>	8.728	0.598	−1.340	−0.549	−1.489	1.543	−0.374
Relevance	Overall	Linear <sup>a</sup>	4.681	0.696	−0.517	−0.785	—	—	—
		Censored <sup>a</sup>	4.038	0.733	−0.514	−0.811	—	—	—
	URM	Linear <sup>a</sup>	4.435	0.697	−0.265	−0.507	1.016	−1.113	−1.186
		Censored <sup>a</sup>	3.799	0.734	−0.268	−0.511	0.952	−1.099	−1.266
	EOP	Linear <sup>a</sup>	4.681	0.696	−0.517	−0.785	—	—	—
		Censored <sup>a</sup>	4.038	0.733	−0.514	−0.811	—	—	—
	First gen	Linear <sup>a</sup>	4.681	0.696	−0.517	−0.785	—	—	—
		Censored <sup>a</sup>	4.038	0.733	−0.514	−0.811	—	—	—
	Women	Linear <sup>a</sup>	4.706	0.682	−0.509	−0.847	0.457	—	—
		Censored <sup>a</sup>	4.067	0.719	−0.507	−0.873	0.445	—	—
Self-efficacy	Overall	Linear <sup>b</sup>	5.371	0.601	−0.441	−1.554	—	—	—
		Censored <sup>b</sup>	4.880	0.637	−0.544	−1.641	—	—	—
	URM	Linear <sup>b</sup>	5.371	0.601	−0.441	−1.554	—	—	—
		Censored <sup>b</sup>	4.880	0.637	−0.544	−1.641	—	—	—
	EOP	Linear <sup>b</sup>	5.371	0.601	−0.441	−1.554	—	—	—
		Censored <sup>b</sup>	4.880	0.637	−0.544	−1.641	—	—	—
	First gen	Linear <sup>b</sup>	5.371	0.601	−0.441	−1.554	—	—	—
		Censored <sup>b</sup>	4.880	0.637	−0.544	−1.641	—	—	—
	Women	Linear <sup>b</sup>	6.445	0.573	−0.476	−1.503	−1.004	—	—
		Censored <sup>b</sup>	6.037	0.607	−0.581	−1.583	−1.082	—	—
ASCI Int. Sat.	Overall	Linear <sup>b</sup>	5.791	0.577	—	—	—	—	—
		Censored <sup>b</sup>	0.545	0.593	—	—	—	—	—
	URM	Linear <sup>b</sup>	5.791	0.577	—	—	—	—	—
		Censored <sup>b</sup>	0.545	0.593	—	—	—	—	—
	EOP	Linear <sup>b</sup>	5.936	0.576	—	—	−0.588	—	—
		Censored <sup>b</sup>	5.700	0.592	—	—	−0.637	—	—
	First gen	Linear <sup>b</sup>	5.692	0.574	—	—	0.393	—	—
		Censored <sup>b</sup>	5.442	0.590	—	—	0.411	—	—
	Women	Linear <sup>b</sup>	6.280	0.564	—	—	−0.590	—	—
		Censored <sup>b</sup>	6.044	0.579	—	—	−0.602	—	—
ASCI Emo. Sat.	Overall	Linear <sup>b</sup>	7.852	0.562	−0.540	−1.248	—	—	—
		Censored <sup>b</sup>	7.658	0.573	−0.552	−1.272	—	—	—
	URM	Linear <sup>b</sup>	7.852	0.562	−0.540	−1.248	—	—	—
		Censored <sup>b</sup>	7.658	0.573	−0.552	−1.272	—	—	—
	EOP	Linear <sup>b</sup>	7.852	0.562	−0.540	−1.248	—	—	—
		Censored <sup>b</sup>	7.658	0.573	−0.552	−1.272	—	—	—
	First gen	Linear <sup>b</sup>	7.852	0.562	−0.540	−1.248	—	—	—
		Censored <sup>b</sup>	7.658	0.573	−0.552	−1.272	—	—	—
	Women	Linear <sup>b</sup>	7.852	0.562	−0.540	−1.248	—	—	—
		Censored <sup>b</sup>	7.658	0.573	−0.552	−1.272	—	—	—

<sup>a</sup> Controlling for CCI and CHEM 142 section. <sup>b</sup> Controlling for SAT, CCI, and CHEM 142 section.



## Appendix 5: instrument validation

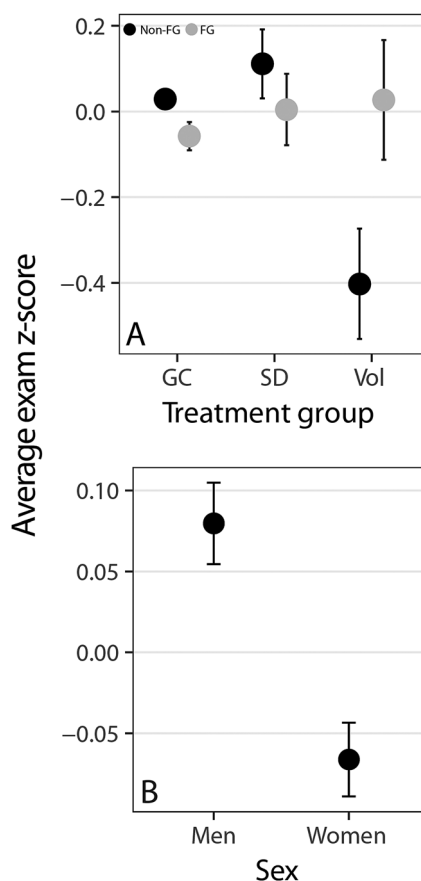
Table 8.

**Table 8** Cronbach's  $\alpha$ , CFI and SRMR of the PERTS and ASCI-V2 instruments. For Cronbach's  $\alpha$ , a value  $>0.7$  is generally considered to be reliable (Murphy and Davidshofer, 2005). For CFA, good fit values are CFI  $> 0.95$  and SRMR  $< 0.08$  (Hu and Bentler, 1999)

		$\alpha$ pre	$\alpha$ post	CFI pre	CFI post	SRMR pre	SRMR post
PERTS $N = 1690$	Mindset	0.91	0.92	0.961	0.957	0.044	0.044
	Belonging	0.82	0.86				
	Relevance	0.83	0.85				
	Self-efficacy	0.90	0.90				
ASCI-V2 $N = 1686$	Intellectual accessibility	0.77	0.79	0.882	0.897	0.071	0.066
	Emotional satisfaction	0.72	0.75				

## Appendix 6: exam z-scores for first generation and women

Fig. 8.



**Fig. 8** (A) First generation (FG) students in the Volunteer group (Vol) show higher scores on exams than non-first generation (non-FG) Volunteers when controlling for SAT, CCI, and term. The non-first generation general chemistry (GC) error bar is subsumed by the point. Exam grades for first generation students in Gen Chem are statistically lower than those of non-first generation students. STEM-Dawgs do not show statistically significant differences in exam scores. (B) Men and women in Gen Chem show a 0.15 standard deviation gap in exam z-scores when controlling for SAT, CCI, and chemistry lecture section. This gap is not present for STEM-Dawgs or Volunteers. Error bars are standard error.

## Appendix 7: raw scores on PERTS and ASCI-V2

Tables 9 and 10.

**Table 9** Average and standard deviation of pre and post scores for all students and students of interest for the PERTS survey instrument

		Mindset		Belonging		Relevance		Self-efficacy	
		Pre	Post	Pre	Post	Pre	Post	Pre	Post
STEM-Dawgs	All students	14.3 ± 2.9	14.3 ± 3.2	18.7 ± 2.7	18.7 ± 3.3	19.8 ± 3.0	18.6 ± 3.6	17.8 ± 3.3	16.3 ± 4.6
	Women	14.2 ± 2.9	14.5 ± 2.9	18.6 ± 2.5	18.3 ± 3.5	20.4 ± 2.4	18.8 ± 3.6	17.1 ± 3.2	15.4 ± 4.6
	Men	14.5 ± 2.9	14.0 ± 3.7	18.8 ± 3.1	19.6 ± 2.7	18.8 ± 3.7	18.3 ± 3.6	19.0 ± 3.4	18.0 ± 4.0
	First gen	14.8 ± 2.5	14.9 ± 2.9	18.5 ± 2.8	18.6 ± 2.9	19.6 ± 3.3	18.5 ± 3.9	17.2 ± 3.6	15.8 ± 4.3
	Non-first gen	13.9 ± 3.2	13.8 ± 3.4	18.8 ± 2.6	18.9 ± 3.7	20.1 ± 2.7	18.7 ± 3.4	18.4 ± 3.0	16.8 ± 4.8
	URM	15.1 ± 2.0	14.7 ± 3.0	19.1 ± 2.5	19.1 ± 3.0	19.7 ± 3.0	19.4 ± 2.5	18.1 ± 2.8	16.4 ± 3.8
	Non-URM	14.0 ± 3.1	14.2 ± 3.3	18.5 ± 2.7	18.6 ± 3.4	19.9 ± 3.0	18.3 ± 3.9	17.7 ± 3.6	16.3 ± 4.8
	EOP	14.9 ± 2.6	14.8 ± 3.0	19.2 ± 2.1	18.8 ± 3.1	20.0 ± 2.8	18.8 ± 3.2	17.7 ± 3.2	15.6 ± 4.6
	Non-EOP	14.0 ± 3.0	14.0 ± 3.3	18.3 ± 3.0	18.7 ± 3.5	19.7 ± 3.1	18.4 ± 3.9	17.8 ± 3.5	16.8 ± 4.5
Volunteers	All students	14.1 ± 2.9	13.9 ± 3.3	18.0 ± 3.2	17.3 ± 3.8	19.2 ± 3.5	17.9 ± 3.3	17.7 ± 3.6	14.3 ± 4.7
	Women	14.4 ± 2.8	13.9 ± 3.2	18.5 ± 2.9	17.0 ± 3.9	20.0 ± 2.5	18.2 ± 3.0	17.5 ± 3.7	13.7 ± 4.5
	Men	13.1 ± 3.2	13.8 ± 3.7	16.5 ± 3.4	18.1 ± 3.3	16.6 ± 4.8	16.9 ± 4.0	18.3 ± 3.6	16.4 ± 5.0
	First gen	14.0 ± 2.9	14.3 ± 3.1	17.7 ± 3.8	17.2 ± 4.2	18.7 ± 3.3	18.2 ± 3.5	16.8 ± 4.3	14.2 ± 5.1
	Non-first gen	14.2 ± 3.0	13.5 ± 3.5	18.3 ± 2.6	17.3 ± 3.4	19.6 ± 3.6	17.6 ± 3.3	18.5 ± 2.8	14.5 ± 4.5
	URM	14.5 ± 2.9	14.3 ± 2.9	17.4 ± 3.9	16.1 ± 5.2	19.0 ± 3.0	18.2 ± 3.4	16.0 ± 4.3	12.5 ± 5.9
	Non-URM	14.0 ± 3.0	13.8 ± 3.4	18.2 ± 2.9	17.6 ± 3.2	19.2 ± 3.7	17.8 ± 3.4	18.2 ± 3.3	14.9 ± 4.2
	EOP	14.4 ± 2.8	14.6 ± 2.9	17.5 ± 3.6	16.6 ± 4.4	19.1 ± 3.3	18.6 ± 3.7	16.6 ± 4.2	13.6 ± 5.7
	Non-EOP	13.9 ± 3.0	13.4 ± 3.5	18.3 ± 2.9	17.7 ± 3.3	19.2 ± 3.7	17.4 ± 3.0	18.4 ± 3.1	14.8 ± 4.0
Gen Chem	All students	13.7 ± 3.2	13.2 ± 3.5	18.7 ± 2.5	18.4 ± 3.3	19.4 ± 2.9	17.7 ± 3.7	18.1 ± 3.3	16.3 ± 4.4
	Women	13.8 ± 3.1	13.3 ± 3.4	18.5 ± 2.4	18.3 ± 3.2	19.8 ± 2.7	18.2 ± 3.4	17.4 ± 3.4	15.4 ± 4.4
	Men	13.5 ± 3.3	12.9 ± 3.6	18.9 ± 2.6	18.6 ± 3.4	18.8 ± 3.0	17.0 ± 4.0	18.9 ± 3.1	17.5 ± 4.1
	First gen	14.0 ± 3.2	13.5 ± 3.5	18.7 ± 2.6	18.2 ± 3.3	19.5 ± 3.0	18.0 ± 3.5	17.9 ± 3.5	15.7 ± 4.5
	Non-first gen	13.5 ± 3.2	13.0 ± 3.5	18.7 ± 2.5	18.5 ± 3.2	19.3 ± 2.8	17.6 ± 3.8	18.2 ± 3.3	16.5 ± 4.4
	URM	14.2 ± 3.0	13.8 ± 3.4	18.8 ± 2.6	18.2 ± 3.6	19.7 ± 2.9	18.1 ± 3.9	18.2 ± 3.5	16.1 ± 4.4
	Non-URM	13.6 ± 3.2	13.1 ± 3.5	18.7 ± 2.5	18.4 ± 3.2	19.3 ± 2.9	17.7 ± 3.7	18.1 ± 3.3	16.3 ± 4.4
	EOP	14.3 ± 3.0	13.9 ± 3.5	18.6 ± 2.7	18.1 ± 3.6	19.6 ± 3.0	17.9 ± 3.8	17.8 ± 3.6	15.7 ± 4.5
	Non-EOP	13.5 ± 3.2	13.0 ± 3.5	18.7 ± 2.5	18.5 ± 3.2	19.3 ± 2.8	17.6 ± 3.7	18.2 ± 3.3	16.4 ± 4.4

**Table 10** Average and standard deviation of pre and post scores for all students and students of interest for the ASCI-V2

		ASCI-V2			
		Intellectual accessibility		Emotional satisfaction	
Student group		Pre	Post	Pre	Post
STEM-Dawgs	All students	12.4 ± 3.9	12.4 ± 4.0	17.8 ± 3.6	17.8 ± 4.1
	Women	11.4 ± 3.5	11.8 ± 3.6	17.6 ± 3.7	17.4 ± 4.0
	Men	14.2 ± 3.8	13.7 ± 4.5	18.0 ± 3.5	18.5 ± 4.2
	First gen	12.0 ± 3.8	12.4 ± 4.3	17.3 ± 3.1	17.5 ± 3.9
	Non-first gen	12.7 ± 4.0	12.5 ± 3.7	18.2 ± 4.0	18.1 ± 4.3
	URM	11.9 ± 3.8	11.7 ± 4.4	17.8 ± 3.3	17.6 ± 3.7
	Non-URM	12.6 ± 3.9	12.7 ± 3.8	17.7 ± 3.7	17.8 ± 4.3
	EOP	11.6 ± 3.8	11.6 ± 4.5	17.7 ± 3.1	17.4 ± 3.7
	Non-EOP	12.9 ± 3.9	13.0 ± 3.5	17.8 ± 3.9	18.0 ± 4.3
Volunteers	All students	11.9 ± 3.2	12.1 ± 4.0	17.4 ± 2.8	16.1 ± 4.0
	Women	11.9 ± 3.2	11.7 ± 4.0	17.4 ± 2.9	15.4 ± 3.9
	Men	11.9 ± 3.3	13.1 ± 4.1	17.4 ± 2.7	18.1 ± 3.8
	First gen	12.0 ± 3.4	12.2 ± 3.8	16.8 ± 3.3	16.2 ± 3.0
	Non-first gen	11.8 ± 3.0	12.0 ± 4.3	17.8 ± 2.4	16.0 ± 4.7
	URM	11.8 ± 3.3	10.8 ± 3.6	16.9 ± 3.1	15.0 ± 3.5
	Non-URM	11.9 ± 3.2	12.4 ± 4.1	17.5 ± 2.8	16.4 ± 4.1
	EOP	11.9 ± 3.8	11.5 ± 3.7	16.3 ± 3.3	15.6 ± 3.5
	Non-EOP	11.9 ± 2.8	12.4 ± 4.3	18.0 ± 2.3	16.4 ± 4.3

Table 10 (continued)

		ASCI-V2			
		Intellectual accessibility		Emotional satisfaction	
Student group		Pre	Post	Pre	Post
Gen Chem	All students	13.1 ± 3.9	13.4 ± 4.3	17.8 ± 4.0	17.5 ± 4.4
	Women	12.4 ± 3.9	12.7 ± 4.3	17.4 ± 4.1	17.1 ± 4.5
	Men	14.0 ± 3.7	14.4 ± 4.0	18.2 ± 3.9	17.9 ± 4.2
	First gen	13.4 ± 3.9	13.5 ± 4.2	17.9 ± 4.0	17.2 ± 4.4
	Non-first gen	13.0 ± 3.9	13.4 ± 4.3	17.1 ± 4.0	17.6 ± 4.4
	URM	12.7 ± 4.4	12.6 ± 4.4	17.7 ± 4.6	16.9 ± 4.8
	Non-URM	13.2 ± 3.8	13.6 ± 4.2	17.8 ± 3.9	17.6 ± 4.3
	EOP	12.9 ± 4.2	12.5 ± 4.3	17.4 ± 4.4	16.6 ± 4.7
	Non-EOP	13.2 ± 3.8	13.7 ± 4.2	17.9 ± 3.9	17.7 ± 4.3

## Acknowledgements

This project was funded by grant 52008126 from the Howard Hughes Medical Institute. We are grateful to members of the University of Washington Chemistry Education Research Group (ChEdR) and Biology Education Research Group (BERG) for suggestions that improved the course design, data analysis, and manuscript. This work would not have been possible without our peer facilitators or the support of Chemistry Department Chair Dr Michael Heinekey, Chemistry Undergraduate Program Committee Chair Dr Gary Drobny, First-year Programs Director LeAnne Jones Wiles, Executive Director of Retention and Academic Support Programs Kristian Wiles, Special Assistant to the Dean of Undergraduate Academic Affairs Anne Browning, Associate Director of the Louis Stokes Alliance for Minority Participation Stephanie R. Gardner, Executive Director of the Dream Project Jenee A. Myers Twitchell, Undergraduate Advisor Ahnya Redman, Director of the Instructional Center Therese F. Mar, and the Director of the STate Academic RedShirt (STARS) Program Sonya Cunningham.

## References

- Adams W. K., *et al.*, (2006), New instrument for measuring student beliefs about physics and learning physics: the Colorado Learning Attitudes about Science Survey, *Phys. Rev. Spec. Top.-Ph.*, **2**(1), 1–14.
- Amaral K. E. and Vala M., (2009), What Teaching Teaches: Mentoring and the Performance Gains of Mentors, *J. Chem. Educ.*, **86**(5), 630.
- Arendale D. R., (1994), Understanding the supplemental instruction model, *New Dir. Teach. Learn.*, **1994**(60), 11–21.
- Barbera J., *et al.*, (2008), Modifying and Validating the Colorado Learning Attitudes about Science Survey for Use in Chemistry, *J. Chem. Educ.*, **85**(10), 1435.
- Barlow A. E. L. and Villarejo M., (2004), Making a difference for minorities: evaluation of an educational enrichment program, *J. Res. Sci. Teach.*, **41**(9), 861–881.
- Barr D. A., Gonzalez M. E., and Wanat S. F., (2008), The leaky pipeline: factors associated with early decline in interest in premedical studies among underrepresented minority undergraduate students, *Acad. Med.*, **83**(5), 503–511.
- Batz Z., *et al.*, (2015), Helping Struggling Students in Introductory Biology: A Peer-Tutoring Approach That Improves Performance, Perception, and Retention, *Cell Biol. Educ.*, **14**(2), 1–12.
- Bauer C. F., (2008), Attitude toward Chemistry: A Semantic Differential Instrument for Assessing Curriculum Impacts, *J. Chem. Educ.*, **85**(10), 1440.
- Becvar J. E., *et al.*, (2008), 'Plus Two' Peer-Led Team Learning improves student success, retention, and timely graduation, *Proceedings—Frontiers in Education Conference*, pp. 15–18.
- Berg A., (2005), Factors related to observed attitude change toward learning chemistry among university students, *Chem. Educ. Res. Pract.*, **6**(1), 1–18.
- Berry D. C., (1983), Metacognitive experience and transfer of logical reasoning, *Quart. J. Exp. Psychol. Sect. A*, **35**(1), 39–49.
- Bjørnebekk G., Diseth Å. and Ulriksen R., (2013), Achievement motives, self-efficacy, achievement goals, and academic achievement at multiple stages of education: a longitudinal analysis, *Psychol. Rep.*, **112**(3), 771–787.
- Blackwell L. S., Trzesniewski K. H. and Dweck C. S., (2007), Implicit theories of intelligence predict achievement across an adolescent transition: a longitudinal study and an intervention, *Child Dev.*, **78**(1), 246–263.
- Brandriet A., *et al.*, (2011), Diagnosing changes in attitude in first-year college chemistry students with a shortened version of Bauer's semantic differential, *Chem. Educ. Res. Pract.*, **12**(2), 271–278.
- Braun V. and Clarke V., (2006), Using Thematic Analysis in Psychology, *Qual. Res. Psychol.*, **3**, 77–101.
- Burnham K. P. and Anderson D. R., (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edn.
- Chan J. Y. K. and Bauer C. F., (2014), Identifying At-Risk Students in General Chemistry via Cluster Analysis of Affective Characteristics, *J. Chem. Educ.*, **91**, 1417–1425.
- Chen X. and Soldner M., (2013), STEM Attrition: College Students' Path Into and Out of STEM Fields (NCES 2014-001).
- Cohen G. L., Steele C. M. and Ross L. D., (1999), The Mentor's Dilemma: Providing Critical Feedback Across the Racial Divide, *Pers. Soc. Psychol. Bull.*, **25**(10), 1302–1318.
- Cohen G. L., *et al.*, (2006), Reducing the racial achievement gap: a social-psychological intervention, *Science*, **313**, 1307–1310.

- Cohen G. L., *et al.*, (2009), Recursive processes in self-affirmation: intervening to close the minority achievement gap, *Science*, **324**, 400–403.
- College Board, (2016), *Concordance Table*, available at: <https://collegereadiness.collegeboard.org/pdf/higher-ed-brief-sat-concordance.pdf>, accessed March 9, 2018.
- Cook E., Kennedy E. and Mcguire S. Y., (2013), Effect of Teaching Metacognitive Learning Strategies on Performance in General Chemistry Courses, *J. Chem. Educ.*, **90**, 961–967.
- Creswell J. and Clark V., (2011), *Designing and conducting mixed-methods research*, pp. 81–86.
- De La Franier B. J., *et al.*, (2016), A First-Year Chemistry Undergraduate “Course Community” at a Large, Research-Intensive University, *J. Chem. Educ.*, **93**(2), 256–261.
- Ding L. and Molloy K., (2015), How College-Level Introductory Instruction Can Impact Student Epistemological Beliefs, *J. Coll. Sci. Teach.*, **44**(4), 21–29.
- Drane D., *et al.*, (2005), The gateway science workshop program: enhancing student performance and retention in the sciences through peer-facilitated discussion, *J. Sci. Educ. Technol.*, **14**(3), 337–352.
- Duckworth A. L., (2011), The significance of self-control, *Proc. Natl. Acad. Sci. U. S. A.*, **108**(7), 2639–2640.
- Duckworth A. L., *et al.*, (2013), From Fantasy to Action: Mental Contrasting With Implementation Intentions (MCII) Improves Academic Performance in Children, *Soc. Psychol. Pers. Sci.*, **4**(6), 745–753.
- Dunlosky J., *et al.*, (2013), Improving Students’ Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology, *Psychol. Sci. Publ. Int.*, **14**(1), 4–58.
- Dweck C. S. and Leggett E. L., (1988), A social-cognitive approach to motivation and personality, *Psychol. Rev.*, **95**(2), 256–273.
- Eddy S. L. and Hogan K. A., (2014), Getting Under the Hood: How and for Whom Does Increasing Course Structure Work? *Cell Biol. Educ.*, **13**(3), 453–468.
- Ellis J., Fosdick B. K. and Rasmussen C., (2016), Women 1.5 times more likely to leave STEM pipeline after calculus compared to men: lack of mathematical confidence a potential culprit, *PLoS One*, **11**(7), 1–14.
- Fullilove R. E. and Treisman P. U., (1990), Mathematics achievement among African American undergraduates at the university of California, Berkeley: An evaluation of the Mathematics Workshop Program, *J. Negro Educ.*, **59**(3), 463–478.
- Gosser D. K., (2011), The PLTL Boost: a critical review of research, *J. Peer-led Team Learn.*, **14**(1), 3–12.
- Haak D. C., *et al.*, (2011), Increased Structure and Active Learning Reduce the Achievement Gap in Introductory Biology, *Science*, **332**, 1213–1216.
- Hall D. M., Curtin-Soydan A. J. and Canelas D. A., (2014), The science advancement through group engagement program: leveling the playing field and increasing retention in science, *J. Chem. Educ.*, **91**(1), 37–47.
- Hanauer D. I., Graham M. J. and Hatfull G. F., (2016), A measure of college student persistence in the sciences (PITS), *CBE Life Sci. Educ.*, **15**(4), 1–10.
- Harackiewicz J. M., *et al.*, (2014), Closing the social class achievement gap for first-generation students in undergraduate biology, *J. Educ. Psychol.*, **106**(2), 375–389.
- Henningsen A., (2010), *Estimating censored regression models in R using the censReg Package*.
- Henningsen A., (2017), *censReg: Censored Regression (Tobit) Models*.
- Hockings S. C., DeAngelis K. J. and Frey R. F., (2008), Peer-Led Team Learning in General Chemistry: Implementation and Evaluation, *J. Chem. Educ.*, **85**(7), 990.
- Holdren J. P. and Lander E., (2010), *Report to the president: Executive Office of the President President’s Council of Advisors*.
- Hu L. and Bentler P. M., (1999), Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria *versus* new alternatives, *Struct. Equ. Modeling*, **6**(1), 1–55.
- Hulleman C. S., *et al.*, (2010), Enhancing interest and performance with a utility value intervention, *J. Educ. Psychol.*, **102**(4), 880–895.
- Ishitani T., (2006), Studying attrition and degree completion behavior among first-generation college students in the United States, *J. High. Educ.*, **77**(5), 861–885.
- Jordt H., *et al.*, (2017), Values affirmation intervention reduces achievement gap between underrepresented minority and white students in introductory biology classes, *CBE Life Sci. Educ.*, **16**(3), 1–10.
- Kapur M. and Bielaczyc K., (2012), Designing for Productive Failure, *J. Learn. Sci.*, **21**(1), 45–83.
- Larson L. M., *et al.*, (2015), Predicting Graduation: The Role of Mathematics/Science Self-Efficacy, *J. Career Assess.*, **23**(3), 399–409.
- Lewis S. E., (2011), Retention and reform: an evaluation of Peer-Led Team Learning, *J. Chem. Educ.*, **88**(6), 703–707.
- Lewis S. E., *et al.*, (2009), Attitude counts: self-concept and success in general chemistry, *J. Chem. Educ.*, **86**(6), 744–749.
- Magnuson K. and Waldfogel J., (2008), *Steady Gains and Stalled Progress: Inequality and the Black-White Test Score Gap*, Russell Sage.
- Matz R. L., *et al.*, (2017), Patterns of Gendered Performance Differences in Large Introductory Courses at Five Research Universities, *AERA Open*, **3**(4), 1–12.
- Meng R., *et al.*, (2015), The nurses’ well-being index and factors influencing this index among nurses in central China: a cross-sectional study, *PLoS One*, **10**(12), 1–11.
- Miyake A., *et al.*, (2010), Reducing the gender achievement gap in college science: a classroom study of values affirmation, *Science*, **330**, 1234–1237.
- Mulford D. R. and Robinson W. R., (2002), An inventory for alternate conceptions among first semester general chemistry students, *J. Chem. Educ.*, **79**(6), 739–744.
- Murphy K. R. and Davidshofer C. O., (2005), *Psychological Testing: Principles and Applications*, 6th edn, Upper Saddle River, NJ: Prentice Hall.
- National Science Foundation, (2017), *Women, Minorities, and Persons with Disabilities in Science and Engineering: 2017, Special Report*.
- Orbe M. P., (2008), *Theorizing Multidimensional Identity Negotiation: Reflections on the Lived Experiences of*



- First-Generation College Students, *New Dir. Child Adolesc. Dev.*, **120**, 81–95.
- Paunesku D., *et al.*, (2015), Mind-Set Interventions Are a Scalable Treatment for Academic Underachievement, *Psychol. Sci.*, 1–10.
- Pekrun R., Goetz T. and Titz W., (2002), Academic Emotions in Students' Self-Regulated Learning and Achievement: A Program to Qualitative and Quantitative Research, *Educ. Psychol.*, **37**(2), 91–105.
- Pekrun R., *et al.*, (2007), The Control-Value Theory of Achievement Emotions. An Integrative Approach to Emotions in Education, *Emot. Educ.*, 13–36.
- PERTS Survey, available at: <https://survey.perts.net/share/dlmooc>, accessed April 19, 2018.
- Pilcher J. J. and Huffcutt A. I., (1996), Effects of sleep deprivation on performance: a meta-analysis, *Sleep*, **19**(4), 318–326.
- R Core Team, (2017), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing.
- Ramirez G. and Beilock S. L., (2011), Writing about testing worries boosts exam per performance in the classroom, *Science*, **331**, 20–23.
- Rask K. and Tiefenthaler J., (2008), The role of grade sensitivity in explaining the gender imbalance in undergraduate economics, *Econ. Educ. Rev.*, **27**(6), 676–687.
- Rath K. A., *et al.*, (2012), Impact of supplemental instruction in entry-level chemistry courses at a mid-sized public university, *J. Chem. Educ.*, **89**(4), 449–455.
- Richter J. and Gast A., (2017), Distributed practice can boost evaluative conditioning by increasing memory for the stimulus pairs, *Acta Psychol.*, **179**, 1–13.
- Roediger H. L., *et al.*, (2011), Test-enhanced learning in the classroom: long-term improvements from quizzing, *J. Exp. Psychol.: Appl.*, **17**(4), 382–395.
- Rohrer D. and Taylor K., (2007), The shuffling of mathematics problems improves learning, *Instruct. Sci.*, **35**(6), 481–498.
- Rosseel Y., (2012), lavaan: An R Package for Structural Equation Modeling, *J. Stat. Softw.*, **48**(2), 1–36.
- Rothkopf E. Z., (1968), Textual constraint as a function of repeated inspection, *J. Educ. Psychol.*, **59**, 20–25.
- Sandi-Urena S., Cooper M. and Stevens R., (2012), Effect of cooperative problem-based lab instruction on metacognition and problem-solving skills, *J. Chem. Educ.*, **89**(6), 700–706.
- Sawtelle V., Brewe E. and Kramer L. H., (2012), Exploring the relationship between self-efficacy and retention in introductory physics, *J. Res. Sci. Teach.*, **49**(9), 1096–1121.
- Schraw G., Crippen K. J. and Hartley K., (2006), Promoting self-regulation in science education: metacognition as part of a broader perspective on learning, *Res. Sci. Educ.*, **36**, 111–139.
- Semsar K., *et al.*, (2011), The colorado learning attitudes about science survey (class) for use in biology, *CBE Life Sci. Educ.*, **10**(3), 268–278.
- Shields S. P., *et al.*, (2012), A transition program for underprepared students in general chemistry: diagnosis, implementation, and evaluation, *J. Chem. Educ.*, **89**(8), 995–1000.
- Slovacek S., *et al.*, (2012), Promoting minority success in the sciences: the minority opportunities in research programs at CSULA, *J. Res. Sci. Teach.*, **49**(2), 199–217.
- Snyder J. J. and Wiles J. R., (2015), Peer led team learning in introductory biology: effects on peer leader critical thinking skills, *PLoS One*, **10**(1), 1–18.
- Snyder J. J., *et al.*, (2016), Peer-Led Team Learning Helps Minority Students Succeed, *PLoS Biol.*, **14**(3), 1–7.
- Stanton J. D., *et al.*, (2015), Differences in Metacognitive Regulation in Introductory Biology Students: When Prompts Are Not Enough, *Cell Biol. Educ.*, **14**(2), 1–12.
- Stephens N. M., *et al.*, (2012), Unseen disadvantage: how American universities' focus on independence undermines the academic performance of first-generation college students, *J. Pers. Soc. Psychol.*, **102**(6), 1178–1197.
- Taylor N., *et al.*, (2015), Near peer teaching in medical curricula: integrating student teachers in pathology tutorials, *Med. Educ. Online*, **20**(1), 20–22.
- Trujillo G. and Tanner K. D., (2014), Considering the role of affect in learning: monitoring students' self-efficacy, sense of belonging, and science identity, *CBE Life Sci. Educ.*, **13**(1), 6–15.
- Vaismoradi M., Turunen H. and Bondas T., (2013), Content analysis and thematic analysis: implications for conducting a qualitative descriptive study, *Nurs. Health Sci.*, **15**(3), 398–405.
- Vishnumolakala V. R., Southam D. C., Treagust D. F., Mocerino M., and Qureshi S., (2017), Students' attitudes, self-efficacy and experiences in a modified process-oriented guided inquiry learning undergraduate chemistry classroom, *Chem. Educ. Res. Pract.*, **18**(2), 340–352.
- Walton G. M. and Cohen G. L., (2007), A question of belonging: race, social fit, and achievement, *J. Pers. Soc. Psychol.*, **92**(1), 82–96.
- Walton G. M. and Cohen G. L., (2011), A Brief Social-Belonging Intervention Improves Academic and Health Outcomes of Minority Students, *Science*, **331**, 1447–1451.
- Warfa A.-R. M., (2016), Using Cooperative Learning To Teach Chemistry: A Meta-analytic Review, *J. Chem. Educ.*, **93**(2), 248–255.
- Wilson S. B. and Varma-Nelson P., (2016), Small Groups, Significant Impact: A Review of Peer-Led Team Learning Research with Implications for STEM Education Researchers and Faculty, *J. Chem. Educ.*, **93**(10), 1686–1702.
- Woloshyn V. E., Pressley M. and Schneider W., (1992), Elaborative interrogation and prior knowledge effects on learning of facts', *J. Educ. Psychol.*, **84**(1), 115–124.
- Xu X. and Lewis J. E., (2011), Refinement of a Chemistry Attitude Measure for College Students, *J. Chem. Educ.*, **81**, 561–568.
- Ye L., *et al.*, (2016), Can they succeed? Exploring at-risk students' study habits in college general chemistry, *Chem. Educ. Res. Pract.*, **17**(4), 878–892.
- Yeager D. S., Duckworth A. L. and Walton G. M., (2014), *Boring but Important: A Self-Transcendent Purpose for Learning Fosters Academic Self-Regulation*.
- Yeager D. S., *et al.*, (2016), Teaching a lay theory before college narrows achievement gaps at scale, *Proc. Natl. Acad. Sci. U. S. A.*, **113**(24), E3341–E3348.
- Zhao N., *et al.*, (2014), Metacognition: An Effective Tool to Promote Success in College Science Learning, *J. Coll. Sci. Teach.*, **43**(4), 48–55.
- Zuur A. F., *et al.*, (2009), *Mixed Effects Models and Extensions in Ecology with R*, Springer.