Contemporary Educational Psychology 70 (2022) 102088

Contents lists available at ScienceDirect

Contemporary Educational Psychology

journal homepage: www.elsevier.com/locate/cedpsych

Investigating gender differences among tutors and students during STEM peer tutoring: Women are as behaviorally engaged as men but experience more negative affect

Oana D. Dumitru [a,*], Katherine R. Thorson [b], Tessa V. West [a]

[a] Dept. of Psychology, New York University, 6 Washington Place, New York, NY 10003, USA
[b] Dept. of Psychology, Barnard College of Columbia, 3009 Broadway, New York, NY 10027, USA

ARTICLE INFO

Keywords:
Peer tutoring
Gender
STEM
Negative affect
Engagement
Social interaction

ABSTRACT

Peer tutoring in STEM has risen in popularity in the past several years and has been proposed as one method of reducing gender disparities in STEM outcomes. Yet, the ways in which students and peer tutors engage with each other remain largely unexplored. In this study, we employed a multi-method approach to investigate whether students' and tutors' engagement behaviors and affective experiences during peer tutoring interactions in STEM fields differed by gender. Sixty unacquainted undergraduate college students formed student-tutor pairs and participated in videotaped thirty-minute tutoring sessions in the lab, all of which covered STEM topics (Biology, Chemistry, Computer Science, Economics, Mathematics, and Physics). We found no consistent gender differences across three measures of behavioral engagement: men and women talked for a similar amount of time, they did not differ in four of five types of questions asked (i.e., "clarification" and "knowledge" questions for tutors, and "feedback" and "more information" questions for students), and they were perceived as equally engaged by outside coders. One behavioral difference emerged: men students asked more "repeat" questions than women students. In contrast, consistent gender differences across four measures of affective experiences were found: women reported more anxiety and less confidence relative to men, they were perceived as less confident by outside coders relative to men, and women tutors evaluated their own performance less positively than men tutors. These findings suggest that despite being similarly engaged as men in peer tutoring interactions, women face psychological barriers in this context that may inhibit them from pursuing advanced degrees or careers in STEM.

1. Introduction

In recent years, peer tutoring has become an increasingly popular strategy for improving students' performance in Science, Technology, Engineering, and Mathematics (STEM) fields (Batz, Olsen, Dumont, Dastoor, & Smith, 2015). During peer tutoring, a novice student works with a more advanced or experienced student, who teaches or reviews material that the student is trying to learn for a particular course or exam (e.g., medical school exams; Mynard & Almarzouqi, 2006). Across the United States, peer tutoring programs in STEM are now offered at many universities, with schools often relying on peer tutoring as a primary method for helping students who are struggling to learn course material (Academic Advising and Support at Caltech, n.d.; Academic Resource Center at Harvard University, n.d.). Peer tutoring programs also allow STEM students to network with each other, which can prevent STEM dropout, especially for students from underrepresented backgrounds who may have relatively few academic connections (Mishra, 2020).

Despite the growing popularity of peer tutoring in higher education, very little attention has been paid to the behavioral and psychological processes that occur during peer tutoring sessions themselves. We propose that gaining a better understanding of these processes—as they occur naturally in real peer tutoring interactions—is an important step in understanding the long-term impacts of peer tutoring. For example, although peer tutoring has been proposed as a potential method for retaining women students in STEM, its ability to do so remains to be seen (Dagley, Georgiopoulos, Reece, & Young, 2016; Good, Halpin, & Halpin, 2000; Savaria & Monteiro, 2017). In addition, although many studies have tried to assess the effectiveness of peer tutoring in improving

* Corresponding author.
E-mail address: oana.dumitru@nyu.edu (O.D. Dumitru).

https://doi.org/10.1016/j.cedpsych.2022.102088

Available online 14 July 2022

performance, the results have thus far been mixed (Alegre, Moliner, Maroto, & Lorenzo-Valentin, 2020; Batz et al., 2015; Lee, Kim, & Yoon, 2004; Suryadarma, Suryahadi, Sumarto, & Rogers, 2006; Thomas, Bonner, Everson, & Somers, 2015; Zhang, 2013). We propose that, to start, scholars should focus on the interpersonal dynamics that unfold during tutoring interactions, as these processes may impact longer-term outcomes, such as performance.

Here, we focus on how one characteristic that has well-established disparities in STEM—gender—shapes how students and tutors experience peer tutoring sessions. At the undergraduate level, gender disparities in many STEM fields continue to exist. For example, in computer science and engineering, approximately 20% of Bachelor's degrees are awarded to women, and, across STEM fields overall, women students consistently report more anxiety and less confidence than men students report (Bloodhart, Balgopal, Casper, Sample McMeeking, & Fischer, 2020; National Science Foundation, National Center for Science and Engineering Statistics (2019), 2019; Schuster & Martiny, 2017; Shapiro & Williams, 2012; Van Veelen, Derks, & Endedijk, 2019). Even in undergraduate STEM fields where women are not underrepresented, such as biology and chemistry (American Physical Society. (2018), 2018), the proportion of women decreases at more advanced levels. For example, in the United States, a similar number of men and women students graduate with Bachelor degrees in chemistry, but men earn more Masters' (55% vs. 45%) and doctoral degrees (61% vs. 39%) and hold more tenure-track positions (82% vs. 18%) relative to women (McMunn, 2017; National Center for Education Statistics. (2018), 2018). Even STEM fields that have typically been less math-intensive, such as psychology, now require high levels of quantitative skills (Connolly, 2020). Given the stereotype that men are better at quantitative work, this shift in focus may begin to discourage women from pursuing advanced degrees and careers in these fields as well (Morrissey, Hallett, Bakhtiar, & Fitzpatrick, 2019; Passolunghi, Ferreira, & Tomasetto, 2014).

In this study, we focus on understanding the role of gender in two processes related to tutoring that are critical for learning and long-term persistence in STEM: people's behavioral engagement in tutoring sessions and their affective experiences following them. We test the question: Do men and women differ in how engaged they are in peer tutoring sessions - that is, in the amount of time they spend talking, the number and types of questions they ask, and the extent to which they appear engaged to outside observers? We further test whether men and women differ in their affective experiences—that is, do they differ in the extent to which they experience anxiety, confidence, and, for tutors, positively evaluate their own teaching performance? To test these questions, we had tutors and students engage in peer tutoring sessions, during which we measured the behaviors of both partners over the course of the sessions and their feelings about the session at its conclusion. We propose that gender may shape engagement and affective experiences during peer tutoring sessions—for both tutors and students—and we elaborate on these ideas in the following sections.

2. Conceptual framework for gender-based differences in behavioral engagement and affective experiences

Students' behavioral engagement with learning is a consistent predictor of remaining in school and remaining in a particular field (Archambault & Dupéré, 2017; Ekstrom, Goertz, Pollack, & Rock, 1986; Rumberger & Rotermund, 2012; Wang & Fredricks, 2014). Behavioral engagement is typically considered to have three primary components: participation in learning environments, like classrooms and office hours; positive conduct (complying with school rules and completing homework on time; Fredricks, Blumenfeld, & Paris, 2004); and engagement in extracurricular activities, such as involvement in student government (Blumenfeld, Modell, Bartko, Secada, Fredricks et al., 2005; Fredricks, Blumenfeld, & Paris, 2004). Although other types of engagement exist (e.g., emotional and cognitive; Fredricks, Blumenfeld, & Paris, 2004; Blumenfeld et al., 2005), here, we study behavioral engagement given its close ties to long-term academic persistence (Finn, 1989; Li & Lerner, 2013). We focus specifically on participation in learning environments given its relevance to dyadic learning (Precourt & Gainor, 2019). Specifically, we measure the amount of time spent talking, the number and types of questions asked, and the extent to which participants were seen as engaged by outside observers.

In other STEM learning environments, gender-based differences in behavioral engagement are well-documented–typically in the direction of men appearing more engaged. For example, men students speak more, ask more questions, and are less likely to be interrupted than women students are (Aguillon et al., 2020; Carter, Croft, Lukas, & Sandstrom, 2018; Crombie, Pyke, Silverthorn, Jones, & Piccinin, 2003; Daly, Kreiser, & Roghaar, 1994; Eddy, Brownell, & Wenderoth, 2014; Sankar, Gilmartin, & Sobel, 2015). In a recent paper, Lee and Mccabe (2021) observed over ninety-five hours of college classroom interactions and found that men spent 1.6 times more time talking in class, compared to women. Furthermore, men students' increased talking time was often a result of asking clarifying questions and carrying on debates with professors (Lee & Mccabe, 2021). This kind of engagement is very important in the classroom for its ability to reinforce existing gender hierarchies. Speaking up, and consequently, the amount of time spent talking, is associated with confidence in the field (Anderson & Kilduff, 2009; Berger, Cohen, & Zelditch, 1972), which is in turn associated with higher status. For women, domination of the "sonic space" (Lee & Mccabe, 2021) by men students not only makes it harder for them to participate, but may also reinforce the idea that they do not belong in the classroom (Cheryan, Plaut, Davies, & Steele, 2009). This may especially be the case in STEM classrooms, where women have pre-existing beliefs about their lack of belongingness (Cheryan et al., 2009; Leaper, 2015; Rainey, Dancy, Mickelson, Stearns, & Moller, 2018).

Gender differences in affective experiences in the classroom are also well-documented: women tend to feel less efficacious, less confident, and more anxious than men (math, Devine, Fawcett, Szűcs, & Dowker, 2012; Else-Quest, Hyde, & Linn, 2010; Voyer, Voyer, & D., 2014; engineering, Cech, Rubineau, Silbey, & Seron, 2011; biology, Pelch, 2018; engineering and computer science, Sterling et al., 2020). When experienced repeatedly over time, these feelings contribute to intentions to drop out of STEM (Finn, 1989; Hascher & Hagenauer, 2010; Korpershoek, Canrinus, Fokkens-Bruinsma, & de Boer, 2020; Suhlmann, Sassenberg, Nagengast, & Trautwein, 2018; Vera et al., 2016). Critically, differences in affect between men and women emerge even in the absence of behavioral differences: women perceive themselves as performing worse than men, even when they perform equally well (Cheryan, Siy, Vichayapai, Drury, & Kim, 2011; Jakobsson, 2012). Thus, in this paper, we examine the extent to which students and peer-tutors experience anxiety, stress, and confidence during the tutoring session. We note that the term "emotional engagement" is often used in this context, but because we are only capturing a subset of components relevant to emotional engagement (others include identification with school and valuing school-related outcomes; Fredricks et al., 2004), we refer to these self-reported measures of anxiety, stress, and confidence (reverse-scored), as "negative affective experiences.".

The present study is the first to our knowledge to provide a snapshot of behavioral engagement and affective experiences in naturally unfolding peer-tutoring interactions across a variety of collegiate STEM courses. We propose several processes that might create gender-based differences in engagement and affect in this context (described below). Although we do not directly test these mechanisms in the present study, we theorize that, if gender differences do emerge in peer-tutoring sessions, these processes could be largely responsible.

3. Potential mechanisms for gender-based differences in peer tutoring

The "gender-STEM" stereotype—the stereotype that men are more competent in STEM fields than women are—underlies many of the

documented gender differences reviewed here, and is thought to be active in nearly all STEM contexts (Dasgupta & Stout, 2014; Greider et al., 2019; Oswald, 2008; Schuster & Martiny, 2017; Spencer, Steele, & Quinn, 1999), potentially including peer tutoring. This stereotype can lead to two gender-based expectations: One, that women will perform worse than men, which can inhibit women's engagement with the field and lead to negative affective experiences, especially when women fear confirming the stereotype (Bedyńska & Żołnierczyk-Zreda, 2015; Cadinu, Maass, Rosabianca, & Kiesner, 2005; Casad, Petzel, & Ingalls, 2019; Johns, Inzlicht, & Schmader, 2008; Schmitt, Branscombe, & Postmes, 2003; Schuster & Martiny, 2017). Two, that women do not belong in STEM learning environments (Deiglmayr, Stern, & Schubert, 2019; Lewis et al., 2017; Master & Meltzoff, 2020). Belonging is defined as the ability to form and maintain positive and lasting interpersonal relationships with others (Baumeister & Leary, 1995). In STEM classrooms, this expectation that women do not belong can be communicated by instructors either explicitly (e.g. by telling women that they would be more successful in other, non-STEM careers) or implicitly (e.g., by only calling on men students when they raise their hands), leading women to internalize this expectation and participate less in STEM classrooms (Gansen, 2019; Good, Sterzinger, & Lavigne, 2018; Urhahne, 2015). Low feelings of belongingness can manifest behaviorally (e.g. women being less assertive than men because they are less comfortable, and in turn not seeking out help when needed; Korpershoek et al., 2020; Wilson et al., 2015) and psychologically (e.g. women feeling less confident than men; Korpershoek et al., 2020; Wilson et al., 2015). Thus, for women, feeling like they do not belong as much as men may also lead to gender differences in behavioral engagement and negative affect in STEM peer tutoring. Given that gender-based expectations have been documented in multiple STEM contexts, it is certainly possible that consistent patterns of gender differences will also emerge in peer tutoring, with men appearing more engaged and reporting more positive affect than women.

However, it is also important to note that certain contextual features of tutoring environments may prevent gender-based differences in engagement and affect from emerging during tutoring sessions. We propose three of these contextual features here: numerical representation, greater focus on learning vs. performance, and interaction with a similar-status peer.

First, given that one-on-one peer tutoring involves dyadic interactions, women cannot be underrepresented numerically, which may help to lessen the activation of the gender-STEM stereotype. Indeed, women's numeric underrepresentation in STEM classrooms is thought to activate the gender-based stereotype, and in turn lead to gender disparities in engagement (Ballen et al., 2019; Murphy, Steele, & Gross, 2007; Sekaquaptewa & Thompson, 2003; Van Veelen et al., 2019). For example, one study found that women were more likely to ask clarification questions from the instructor when the class was predominantly (80%) women (Lee & Mccabe, 2021). Furthermore, when men and women are equally represented, they tend to feel similarly in terms of their identification with STEM (Niler et al., 2020). In STEM contexts, women students are also less likely to experience negative affect and less likely to feel insecure when they are part of smaller classrooms (i.e. under 15 students) compared to larger classrooms, where the numeric underrepresentation of women can be more extreme (Wegner, Strehlke, Weber ClaasWegner, & Weber, 2014). Thus, even if women are paired with men in peer-tutoring sessions, the gender-STEM stereotype may be less salient simply by virtue of having no other students in the room, and thus have a weaker impact on their engagement and affective experiences, compared to the classroom context where women are numerically underrepresented.

Second, in peer tutoring environments, the most salient goal is learning–to get better at the topic, not necessarily to perform well (in fact, performance is rarely evaluated; Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Cohen, Kulik, & Kulik, 1982). Other STEM learning environments often involve performing under evaluative conditions (e.g., test-taking in the classroom)—which can activate the gender stereotype (Appel & Kronberger, 2012; Duke, Krishnan, Faith, & Storch, 2006; Maresh, Teachman, & Coan, 2017; Spencer et al., 1999). When people focus on learning (as opposed to performance) goals, they may also be less concerned about confirming negative performance stereotypes (Park, Schmidt, Scheu, and DeShon, 2007). Thus, for women, the learning goals present in peer tutoring may reduce concerns about confirming the gender-STEM stereotype relative to other STEM learning environments, weakening the likelihood that gender-based differences in behavior and affect will emerge.

Third, in peer tutoring, students are in the presence of peers, rather than higher-status experts. Many studies demonstrating gender differences in STEM outcomes involve the presence of high-status experts, such as more senior evaluators or teachers, who have control over the students' educational outcomes (e.g. Beasley & Fischer, 2012; Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012; Pomerantz, Altermatt, & Saxon, 2002; Wegner et al., 2014). It remains to be seen whether learning from a peer, who is closer in status to students, and who, at least in this context, does not have power over students (they do not control students' outcomes, like grades, Colvin, 2007) weakens gender-based differences in the outcomes we plan to measure.

In sum, these three contextual features in peer tutoring environments—in contrast to other STEM learning environments—may result in similar engagement and affect for women relative to men.

4. Role of student vs. tutor

Lastly, a final goal of this research is to test whether gender differences in behavioral engagement and affect (e.g., anxiety and confidence regarding STEM) emerge for both students and tutors. Undergraduate students who are most likely to request peer tutoring in STEM are typically in the first two years of college, completing coursework for large introductory courses (i.e. "gatekeeper" courses; Dempsey, 2016). These students are often the focus of gender research in STEM, particularly in studies that have shown gender differences in engagement and affect in the classroom (e.g. Banchefsky, Lewis, & Ito, 2019; Eddy et al., 2014).

In contrast, less research has examined the factors that contribute to STEM dropout for women who are more advanced in the major. Students who work as peer tutors have not only performed well in STEM classes, but also, are much closer (relative to students, on average) to completing STEM majors. Such students have made it past the point at which they are most likely to drop out of STEM fields—that is, past the "gatekeeper" courses that constitute strong barriers to pursuing a major in STEM and which women are more likely to drop out of than men (Crisp, Nora, & Taggart, 2009; Dempsey, 2016). If gender-based differences emerge among tutors, there could be implications for gender disparities at higher levels in STEM. For example, peer tutors who leave the interaction feeling like they performed poorly may be less likely to pursue STEM further (e.g. to attend graduate school or take on a leadership role in a STEM industry position). Thus, studying how tutors feel with regard to their tutoring experiences and how engaged they are may provide important insights into how working as a tutor contributes to gender differences in STEM education and occupations.

5. Research overview

The current study investigates differences in behavioral engagement and affective experiences of men and women in STEM peer tutoring sessions. In sixty videotaped peer tutoring interactions ($N = 113$),[1] college students conducted thirty minute free-form tutoring sessions in

[1] Note that the total number of participants is odd because two tutors participated in the study on two occasions, with different students, and one student also participated as a tutor in a different topic.

various STEM subjects (see Table 1 in the Supplementary Material). These tutoring sessions were explicitly focused on students learning course material. As such, they differ from "peer mentoring" programs, which also involve connecting with peers in the field but instead focus on developing positive academic attitudes in and outside of the learning environment, rather than on learning (Kowalewski, Massen, & Mullins, 2010).

5.1. Measures of behavioral engagement

To examine participants' behavioral engagement, we measured the amount of time they spent talking, the frequency with which they asked different types of questions, and the extent to which they appeared engaged in the tutoring sessions (as rated by outside observers). Talk time is an overt marker of engagement in educational settings; it is critical for learning during peer tutoring because it can help students communicate what is unclear to them (Fritschner, 2000; Graesser & Person, 1994; Tanner, 2009) and can help tutors determine what concepts students need covered (Gasiewski, Eagan, Garcia, Hurtado, & Chang, 2012). In this study, we measured the ratio of one person's talking time relative to the total amount of time that the dyad spent talking (Hagiwara et al., 2013; Schoenthaler, Basile, West, & Kalet, 2018). This is a common approach used when studying talking time in dyadic interactions because it adjusts for dyad-level variation in talking time (because some dyads might be more talkative than others) by standardizing the amount of time that one participant spends talking by the total amount of speaking time in the interaction.

Question asking is also an overt marker of engagement; it is a direct way to ask for help and facilitate learning (Dillon, 1982; Karabenick & Dembo, 2011). For example, question asking is positively associated with reading comprehension in science topics and better final grades in science courses (Cano, García, Berbén, & Justicia, 2014). Here, we investigated the types of questions that we reasoned were most likely to signal engagement in the tutoring session, adapting taxonomies that have been used in previous studies (Hawkins & Power, 1999; Pearson & West, 1991). For students, we measured questions that asked for more information from the tutor ("more information" questions), questions that asked for feedback from the tutor ("feedback" questions), and questions that had been asked before ("repeat" questions). For tutors, we examined the number of questions that were meant to clarify what the student needed ("clarification" questions) and questions that were meant to test the student's knowledge or understanding ("knowledge" questions). These categories of questions accounted for over 80% of all questions asked.

Lastly, coders rated participants' overall levels of engagement. We provided coders with specific examples of behavioral engagement, including being attentive and making eye contact. As is the case with confidence, appearing engaged is important for women's interactions in STEM because it can determine how they are treated by others. For instance, students who show engagement receive more support from instructors, which can affect performance and persistence in STEM (Simon, Aulls, Dedic, Hubbard, & Hall, 2015; Skinner & Belmont, 1993).

5.2. Measures of affect

We examine students' and tutors' self-reported confidence and anxiety after their tutoring sessions given that these affective states are associated with persistence and retention in STEM (Banchefsky, Lewis, & Ito, 2019; Chipman, Krantz, & Silver, 1992; Schuster & Martiny, 2017; Tellhed, Bäckström, & Björklund, 2017). We also examine confidence through external coders' ratings of participants during the session to test the possibility that affective experiences of confidence "leak" out behaviorally and are picked up on by others. Being seen as confident has direct implications for doing well in STEM fields, given that people who are perceived as confident by others are more likely to influence group decision-making and to be promoted (Anderson & Kilduff, 2009; Guillén, Mayo, & Karelaia, 2018). Lastly, we examine tutors' evaluations of their own performance at the end of the session, given that perceived performance is a strong predictor of whether women persist in STEM education (Bench, Lench, Liew, Miner, & Flores, 2015; Cheryan et al., 2011). We did not measure students' perceived performance given that performing well on the material was not the goal of the session for students, but performing well as a tutor (i.e., improving students' understanding) was a goal for tutors.

Lastly, to ensure that none of the potential gender differences in behavioral engagement or affect were due to pre-existing differences in academic achievement (i.e. GPA) or identification with STEM fields (i.e. responses to a field identification scale adapted from Crocker, Luhtanen, Cooper, & Bouvrette, 2003) between men and women, we also collected data on and analyzed gender differences for these variables.

To our knowledge, this study is one of the first to take a multimethod approach to studying students' and tutors' behaviors with each other in STEM tutoring sessions. We consider this study to be multi-methods because it uses self-report and observational measures (behavioral coding) to obtain a broader view of behavioral engagement and negative affect in tutoring sessions (Hunter & Brewer, 2015; Gajda, Beghetto, & Karwowski, 2017). By studying how gender influences engagement and affective experiences, we aim to better understand the potential of peer tutoring to influence performance and retention in STEM for both students and tutors.

6. Method

Additional methodological details and results are provided in the Supplementary Material (SM); study materials, data, and analysis syntax are available at https://osf.io/ygz2d.

6.1. Participants

We recruited 60 student-tutor dyads from New York University's campus. Research assistants posted flyers around campus and sent emails to STEM classes within the university informing students about the possibility of participating in the study as either tutors or students in their subject of choice. We asked interested people to complete a screening survey in which they indicated whether they wanted to participate as a student or a tutor. If participating as a student, we asked what subject they needed help with. To participate in the study, the tutoring session needed to be for a class in a Science, Technology, Engineering, and Mathematics (STEM) discipline (for a list of tutoring session subjects and undergraduate majors for students and tutors, see Tables 1, 2, and 3, respectively, in the Supplementary Material). Overall, students sought help for nine academic subjects, with the most frequent ones being computer science (20 sessions) and mathematics (16 sessions). We screened potential tutors by a) first asking what subject they could tutor in, b) what class they could tutor in, and c) what grade they had received in the class. If they had gotten an A, they qualified as tutors for the study. The majority of tutors reported prior experience with tutoring, ranging from three months to six years ($M = 10.75$ months, $SD = 15.67$). Only six tutors reported no prior experience with tutoring, but all of these tutors reported GPAs between 3.5 and 4 (corresponding to the US letter grade "A") in the topic in which they would be tutoring.

A research assistant had access to the prescreening survey and matched students and tutors according to subject. Participants also had the option to invite a student or tutor they were already acquainted with to participate in the study. Overall, seven student-tutor pairs knew each other beforehand (11.6%), and 53 were matched by our research assistants.

We ran a power analysis for an independent samples *t*-test using

G*Power specifying a medium Cohen's *d* of 0.5 (Faul, Erdfelder, Lang, & Buchner, 2007). This effect size was based on prior research examining gender differences in anxiety within STEM (Goetz, Bieg, Lüdtke, Pekrun, & Hall, 2013; Pomerantz et al., 2002; Tapia & Marsh, 2004).[2] We specified 80% power at an alpha level of 0.05, which yielded a suggested sample size of 51 participants in each group, for a total of 102 participants. We used this sample size as our initial stopping point in recruitment, but we continued data collection until the end of the semester, which resulted in 120 total participants. We chose to do our power analysis on single-outcome dependent variables (namely, gender differences in self-reported negative affect), given that repeated measures analyses (i.e. talking time, objective coders' ratings of confidence and anxiety, and types of questions asked), involve multiple time points and are higher in power when the effect of the predictor variable (in this case, gender) is consistent across time points (Bolger & Laurenceau, 2013). We hypothesized that our effects would be consistent across time points, so we used a more conservative approach to determine the targeted sample size by focusing on analyses with one data point per participant.

Participants self-identified their gender (students: women [72.4%] and men [27.6%]; tutors: women [39.6%] and men [60.4%]). Two students and one tutor who identified their gender as "other" participated in the study, but dyads involving these participants are not included in the analysis dataset. Students were predominantly undergraduates (91.4%) between the ages of 18 and 34 ($M = 19.46$, $SD = 2.40$). Students who were not undergraduates were either Masters' students studying for an exam (5.2%), or non-matriculated students with Bachelor's degrees enrolled in private classes (3.4%)[3]. Ten students identified as White (17.2%), four identified as Black or African American (6.9%), 25 identified as Asian (43.1%), six identified as South Asian (10.3%), two identified as "Other" (3.4%), and eleven chose not to respond to the racial identity question (19%).

Thirty-four of the tutors (64.2%) were undergraduates, 17 were graduate students (32.1%), and one was not currently a student (1.9% — a student who was currently taking additional courses in computer science post-graduation), all between the ages of 18 and 34 ($M = 22.04$, $SD = 3.30$). Among the tutors, five identified as White (9.4%), two identified as Black or African American (3.8%), one identified as American Indian or Alaska Native (1.9%), 24 identified as Asian (45.3%), seven identified as South Asian (13.2%), one identified as "Other" (1.9%), and thirteen chose not to respond (24.5%). The demographic make-up of this sample closely mirrors that of STEM degrees obtained at the undergraduate level, with a slight overrepresentation of Asian students (National Center for Education Statistics. (2019), 2019).

Six of the tutors participated in the study on two occasions but with different students. One student also participated as a tutor on a different subject. Twenty-five dyads had same-gender students and tutors, with 15 women dyads and 10 men dyads, and 32 of the dyads had students and tutors that differed in gender, with 25 dyads having a man tutor and woman student, and 7 dyads having a woman tutor and man student (see Table 4 in the Supplementary Material for this breakdown). We did not have the statistical power to examine how different gender combinations between tutors and students might affect outcomes, but we elaborate on this research question in the Discussion.

[2] Indeed, in the current data, we replicated past work and found, on average, a medium effect size of gender on affective experiences.

[3] Results reported in the paper are largely consistent when excluding the 5 students who were not undergraduates, with one exception: we found a significant role by gender interaction for talking time, qualified by a simple effect of gender for tutors, such that men tutors talked for a higher percentage of time compared to women tutors.

6.2. Procedure

Participants were scheduled to come to the lab for a 30-minute tutoring session. Upon arrival, each participant was welcomed by a research assistant and taken into a private room. Another goal of this larger study was to measure physiological responses of tutors and students, and therefore, experimenters placed physiological sensors on participants' bodies and recorded a five minute baseline physiological recording for both participants. These measures are not the focus of this paper, but more details regarding them can be found in the Supplementary Material. Next, the student moved into the tutor's room, which was equipped with a whiteboard, desk, and materials for the tutoring session. The room was also equipped with video cameras, and we took audio and video recordings of both participants separately by using two different cameras. A research assistant introduced the student and tutor and explained that the two of them were matched based on subject. The student and tutor were allowed to chat briefly before the session and figure out the specifics of the tutoring session, such as specific concepts that the student would most want to learn about. Tutors were not instructed to use a specific pedagogical approach, but students and tutors were instructed to remain on topic. We also placed an intercom system in the room and told participants that we would communicate with them through the intercom. Thus, participants were aware that we were listening to their conversation and ensuring that tutoring was taking place.

Once they were ready to begin the session, participants were given the following instructions:

> *"You have thirty minutes to complete your session. There is an intercom system in the room, so if you need any assistance, just let us know and we will hear you. You may begin your session now."*

At the end of the tutoring session, the student was taken into a separate room to allow privacy while filling out measures about the session. Participants received $20 for participation.

6.3. Measures

6.3.1. Academic achievement and field identification

To obtain a measure of students' and tutors' academic achievement, we asked them to report their Grade Point Average (GPA). To measure participants' identification with their subject of study (major), we asked participants to rate the extent to which they agreed with the following four statements, on a 7-point Likert scale (1 = "Strongly agree"; 7 = "Strongly disagree"; Cronbach's α = 0.88 for tutors, 0.80 for students): "It is important for me to be good at tasks that require the use of [topic]," "I feel good about myself when I do well on tasks that involve [topic]," "I feel a sense of pride in doing well on tasks that involve [topic]," and "I like tasks that involve [topic]." These items were adapted from Crocker et al. (2003). Both GPA and subject identification were measured in the pre-screening questionnaire, before participants took part in the tutoring session.

6.3.2. Behavioral engagement

For all coded behaviors, women research assistants underwent a training process, further detailed below for each kind of coding. We intentionally kept coder gender consistent across all coding to avoid bias that could arise from gender differences in perceptions of behavior.

Talking Time. To determine the amount of time that each participant spent talking during the tutoring session, three trained research assistants indicated when participants started and stopped talking throughout the tutoring session. They did this using Datavyu, a behavioral coding platform that allows coders to simultaneously watch videos and mark when certain events occur during the interaction (https://www.datavyu.org). To train, coders first learned how to use the Datavyu platform and how to mark talking time in the interaction. For example, they were taught to mark a pause in speech if it lasted for at

least one second, to count interjections as speech only if they had meaning in the context of tutoring (e.g. "hmm" or "ok" as opposed to "ugh" or "ouch"), and to discount laughter. All coders completed the same set of five practice videos independently, then met and resolved discrepancies. Most differences between coders were due to omitting breaks in speech or mistakenly categorizing interjections as meaningful. After the training phase, two of the coders were each assigned a number of participants to code, with a minor amount of overlap, while the third coder served as the "lead coder" and overlapped on 25% of videos with each of the other two coders. We analyzed the amount of time that participants spent talking in 30-second intervals throughout the task (for a total of 60 segments per person). We estimated reliability using a two-way mixed absolute agreement single-measures ICC model between the lead coder and each of the other coders. The ICC values for the single-measures were in the "excellent" range (ICCs from 0.90 to 0.96; Cichetti, 1994). We averaged the scores of two coders when there was overlap.

To obtain the talk time ratio, we computed a variable to represent the ratio of one dyad member's talking time to the total amount of talking time—that is, the sum of one person's talking time and their interaction partner's talking time. This approach allowed us to account for the possibility that the two participants overlapped in their talking time. For instance, if a participant did not talk at all during a 30-second interval and their partner talked the whole time, they would be marked as having talked 0% of the time and 100% of the time, respectively. On the other hand, if both participants talked for the whole 30-second interval, they would each be assigned a score of 50%.

Questions Asked. For both students and tutors, we coded the type of questions asked. For students, the categories were feedback questions (e.g. "Is this correct?"), questions that had been asked before (e.g. "Can you go over that again?"), and questions that asked for more information (e.g. "What does this mean?"). For tutors, the categories were clarification questions (e.g. "Is this what you need help with?) or knowledge questions, meant to test the student's knowledge (e.g., "Do you know what a differential is?").

Again, three women research assistants served as coders. They first trained on the same set of five videos and resolved discrepancies when applicable. This type of coding also involved using the Datavyu software and coders were trained on how to use the software to mark when a question was asked. Research assistants were trained on what kind of sentences should be counted as questions. For instance, they were instructed that only questions related to the tutoring session content should be counted (e.g. questions about their personal life were not counted) and that words or interjections like "what?" or "huh?" only counted as questions if they were used to clarify content (and not because they did not hear the other person). Discrepancies between coders in the practice set of videos were primarily due to subjective interpretations of intonation. After training, the research assistants first marked every instance when a question was asked. As before, we also assigned a lead coder 25% of each of the other coders' videos, to check for reliability. Using a two-way mixed absolute agreement single measures ICC model, we obtained values that were all in the "excellent" range (ICCs from 0.8 to 1). We then had one separate coder identify and resolve the discrepancies between coders to obtain 100% agreement.

Next, we had coders examine each instance when a question was asked and mark what kind of question it was. Here again, we provided them with instructions and examples of each type of question. We once again used a lead coder who overlapped on 25% of the other three coders' files. We then checked reliability for these ratings using a two-way mixed absolute agreement average measures ICC (Hallgren, 2012; Shrout & Fleiss, 1979). The ICC values were all in the excellent range (ICCs from 0.89 to 1; Cichetti, 1994). Whenever there were discrepancies between two coders, we retained the "lead" coder's rating. We summed the number of questions for each time point and obtained one measure for each type of question.

Coders' Ratings of Engagement. Two trained research assistants assessed the degree to which each participant appeared engaged. We took a "thin-slice" approach to this coding procedure, which relies on having coders rate a few selected segments of an individual's behavior, as opposed to providing a rating for the entire length of the behavior. This method yields comparable results to ratings of the entire length of individuals' behaviors (Carney, Colvin, & Hall, 2007; Murphy et al., 2015). When judging engagement, the two coders were provided with the following instructions: "Overall, how engaged is the participant in the segment you just watched? (behaviors that might suggest engagement could include making eye contact, paying attention, asking questions, etc.)." To train, coders first coded eight participants together and discussed their ratings. To ensure that they would use the same criteria to judge engagement moving forward, coders discussed why they gave a particular rating and converged on several behaviors that indicated engagement. Some of the most common engagement behaviors were making eye contact, responding to questions quickly, nodding their head, and stopping to let the other person speak. Keeping these discussed behaviors in mind, coders next coded individually. They watched three sixty-second segments of the interaction: the first segment comprised the first 60 s of the interaction, the second was fifteen minutes after the start of the interaction, and the last was twenty-five minutes after the start. We chose a 60-second segment because previous studies have shown that this duration of time is optimal for capturing similar types of behaviors, such as how extroverted participants look (Carney et al., 2007). For each segment, coders were given instructions for what to look for when making their rating. The two coders overlapped for all of the videos, which allowed us to assess reliability for each construct. Reliability was assessed as above, using a two-way mixed consistency average-measures ICC. The ICC value was once again in the "good" range (ICC = 0.69; Cichetti, 1994).

6.3.3. *Affective experiences during the session*

Anxiety. Participants rated the extent to which they felt stressed, tense, and anxious, during the tutoring session on 7-point scales (1 = "Not at all", 7 = "Very much"; Cronbach's α = 0.83 for students, 0.94 for tutors). We averaged the three items to create an overall anxiety composite (Bosson, Haymovitz, & Pinel, 2004; West, Dovidio, & Pearson, 2014).

Confidence. Participants rated the extent to which they felt confident, secure, and uncertain (reverse-scored) during the tutoring session on 7-point scales (1 = "not at all", 7 = "very much"; Cronbach's α = 0.67 for students, 0.83 for tutors). We averaged all three items together to create a confidence composite.

Coders' Ratings of Confidence. For this measure, we used the same thin-slice approach as when assessing engagement, with two research assistants watching three 60-second segments of the interaction and rating participants' confidence. When assessing confidence, the two coders were provided with the following instructions: "Overall, how confident is the participant in the segment you just watched? (behaviors that suggest confidence include being assertive, not hesitating, etc.)." As was the case with engagement, coders first trained by coding the same eight participants. They then discussed discrepancies and converged on the types of behaviors indicating confidence. Some of the most common confidence behaviors included initiating the conversation at the start of the session and asking and answering questions without hesitation. With these discussed behaviors in mind, coders then coded all videos individually, with 100% overlap. We assessed reliability by using two-way mixed consistency average measures ICC (Hallgren, 2012; Shrout & Fleiss, 1979). The ICC was in the "good" range (ICC = 0.72; Cichetti, 1994).

Tutors' Ratings of Their Own Performance. Tutors were asked to indicate, on 7-point Likert scales (1 = "Strongly disagree"; 7 = "Strongly agree"), the extent to which they agreed with the following statements: "I successfully addressed my students' concerns," "The tutoring session went smoothly," "I understood my student's needs," and "I was confident with the information I provided during the session" (Cronbach's α

= 0.81). We averaged these scores and created a composite to represent the tutors' overall evaluation of their own performance during the tutoring session.

6.4. Analytic strategy

To examine the influence of participants' gender on our outcomes of interest, we conducted Actor-Partner Interdependence Models (APIM; Cook & Kenny, 2005). These models allowed us to investigate how one dyad member's gender ("actor gender") influences their own responses (e.g., "Does a person's gender influence one's own anxiety?"), while adjusting for the influence of their partner's gender ("partner gender") on their responses (e.g., "Does the gender of one's partner influence one's own anxiety?"). Because they are not the focus of this paper and because, in general, we did not find any significant partner gender effects (with the exception of talk time ratio), we report these results in the Supplementary Material only. We note that removing partner effects from the models does not affect the results reported. Where relevant, we report all covariance parameters (used to adjust for multiple forms of nonindependence) in the Supplementary Material as well. Where possible, we report effect sizes using the approach suggested by Edwards et al. (2008) for calculating partial R^2 for multilevel models (R_β^2).

6.4.1. Univariate outcomes measured at one time point for tutors only

To analyze tutors' evaluations of their own performance, we ran a linear regression investigating the effect of actor gender (i.e., tutor gender), while adjusting for partner gender (i.e., student gender).

6.4.2. Univariate outcomes measured at one time point for both dyad members

For self-reported anxiety and self-reported confidence, we used multilevel modeling (MIXED in SPSS) to account for nonindependence between dyadic combinations with the same tutor (given that some tutors tutored multiple students) and between dyad members. We investigated the effects of actor gender, role (student or tutor), and the interaction between actor gender and role, while adjusting for partner gender and the interaction between partner gender and role.

6.4.3. Multivariate outcomes measured at one time point for both dyad members

We coded multiple types of questions asked for tutors and students. These outcomes were overdispersed (meaning that the variance was larger than the mean), and so we used a Generalized Linear Mixed Model (PROC GLIMMIX in SAS) with a negative binomial distribution and logarithmic function (Rodrıguez, 2013). We accounted for nonindependence between dyadic combinations with the same tutor and between dyad members. Because the question types we coded were different for students and tutors, we ran separate models for students and tutors. Both models predicted the number of questions asked from actor gender, question type, and the interaction between actor gender and question type, while adjusting for partner gender and the interaction between partner gender and question type.

6.4.4. Univariate outcomes measured at multiple time points for both dyad members

To analyze talk time ratio (sixty time points), coders' ratings of anxiety, and coders' ratings of confidence (each coded at three time points), we used multilevel modeling (PROC MIXED in SAS) to account for nonindependence between dyadic combinations with the same tutor, between dyad members, and within-person across time. We examined the effects of actor gender, role, and the interaction between actor gender and role, while adjusting for partner gender and the interaction between partner gender and role.

7. Results

To rule out the possibility that pre-existing differences between men and women led to differences in engagement and affective experiences, we first examined whether there were gender differences in participants' reported academic performance and identification with their field of study.

7.1. Field identification

We did not find any differences in domain identification between men and women, $b = 0.08$, $SE = 0.10$, $t(101) = 0.82$, $p = .414$, 95% CI [5.87, 6.25], $R_\beta^2 = 0.007$. As expected, tutors ($M = 6.31$, $SD = 0.72$) identified significantly more with their field compared to students ($M = 5.75$, $SD = 1.05$), $b = 0.25$, $SE = 0.10$, $t(101) = 2.58$, $p = .011$, 95% CI [0.06, 0.44], $R_\beta^2 = 0.06$, and this effect did not vary by gender, $b = -0.07$, $SE = 0.10$, $t(101) = -0.67$, $p = .50$, 95% CI [−0.26, 0.13], $R_\beta^2 = 0.004$.

7.2. Academic performance

We did not find any differences in academic performance, as measured by GPA, between men and women, $b = 0.10$, $SE = 0.07$, $t(89) = 1.50$, $p = .138$, 95% CI [−0.03, 0.23], $R_\beta^2 = 0.02$. As expected, we found that tutors had a higher GPA ($M = 3.70$, $SD = 0.75$) compared to students ($M = 3.35$, $SD = 0.41$), $b = 0.16$, $SE = 0.07$, $t(89) = 2.43$, $p = .017$, 95% CI [0.03, 0.29], $R_\beta^2 = 0.06$; this effect did not vary by gender (no gender by role interaction), $b = 0.05$, $SE = 0.07$, $t(89) = 0.82$, $p = .42$, 95% CI [−0.08, 0.18], $R_\beta^2 = 0.007$.

7.3. Behavioral measures of engagement

7.3.1. Talk time ratio

There was no main effect of actor gender on the percentage of time participants spent talking relative to their partners, $b = -1.03$, $SE = 1.69$, $t(68.9) = -0.61$, $p = .55$, 95% CI [−4.40, 2.35], $R_\beta^2 = 0.005$. There was a significant main effect of role, $b = 17.33$, $SE = 3.07$, $t(54.2) = 5.65$, $p < .001$, 95% CI [11.18, 23.47], $R_\beta^2 = 0.37$ (see Table 1 and Fig. 1), such that tutors ($M = 68.29\%$, $SD = 30.43$) spoke for a higher percentage of time compared to students ($M = 28.96\%$, $SD = 28.50$). The interaction effect between actor gender and role was not significant, $b = -1.26$, $SE = 1.72$, $t(71.5) = -0.73$, $p = .47$, 95% CI [−4.70, 2.18], $R_\beta^2 = 0.007$.

Table 1
Talking Time Ratio and Coded Engagement by Gender and Role.

Rating	Talking Time Ratio			Coded Engagement		
	M	*SD*	95% CI	*M*	*SD*	95% CI
Gender						
Woman	42.02	34.88	[40.83, 43.21]	5.59	1.05	[5.44, 5.74]
Man	58.18	33.99	[56.91, 59.46]	5.65	1.03	[5.49, 5.81]
Role						
Student	29.25	28.66	[28.22, 30.28]	5.54	1.09	[5.37, 5.70]
Tutor	68.56	30.19	[67.49, 69.62]	5.7	0.98	[5.55, 5.85]
Role × Gender						
Man tutor	69.29	29.33	[67.94, 70.64]	5.62	1.03	[5.42, 5.81]
Man student	35.75	31.60	[33.68, 37.81]	5.73	1.03	[5.43, 6.03]
Woman tutor	67.49	31.36	[65.76, 69.23]	5.84	0.89	[5.62, 6.06]
Woman student	26.41	26.79	[25.25, 27.57]	5.46	1.10	[5.26, 5.66]

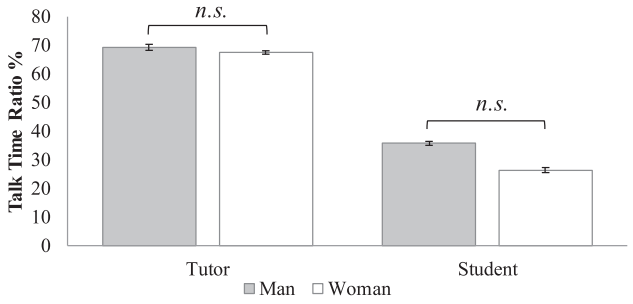

Fig. 1. Talking Time Ratio as a Function of Role and Actor Gender. *Note.* Error bars represent the standard error of the mean. *ns* indicates $p > .05$.

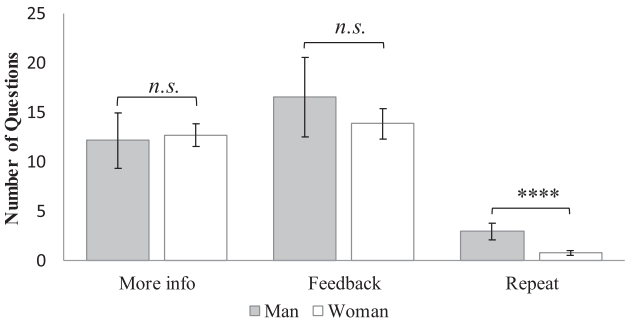

Fig. 2. Questions Asked by Students as a Function of Gender. *Note.* Error bars represent the standard error of the mean. *** indicates $p < .001$. *ns* indicates $p > .05$.

7.3.2. Types of questions asked

Questions Asked by Students. When analyzing questions asked by students, we found that 4.6% of the questions were "repeat" questions, 50.9% were "feedback" questions, and 43.3% were "more information" questions. We found a significant main effect of question type, $F(2, 106) = 95.07$, $p < .001$, $R_\beta^2 = 0.64$. Students asked more questions requesting more information than repeat questions (see Table 2), $b = 2.06$, $SE = 0.17$, $t(106) = 12.11$, $p < .001$, 95% CI [1.72, 2.40], $R_\beta^2 = 0.58$. Students also asked more feedback questions than repeat questions, $b = 2.24$, $SE = 0.17$, $t(106) = 13.36$, $p < .001$, 95% CI [1.92, 2.58], $R_\beta^2 = 0.63$. There was no significant difference between the number of feedback questions students asked and the number of questions requesting more information, $b = 0.19$, $SE = 0.12$, $t(106) = 1.54$, $p = .13$, 95% CI [0.06, 0.44], $R_\beta^2 = 0.02$.

We also found a significant main effect of gender across all question types, such that men students ($M = 10.53$, $SD = 12.29$) asked more questions than women students ($M = 9.09$, $SD = 9.32$), $F(1,106) = 5.73$, $p = .018$, $R_\beta^2 = 0.05$; however, this was qualified by a significant interaction between question type and actor gender, $F(2,106) = 9.62$, $p < .001$, $R_\beta^2 = 0.15$. We found a significant effect of gender on the number of repeat questions that students asked, $b = 0.70$, $SE = 0.17$, $t(106) = 4.15$, $p < .001$, 95% CI [0.36, 1.03], $R_\beta^2 = 0.14$, such that men students asked more repeat questions than women students (see Table 2 and Fig. 2). We did not find a significant simple gender effect on the number of feedback questions asked, $b = 0.11$, $SE = 0.12$, $t(106) = 0.88$, $p = .38$, 95% CI [−0.13, 0.35], $R_\beta^2 = 0.007$, or on the number of questions asking for more information, $b = -0.03$, $SE = 0.12$, $t(106) = -0.29$, $p = .77$, 95% CI [−0.27, 0.21], $R_\beta^2 = 0008$. Thus, the only category of questions for which we found a gender difference was repeat questions (men asked more).

Questions Asked by Tutors. When analyzing questions asked by tutors, we found that 21.3% were "clarification" questions and 58.5% were "knowledge" questions. We found a main effect of question type, $F(1, 52) = 40.87$, $p < .001$, such that tutors asked more knowledge questions compared to clarification questions, $b = 0.88$, $SE = 0.14$, $t(57) = 6.39$, $p < .001$, 95% CI [0.60, 1.15], $R_\beta^2 = 0.44$ (see Table 2 and Fig. 3). We did not find an effect of actor gender, $F(1, 52) = 2.55$, $p = .12$, $R_\beta^2 = 0.05$, nor an interaction between actor gender and question type, $F(1, 52) = 0.86$, $p = .36$, $R_\beta^2 = 0.02$.

Table 2
Means for question types for students and tutors.

	M	*SD*	95% CI
Student gender	More information Questions		
Woman	12.69	7.44	[10.37, 15.01]
Man	12.13	10.85	[6.13, 18.14]
	Feedback Questions		
Woman	13.83	9.97	[10.73, 16.94]
Man	16.53	15.59	[7.90, 25.17]
	Repeat Questions		
Woman	0.76	1.57	[0.27, 1.25]
Man	2.93	3.28	[1.12, 4.75]
Tutor gender	Clarification Questions		
Woman	9.52	9.22	[5.33, 13.72]
Man	10.03	7.15	[7.61, 12.45]
	Knowledge Questions		
Woman	22.95	27.14	[11.60, 36.30]
Man	29.89	21.79	[22.51, 37.26]

7.3.3. Coders' ratings of engagement

There was no main effect of actor gender, $b = -0.003$, $SE = 0.07$, $t(115) = -0.05$, $p = .96$, 95% CI [−0.14, 0.14], $R_\beta^2 < 0.001$, or role, $b = 0.05$, $SE = 0.05$, $t(90.9) = 1.01$, $p = .32$, 95% CI [−0.05, 0.16], $R_\beta^2 = 0.01$, on ratings of engagement. There was also no significant interaction between actor gender and role, $b = -0.12$, $SE = 0.07$, $t(117) = -1.80$, $p = .074$, 95% CI [−0.15, 0.12], $R_\beta^2 = 0.03$ (see Table 2 and Fig. 4).

In sum, women and men participants were engaged with the tutoring sessions to a similar degree: although men students asked more "repeat questions" compared to women students, men and women students and tutors did not significantly differ in any other engagement measures. They spent a similar amount of time talking relative to their partners, they asked a similar number of "feedback" and "more information" questions (for students) and "clarification" and "knowledge" questions (for tutors), and they appeared to be engaged to a similar degree (as coded by outside raters).

7.4. Affective experiences during the session

7.4.1. Anxiety

We found a main effect of actor gender on the amount of anxiety that participants experienced during the session, such that women reported higher levels of anxiety compared to men, $b = -0.41$, $SE = 0.14$, $t(67.57) = -2.93$, $p = .005$, 95% CI [-0.70, −0.13], $R_\beta^2 = 0.11$ (see Table 3 and Fig. 5). We did not find a main effect of role, $b = 0.02$, $SE = 0.14$, t

Fig. 3. Questions Asked by Tutors as a Function of Gender. *Note.* Error bars represent the standard error of the mean. *ns* indicates $p > .05$.

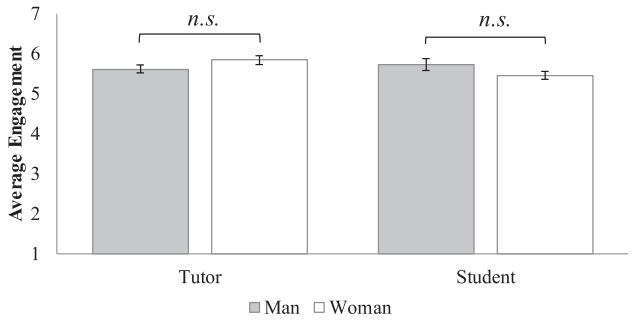

Fig. 4. Coders' Ratings of Engagement by Actor Gender. *Note.* Error bars represent the standard error of the mean. *ns* indicates $p > .05$.

(37.87) = 0.15, p =.88, 95% CI [−0.27, 0.31], $R_\beta^2 = 0.0006$, nor did we find a significant actor gender by role interaction, $b = -0.15$, $SE = 0.14$, $t(67.91) = -1.04$, p =.30, 95% CI [−0.43, 0.14], $R_\beta^2 = 0.02$. Thus, all women, regardless of whether they were a tutor or a student, felt more anxious than men.

7.4.2. Confidence

We also found a main effect of actor gender on the amount of confidence that participants reported during the session, such that men reported significantly more confidence compared to women, $b = 0.46$, $SE = 0.12$, $t(98.93) = 3.77$, $p < .001$, 95% CI [0.22, 0.70], $R_\beta^2 = 0.13$ (see Table 3 and Fig. 6). There was also a main effect of role, such that tutors reported feeling more confident compared to students, $b = 0.36$, $SE = 0.14$, $t(62.38) = 2.66$, p =.010, 95% CI [0.09, 0.64], $R_\beta^2 = 0.10$. We did not find a significant interaction between actor gender and role, $b = 0.17$, $SE = 0.12$, $t(99.12) = 1.37$, p =.18, 95% CI [−0.08, 0.41], $R_\beta^2 = 0.02$.

7.4.3. Coders' ratings of confidence

There was a main effect of actor gender on how confident participants were seen, such that men participants were seen as more confident compared to women participants, $b = 0.25$, $SE = 0.08$, $t(120) = 3.02$, p =.003, 95% CI [0.08, 0.41], $R_\beta^2 = 0.07$ (see Table 3 and Fig. 7 for means). There was also a main effect of role, such that tutors were seen as more confident than students, $b = 0.41$, $SE = 0.08$, $t(91.2) = 5.69$, $p < .001$, 95% CI [0.27, 0.56], $R_\beta^2 = 0.26$. There was no interaction between actor gender and role, $b = -0.04$, $SE = 0.08$, $t(122) = -0.53$, p =.60, 95% CI [−0.20, 0.12], $R_\beta^2 = 0.002$.

7.4.4. Tutors' ratings of their own performance

We found a main effect of tutor gender on their evaluation of their own performance during the session, such that men tutors rated their own performance significantly more positively compared to women tutors (see Fig. 8), $\beta = 0.28$, $SE = 0.09$, $t(54) = 3.04$, p =.004, 95% CI [0.10, 0.47], $R_\beta^2 = 0.15$.

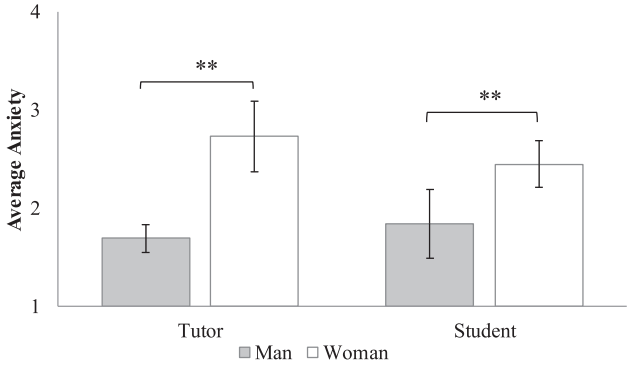

Fig. 5. Self-reported Anxiety by Actor Gender. *Note.* Error bars represent the standard error of the mean. ** indicates $p < .01$.

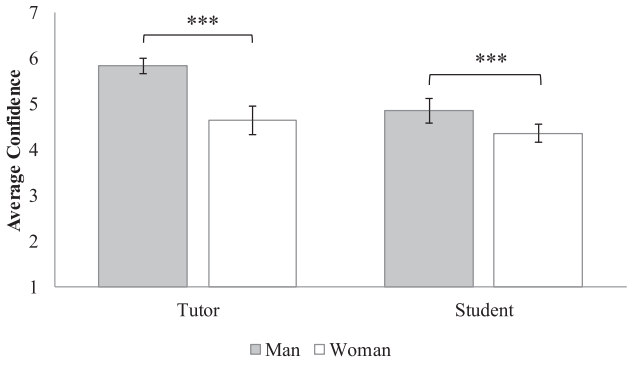

Fig. 6. Self-reported Confidence by Actor Gender. *Note.* *** indicates $p < .001$. Error bars represent the standard error of the mean.

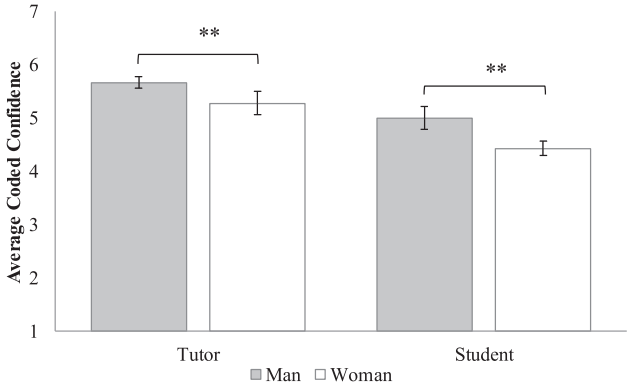

Fig. 7. Coders' Ratings of Confidence by Actor Gender. *Note.* Error bars represent the standard error of the mean. ** indicates $p < .01$.

Table 3
Means for Subjective Ratings of the Tutoring Session.

Rating	Anxiety			Confidence			Coded confidence			Evaluation of own performance		
	M	*SD*	95% CI	*M*	*SD*	95% CI	*M*	*SD*	95 %CI	*M*	*SD*	95 %CI
Gender												
Woman	2.55	1.58	[2.16, 2.95]	4.39	1.26	[4.08, 4.71]	4.73	1.00	[4.48, 4.98]			
Man	1.74	1.05	[1.45, 2.03]	5.58	1.19	[5.25, 5.91]	5.46	0.77	[5.23, 5.68]			
Role												
Student	2.28	1.53	[1.88, 2.69]	4.43	1.26	[4.10, 4.76]	4.59	1.30	[4.40, 4.79]			
Tutor	2.10	1.33	[1.76, 2.45]	5.41	1.29	[5.07, 5.74]	5.52	1.11	[5.35, 5.69]			
Role × Gender												
Man tutor	1.69	0.85	[1.41, 1.98]	5.90	1.06	[5.54, 6.26]	5.67	0.96	[5.45, 5.89]	5.99	0.83	[5.75, 6.22]
Woman tutor	2.73	1.68	[2.01, 3.46]	4.64	1.24	[4.10, 5.18]	5.28	1.30	[4.83, 5.74]	5.66	1.03	[5.41, 5.92]
Man student	1.84	1.44	[1.07, 2.61]	4.87	1.16	[4.26, 5.49]	5.00	1.22	[4.54, 5.46]	–	–	–
Woman student	2.45	1.54	[1.97, 2.93]	4.26	1.27	[3.97, 4.76]	4.43	1.30	[4.16, 4.70]	–	–	–

In sum, in contrast to the overall patterns of results for engagement, we found a consistent gender difference in students' and tutors' affective experiences related to the tutoring sessions. Women students and tutors reported more anxiety, less confidence, and were perceived as less confident by outside observers compared to men. In addition, women tutors rated their own performance significantly more negatively than men did.

8. Discussion

We examined whether the behavioral and psychological processes that unfold for students and tutors during STEM peer tutoring sessions differ by gender. We observed tutors and students as they participated in peer tutoring sessions, and we measured engagement behaviors of both partners during the session and their affective experiences after it. Our work reveals two primary findings. One, women and men did not differ from each other across several measures of engagement: the amount of time they spent talking, the number of feedback and more information questions they asked (for students), the number of clarification and knowledge questions they asked (for tutors), and how engaged they were as perceived by outside observers. Furthermore, gender effects for engagement behaviors were consistent across tutors and students. We found one gender difference in behavior: men students were more likely to ask questions they had asked before, compared to women students.

Two, although men and women showed similar levels of engagement in the tutoring sessions, they differed in their affective responses to the sessions. Women experienced more anxiety, less confidence, and were perceived as less confident by outside observers, compared to men; these findings held for both students and tutors. In addition, women tutors evaluated their own performance less positively compared to men tutors.

Our finding that men and women did not differ in their levels of engagement contrasts with research that has found gender differences in classroom engagement. This finding suggests that the structural differences between classroom and peer tutoring learning may be effective at reducing gender differences in engagement (Aguillon et al., 2020; Carter et al., 2018; Crombie et al., 2003; Daly et al., 1994; Eddy et al., 2014; Sankar et al., 2015). For women, not being numerically outnumbered might reduce activation of the gender-STEM stereotype, and in turn, decrease expectations that women do not belong and perform worse than men in STEM. With learning as the main goal, no formal evaluation immediately at stake, and a same-status instructor, women may feel more comfortable talking and asking questions during these tutoring sessions. Our finding also aligns with research suggesting that peer tutoring may be a helpful tool for encouraging women students' engagement in STEM (Dagley et al., 2016; Good et al., 2000; Savaria & Monteiro, 2017).

In contrast to our findings on engagement, however, our findings on negative affect align with the existing literature on men and women's experiences in STEM classrooms, where women tend to have less positive experiences compared to men (Cech et al., 2011; Pelch, 2018; Schuster & Martiny, 2017; Sterling et al., 2020). This finding is surprising considering existing evidence that emotion and engagement are typically linked in academic contexts (Dettmers et al., 2011; Pekrun, Goetz, Titz, & Perry, 2002) and can have a reciprocal relationship: negative emotions lead to disengagement, which in turn fuels negative outcomes (such as poor performance), ultimately reinforcing the negative emotions (Kahu, Stephens, Leach, & Zepke, 2015). However, it is possible that women who are still interested in STEM at the college level have somewhat adjusted to experiencing negative affect in STEM environments and have learned to cope with it in order to succeed. In other words, although women may feel worse than men in these tutoring sessions, they may understand that engagement during these sessions is required in order to learn. Thus, they persist in talking and asking questions even though they feel anxious and uncertain. To the extent that the accumulation of these negative affective experiences contribute to women's decision to leave STEM fields, peer tutoring interactions may not be entirely effective at reducing women's risk of dropping out of STEM fields.

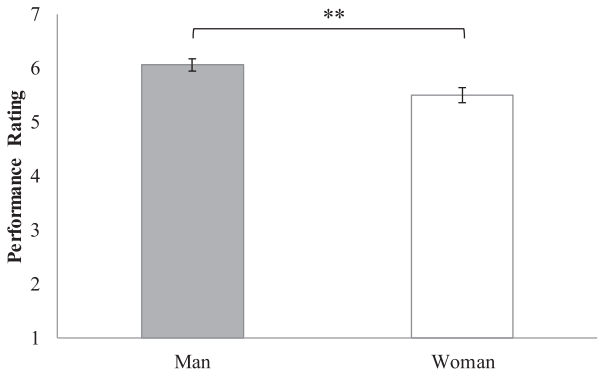

Fig. 8. Tutors' Evaluations of Their Own Performance by Gender. *Note.* Error bars represent the standard error of the mean. ** indicates $p < .01$.

Our findings have important implications for understanding how affective experiences in STEM may contribute to gender disparities in participation down the road, for two reasons. First, in the long run, consistently experiencing negative affect in peer tutoring interactions could lead women to experience burnout and eventually disengage from the field (Bedyńska & Żołnierczyk-Zreda, 2015; Bumbacco & Scharfe, 2020a,b; Jensen & Deemer, 2019; Pedersen & Minnotte, 2017). Although we did not find gender differences in engagement during one tutoring session, repeatedly experiencing negative affect during tutoring may eventually lead to disengagement—both during STEM interactions, and from STEM careers more generally. Indeed, studies have found that negative emotions experienced during learning negatively affect students' self-reported academic effort and motivation (Dettmers et al., 2011; Pekrun et al., 2002). Burnout may not only occur for women at early career stages, but also for women who are more advanced in the field (Pedersen & Minnotte, 2017), such as the peer-tutors in our study. For example, negative experiences might prevent tutors from pursuing future tutoring opportunities or other academic endeavors in STEM in the future. Future research should take a longitudinal approach to studying peer learning interactions to test whether there is a critical period during which negative emotions lead to behavioral disengagement during peer learning. Understanding the interdependence between negative emotions and disengagement is an important direction for future research and for developing interventions aimed at improving engagement more broadly.

Second, experiencing negative affect during peer tutoring could hinder the learning process and impact women's subsequent performance (Cassady, 2004; Cassady & Johnson, 2002). For students, experiencing negative affect might interfere with their learning during the session, for example, by shortening the span of working memory (Ashcraft & Kirk, 2001; Lavric, Rippon, & Gray, 2003; Mangels, Good, Whiteman, Maniscalco, & Dweck, 2012). For tutors, such experiences may prevent them from strengthening their understanding of concepts they already know (Good et al., 2000), thus interfering with the learning process and performance. These negative affective experiences could also indirectly influence retention for both students and peer-tutors.

Our study also has implications for the quality of peer relations in STEM. Interactions with peers, both early and later in individuals' career stages, can influence people's sense of belonging in the field (Hall, Schmader, Aday, & Croft, 2019; Walton, Logel, Peach, Spencer, & Zanna, 2015), and, for women especially, can determine whether they remain in the field (Ito & McPherson, 2018; Lewis et al., 2017). Furthermore, networking with peers is a critical component of persistence and success in STEM (Spurk, Kauffeld, Barthauer, & Heinemann, 2015; Xu & Martin, 2011). Peer tutoring has the potential to encourage stronger relationships within STEM because it provides structural

advantages compared to the classroom. However, if individuals consistently experience negative affect in these interactions, this could preclude them from forming such relationships.

Although we did not find differences in engagement between men and women on most of our outcomes of interest, there was one exception—the number of "repeat" questions that students asked during the session. We interpret these findings with caution primarily because the overall occurrence of "repeat" questions was infrequent (4.9%). Nevertheless, repeating a formerly-asked question could generate concerns about being perceived as incompetent because it indicates that the student did not understand the tutor's explanation the first time the question was asked. These concerns may affect women more than men because women may already be concerned about confirming the gender-STEM stereotype (Shapiro & Williams, 2012). Therefore, men students might be more comfortable asking these types of questions. However, further research is needed to explore this hypothesis, as there are other potential reasons for repeating a question such as, for instance, being inattentive.

8.1. Limitations and future directions

One unexplored aspect of this work is whether gender-matching in the tutoring pairs affects the behaviors and affective experiences we investigated. For women, interacting with women role models in STEM can lead to more positive experiences and increased retention (e.g. Dennehy & Dasgupta, 2017). Thus, it is possible that having a woman tutor could reduce the negative affective experiences that women students reported in this study. Furthermore, patterns of engagement might also differ for women when they are part of cross- versus same-gender dyads. For instance, one study showed that when women experienced stereotype threat prior to working on a mathematical task with a partner, they were more likely to be engaged when the partner was a woman, but not when the partner was a man (Thorson, Forbes, Magerman, & West, 2019). We did not have the statistical power to examine whether gender-matching influenced outcomes in this study, but future work with a larger sample size could answer this question. To that end, we also note that a qualitative analysis of the tutoring session recordings could provide further insight into the patterns of engagement and affect that men and women students demonstrate. Although it is beyond the scope of this paper, the quality of these interactions is important in its own right. For example, it is possible that the kind of language (e.g. positive or negative words) that tutors and students used could impact engagement and affective experiences.

In addition, although our findings suggest that negative affect is experienced disproportionately by both women students and women tutors, relative to their men counterparts, many women students in the study majored in subjects that are less math-intensive than traditional STEM courses (such as Psychology) and were tutored in a math-intensive course (such as Calculus). This distinction is an important avenue for future research, considering the vast amount of research demonstrating gender differences in mathematical self-concept (Devine, Fawcett, Szűcs, & Dowker, 2012; Else-Quest, Hyde, & Linn, 2010; ; Kim & Sax, 2018; Voyer et al., 2014). One the one hand, it is possible that math-intensive courses are more threatening for women who are not majoring in a math-intensive STEM field. On the other hand, students who major in STEM fields that are not very math-intensive may identify with STEM less, and may be less susceptible to the negative effects of stereotypes about women in STEM (Spencer et al., 1999). Thus, students' and tutors' intentions to pursue math-intensive STEM majors could play a role in the patterns of engagement and affect that we observed, and should be investigated in future research.

Another important aspect of our recruitment process is that we selected students for tutoring sessions based on their self-reported need and not on their objective performance in a class (e.g. grades). Although this process is similar to that of the university's tutoring center, an important direction for future research would be to examine whether there is a difference in behavioral and affective patterns between students who are objectively in need of tutoring, as demonstrated by their class performance, and students who self-selected into the tutoring session. It is possible that students who self-select into tutoring may experience more negative affect compared to those who do not, having already acknowledged that they are falling behind in the class. However, it is important to note that we found differences in affect for tutors as well, suggesting that these patterns overall may not be a result of selection bias due to anxiety in the domain.

As mentioned above, future research could also use a longitudinal design (e.g., where the same tutor and student meet multiple times over the course of one semester) to investigate the extent to which the patterns we observed persist once students and tutors become more acquainted with each other or whether feelings of uncertainty and anxiety decline over time, as they have been shown to do in previous research (Reis, Maniaci, Caprariello, Eastwick, & Finkel, 2011; Vittengl & Holt, 2000). Moreover, because our design was cross-sectional, we do not have a baseline measure of women's negative affect outside of the tutoring session that would allow us to compare whether the tutoring session made them experience more or less negative affect compared to their baseline experience in STEM, or in academic environments more generally. By directly comparing peer tutoring experiences in STEM to classroom experiences, for example, we may find that peer tutoring sessions generate less negative affect compared to other STEM learning contexts, considering the theorized differences in evaluative concerns. This pattern, coupled with our findings from this study, would suggest that despite inducing less anxiety overall, peer tutoring may still be less advantageous for women than it is for men, at least when it comes to their affective experiences.

We also note that in this study we focused on measures of self-report and outside coders' perceptions, and thus did not examine self-other perception differences in these dyads. For instance, with regard to the tutors' evaluation of their own performance during the session, it is possible that rather than women underestimating their performance, men may be overestimating theirs. Investigating this difference is an important avenue of future research, as students' evaluations of their tutors, and vice-versa, could also influence decisions to persist in the field.

Lastly, an important limitation of this study is that we did not ask students whether they had any learning disabilities or academic accommodations. Students with disabilities can experience more negative affect and show less engagement during learning interactions compared to students without disabilities (Sideridis, 2005). However, one-on-one peer tutoring interactions may actually be easier to navigate for students with disabilities compared to interactions in large classrooms. Therefore, it is possible that women students with disabilities may experience less negative affect in peer-tutoring compared to other learning environments. Future research could investigate whether the intersectionality of learning disabilities and gender could change the dynamics observed in this study. Furthermore, individuals from different cultures and across the neurodiversity spectrum might show different patterns of engagement and affective experience than the ones we considered in the present study. Since our coders were all from Western cultures, it is possible that their interpretations of engagement and confidence were biased. For instance, there are cultural differences in the kinds of behavior that are considered polite (Yu, 2011). It is possible that students and tutors from different cultures may perceive behaviors such as asking questions and talking for long periods of time as impolite, and they may try to regulate the frequency of these behaviors. Future research could also investigate factors, other than gender, that could influence engagement and affective experiences in peer tutoring. Within STEM, racial minorities and first-generation college students are also underrepresented and can show different patterns of classroom engagement and affective experiences relative to their peers (e.g. Flynn, 2016). Thus, examining whether our findings extend to other underrepresented groups in STEM and understanding how the

intersectionality of multiple underrepresented identities might impact engagement and negative affect are other important next steps in this line of research.

8.2. Conclusion

Despite being similarly engaged in peer tutoring sessions across several behavioral measures, we found that women experience more anxiety, feel less confident, are perceived as less confident, and evaluate their performance more poorly (as tutors). To our knowledge, this is the first study to directly investigate how students and tutors interact during peer tutoring sessions, using a multi-method approach for studying men and women's engagement and affective experiences. Overall, our findings highlight affective experiences during STEM tutoring interactions as underexplored factors that could contribute to previously documented gender disparities across multiple stages of the STEM pipeline.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by National Science Foundation, United States (grant number DRL1535414, awarded to Chad E. Forbes and Tessa V. West).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cedpsych.2022.102088.

References

Academic Advising and Support at Caltech. (n.d.). Peer Tutor Information. https://www.deans.caltech.edu/AcademicSupport/peer-tutor-information.

Academic Resource Center at Harvard University. (n.d.) Peer Tutoring. https://academicresourcecenter.harvard.edu/peer-tutoring.

Aguillon, S. M., Siegmund, G. F., Petipas, R. H., Drake, A. G., Cotner, S., & Ballen, C. J. (2020). Gender differences in student participation in an active-learning classroom. *CBE Life Sciences Education, 19*(2), 1–10. https://doi.org/10.1187/cbe.19-03-0048

Alegre, F., Moliner, L., Maroto, A., & Lorenzo-Valentin, G. (2020). Academic achievement and peer tutoring in mathematics: a comparison between primary and secondary education. *SAGE Open, 10*(2). https://doi.org/10.1177%2F2158244020929295.

American Physical Society. (2018). Bachelor's Degrees Earned by Women, by Major. https://www.aps.org/programs/education/statistics/womenmajors.cfm.

Anderson, C., & Kilduff, G. J. (2009). Why do dominant personalities attain influence in face-to-face groups? The competence-signaling effects of trait dominance. *Journal of Personality and Social Psychology, 96*(2), 491. https://psycnet.apa.org/doi/10.1037/a0014201.

Appel, M., & Kronberger, N. (2012). Stereotypes and the achievement gap: Stereotype threat prior to test taking. *Educational Psychology Review, 24*(4), 609-635. https:/doi.org/10.1007/s10648-012-9200-4. https://doi.org/10.1007/s10648-012-9200-4.

Archambault, I., & Dupéré, V. (2017). Joint trajectories of behavioral, affective, and cognitive engagement in elementary school. *Journal of Educational Research, 110*(2), 188–198. https://doi.org/10.1080/00220671.2015.1060931

Ashcraft, M. H., & Kirk, E. P. (2001). The relationships among working memory, math anxiety, and performance. *Journal of Experimental Psychology: General, 130*(2), 224. https://psycnet.apa.org/doi/10.1037/0096-3445.130.2.224.

Ballen, C. J., Aguillon, S. M., Awwad, A., Bjune, A. E., Challou, D., Drake, A. G., … Cotner, S. (2019). Smaller classes promote equitable student participation in STEM. *BioScience, 69*(8), 669–680. https://doi.org/10.1093/biosci/biz069

Banchefsky, S., Lewis, K. L., & Ito, T. A. (2019). The role of social and ability belonging in men's and women's pSTEM persistence. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.02386

Batz, Z., Olsen, B. J., Dumont, J., Dastoor, F., & Smith, M. K. (2015). Helping struggling students in introductory biology: A peer-tutoring approach that improves performance, perception, and retention. *CBE—Life Sciences Education, 14*(2), ar16. https://doi.org/10.1187/cbe.14-08-0120

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Blletin, 117*(3), 497.

Beasley, M. A., & Fischer, M. J. (2012). Why they leave: The impact of stereotype threat on the attrition of women and minorities from science, math and engineering majors. *Social Psychology of Education, 15*(4), 427–448. https://doi.org/10.1007/s11218-012-9185-3

Bedyńska, S., & Żołnierczyk-Zreda, D. (2015). Stereotype threat as a determinant of burnout or work engagement. Mediating role of positive and negative emotions. *International Journal of Occupational Safety and Ergonomics, 21*(1), 1–8. https://doi.org/10.1080/10803548.2015.1017939

Bench, S. W., Lench, H. C., Liew, J., Miner, K., & Flores, S. A. (2015). Gender gaps in overestimation of math performance. *Sex Roles, 72*(11), 536–546. https://doi.org/10.1007/s11199-015-0486-9

Berger, J., Cohen, B. P., & Zelditch, M., Jr. (1972). Status characteristics and social interaction. *American Sociological Review*, 241–255.

Bloodhart, B., Balgopal, M. M., Casper, A. M. A., Sample McMeeking, L. B., & Fischer, E. V. (2020). Outperforming yet undervalued: Undergraduate women in STEM. *PLOS ONE, 15*(6). https://doi.org/10.1371/journal.pone.0234685

Blumenfeld, P., Modell, J., Bartko, W. T., Secada, W. G., Fredricks, J. A., Friedel, J., & Paris, A. (2005). School engagement of inner-city students during middle childhood. *Developmental pathways through middle childhood: Rethinking contexts and diversity as resources,* 145-170.

Bolger, N., & Laurenceau, J. P. (2013). Intensive longitudinal methods: An introduction to diary and experience sampling research. Guilford Press.

Bosson, J. K., Haymovitz, E. L., & Pinel, E. C. (2004). When saying and doing diverge: The effects of stereotype threat on self-reported versus non-verbal anxiety. *Journal of Experimental Social Psychology, 40*(2), 247–255. https://doi.org/10.1016/S0022-1031(03)00099-4

Bumbacco, C., & Scharfe, E. (2020a). Why Attachment Matters: First-Year Post-secondary Students' Experience of Burnout, Disengagement, and Drop-Out. *Journal of College Student Retention: Research, Theory & Practice.* https://doi.org/10.1177%2F1521025120961012.

Bumbacco, C., & Scharfe, E. (2020b). Why Attachment Matters: First-Year Post-secondary Students' Experience of Burnout, Disengagement, and Drop-Out. *Journal of College Student Retention: Research, Theory and Practice, 1–14*. https://doi.org/10.1177/1521025120961012

Cadinu, M., Maass, A., Rosabianca, A., & Kiesner, J. (2005). Why do women underperform under stereotype threat? Evidence for the role of negative thinking. *Psychological Science, 16*(7), 572–578. https://doi.org/10.1111/j.0956-7976.2005.01577.x

Cano, F., García, Á., Berbén, A. B. G., & Justicia, F. (2014). Science learning: A path analysis of its links with reading comprehension, question-asking in class and science achievement. *International Journal of Science Education, 36*(10), 1710–1732. https://doi.org/10.1080/09500693.2013.876678

Carney, D. R., Colvin, C. R., & Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality, 41*(5), 1054–1072. https://psycnet.apa.org/doi/10.1016/j.jrp.2007.01.004.

Carter, A. J., Croft, A., Lukas, D., & Sandstrom, G. M. (2018). Women's visibility in academic seminars: Women ask fewer questions than men. *PLoS ONE, 13*(9), 1–22. https://doi.org/10.1371/journal.pone.0202743

Casad, B. J., Petzel, Z. W., & Ingalls, E. A. (2019). A model of threatening academic environments predicts women STEM majors' self-esteem and engagement in STEM. *Sex Roles, 80*(7), 469–488. https://doi.org/10.1007/s11199-018-0942-4

Cassady, J. C. (2004). The influence of cognitive test anxiety across the learning–testing cycle. *Learning and Instruction, 14*(6), 569–592. https://doi.org/10.1006/ceps.2001.1094

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27*(2), 270–295. https://doi.org/10.1006/ceps.2001.1094

Cech, E., Rubineau, B., Silbey, S., & Seron, C. (2011). Professional role confidence and gendered persistence in engineering. *American Sociological Review, 76*(5), 641–666. https://doi.org/10.1177%2F0003122411420815.

Cheryan, S., Plaut, V. C., Davies, P. G., & Steele, C. M. (2009). Ambient belonging: How stereotypical cues impact gender participation in computer science. *Journal of Personality and Social Psychology, 97*(6), 1045. https://doi.org/10.1037/a0016239

Cheryan, S., Siy, J. O., Vichayapai, M., Drury, B. J., & Kim, S. (2011). Do female and male role models who embody STEM stereotypes hinder women's anticipated success in STEM? *Social Psychological and Personality Science, 2*(6), 656–664. https://doi.org/10.1177%2F1948550611405218.

Chi, M. T., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*(4), 471–533.

Chipman, S. F., Krantz, D. H., & Silver, R. (1992). Mathematics anxiety and science careers among able college women. *Psychological Science, 3*(5), 292–296. https://doi.org/10.1111%2Fj.1467-9280.1992.tb00675.x.

Cichetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284. https://psycnet.apa.org/doi/10.1037/1040-3590.6.4.284.

Cohen, P. A., Kulik, J. A., & Kulik, C. L. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19*(2), 237–248. https://doi.org/10.3102%2F00028312019002237.

Colvin, J. W. (2007). Peer tutoring and social dynamics in higher education. *Mentoring & Tutoring, 15*(2), 165–181. https://doi.org/10.1080/13611260601086345

Connolly, R. (2020). Why computing belongs within the social sciences. *Communications of the ACM, 63*(8), 54–59. https://doi.org/10.1145/3383444

Cook, W. L., & Kenny, D. A. (2005). The actor–partner interdependence model: A model of bidirectional effects in developmental studies. *International Journal of Behavioral Development, 29*(2), 101–109. https://doi.org/10.1080%2F01650250444000405.

Crisp, G., Nora, A., & Taggart, A. (2009). Student characteristics, pre-college, college, and environmental factors as predictors of majoring in and earning a STEM degree: An analysis of students attending a Hispanic serving institution. *American Educational Research Journal, 46*(4), 924–942. https://doi.org/10.3102%2F0002831209349460.

Crocker, J., Luhtanen, R. K., Cooper, M. L., & Bouvrette, A. (2003). Contingencies of self-worth in college students: Theory and measurement. *Journal of Personality and Social Psychology, 85*(5), 894. https://doi.org/10.1037/0022-3514.85.5.894

Crombie, G., Pyke, S. W., Silverthorn, N., Jones, A., & Piccinin, S. (2003). Students' perceptions of their classroom participation and instructor as a function of gender and context. *The Journal of Higher Education, 74*(1), 51–76. https://doi.org/10.1080/00221546.2003.11777187

Dagley, M., Georgiopoulos, M., Reece, A., & Young, C. (2016). Increasing retention and graduation rates through a STEM learning community. *Journal of College Student Retention: Research, Theory & Practice, 18*(2), 167–182. https://doi.org/10.1177%2F1521025115584746.

Daly, J. A., Kreiser, P. O., & Roghaar, L. A. (1994). Question-asking comfort: Explorations of the demography of communication in the eighth grade classroom. *Communication Education, 43*(1), 27–41. https://doi.org/10.1080/03634529409378959

Dasgupta, N., & Stout, J. G. (2014). Girls and women in science, technology, engineering, and mathematics: STEMing the tide and broadening participation in STEM careers. *Policy Insights from the Behavioral and Brain Sciences, 1*(1), 21–29. https://doi.org/10.1177/2372732214549471

Deiglmayr, A., Stern, E., & Schubert, R. (2019). Beliefs in "brilliance" and belonging uncertainty in male and female STEM students. *Frontiers in Psychology, 10*(5), 1–7. https://doi.org/10.3389/fpsyg.2019.01114

Dempsey, M. A. (2016). Assessing the effect of peer tutoring in STEM gatekeeper courses with respect to gender and ethnicity (Doctoral dissertation).

Dennehy, T. C., & Dasgupta, N. (2017). Female peer mentors early in college increase women's positive academic experiences and retention in engineering. *Proceedings of the National Academy of Sciences, 114*(23), 5964–5969. https://doi.org/10.1073/pnas.1613117114

Dettmers, S., Trautwein, U., Lüdtke, O., Goetz, T., Frenzel, A. C., & Pekrun, R. (2011). Students' emotions during homework in mathematics: Testing a theoretical model of antecedents and achievement outcomes. *Contemporary Educational Psychology, 36*(1), 25–35. https://doi.org/10.1016/j.cedpsych.2010.10.001

Devine, A., Fawcett, K., Szűcs, D., & Dowker, A. (2012). Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety. *Behavioral and Brain Functions, 8*(1), 1–9.

Dillon, J. T. (1982). Cognitive correspondence between question/statement and response. *American Educational Research Journal, 19*(4), 540–551. https://doi.org/10.3102%2F00028312019004540.

Duke, D., Krishnan, M., Faith, M., & Storch, E. A. (2006). The psychometric properties of the Brief Fear of Negative Evaluation Scale. *Journal of Anxiety Disorders, 20*(6), 807–817. https://doi.org/10.1016/j.janxdis.2005.11.002

Eddy, S. L., Brownell, S. E., & Wenderoth, M. P. (2014). Gender gaps in achievement and participation in multiple introductory biology classrooms. *CBE—Life Sciences Education, 13*(3), 478–492. https://doi.org/10.1187/cbe.13-10-0204

Edwards, L. J., Muller, K. E., Wolfinger, R. D., Qaqish, B. F., & Schabenberger, O. (2008). An R2 statistic for fixed effects in the linear mixed model. *Statistics in Medicine, 27*(29), 6137–6157. https://doi.org/10.1002/sim.3429

Ekstrom, R. B., Goertz, M. E., Pollack, J. M., & Rock, D. A. (1986). Who drops out of high school and why? Findings from a national study. *Teachers College Record, 87*(3), 356–373.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-National Patterns of Gender Differences in Mathematics: A Meta-Analysis. *Psychological Bulletin, 136*(1), 103–127. https://doi.org/10.1037/a0018053

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: a meta-analysis. *Psychological bulletin, 136*(1), 103.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Finn, J. D. (1989). Withdrawing from school. *Review of educational research, 59*(2), 117–142.

Flynn, D. T. (2016). STEM field persistence: The impact of engagement on postsecondary STEM persistence for underrepresented minority students. *Journal of Educational Issues, 2*(1), 185–214. https://doi.org/10.5296/jei.v2i1.9245

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59–109.

Fritschner, L. M. (2000). Inside the undergraduate college classroom: Faculty and students differ on the meaning of student participation. *The Journal of Higher Education, 71*(3), 342–362. https://doi.org/10.1080/00221546.2000.11780826

Gajda, A., Beghetto, R. A., & Karwowski, M. (2017). Exploring creative learning in the classroom: A multi-method approach. *Thinking Skills and Creativity, 24*, 250–267. https://doi.org/10.1016/j.tsc.2017.04.002

Gansen, H. M. (2019). Push-Ups Versus Clean-Up: Preschool Teachers' Gendered Beliefs, Expectations for Behavior, and Disciplinary Practices. *Sex Roles, 80*(7–8), 393–408. https://doi.org/10.1007/s11199-018-0944-2

Gasiewski, J. A., Eagan, M. K., Garcia, G. A., Hurtado, S., & Chang, M. J. (2012). From gatekeeping to engagement: A multicontextual, mixed method study of student academic engagement in introductory STEM courses. *Research in Higher Education, 53*(2), 229-261. https:/doi.org/10.1007/s11162-011-9247-y.

Goetz, T., Bieg, M., Lüdtke, O., Pekrun, R., & Hall, N. C. (2013). Do girls really experience more anxiety in mathematics? *Psychological Science, 24*(10), 2079–2087. https://doi.org/10.1177%2F0956797613486989.

Good, J. M., Halpin, G., & Halpin, G. (2000). A promising prospect for minority retention: Students becoming peer mentors. *Journal of Negro Education, 375–383*. https://doi.org/10.2307/2696252

Good, T. L., Sterzinger, N., & Lavigne, A. (2018). Expectation effects: Pygmalion and the initial 20 years of research1. *Educational Research and Evaluation, 24*(3–5), 99–123. https://doi.org/10.1080/13803611.2018.1548817

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal, 31*(1), 104–137. https://doi.org/10.3102%2F00028312031001104.

Greider, C. W., Sheltzer, J. M., Cantalupo, N. C., Copeland, W. B., Dasgupta, N., Hopkins, N., … Wong, J. Y. (2019). Increasing gender diversity in the STEM research workforce. *Science, 366*(6466), 692–695. https://doi.org/10.1126/science.aaz0649

Guillén, L., Mayo, M., & Karelaia, N. (2018). Appearing self-confident and getting credit for it: Why it may be easier for men than women to gain influence at work. *Human Resource Management, 57*(4), 839–854. https://doi.org/10.1002/hrm.21857

Hagiwara, N., Penner, L. A., Gonzalez, R., Eggly, S., Dovidio, J. F., Gaertner, S. L., West, T. V., & Albrecht, T. L. (2013). Racial attitudes, physician–patient talk time ratio, and adherence in racially discordant medical interactions. *Social Science & Medicine, 87*, 123–131. https://doi.org/10.1016/j.socscimed.2013.03.016

Hall, W., Schmader, T., Aday, A., & Croft, E. (2019). Decoding the dynamics of social identity threat in the workplace: A within-person analysis of women's and men's interactions in STEM. *Social Psychological and Personality Science, 10*(4), 542–552. https://doi.org/10.1177%2F1948550618772582.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods For Psychology, 8*(1), 23. https://doi.org/10.20982/tqmp.08.1.p023.

Hascher, T., & Hagenauer, G. (2010). Alienation from school. *International Journal of Educational Research, 49*(6), 220–232.

Hawkins, K., & Power, C. B. (1999). Gender differences in questions asked during small decision-making group discussions. *Small Group Research, 30*(2), 235–256. https://doi.org/10.1177%2F104649649903000205.

Hunter, A., & Brewer, J. D. (2015). Designing multimethod research. In *The Oxford handbook of multimethod and mixed methods research inquiry*.

Ito, T. A., & McPherson, E. (2018). Factors influencing high school students' interest in pSTEM. *Frontiers in Psychology, 9*, 1535. https://doi.org/10.3389/fpsyg.2018.01535

Jakobsson, N. (2012). Gender and confidence: Are women underconfident? *Applied Economics Letters, 19*(11), 1057–1059. https://doi.org/10.1080/13504851.2011.613745

Jensen, L. E., & Deemer, E. D. (2019). Identity, Campus Climate, and Burnout Among Undergraduate Women in STEM Fields. *Career Development Quarterly, 67*(2), 96–109. https://doi.org/10.1002/cdq.12174

Johns, M., Inzlicht, M., & Schmader, T. (2008). Stereotype threat and executive resource depletion: Examining the influence of emotion regulation. *Journal of Experimental Psychology: General, 137*(4), 691. https://doi.org/10.1037/a0013834

Kahu, E., Stephens, C., Leach, L., & Zepke, N. (2015). Linking academic emotions and student engagement: Mature-aged distance students' transition to university. *Journal of Further and Higher Education, 39*(4), 481–497. https://doi.org/10.1080/0309877X.2014.895305

Karabenick, S. A., & Dembo, M. H. (2011). Understanding and facilitating self-regulated help seeking. *New Directions for Teaching and Learning, 126*, 33–43. https://doi.org/10.1002/tl.442

Kim, Y. K., & Sax, L. J. (2018). *The Effect of Positive Faculty Support on Mathematical Self-Concept for Male and Female Students in STEM Majors. Research in Higher Education* (Vol. 59). Springer Netherlands. https://doi.org/10.1007/s11162-018-9500-8.

Korpershoek, H., Canrinus, E. T., Fokkens-Bruinsma, M., & de Boer, H. (2020). The relationships between school belonging and students' motivational, social-emotional, behavioural, and academic outcomes in secondary education: A meta-analytic review. *Research Papers in Education, 35*(6), 641–680. https://doi.org/10.1080/02671522.2019.1615116

Kowalewski, B., Massen, A., & Mullins, S. (2010). Preparing to Serve: Online Training Modules. *Weber State University*. Retrieved from https://docplayer.net/21091051-Preparing-to-serve-online-training-modules.html.

Lavric, A., Rippon, G., & Gray, J. R. (2003). Threat-evoked anxiety disrupts spatial working memory performance: An attentional account. *Cognitive Therapy and Research, 27*(5), 489–504. https://doi.org/10.1023/A:1026300619569

Leaper, C. (2015). Do I belong?: Gender, peer groups, and STEM achievement. *International Journal of Gender, Science and Technology, 7*(2), 166–179.

Lee, J. J., & Mccabe, J. M. (2021). Who speaks and who listens: Revisiting the chilly climate in college classrooms. *Gender & society, 35*(1), 32–60. https://doi.org/10.1177/0891243220977141

Lee, J. T., Kim, Y. B., & Yoon, C. H. (2004). The effects of pre-class tutoring on student achievement: Challenges and implications for public education in Korea. *KEDI Journal of Educational Policy, 1*(1).

Lewis, K. L., Stout, J. G., Finkelstein, N. D., Pollock, S. J., Miyake, A., Cohen, G. L., & Ito, T. A. (2017). Fitting in to move forward: Belonging, gender, and persistence in the physical sciences, technology, engineering, and mathematics (pSTEM). *Psychology of Women Quarterly, 41*(4), 420–436. https://doi.org/10.1177/0361684317720186

Li, Y., & Lerner, R. M. (2013). Interrelations of behavioral, emotional, and cognitive school engagement in high school students. *Journal of Youth and Adolescence, 42*(1), 20–32.

Mangels, J. A., Good, C., Whiteman, R. C., Maniscalco, B., & Dweck, C. S. (2012). Emotion blocks the path to learning under stereotype threat. *Social Cognitive and Affective Neuroscience, 7*(2), 230–241. https://doi.org/10.1093/scan/nsq10

Maresh, E. L., Teachman, B. A., & Coan, J. A. (2017). Are you watching me? Interacting effects of fear of negative evaluation and social context on cognitive performance.

Journal of Experimental Psychopathology, 8(3), 303–319. https://doi.org/10.5127/jep.059516
Master, A., & Meltzoff, A. N. (2020). Cultural Stereotypes and Sense of Belonging Contribute to Gender Gaps in STEM. *International Journal of Gender, Science and Technology, 12*(1), 152–198.
McMunn, A. (2017). Gender differences. *The Routledge International Handbook of Psychosocial Epidemiology, 188–212*. https://doi.org/10.4324/9781315673097
Mishra, S. (2020). Social networks, social capital, social support and academic success in higher education: A systematic review with a special focus on 'underrepresented' students. *Educational Research Review, 29*, 100307. https://doi.org/10.1016/j.edurev.2019.100307
Morrissey, K., Hallett, D., Bakhtiar, A., & Fitzpatrick, C. (2019). Implicit math-gender stereotype present in adults but not in 8th grade. *Journal of adolescence, 74*, 173–182. https://doi.org/10.1016/j.adolescence.2019.06.003
Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences, 109*(41), 16474–16479. https://doi.org/10.1073/pnas.1211286109
Murphy, M. C., Steele, C. M., & Gross, J. J. (2007). Signaling threat: How situational cues affect women in math, science, and engineering settings. *Psychological Science, 18*(10), 879–885. https://doi.org/10.1111%2Fj.1467-9280.2007.01995.x.
Murphy, N. A., Hall, J. A., Schmid Mast, M., Ruben, M. A., Frauendorfer, D., Blanch-Hartigan, D., … Nguyen, L. (2015). Reliability and validity of nonverbal thin slices in social interactions. *Personality and Social Psychology Bulletin, 41*(2), 199–213. https://doi.org/10.1177%2F0146167214559902.
Mynard, J., & Almarzouqi, I. (2006). Investigating peer tutoring. *Elt Journal, 60*(1), 13-22. https://doi.org/10.1093/elt/cci077.
National Center for Education Statistics. (2018). Bachelor's, master's, and doctor's degrees conferred by postsecondary institutions, by sex of student and discipline division: 2017-18. https://nces.ed.gov/programs/digest/d19/tables/dt19_318.30.asp.
National Center for Education Statistics. (2019). Status and Trends in the Education of Racial and Ethnic Groups. https://nces.ed.gov/programs/raceindicators/indicator_REG.asp.
National Science Foundation, National Center for Science and Engineering Statistics (2019). Women, Minorities, and Persons with Disabilities in Science and Engineering: 2019. *Special Report NSF 19-304*. Alexandria, VA. Available at https://www.nsf.gov/statistics/wmpd.
Niler, A. A., Asencio, R., & DeChurch, L. A. (2020). Solidarity in STEM: How gender composition affects women's experience in work teams. *Sex Roles, 82*(3), 142–154. https://doi.org/10.1007/s11199-019-01046-8
Oswald, D. L. (2008). Gender stereotypes and women's reports of liking and ability in traditionally masculine and feminine occupations. *Psychology of Women Quarterly, 32*(2), 196–203. https://doi.org/10.1111/j.1471-6402.2008.00424.x
Park, G., Schmidt, A. M., Scheu, C., & DeShon, R. P. (2007). A process model of goal orientation and feedback seeking. *Human Performance, 20*(2), 119–145. https://doi.org/10.1080/08959280701332042
Passolunghi, M. C., Ferreira, T. I. R., & Tomasetto, C. (2014). Math–gender stereotypes and math-related beliefs in childhood and early adolescence. *Learning and Individual Differences, 34*, 70–76. https://doi.org/10.1016/j.lindif.2014.05.005
Pearson, J. C., & West, R. (1991). An initial investigation of the effects of gender on student questions in the classroom: Developing a descriptive base. *Communication Education, 40*(1), 22–32. https://doi.org/10.1080/03634529109378823
Pedersen, D. E., & Minnotte, K. L. (2017). Workplace Climate and STEM Faculty Women's Job Burnout. *Journal of Feminist Family Therapy, 29*(1–2), 45–65. https://doi.org/10.1080/08952833.2016.1230987
Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist, 37*(2), 91–105. https://doi.org/10.1207/S15326985EP3702_4
Pelch, M. (2018). Gendered differences in academic emotions and their implications for student success in STEM. *International Journal of Stem Education, 5*(1), 1–15. https://doi.org/10.1186/s40594-018-0130-7
Pomerantz, E. M., Altermatt, E. R., & Saxon, J. L. (2002). Making the grade but feeling distressed: Gender differences in academic performance and internal distress. *Journal of Educational Psychology, 94*(2), 396. https://doi.org/10.1037//0022-0663.94.2.396
Precourt, E., & Gainor, M. (2019). Factors affecting classroom participation and how participation leads to a better learning. *Accounting Education, 28*(1), 100–118.
Rainey, K., Dancy, M., Mickelson, R., Stearns, E., & Moller, S. (2018). Race and gender differences in how sense of belonging influences decisions to major in STEM. *International Journal of STEM Education, 5*(1), 1–14. https://doi.org/10.1186/s40594-018-0115-6
Reis, H. T., Maniaci, M. R., Caprariello, P. A., Eastwick, P. W., & Finkel, E. J. (2011). Familiarity does indeed promote attraction in live interaction. *Journal of Personality and Social Psychology, 101*(3), 557. https://doi.org/10.1037/a0022885
Rodriguez, G. (2013). Models for count data with overdispersion. *Addendum to the WWS, 509*.
Rumberger, R. W., & Rotermund, S. (2012). *The relationship between engagement and high school dropout. In Handbook of research on student engagement* (pp. 491–513). Boston, MA: Springer.
Sankar, P., Gilmartin, J., & Sobel, M. (2015). An examination of belongingness and confidence among female computer science students. *Acm Sigcas Computers and Society, 45*(2), 7–10.
Savaria, M. C., & Monteiro, K. A. (2017). A critical discourse analysis of engineering course syllabi and recommendations for increasing engagement among women in STEM. *Journal of STEM Education: Innovations and Research, 18*(1). https://doi.org/10.1145/2809957.2809960
Schmitt, M. T., Branscombe, N. R., & Postmes, T. (2003). Women's emotional responses to the pervasiveness of gender discrimination. *European Journal of Social Psychology, 33*(3), 297–312. https://doi.org/10.1002/ejsp.147
Schoenthaler, A., Basile, M., West, T. V., & Kalet, A. (2018). It takes two to tango: A dyadic approach to understanding the medication dialogue in patient-provider relationships. *Patient Education and Counseling, 101*(8), 1500–1505. https://doi.org/10.1016/j.pec.2018.02.009
Schuster, C., & Martiny, S. E. (2017). Not feeling good in STEM: Effects of stereotype activation and anticipated affect on women's career aspirations. *Sex Roles, 76*(1–2), 40–55. https://doi.org/10.1007/s11199-016-0665-3
Sekaquaptewa, D., & Thompson, M. (2003). Solo status, stereotype threat, and performance expectancies: Their effects on women's performance. *Journal of Experimental Social Psychology, 39*(1), 68–74. https://doi.org/10.1016/S0022-1031(02)00508-5
Shapiro, J. R., & Williams, A. M. (2012). The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields. *Sex Roles, 66*(3–4), 175–183. https://doi.org/10.1007/s11199-011-0051-0
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420. https://doi.org/10.1037/0033-2909.86.2.420
Sideridis, G. D. (2005). Classroom goal structures and hopelessness as predictors of day-to-day experience at school: Differences between students with and without learning disabilities. *International Journal of Educational Research, 43*(4–5), 308–328. https://doi.org/10.1016/j.ijer.2006.06.008
Simon, R. A., Aulls, M. W., Dedic, H., Hubbard, K., & Hall, N. C. (2015). Exploring student persistence in STEM programs: A motivational model. *Canadian Journal of Education, 38*(1), n1.
Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology, 85*(4), 571. https://doi.org/10.1037/0022-0663.85.4.571
Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*(1), 4–28. https://doi.org/10.1006/jesp.1998.1373
Spurk, D., Kauffeld, S., Barthauer, L., & Heinemann, N. S. (2015). Fostering networking behavior, career planning and optimism, and subjective career success: An intervention study. *Journal of Vocational Behavior, 87*, 134–144. https://doi.org/10.1016/j.jvb.2014.12.007
Sterling, A. D., Thompson, M. E., Wang, S., Kusimo, A., Gilmartin, S., & Sheppard, S. (2020). The confidence gap predicts the gender pay gap among STEM graduates. *Proceedings of the National Academy of Sciences, 117*(48), 30303–30308. https://doi.org/10.1073/pnas.2010269117
Suhlmann, M., Sassenberg, K., Nagengast, B., & Trautwein, U. (2018). Belonging mediates effects of student-university fit on well-being, motivation, and dropout intention. *Social Psychology, 49*(1), 16.
Suryadarma, D., Suryahadi, A., Sumarto, S., & Rogers, F. H. (2006). Improving student performance in public primary schools in developing countries: Evidence from Indonesia. *Education Economics, 14*(4), 401–429. https://doi.org/10.1080/09645290600854110
Tanner, K. D. (2009). Talking to learn: Why biology students should be talking in classrooms and how to make it happen. *CBE—Life Sciences Education, 8*(2), 89–94. https://doi.org/10.1187/cbe.09-03-0021
Tapia, M., & Marsh, G. E. (2004). The relationship of math anxiety and gender. *Academic Exchange Quarterly, 8*(2), 130–134.
Tellhed, U., Bäckström, M., & Björklund, F. (2017). Will I fit in and do well? The importance of social belongingness and self-efficacy for explaining gender differences in interest in STEM and HEED majors. *Sex Roles, 77*(1), 86–96. https://doi.org/10.1007/s11199-016-0694-y
Thomas, A. S., Bonner, S. M., Everson, H. T., & Somers, J. A. (2015). Leveraging the power of peer-led learning: Investigating effects on STEM performance in urban high schools. *Educational Research and Evaluation, 21*(7–8), 537–557. https://doi.org/10.1080/13803611.2016.1158657
Thorson, K. R., Forbes, C. E., Magerman, A. B., & West, T. V. (2019). Under threat but engaged: Stereotype threat leads women to engage with female but not male partners in math. *Contemporary Educational Psychology, 58*, 243–259. https://doi.org/10.1016/j.cedpsych.2019.03.012
Urhahne, D. (2015). Teacher behavior as a mediator of the relationship between teacher judgment and students' motivation and emotion. *Teaching and Teacher Education, 45*, 73–82. https://doi.org/10.1016/j.tate.2014.09.006
Van Veelen, R., Derks, B., & Endedijk, M. D. (2019). Double trouble: How being outnumbered and negatively stereotyped threatens career outcomes of women in STEM. *Frontiers in Psychology, 10*, 150. https://doi.org/10.3389/fpsyg.2019.00150
Vera, E., Shriberg, D., Alves, A., de Oca, Reker, K., … Rau, E. (2016). Evaluating the impact of a summer dropout prevention program for incoming freshmen attending an under-resourced high school. *Preventing School Failure, 60*(2), 161–171. https://doi.org/10.1080/1045988X.2015.1063039
Vittengl, J. R., & Holt, C. S. (2000). Getting acquainted: The relationship of self-disclosure and social attraction to positive affect. *Journal of Social and Personal Relationships, 17*(1), 53–66. https://doi.org/10.1177%2F0265407500171003.
Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin, 140*(4), 1174–1204. https://doi.org/10.1037/a0036620
Wang, M. T., & Fredricks, J. A. (2014). The reciprocal links between school engagement, youth problem behaviors, and school dropout during adolescence. *Child development, 85*(2), 722–737.

Walton, G. M., Logel, C., Peach, J. M., Spencer, S. J., & Zanna, M. P. (2015). Two brief interventions to mitigate a "chilly climate" transform women's experience, relationships, and achievement in engineering. *Journal of Educational Psychology, 107*(2), 468. https://doi.org/10.1037/a0037461

Wegner, C., Strehlke, F., Weber ClaasWegner, P., & Weber, P. (2014). Investigating the differences between girls and boys regarding the factors of frustration, boredom and insecurity they experience during science lessons. *Themes in Science & Technology Education, 7*(1), 35–45.

West, T. V., Dovidio, J. F., & Pearson, A. R. (2014). Accuracy and bias in perceptions of relationship interest for intergroup and intragroup roommates. *Social Psychological and Personality Science, 5*(2), 235–242. https://doi.org/10.1177%2F1948550613490966.

Wilson, D., Jones, D., Bocell, F., Crawford, J., Kim, M. J., Veilleux, N., Floyd-Smith, T., Bates, R., & Plett, M. (2015). Belonging and academic engagement among undergraduate STEM students: A multi-institutional study. *Research in Higher Education, 56*(7), 750–776. https://doi.org/10.1007/s11162-015-9367-x

Xu, Y. J., & Martin, C. L. (2011). Gender differences in STEM disciplines: From the aspects of informal professional networking and faculty career development. *Gender Issues, 28*(3), 134–154. https://doi.org/10.1007/s12147-011-9104-5

Yu, K. (2011). Culture-specific concepts of politeness: Indirectness and politeness in English, Hebrew, and Korean requests. *Intercultural Pragmatics, 8*, 385–409. https://doi.org/10.1515/iprg.2011.018

Zhang, Y. (2013). Does private tutoring improve students' National College Entrance Exam performance?—A case study from Jinan, China. *Economics of Education Review, 32*, 1–28. https://doi.org/10.1016/j.econedurev.2012.09.008