# Reducing STEM Gender Bias with VIDS (Video Interventions for Diversity in STEM)

Corinne A. Moss-Racusin
Skidmore College
Evava S. Pietri
Indiana University-Purdue University Indianapolis
Erin P. Hennes
Purdue University
John F. Dovidio, Victoria L. Brescoll, Gina Roussos
Yale University
Jo Handelsman
University of Wisconsin-Madison

Corresponding Author:
Corinne A. Moss-Racusin
815 North Broadway
Saratoga Springs, NY 12866
cmossrac@skidmore.edu
Phone: (518) 580-8329
Fax: (518) 580-531

REDUCING STEM GENDER BIAS WITH VIDEO INTERVENTIONS

**Abstract**

Gender biases contribute to the underrepresentation of women in STEM. In response, the scientific community has called for methods to reduce bias, but few validated interventions exist. Thus, an interdisciplinary group of researchers and filmmakers partnered to create VIDS (Video Interventions for Diversity in STEM), which are short videos that expose participants to empirical findings from published gender bias research in one of three conditions. One condition illustrated findings using *narratives* (compelling stories), and the second condition presented the same results using *expert interviews* (straightforward facts). A *hybrid* condition included both narrative and expert interview videos. Results of two experiments revealed that relative to controls, VIDS successfully reduced gender bias and increased awareness of gender bias, positive attitudes towards women in STEM, anger, empathy, and intentions to engage in behaviors that promote gender parity in STEM. The narratives were particularly impactful for emotions, while the expert interviews most strongly impacted awareness and attitudes. The hybrid condition reflected the strengths of both the narratives and expert interviews (though effects were sometimes slightly weaker than the other conditions). VIDS produced substantial immediate effects among both men and women in the general population and STEM faculty, and effects largely persisted at follow-up.

Public Significance Statement

This study suggests that Video Interventions for Diversity in STEM (VIDS) successfully reduce gender bias on the part of both men and women in the general public, as well as those in the STEM community. These results are important because gender biases favoring men contribute to women's underrepresentation in STEM fields, and it is thus critical to develop and implement interventions that reduce gender bias and clear the pathways to women's full participation.

Reducing STEM Gender Bias with VIDS (Video Interventions for Diversity in STEM)

"We want to increase the diversity of STEM programs…We get the most out of *all* our nation's talent—and that means reaching out to boys and girls, men and women of all races and backgrounds. Science is for all of us. And we want our classrooms and labs and workplaces and media to reflect that."
President Barack Obama (2015)

Despite ongoing efforts to obtain gender parity, women remain underrepresented across many science, technology, engineering, and mathematics (STEM) fields (American Association of University Women [AAUW], 2010; 2015; National Science Foundation [NSF], 2017). One factor contributing to this lingering disparity is gender bias, which results in preferential evaluation and treatment of men relative to equally qualified women (e.g., Foschi, 1996; Heilman, Wallen, Fuchs, & Tamkins, 2004; Nosek, Banaji, & Greenwald, 2002; Rudman, Moss-Racusin, Glick, & Phelan, 2012). Indeed, despite their rigorous training in scientific objectivity, recent experiments reveal that male and female STEM professionals express the same gender biases favoring men that have been previously observed among many other populations (e.g., Knoblock-Westerwick, Glynn, & Huge, 2013; Milkman, Akinola, & Chug, 2012; 2015; Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012; Steinpreis, Anders, & Ritzke, 1999). Evidence-based interventions are thus needed to raise awareness of and ultimately reduce gender biases in STEM, so that the most talented individuals—regardless of demographic background—can contribute to scientific

progress.

Indeed, many in the STEM community have called for the wide implementation of validated diversity interventions in response to prominent instances of gender bias (Al-Gazali et al., 2013). For example, in an open letter criticizing gender-biased content published in *Science*, more than 600 co-signers from the STEM community wrote, "We suggest STEM diversity training for *Science* editorial staff" (Ferguson & Ghorayshi, 2015). Similarly, a 2015 *Nature* editorial condemning sexism in science called for "Gender-bias training for scientists" (Nature, 2015). Unfortunately, the common perception that validated diversity trainings are readily available for implementation is largely inaccurate. In reality, subtle biases are often extremely resistant to change (Lai et al., 2014), and very few successful evidence-based interventions exist (Moss-Racusin, van der Toorn, Dovidio, Brescoll, Graham, & Handelsman, 2014). The current research sought to address this gap by testing the efficacy of VIDS (Video Interventions for Diversity in STEM), novel evidence-based interventions designed by an interdisciplinary group of researchers and artists to ameliorate pernicious STEM gender biases.

**The Need for STEM Gender Bias Interventions**

Results of both correlational and experimental research suggest that gender bias remains problematic within STEM fields. Early evidence reflecting this idea includes an analysis of peer review scores for Swedish postdoctoral fellowship applications through

the late 1990s, which revealed that female applicants needed three additional papers in top outlets (e.g., *Nature*, *Science*) or 20 additional papers in strong specialized journals in order to receive the same competence rating as similar male colleagues (Wenneras & Wold, 1997). These biases have persisted over time, with more recent work indicating that elite male faculty in the biological sciences are less likely to hire female postdoctoral researchers and graduate students in their laboratories (Sheltzer & Smith, 2014), and articles with female first authors are less likely to be cited than those with male first authors (Lariviere, Ni, Gingras, Cronin, & Sugimoto, 2013). Moreover, women in STEM report encountering frequent and severe instances of gender bias (e.g., Robnett, 2016; Steele, James, & Barnett, 2002; Williams, Phillips, & Hall, 2016).

Building upon this type of correlational evidence, experiments have compared responses to identical qualifications randomly attributed to either a male or female target, thereby isolating the causal impact of target gender. These experiments consistently reveal a preference for men in STEM, both on the part of general population and STEM faculty participants. For example, STEM faculty were more likely to judge as competent, hire, mentor, and pay an equitable salary to a male student lab manager applicant than an identically qualified female (Moss-Racusin et al., 2012). Similarly, a male candidate was more likely to be hired for a mathematics position relative to an identically qualified female candidate, even when objective performance information was provided (Reuben, Sapienza, & Zingales, 2014). STEM faculty were more likely to agree to meet with and mentor a male prospective doctoral student relative to the identical female student (Milkman et al., 2012; 2015). Psychologists were more likely to hire a male applicant (relative to the identical female applicant) for a faculty position (Steinpreis et al., 1999).

Finally, graduate students judged a conference abstract more negatively when it was attributed to a female (vs. a male) author (Knoblock-Westerwick et al., 2013).

It is important to note that there were no participant gender differences across these experiments, suggesting that women are just as likely as men to display pernicious gender biases. This reflects the fact that gender biases are often unintentional, and likely stem from equal exposure to pervasive cultural gender stereotypes (Rudman & Phelan, 2008). However, problems associated with gender bias may be exacerbated by the issue of gendered skepticism of gender bias research, such that men tend to be more resistant to experimental evidence of gender bias than are women. For example, after reading a news article describing published experimental evidence of gender bias in STEM (i.e., the results of Moss-Racusin et al., 2012), men were more likely than women to post negative online comments (e.g., denying the evidence, justifying the existence of gender bias, and personally criticizing the researchers). Further, in a laboratory experiment, men evaluated research revealing gender bias more poorly than did women (Handley, Brown, Moss-Racusin, & Smith, 2015). Thus, although men and women may be equally likely to exhibit gender bias, men may be more resistant than women to evidence of its existence. Thus, it is critical to ensure that any STEM gender bias intervention is effective for both male and female participants.

Although the vast majority of experimental research reveals evidence of gender bias targeting women, one recent paper suggests that there may be situations in which gender bias favors women (Williams & Ceci, 2015). Results from this experiment revealed a preference for hiring female STEM faculty candidate relative to similarly qualified male candidates. At face value, these results are heartening for those committed

to gender equity in STEM faculty hiring. However, elements of the methodology of these experiments may limit their generalizability. For example, this pattern emerged when women applicants were clearly and uniformly excellent, a situation that is rare in real assessment situations and that is well-known to suppress the expression of biases that leak out in more realistic, nuanced situations (Heilman et al., 2004; Moss-Racusin et al., 2012). Additionally, the study was limited to the faculty hiring stage, and the existing experimental evidence suggests that bias also functions in multiple other contexts that could powerfully undermine women's progress.

Thus, while future research is certainly needed to identify the scope and boundary conditions of gender bias within STEM, the confluence of correlational and experimental evidence suggests that it remains an unfortunate problem that undermines meritocracy and access to talented professionals. Of importance, the presence (e.g., hiring) and success of some women in STEM *does not demonstrate the absence of bias against them*. Rather, these women may persist *in spite of* gender biases that undermine their productivity and contribute to their differential attrition (and serious underrepresentation) in many STEM fields (AAUW, 2010; 2015; NSF, 2017). Thus, evidence highlights the need for interventions that can effectively target gender bias in STEM.

**Existing Interventions**

Despite repeated calls for the implementation of gender bias trainings (e.g., Al-Gazali et al., 2013; Ferguson & Ghorayshi, 2015; Nature, 2015), very few empirically supported interventions actually exist (Moss-Racusin et al., 2014). Many existing interventions were developed without a clear theoretical rationale for why they ought to

work, and most are not evaluated empirically to determine whether they are effective (Paluck, 2006; Paluck & Green, 2009). Interventions that specifically target gender bias within STEM have been particularly understudied: a systematic review of the existing literature uncovered zero double-blind randomized controlled trials (RCTs) of STEM gender bias interventions (Moss-Racusin et al., 2014). Of concern, interventions that have been assessed demonstrate dramatically mixed levels of success, with some ironically worsening bias rather than improving it (Dobbin, Schrage, & Kalev, 2015). Interventions that —intentionally or unintentionally—imply that participants are at fault appear to be particularly ineffective. For example, one intervention that stressed societal requirements not to express prejudice ironically resulted in higher levels of prejudice than did a non-intervention control (Legault, Gutsell, & Inzlicht, 2011). Thus, further work is needed to develop interventions that can effectively reduce gender bias.

However, some existing research suggests that carefully designed interventions can effectively raise awareness of and reduce gender bias. For example, the Workshop Activity for Gender Equity Simulation (WAGES) program consists of a board game that allows participants to experience the cumulative effects of gender bias (Shields, Zawadzki, & Johnson, 2011). WAGES has been shown to increase knowledge and retention of gender bias issues (Shields et al., 2011), reduce feelings of reactance and promote self-efficacy

(Zawadzki, Danube, & Shields, 2012), reduce the endorsement of sexist beliefs (Zawadzki, Shields, Danube, & Swim, 2014), and increase the recognition of sexism as harmful (Cundiff, Zawadzki, Danube, & Shields, 2014). Similarly, a "Scientific Diversity" course facilitating exposure to published empirical data on gender bias was associated with academic scientists' increased awareness of gender bias, reduced expression of bias, and enhanced willingness to engage in actions that combat gender bias (Moss-Racusin, van der Toorn, Dovidio, Brescoll, Graham, & Handelsman, 2016).

Two theory-driven interventions evaluated in single-blind experiments with STEM participants showed particularly promising results. One educational intervention designed to break the "habit" of gender bias increased faculty's gender bias literacy (or awareness, knowledge, and intentions to confront gender bias; Carnes et al., 2012), as well as their internal motivation and self-efficacy to address gender bias (Carnes et al., 2015). Because these data were gathered from participants at only one university, the extent to which results might generalize to other contexts remains unclear. Another educational intervention, implemented at four U.S. universities, utilized an evidence-based, non-confrontational pedagogical approach to teaching faculty about gender bias (Jackson, Hillard, & Schneider, 2014). Results indicated that the intervention successfully reduced implicit gender bias among male (but not female) STEM faculty.

These results are promising, in that they suggest that rigorous interventions can effectively target STEM gender bias. However, they require participants to attend lengthy in-person meetings. Further, they necessitate the systematic, time-and-resource-consuming training of facilitators in order to deliver a consistent product. As a result, they may not be readily implemented on a wide scale (particularly among STEM communities, in which time and resources are frequently at a premium). Thus, while this existing research demonstrates that STEM communities at some institutions can respond to well-designed in-person interventions, the development and experimental testing of flexible, scalable interventions is necessary in order to more effectively target STEM gender bias on a large scale.

**Video Interventions for Diversity in STEM (VIDS)**

To address this need, we formed an interdisciplinary team of researchers and artists to develop Video Interventions for Diversity in STEM (VIDS), a novel evidence-based STEM gender bias intervention. Academic psychologists and biologists partnered with a professional playwright, actors, and filmmakers to create a series of high-quality videos portraying the results of published gender bias research (all videos are freely available on the VIDS website, https://academics.skidmore.edu/blogs/vids/). Thus, in contrast to the interventions described above, VIDS can be easily and affordably disseminated across institutions, is brief and convenient for

participants to complete at their own computers, and does not fluctuate as a function of the quality or conduct of individual instructors or other course attendees. Moreover, the design enabled the first (to our knowledge) double-blind RCT testing a STEM gender bias intervention.

VIDS consists of two sets of six five-minute videos that each describe the results of one of six high-quality, peer-reviewed, published studies on gender bias. Both sets of videos presented the same underlying evidence of gender bias, but did so via different communication styles. Because a single communication style might not target all audiences effectively (Paluck & Green, 2009), we tested the efficacy of video interventions utilizing two persuasive communication styles drawn from the existing literature. The first communication style relies on entertaining and emotionally engaging narratives, which present information via stories with an identifiable beginning, middle, and end in which compelling characters grapple with and attempt to resolve conflicts (Green & Brock, 2000). Instead of explicitly presenting a straightforward set of facts, narratives rely on a more engrossing, indirect route to persuasion (Green & Brock, 2000; Johnson, Jasper, Griffin, & Huffman, 2013). Specifically, the six *narrative condition* videos, written by a professional playwright, illustrated published empirical evidence of gender bias through scripted, television-style stories that were immersing and engaging (see Method section for a more detailed description of the videos).

In contrast, the second communication style utilized expert interviews to deliver a direct persuasive argument relying on a strong set of facts presented by a credible source (Green, Garst, Brock, & Chung, 2006; Slater & Rouner, 1996). An expert interview involves the introduction and support of a clearly identified argument using concrete evidence. The six videos in the *expert interview* condition communicated the same empirical evidence as the narrative condition, but utilized an interesting interview format (Petty & Cacioppo, 1986). Our objective was to provide multiple versions of an efficacious video intervention that could be further tested with different populations, in order to expand a theoretical understanding of communication styles and bias reduction, as well as to develop readily accessible materials and equip practitioners with a deep diversity intervention toolkit.

We chose to use visual media because it has been shown to improve attitudes towards stigmatized groups (Mazziotta, Mummendey, & Wright, 2011; Ortiz & Harwood, 2007), improve behavior towards outgroup members (Paluck, 2009), and enable engaged active learning (Handelsman, Miller, & Pfund, 2007; for a detailed discussion on the use of visual media to address bias, see Pietri et al., 2017). However, the majority of existing media intervention studies have examined reactions to racial outgroups. Of importance, gender relations (and thus, gender bias) differ from relationships between other social groups (and thus, other forms of bias) in critical ways. Many people do not come into frequent contact with members of most stigmatized groups. In contrast, men and women interact frequently and, in the case

of heterosexual romantic relationships, intimately (Glick & Fiske, 1996; 1999). These close relationships (and accompanying generally positive attitudes towards women) co-exist paradoxically with the subjugation of women and negative stereotypes regarding women's competence. This leads many people to endorse "benevolently sexist" ideologies that undermine women's autonomy with seemingly positive, paternalistic praise (Glick & Fiske, 2001). Additionally, women and men often demonstrate the same biases about women (Rudman & Phelan, 2008). Thus, it is critical to move beyond the existing work on other stigmatized groups by examining the ways in which media interventions may effectively target gender bias in STEM.

Despite these important differences, we note that a number of interventions have been developed to reduce racial bias, and were informed by the literature evaluating these interventions when designing VIDS. For example, interventions that present counter-stereotypic exemplars and offer strategies to combat bias were effective in reducing implicit racial bias, while those that facilitated perspective taking, encouraged the adoption of egalitarian values, or induced positive emotions were ineffective (Lai et al., 2014). Consistent with this, one habit-breaking intervention that emphasized stereotype replacement, counter-stereotypic imaging, individuation, and ways to increase opportunities for contact with outgroup members produced long-term reductions in undergraduate's implicit racial biases (although this intervention also encouraged perspective-taking; Devine, Forscher, Austin, & Cox, 2012).

Relatedly, interventions derived from Intergroup Contact Theory (Allport, 1954) that facilitate positive contact between outgroup members have long been shown to reduce explicit racial bias towards people of color (e.g., Pettigrew & Tropp, 2006). As noted above, intergroup contact may be effective even when it occurs vicariously (e.g., via exposure to visual media portrayals of stigmatized group members; Mazziotta et al., 2011; Ortiz & Harwood, 2007). Thus, while remaining cognizant of the important differences between racial and gender biases, we built upon the existing literature by integrating several elements of successful racial bias interventions (e.g., repeated exposure to counter-stereotypic exemplars/images, stereotype replacement, individuation, and vicarious contact with stigmatized group members—i.e., women in STEM) in VIDS. Further, we were careful to avoid unsuccessful practices, such as explicitly requiring perspective taking, the adoption of egalitarian values, or attempts to induce positive emotions.

In addition to drawing upon this existing literature on racial bias interventions when developing VIDS, we also followed an evidence-based framework (drawn from a systematic review of existing diversity interventions) that identified four elements shared by successful interventions (Moss-Racusin et al., 2014). VIDS adhered to each of the four elements, in that it: (1) was grounded in existing theory and evidence, (2) employed an engrossing approach designed to stimulate

active learning, (3) avoided assigning blame for current diversity challenges, and (4) was assessed experimentally and longitudinally to determine effectiveness with several participant groups. Indeed, a notable strength of VIDS is that it is evidence-based in multiple ways. First, in conforming to this framework for developing successful interventions (Moss-Racusin et al., 2014), we employed tested bias-reduction strategies (Lai et al., 2014; Paluck, 2009), avoided blaming participants for diversity challenges (Dobbin et al., 2015; Legault et al., 2011), and utilized communication styles drawn from the research literature (Green & Brock, 2000; Petty & Cacioppo, 1986). In addition, VIDS directly communicates empirical evidence of gender bias, and thus has the additional benefit of disseminating the results of published psychological research to broader audiences.

Preliminary research provided a detailed presentation of the theoretical underpinning of VIDS, as well as a full description of the content, development, and pilot-testing of the videos (Pietri et al., 2017). Of importance, this preliminary research suggested that the narrative and expert interview videos each successfully utilized their intended communication style. Further, relative to controls, both VIDS conditions increased general population adult participants' bias literacy, as characterized by (1) awareness of bias, (2) knowledge of gender inequity, (3) feelings of efficacy at being able to notice bias, and (4) recognition and confrontation of bias across situations (Pietri et

al., 2017). Thus, initial evidence suggests that VIDS may be a promising tool for raising bias literacy among adults from the general population. However, additional systematic testing is needed to determine whether VIDS reduces gender bias itself, impacts other outcomes critical to promoting gender parity in STEM, functions similarly with the relevant population of STEM professionals, and is effective over time.

**The Current Research**

Across two experiments, the primary goal of the current research was to provide an initial test of VIDS' ability to reduce viewers' gender biases and impact other critical outcome variables (see below). To do so, we expanded upon preliminary research (Pietri et al., 2017) in five key ways. First, the prior research only compared the narrative and expert interview videos to the *video control condition* used in the current research, in which participants viewed science documentary clips that were matched to the intervention videos on key characteristics but contained no bias-related information. For the purposes of the current research, we developed two novel experimental conditions. A new *hybrid condition* was created because we expected that both versions of the intervention would offer different strengths. In the hybrid condition, participants viewed three narrative videos and three corresponding expert interview videos (i.e., the expert interview that covered the same research evidence on gender

bias). In the new *non-intervention control condition*, participants completed the same outcome measures (described below) included in the other conditions, but did not view videos. Including both a video control condition and a non-intervention control condition allowed us to isolate any impact associated with mere exposure to women scientists (in the video control condition), as well as the mere passage of time (in the non-intervention control condition). Second, participants in the prior research were each only exposed to one video (narrative, expert interview, or control). In order to more fully explore the potential impact of VIDS, we exposed participants to six videos (for thirty total minutes of viewing time).

Third, prior research only recruited adults from the general population. Although positive effects among the general population are impactful and heartening, it is critical to determine the extent to which results generalize to STEM professionals (i.e., the people most capable of either interrupting or perpetuating STEM gender biases). Thus, while Experiment 1 recruited a diverse group of general population adults via Amazon Mechanical Turk, Experiment 2 recruited professionals from a variety of STEM disciplines. Fourth, preliminary research was cross-sectional. As a result, the extent to which VIDS' effects might endure over time was unclear. Thus, Experiment 2 included three measurement time points: baseline (one week before the intervention), post-test (immediately following the intervention), and follow-up (one

week after the intervention). This design allowed us to directly assess both between-person effects of condition and within-person change over time, as well as whether our intervention had similar effects on participants regardless of their initial levels of bias.

Fifth, the current research greatly expanded the outcomes assessed. Preliminary research focused on VIDS' impact on bias literacy, an important predictor of bias itself and related outcomes (Carnes et al., 2012; Zawadzki et al., 2012). However, VIDS' impact on other important outcomes remained unexamined. Thus, in Experiment 1, we first sought to replicate prior research by measuring awareness of bias, one key indicator of bias literacy (Pietri et al., 2017). More importantly, we added numerous additional outcomes, reflecting a range of constructs critical to advancing gender parity in STEM. Specifically, existing experimental evidence suggests that gender bias is often related to negative attitudes towards women in STEM, particularly assumptions of incompetence (Moss-Racusin et al., 2012; Reuben et al., 2014). Thus, we assessed VIDS' impact on participants' attitudes towards women in STEM. Of importance, we also explored VIDS' ability to reduce participants' gender bias itself. To assess this critical variable, we utilized the modern sexism scale, a well-validated instrument frequently employed to measure subtle, contemporary forms of bias against women (Moss-Racusin et al., 2012; Swim, Aikin, Hall, & Hunter, 1995).

Additionally, we explored the extent to which VIDS might engage participants' action-oriented emotions. Specifically, we were interested in effects on anger and empathy, emotions that have been particularly linked to action in past research (Batson, Chang, Orr, & Rowland, 2002; Carver & Harmon-Jones, 2009). For example, experiencing empathy towards a target increased altruistic helping behavior directed towards the target's group (Batson, et al., 2002). Similarly, anger directed towards an outgroup predicted willingness to take action against unjust behaviors enacted by that group against another (Gordijn, Yzerbyt, Wigboldus, & Dumont, 2006). Thus, we were interested in determining whether VIDS might promote these action-oriented emotional states among viewers.

Finally, in Experiment 2, we sought to move beyond attitude and emotion outcomes by measuring relevant behavioral intentions[1]. Of importance, a large body of research has identified the importance of measuring people's behavioral intentions (Ajzen, 1991) as the critical link between attitudes and behaviors (e.g., Godin & Kok, 1995; Kollmuss & Agymen, 2002). Indeed, because behavioral intentions reflect the motivational factors that shape behaviors, they often serve as particularly reliable predictors of subsequent behavior (Ajzen, 1985;

---

1 Although measuring objective behaviors relevant to STEM gender bias would be a particularly compelling outcome, strong concerns about participant recruitment, our tight data collection timeline (as described in Experiment 2 below), and the importance of ensuring faculty participants' anonymity did not permit us to measure bias-related behaviors directly.

1987; 1991). For example, intentions to ride the bus to campus (rather than drive a car) predicted students' actual bus-riding behavior, while their prior bus-riding behavior did not (Bamberg, Ajzen, & Schmidt, 2003). Thus, consistent with several other studies investigating STEM gender bias interventions (e.g., Carnes et al., 2015; Cundiff et al., 2014; Moss-Racusin et al., 2016), we assessed participants' intentions to engage in behaviors that promote gender parity in STEM.

In sum, Experiment 1 provided the first test of the extent to which viewing six narrative or expert interview VIDS (or three of each, in the hybrid condition) increased general population adults' awareness of gender bias, improved attitudes toward women in STEM, reduced gender bias, and engaged action-oriented emotions (relative to video and non-intervention controls). Experiment 2 recruited STEM faculty participants who were scheduled to attend a workshop on STEM education. We added a direct measure of participants' propensity to engage in behaviors that increase gender parity. We also added baseline and follow-up measurements to assess VIDS' effectiveness over time.

Our primary hypothesis was that VIDS would result in superior outcomes relative to the video control condition. We did not make a-priori predictions regarding which set of videos (i.e., which communication style) would be most effective. Because both narratives (Green & Brock, 2000) and expert interviews (Petty &

Cacioppo, 1986) have demonstrated effectiveness in the previous literature and because we were the first to test a STEM gender diversity intervention using these communication styles, there was insufficient evidence to warrant a clear prediction that one communication style should be superior in this context. Instead, we sought to develop two sets of evidence-based videos that could serve as effective STEM diversity interventions for flexible, strategic use with different audiences and contexts. However, the experimental design enabled us to explore whether the VIDS conditions were differentially effective, and whether these patterns varied for different audiences (e.g., adults from the general population vs. STEM faculty). For example, although narratives may appeal to a general audience, we investigated whether STEM faculty—who are trained to rigorously analyze empirical results— might be more influenced by the expert interview relative to narrative videos. Further, the longitudinal design of Experiment 2 allowed us to perform a first test of the persistence of VIDS' effects over time.

## Experiment 1

The primary aim of Experiment 1 was to provide an initial test of VIDS' impact on gender bias, attitudes towards women in STEM, and action-oriented emotions. Moreover, in contrast to prior work (Pietri et al., 2017), we included hybrid and non-intervention control conditions, and sought to determine whether viewing six five-minute videos (for a total of 30 minutes of exposure) may result in increasingly robust effects.

**Method**

**Power assessment.** We conservatively estimated medium effect sizes and sought to recruit at least 75 participants in each of the 5 experimental conditions (significantly exceeding recommendations; e.g., Simmons, Nelson, & Simonsohn, 2011), for a target of 375 participants. Our final sample size ($N = 450$) surpassed this target.

**Participants and recruitment.** Participants completed the experiment in exchange for $3.00 payment on Amazon's Mechanical Turk website. Mechanical Turk is a crowdsourcing website that provides listings of "Human Intelligence Tasks" (or HITs) available for participants to complete in exchange for financial compensation (Mason & Suri, 2012). Participants recruited through Mechanical Turk have the benefit of being more diverse than traditional university samples, particularly in ethnic background, age, and socioeconomic status (Behrend, Sharek, Meade, & Wiebe, 2011; Buhrmester, Kwang, & Gosling, 2011; Paolacci, Chandler, & Ipeirotis, 2010). The experiment was advertised as a study on "Movies and Memory," and participants were told that they would watch six short videos, complete a test of their ability to recall details from the videos, and fill out surveys related to their impressions of the videos.

We originally recruited 501 participants. However, thirty-six (7.2%) participants incorrectly answered more than three (of 18) easy attention check questions (see Materials below) about video content, suggesting that they were not carefully attending to the videos. They were thus excluded from analyses, along with fifteen participants (3%) who completed the experiment twice as a result of computer error. This

resulted in a final sample of 450 participants (54% women). Of those, 74% were White, 10% were Black, 7% were Hispanic, 5% were Asian, and 5% identified as another race. Participants ranged in age from 18 to 68 ($M$ = 33.99, $SD$ = 11.20). The plurality of participants had a college degree (13% High School/GED, 29% some college, 11% 2-year college degree, 37% 4-year college degree, 9% Master Degree, and less than 1% each for Doctorate and Professional degree). On a scale of 1 (*strongly liberal*) to 7 (*strongly conservative*), participants' average political orientation was 3.38 ($SD$ = 1.70).

**Materials.**

*Intervention videos.* As described in the introduction above, the different versions of VIDS were derived from existing theory and evidence from research on attitude change, persuasion, the efficacy of diversity interventions, and media communications (e.g., Green & Brock, 2000; Legault et al., 2011; Moss-Racusin et al, 2014; Paluck, 2009; Paluck & Green, 2009; Petty & Cacioppo, 1986). VIDS was developed by an interdisciplinary team of psychologists and biological scientists, in partnership with a professional playwright and filmmaking crew (see Pietri et al., 2017 for a detailed description of the theoretical foundation, development, pre-testing, and validation of VIDS).

Participants in the current experiment were randomly assigned to one of five conditions: narrative, expert interview, hybrid, video control, or non-intervention control. There were six videos (each approximately five minutes long) in each video condition. In the

*narrative condition,* participants learned about the published research *via* engaging stories portraying interactions between women and men graduate students and faculty in a science department. A professional playwright wrote the script for these videos, in close consultation with the authors and two expert biological sciences consultants. Prior to writing the script, the playwright carefully read and discussed the empirical gender bias papers with the authors, and also conducted structured interviews with female and male graduate students, postdoctoral associates, and faculty in STEM to learn about their everyday experiences. These experiences were utilized to generate ideas for specific stories that would illustrate the underlying empirical evidence of bias drawn from each published paper.

For example, one published paper (Rudman & Glick, 1999) revealed evidence of backlash, or social and economic penalties targeting women who violate gender stereotypes even when these violations are necessary for professional success. To illustrate these empirical results, we utilized an emblematic real-world experience shared by one postdoctoral researcher. After behaving in a stereotypically feminine, self-effacing way during a practice conference talk, she was told by her advisor that she needed to be "more confident" in order to be viewed as competent. However, when she subsequently violated female gender stereotypes by behaving more confidently during the next practice talk, she was chastised by faculty for being "too aggressive." Of importance, her male graduate student co-presenter's similarly confident style was in keeping with male gender stereotypes, and he was highly praised by faculty. Thus, the narrative video portraying the results of Rudman and Glick (1999) follows this storyline. In this way, we utilized real-

world experiences drawn from the lives of STEM professionals to illustrate the published

empirical evidence of gender bias.

The *expert interview* condition contained videos that portrayed the results of the

same six gender bias papers using an interesting interview with an expert "psychology

professor." Professional actors played the roles of the professor and interviewer in order

to ensure that performance quality and presentation fluency remained consistent across

experimental conditions. In collaboration with the playwright who wrote the narrative

scenes, we scripted the expert interview scenes to contain a straightforward presentation

of information from the six selected papers. For instance, in the expert interview video

about Rudman and Glick's (1999) research corresponding to the narrative described

above, the psychology professor described the articles' findings in an evening news-type

interview format. The professional film crew also filmed these videos to ensure that the

production quality did not differ across the narrative and expert interview conditions.

The *hybrid* condition consisted of three narrative videos and the

three corresponding expert interviews (i.e., the expert interview that

covered the same gender bias research evidence). We

counterbalanced whether participants saw the narrative or expert

interviews first, and which three narratives and expert interviews

participants watched. There were no effects associated with

counterbalancing (all $p$s > 0.24).

The *video control* condition was comprised of six control videos

of existing documentaries on research in the basic sciences. These

videos were matched to the intervention conditions, in that they were

rated as equally entertaining as the narratives and equally informative as the expert interviews and portrayed similar numbers of male and female scientists (Pietri et al., 2017). However, the control videos contained no mention of bias-related information. In the *non-intervention control* condition, participants completed all measures but did not view any videos. The two control conditions allowed us to assess the efficacy of the interventions above and beyond any effects associated with watching *any* science-related video and mere exposure to women scientists, and/or the mere passage of time (respectively).

*Attention checks.* For participants in the three VIDS conditions, we asked 18 easy questions about information clearly addressed in the videos (e.g., a question for one of the expert interview videos was "According to the professor, what does data suggest about men and women's performance in math and science?"). Participants were presented with four potential answers and had to choose the correct one (e.g., "The data is inconclusive," "There is a very small difference between men and women's performance," "Women perform much better than men," or "Men perform much better than women"). These were the same attention check questions employed in Pietri et al. (2017).

*Manipulation checks.* We sought to replicate initial results (Pietri et al., 2017) indicating that the narrative and expert interview formats were differentially perceived by participants as intended. Specifically, prior work suggests that narratives should result in participants' feeling more transported (i.e., immersed and engaged by a story; Green & Brock, 2000), whereas the expert interviews' strong arguments should stimulate

participants' logical thinking and ability to think critically about the message (Petty & Caccioppo, 1986). Thus, consistent with Pietri et al. (2017), we measured transportation using five items assessing the extent to which participants felt immersed and emotionally involved when watching the videos (e.g., "I was mentally involved in the video while I was watching it," "The videos affected me emotionally;" Green & Brock, 2000). Responses were provided on a scale of 1 (*strongly disagree*) to 5 (*strongly agree*) and were averaged to form the transportation index, with higher numbers reflecting greater amounts of psychological transportation ($M = 3.50$, $SD = 0.72$, $\alpha = .53$).

Also consistent with Pietri et al. (2017), we measured logical thinking using two items assessing the extent to which participants were engaged in logical thinking ("This movie made its point clearly with evidence and logic," "This movie presented its evidence and facts in a clear and logical manner"). Responses were provide on a scale of 1 (*strongly disagree*) to 5 (*strongly agree*) and were averaged to form the logical thinking index, with higher scores indicating more logical thinking ($M = 4.30$, $SD = 0.69$, $r(279) = .85$, $p < .001$).

***Awareness of gender bias in STEM.*** To assess awareness of STEM gender bias, we again utilized a scale from previous research (Pietri et al., 2017). Supporting the utility of this scale, it has been found to correlate with relevant measures of gender bias literacy, including knowledge of gender inequality, awareness of male privilege, and ability to detect gender bias in new situations (Pietri et al., 2017). Participants responded to eight items (on a scale of 1 = *strongly disagree* to 5 = *strongly agree*), including, "In my opinion, women in STEM often are not taken as seriously as their male colleagues," and "In my opinion, people who work in STEM often do not want to hire women because

they worry that the women might become pregnant and be unable to do their job adequately." Items were averaged to create the awareness of bias index, with higher scores indicating more awareness of bias. Descriptive and reliability statistics for all scales (for both Experiments) are presented in Table 1. Bivariate correlations between all variables (for both Experiments) are presented in Table 2.

*Attitudes toward women in STEM*. Using a previously validated scale (Stake, 2003), participants responded to six items assessing their attitudes towards women in STEM and their capabilities (e.g., "Women can make important scientific discoveries," "Women have the innate ability to be as good in STEM careers as men"). Responses were provided on a scale of 1 (*strongly disagree*) to 5 (*strongly agree*). Items were averaged to form the attitudes toward women in STEM index, with higher scores indicating more positive attitudes.

*Gender bias.* Gender bias was assessed using the modern sexism scale, a well-validated instrument frequently employed to measure subtle, contemporary forms of bias against women (Swim et al., 1995). Participants completed the modern sexism scale by rating their agreement (1 = *strongly disagree*, 5 = *strongly agree*) with eight statements assessing their levels of subtle gender bias (e.g., "Discrimination against women is no longer a problem in the United States," "It is rare to see women treated in a sexist manner on television"). Items were averaged to create the modern sexism scale, with higher scores indicating higher levels of gender bias.

*Empathy.* Using an existing scale (Batson, Early, & Salvarani, 1997), participants rated the degree to which they felt compassionate, sympathetic, and concerned (1 = *not at*

*all* to 5 = *to a very high degree*). Items were averaged to form the empathy index, with greater numbers reflecting higher levels of empathy[2].

**Anger.** Using an existing scale (Okimoto & Brescoll, 2010), participants rated the degree to which they felt anger, disgust, and outrage (1 = *not at all* to 5 = *a very high degree*). Items were averaged to comprise the anger index, with higher scores indicating greater amounts of anger.

**Procedure.** The experiments were administered online (utilizing participants' own computers) via Mechanical Turk and the online survey program Qualtrics. Participants first indicated their informed consent, and were then randomly assigned to experimental condition. Each condition began with a brief instructional video that administered the detailed cover story. Specifically, an actress informed participants that we were interested in how they reacted to and remembered a variety of videos. She provided a list of potential topic videos that included both the real video topics that participants could be randomly assigned to view ("gender bias and discrimination" in the experimental conditions, and "general issues in basic science" in the video control condition) and a list of distractor topics that were never actually assigned but were included to disguise the true purpose of the experiment ("modern American history," "racial tensions in the modern era," "the influence of reality television," "general misconceptions about common diseases"). She then told participants

_____

2 Because the video content differed across conditions, we simply asked participants to report the extent to which they felt each emotion and did not specify or constrain the emotional target. For example, had we asked participants to report their anger/empathy in response to thinking about gender bias, this would have made sense to participants in the VIDS conditions but not in either control condition.

that "you will now be randomly assigned to your topic condition," and a black screen appeared indicating the participant's condition.

Participants in the narrative condition were then informed that they would watch videos showcasing interactions among people in the sciences. They were assured that every scene was based on the results of psychological research, as well as real-world experiences of STEM professionals. Participants in the expert interview condition were told that they would view interviews with a psychology professor who was an expert in gender bias research, and who would discuss the results of psychological research. Participants in the hybrid condition were given the information about both the narratives and expert interviews.

Participants in the video control condition were told that they would watch clips discussing basic science research, and that all research featured in the videos was based on scientific studies. Participants in the non-intervention control condition were told that they would watch videos about basic science research but would first complete a variety of measures. In actuality, they completed all measures but did not view any videos.

Next, all participants (except those in the non-intervention control condition) watched six films corresponding to their experimental condition. They then completed the scales measuring the dependent variables, attention and manipulation checks, and demographics. All scales were presented in a random order, and items

within each measure were also randomized. Finally, participants were fully debriefed and compensated.

## Results

**Manipulation checks.** We began by examining the impact of the experimental videos on the manipulation checks (transportation and logical thinking, respectively). As expected, there was a significant effect of condition on transportation, $F(3,273) = 12.09$, $MSE = 0.46$, $p < .001$, $\eta^2 = .12$. In particular, the narrative videos ($M = 3.79$, $SE = 0.07$) were perceived as significantly more transporting than were the expert interview videos ($M = 3.32$, $SE = 0.09$), $t(273) = 4.05$, $p < .001$. There was also a significant effect of condition on logical thinking, $F(3,277) = 9.58$, $MSE = 0.44$, $p < .001$, $\eta^2 = .09$. As expected, the expert interview videos ($M = 4.33$, $SE = 0.09$) stimulated significantly more logical thinking than did the narrative videos ($M = 4.03$, $SE = 0.07$), $t(277) = -2.58$, $p = .011$. These results are consisted with preliminary findings reported elsewhere (Pietri et al., 2017), and further bolster the success of the VIDS manipulation.

**Primary analyses.** In our primary analyses, we assessed the efficacy of VIDS by testing whether the five outcome variables differed by experimental condition using a one-way analysis of variance. These effects are illustrated in Figure 1. Because of the large number of possible comparisons between the five conditions, we describe a subset of the most theoretically important comparisons in the main text. We first note whether the non-intervention control differed from the video control condition. We then note whether the VIDS conditions differed significantly from the non-intervention control, and then (most importantly) whether the VIDS conditions differed from the video control as

hypothesized. Finally, we report whether the VIDS conditions differed significantly from each other. All means and standard errors for Experiment 1, along with the statistical significance of all possible pairwise comparisons, are presented in Table 3. Effect sizes (*d*) were calculated comparing VIDS to the video control condition using the omnibus standard deviation (all effect sizes for both Experiments are presented in Table 4).

*Awareness of bias.* As expected, there was a significant effect of condition on awareness of bias, $F(4,440) = 12.60$, $MSE = 0.51$, $p < .001$, $\eta^2 = .10$. Participants in the non-intervention control condition did not significantly differ from those in the video control condition. Participants in each of the three VIDS conditions reported significantly greater awareness of bias than did participants in the non-intervention control condition. Replicating previous effects (Pietri et al., 2017) and supporting our hypothesis, pairwise comparisons revealed that participants in the narrative ($d = 0.65$), hybrid ($d = 0.75$), and expert ($d = 0.82$) conditions each reported significantly more awareness of bias against women in science than did participants in the video control condition. The VIDS conditions did not differ significantly from each other, suggesting that the VIDS conditions were each successful in improving awareness of bias.

*Attitudes towards women in STEM.* As expected, there was a significant effect of condition on attitudes towards women in STEM, $F(4,440) = 2.68$, $MSE = 0.34$, $p = .031$, $\eta^2 = .02$. Participants in the video control condition reported marginally more positive attitudes than did those in the non-intervention control condition. Participants in the VIDS conditions all expressed significantly more positive attitudes toward women in STEM than did participants in the non-intervention control condition. Contrary to expectations, pairwise comparisons did not reveal significant differences between any of

the intervention conditions and the video control condition (narrative $d = 0.03$; hybrid $d = 0.15$; expert $d = 0.15$), or between VIDS conditions. However, because each of the VIDS conditions did significantly differ from the non-intervention control (as noted above), it appears that viewing the female scientists in the video control may have also encouraged more positive attitudes towards women in STEM. Moreover, the grand mean on this scale was 4.41 out of 5.00, with a standard deviation of 0.58, indicating evidence of a ceiling effect. In other words, effects concerning this outcome may have been obscured by the fact that this population generally possessed fairly positive explicit attitudes towards women in STEM as measured by this scale.

*Gender bias.* As expected, there was a significant effect of condition on gender bias, $F(4,440) = 3.55$, $MSE = 0.56$, $p = .007$, $\eta^2 = .03$. Participants in the video control condition did not significantly differ from those in the non-intervention control condition. Participants in the expert and hybrid conditions expressed significantly lower, and those in the narrative condition marginally lower, gender bias than did participants assigned to the non-intervention control condition. As predicted, pairwise comparisons revealed that participants in the hybrid ($d = 0.31$) and expert ($d = 0.41$) conditions each expressed significantly lower gender bias post-intervention than did participants in the video control condition. However, participants in the narrative condition did not differ significantly from those in the video control condition ($d = 0.21$), nor the expert and hybrid conditions (which also did not significantly differ from each other). These results suggest that the hybrid and expert interview conditions may be somewhat more useful tools than the narrative condition for reducing gender bias among adults in the general population (i.e., because only the hybrid and expert interview conditions produced significantly less

gender bias than the video control condition).

**Empathy.** As expected, there was a significant effect of condition on empathy, $F(4,375) = 27.25$, $MSE = 1.19$, $p < .001$, $\eta^2 = .23$. Participants in the video control condition did not significantly differ from those in the non-intervention control condition. Participants in the VIDS conditions expressed significantly greater empathy than did participants assigned to the non-intervention control condition. As hypothesized, pairwise comparisons revealed that participants in the narrative ($d = 1.13$), hybrid ($d = 0.94$), and expert ($d = 0.79$) conditions each reported significantly more empathy than did participants in the video control condition. Participants in the narrative condition also reported significantly more empathy than did those in the expert interview condition. The hybrid condition did not significantly differ from either of the other VIDS conditions. This suggests that each of the VIDS conditions successfully heightened participants' empathy, and that the narrative condition appeared to do so particularly effectively.

**Anger.** As expected, there was a significant effect of condition on anger, $F(4,366) = 43.11$, $MSE = 0.96$, $p < .001$, $\eta^2 = .32$. Participants in the video control condition did not significantly differ from those in the non-intervention control condition. Participants in the VIDS conditions reported significantly greater anger than did participants in the non-intervention control condition. As expected, pairwise comparisons revealed that participants in the narrative ($d = 1.43$), hybrid ($d = 1.14$), and expert ($d = 0.65$) conditions each reported significantly more anger than did participants in the video control condition, and the intervention conditions also each differed significantly from each other. Again, this suggests that each of the VIDS conditions actively engaged participants' anger, and that the narrative condition was again particularly likely to do so

**Secondary analyses.** Some previous research has found gender bias interventions to be more effective for men than women (e.g., Jackson et al., 2014). This is perhaps related to the finding that men often express more skepticism about experimental findings of gender bias than do women (Handley et al., 2015; Moss-Racusin, Molenda, & Cramer, 2015). However, consistent with preliminary work on VIDS (Pietri et al., 2017), there was no moderating effect of gender on any outcome in the current research, $p$s > .46. Of importance, this suggests that VIDS was equally effective for men and women[3].

**Discussion**

Evidence from Experiment 1 provided novel support for the efficacy of VIDS as an intervention targeting gender bias in STEM, as well as related awareness, attitudes, and emotions. Of importance, all three VIDS conditions improved general population adults' awareness of gender bias, attitudes towards women in STEM, modern sexism, and action-oriented emotions relative to the non-intervention control. As hypothesized, VIDS also largely produced superior outcomes relative to the video control. Effect sizes for these comparisons were generally medium to large (maximum $d = 1.43$, average $d = .64$).

More specifically, the narrative, expert interview, and hybrid conditions all raised awareness of gender bias relative to both control conditions, bolstering initial support for the efficacy of VIDS in promoting bias literacy (Pietri et al., 2017). Novel to the current research, VIDS significantly improved attitudes towards women in STEM (relative to the

3 We also conducted exploratory analyses to examine whether the intervention was differentially effective on any of the five dependent variables based on membership in any other demographic groups (i.e., age, race, level of education, or political orientation). Of the 20 analyses of variances that we estimated, the interaction effect was non-significant in 17 cases. Of the remaining three (the moderating effect of age on awareness of bias and the moderating effect of political orientation on both awareness of bias and modern sexism) which were not hypothesized, none reached statistical significance using a Bonferroni-corrected alpha of 0.0025. Taken together, these findings suggests that the interventions tended to be equivalently effective in improving bias outcomes, regardless of participants' demographic groups.

non-intervention control). The expert interview and hybrid conditions also reduced modern sexism and heightened empathy and anger, relative to both control conditions. The narrative condition followed this pattern as well, with the exception that it marginally reduced gender bias relative to the non-intervention control, and did not produce significantly less gender bias than the video control. Thus, results broadly supported the efficacy of all three VIDS conditions, supporting our hypothesis.

However, there were some important differences between the narrative, expert interview, and hybrid conditions. Notably, the hybrid and expert interview conditions seemed to be somewhat more effective in reducing modern sexism than the narrative condition. Conversely, the narrative condition appeared to most effectively engage participants' empathy and anger. Thus, while evidence emerged to suggest that all three VIDS conditions may be powerful tools, the expert interview condition may be particularly effective at targeting modern sexism, whereas the narrative condition may more effectively engage action-oriented emotions. These results reflect critical information for practitioners selecting interventions for use with various populations, with the goal of targeting different outcomes. Indeed, the different VIDS conditions may be most appropriate under different circumstances, suggesting that VIDS is a flexible, effective tool for diversity practitioners.

The video control condition differed from the non-intervention control in only one instance—participants in the video control reported marginally more positive attitudes towards women in STEM than those in the non-intervention control condition, suggesting that viewing the female scientists in the basic science control videos may have

inadvertently worked to improve attitudes towards women in STEM. In general, however, these findings suggest that including both a non-intervention and video control condition may not be necessary in future research, and that the video control may be a slightly more conservative comparison to VIDS.

Taken together, these results suggest that VIDS had a strong effect on bias outcomes. These findings are notable in that they highlight the potential for even a very brief online intervention to create significant improvement across a host of outcomes related to gender bias in STEM. VIDS' effectiveness is good news for institutions balancing a desire to utilize validated diversity interventions against the realities of budgetary and time restrictions. Indeed, VIDS is fast, consistent, inexpensive, and easy to implement, particularly because it can be completed at a time of participants' choosing using their own computers.

The current experiment also raised additional questions. Although results were extremely promising, it was not clear that they would generalize to the relevant STEM population. Additionally, it was unclear whether the observed effects would persist over time. Finally, although our selected outcomes reflect a range of established approaches to assessing gender bias, it was not feasible to assess these participants' intentions to engage in behaviors targeting gender bias in STEM (i.e., because many lay people may generally not have the opportunity to do so). Thus, it was not possible to determine whether VIDS might impact participants' behavioral tendencies.

## Experiment 2

Experiment 2 had four primary goals. First, we sought to determine whether the results of Experiment 1 would generalize to the

relevant population of STEM professionals. Thus, we recruited academic scientists as participants. Second, we sought to expand our outcome measures by considering variables beyond attitudes and emotional responses. Although objective measures of actual behavior would be ideal, it was not possible to obtain this information while maintaining STEM participants' anonymity and ensuring that all measurements were completed before they attended the National Academies Summer Institute (see below). To address this issue, past research testing diversity interventions (e.g., Carnes et al., 2015; Cundiff et al., 2014; Moss-Racusin et al., 2016) has measured participants' behavioral intentions, which have been identified as a critical link between attitudes and subsequent behaviors (e.g., Ajzen, 1991). Thus, we added a measure of participants' intentions to engage in behaviors promoting STEM gender parity.

Third, we adopted a longitudinal design (including a baseline measurement one week before the intervention, an immediate post-intervention measurement, and a follow-up one week after the intervention) in order to assess the lasting impact of VIDS. Additionally, this design allowed us to analyze participants' change from baseline measurements and assess whether the effectiveness of our intervention varied as a function of baseline levels of bias. Thus, all results in Experiment 2 reflect change scores from baseline. Finally, we again tested the efficacy of the narrative, expert interview, and hybrid

videos relative to the control videos. However, we eliminated the non-intervention control condition because we added a measure of participants' baseline scores (and because the non-intervention control largely did not differ from the video control condition in Experiment 1). As a result, Experiment 2 served as a critical test of the interventions, because results provide information about change in participants' awareness, attitudes, emotions, and behavioral intentions over time.

**Method**

   **Power assessment.** Prior to beginning participant recruitment for Experiment 2, we conducted a power analysis in SAS version 9.4 (Lane & Hennes, 2017). To estimate the number of participants needed to have 80% power to detect significant effects, we simulated data using the results of Experiment 1. Because there were ceiling effects for attitudes towards women in STEM in Experiment 1, we were primarily interested in powering our study to detect significant effects of awareness of bias and modern sexism. We were agnostic about the relative pattern of effects for the three intervention videos, so we estimated the effect of condition post-intervention using the average of the three experimental condition post-intervention effect sizes in Experiment 1. We estimated no decay and anticipated less random noise in our faculty sample than in MTurk (50% of the observed residual from Study 1). These simulations determined that we would need 140 participants to power omnibus effects of the intervention on awareness of bias and modern sexism at Time 2. After accounting for potential attrition (estimated at 20%), we sought to recruit 175 participants in Study 2. Our final sample of $N = 173$ at Time 1, $N =$

148 at Time 2, and $N = 142$ at Time 3 exceeded the $N = 140$ suggested by the power analysis.

**Participants and recruitment.** We recruited participants for Experiment 2 in collaboration with the National Academies Summer Institute (NASI) on undergraduate education (Wood & Handelsman, 2004). NASI focuses on training faculty to create more engaging and active science classrooms, with the ultimate goals of improving science education and student learning. Each summer, multiple NASIs are held across the country. During each NASI, attendees participate in workshops on teaching and work in groups to develop a short in-class exercise designed to effectively teach a scientific topic. Because the NASIs recruited a sample of STEM faculty from across the United States, they provided an excellent opportunity to identify participants for our research.

During the summer of data collection for Experiment 2 (2014), six NASIs were held throughout the country, representing a diversity of geographic regions including the Northeast, Southwest, West Coast, Mountain West, Southeast, and Gulf Coast. We also recruited from three university-specific Summer Institutes that had the same format as the regional NASIs. The organizers of each NASI informed the attendees (via an email originating from their institutional email address) that we would be contacting them to offer the opportunity to voluntarily participate in the experiment. We then sent a recruitment email (originating from the generic NASIresearch@yale.edu email address) inviting attendees to participate in our experiment. The recruitment email stated that participation was voluntary and not required for the NASI. Similar to Experiment 1, the recruitment email informed participants that, "This research looks at how individuals react to and remember information from videos."

All 331 academic scientists who were scheduled to attend a NASI during Summer 2014 were invited to take part in our experiment in exchange for Amazon.com gift certificates. Of those, one hundred seventy-three (52% of eligible participants) completed the experiment. This response rate is higher than those obtained in similar research utilizing STEM faculty participants (Moss-Racusin et al., 2012) and the power analysis reported above indicated that this sample size afforded sufficient statistical power to test our primary hypothesis. Of importance, attrition rates across the three measurement time points were unrelated to experimental condition and were quite low, comparing favorably to those frequently obtained in longitudinal research (Capaldi & Paterson, 1987). Specifically, 148 participants (86% of the original sample) took part in the Time 2 (post-intervention) session, and 142 participants (82.1% of the original sample) completed the Time 3 (one week follow-up) session. At Time 2, four participants experienced technical difficulties when watching the videos and accidentally watched two different sets of videos. As a result, only their Time 1 (pre-intervention) data was used, and the reported retention rates excluded these individuals. Participants were compensated with a $25 Amazon.com gift certificate after completing the Time 1 session, another $25 Amazon.com gift certificate after finishing the Time 2 session, and a final $50 Amazon.com gift certificate after completing the Time 3 session.

Our final sample reflected the diversity of the underlying population of NASI attendees. Faculty participants (54% female; age: $M = 42.77$, $SD = 10.43$, range = 25-69) were 72% White, 3% Black, 3% Hispanic, 11% Asian, 8% identified as another race, and 3% did not report their race or ethnicity. On a scale of 1 (*strongly liberal*) to 7 (*strongly conservative*), participants' average political orientation was 2.50 ($SD = 1.41$). Participants were from diverse institution types, with the majority from Research I universities (60% Research I universities, 25% Research II universities, 3% liberal arts colleges, 7% primarily teaching colleges, 5% 2-year community colleges). Participants were in various career stages, with the majority being tenured or tenure-track professors (5% dean or department chair, 12% full professors, 18% associate professors, 29% assistant professors, 23% lecturers, 11% postdoctoral fellows or graduate students, and 2% other), and had taught for an average of 11 years ($SD = 10.33$, range $0 – 60$ years). The majority of the participants worked in the biological sciences, but other STEM departments were also represented (68% biological sciences, 9% biomedical sciences, 5% chemistry, 5% physics, 4% engineering, 2% mathematics, 2% psychology, 1% computer science, 4% other)[4].

**Materials.**

---

4 To ensure that faculty who chose to participate in our experiment did not meaningfully differ from those who did not choose to participate, we compared our participants' demographic information to the information of the other attendees who did not partake in our experiment. Demographic information for these non-participants was obtained in partnership with the NASI leaders. There was no difference on any demographic characteristics between participants and non-participants (gender, race, age, job rank, and experience with behavioral science research), $p$s > .25, with one exception: The participant sample ($M = 9.24$ years, $SD = 9.02$ years) had been in their current job for slightly longer than had those in the nonparticipant sample ($M = 6.89$ years, $SD = 7.29$ years), $t(198) = 2.04$, $p = .04$. Thus, results support the idea that our participants largely did not systematically differ from potential participants who chose not to participate, facilitating generalizability of the current results.

*Intervention videos.* We utilized the same three versions of VIDS (narrative, expert interview, and hybrid) as in Experiment 1, as well as the same control videos.

*Attention and manipulation checks.* Participants in the VIDS conditions responded to the same 18 attention check questions used in Experiment 1. In Experiment 2, no participants failed more than 3 attention check questions, and thus all were included in the final sample. We utilized the same measures of transportation ($M = 3.22$, $SD = 0.75$, $\alpha = .66$) and logical thinking ($M = 4.12$, $SD = 0.72$, $\alpha = .82$) as in Experiment 1.

*Outcome measures.* We administered the same measures of awareness of gender bias, attitudes towards women in STEM, gender bias, and action-oriented emotions as in Experiment 1.

*Behavioral intentions.* Novel to Experiment 2, we measured participants' propensity to take action on critical issues related to addressing gender bias. Participants rated their level of agreement (1 = *strongly disagree* to 5 = *strongly agree*) with 13 statements assessing their likelihood of engaging in actions that would help increase gender parity in STEM. The items specifically addressed participants' likelihood of collaborating with, mentoring, sharing resources with, and generating support for female scientists, actions identified by prior research as particularly critical to addressing STEM gender inequalities (Prochaska et al., 2006; see Table 5 for all items). These 13 items demonstrated adequate reliability and were therefore averaged to create the behavioral intentions scale, with higher numbers reflecting greater intentions to engage in gender bias-reducing behaviors.

**Procedure.** Participants were contacted via email and invited to take part in the online study. After completing the baseline measures,

participants were informed that we would contact them again in one week to complete the post-intervention measures, and again after another week to complete the follow-up measures. Because the NASI workshop content could have influenced results, it was important to ensure that faculty completed all measures prior to attending the NASI. Because the list of attendees was finalized only several weeks before each NASI, this allowed for only a several week period in which all measures could be administered, and necessitated that the follow-up session occurred one week after viewing the videos.

Participants were randomly assigned to either the narrative, expert interview, hybrid, or video control condition and then completed all measures in a random order (and items within each measure were also randomized). Participants were partially debriefed after completing the baseline and post-intervention measures, and were fully debriefed after completing the follow-up measures. Participants who elected to suspend their participation after the baseline or post-intervention measures were fully debriefed and compensated at that point.

**Results**

**Manipulation check.** As expected, there was a significant effect of condition on transportation, $F(3,138) = 11.81$, $MSE = 0.45$, $p < .001$, $\eta^2 = .20$. The narrative condition videos ($M = 3.39$, $SE = 0.11$) were again perceived to be more transporting than were the expert interview videos ($M = 3.19$, $SE = 0.11$), but this

difference was not statistically significant among the faculty sample, $t(138) = 1.25$, $p = .$

21. Consistent with Experiment 1 and prior research (Pietri et al., 2017), there was also a

significant effect of condition on logical thinking, $F(3,138) = 3.86$, $MSE = 0.46$, $p$

$= .011$, $\eta^2 = .08$. In particular, the expert interview videos ($M = 4.40$, $SE = 0.12$)

stimulated significantly more logical thinking than did our narrative videos ($M = 3.96$, $SE$

$= 0.12$), $t(138) = -2.67$, $p < .01$.

**Analytic strategy.** Because our design was longitudinal, we analyzed our data

using multilevel modeling (Raudenbush & Bryk, 2002). To investigate the effects of the

videos, we estimated multilevel regression models using the mixed procedure in SAS

version 9.4. Because we included baseline responses in Experiment 2, we were now able

to model change. To do so, we adopted the change-regression approach recommended by

McArdle (2009) for models with baseline measurement and a Time 2 experimental

manipulation. This resulted in a stacked dataset in which Times 2 (post-intervention) and

3 (follow-up) were modeled as repeated measures with a Time 1 (baseline) covariate[5].

Difference scores on the outcome measures (which were measured on 5 point scales)

could range from -4 to 4. We then fit models in which change from baseline was

predicted by experimental condition (dummy coded, with the video control condition as

the reference group), time, and the interaction of condition and time, adjusting for

---

5 We estimated a random person intercept and allowed the residual variation at Times 2 and 3 to be freely
estimated but constrained the residual covariance to be 0 (i.e., a variance components matrix). We initially
allowed the residual covariance to be freely estimated, but the estimate was negligible and prohibited
model convergence.
Because people in different parts of the country may have different attitudes toward women in STEM, we
also initially modeled a random site intercept (e.g., which of the workshop locations the participant was
scheduled to attend) to statistically account for the potential tendency for our effects to vary systematically
based on the location of participants' upcoming workshop. However, very little of the variability in
outcomes could be attributed to variability by site, again prohibiting model convergence. Therefore, this
intercept was not modeled in the final analyses.

baseline score. We also output the least square means in each analysis, allowing us to estimate mean differences between conditions.

Below we report the effect of experimental condition on changes from baseline in the six outcome variables first at post-intervention and then at follow-up. At each time point, we first note whether each VIDS condition differed from control and then report any inter-VIDS condition differences. New to Experiment 2, we are now also able to report whether changes from baseline were significant at each time point, and whether effects significantly decayed from post-intervention to follow-up. Full regression results for Experiment 2 are presented in Table 6, means and standard errors for Experiment 2 are presented in Table 7, bivariate correlations between all variables for both experiments are presented in Table 2, effect sizes for both experiments appear in Table 4, and graphs of Experiment 2's findings are illustrated in Figure 2.

**Awareness of bias.** We began by examining post-intervention effects on awareness of bias. Consistent with expectations and results for the general population in Experiment 1, STEM faculty participants in the narrative ($d = 0.55$), hybrid ($d = 0.96$), and expert interview ($d = 1.16$) conditions each expressed greater awareness of gender bias than did those in the control condition. The hybrid and expert conditions both resulted in significantly more awareness of bias than did the narrative condition, but did not significantly differ from each other. Further, participants in the hybrid and expert interview (but not the narrative) conditions each expressed a significant increase in awareness of bias post-intervention relative to baseline, while those in the narrative reported no significant change from baseline.

Of importance, participants in the control condition reported a significant *decrease* in awareness of bias relative to baseline. This is likely reflective of longitudinal attenuation effects observed elsewhere in the literature; responses to a variety of psychological measures tend to be artificially elevated at Time 1 due to participants' initial biases to respond in socially desirable ways (Knowles, Coker, Scott, & Cook, 1996; Shrout et al., 2017). This pressure to respond desirably often weakens over time, resulting in later scores that are likely more reflective of participants' true levels of awareness (e.g., the decrease in reported awareness of bias observed in the control condition here). Of importance, positive effects over time in the expert interview and hybrid conditions *reflect improvements in bias outcomes over and above this general attenuation tendency*, further signaling the impact of the intervention (in addition to the fact that each VIDS condition resulted in significantly more awareness of bias than did the control condition post-intervention).

We then examined effects on awareness of bias at follow-up. There was no significant time by condition interaction, indicating no decay over time, and participants in the narrative ($d = 0.64$), hybrid ($d = 0.98$), and expert ($d = 1.13$) conditions still reported significantly greater awareness of bias one week later than did participants in the control condition. The difference between the narrative and expert conditions was also still significant (though the hybrid condition did not differ significantly from the narrative or expert interview). Further, faculty in all VIDS conditions expressed significantly

greater awareness of bias at follow-up compared to baseline (while those in the control condition continued to express significantly *less* awareness of bias than at baseline). These results provided promising first evidence of VIDS' effectiveness over time, and suggested that the expert interviews may be particularly effective for STEM faculty.

**Attitudes towards women in STEM.** We next examined the effect of experimental condition on academic scientists' attitudes toward women in STEM. As in Experiment 1, there was evidence of a ceiling effect on this measure, with academic scientists reporting an average score of 4.70 out of 5.00 ($SD = 0.36$) at baseline. Nonetheless, despite the fact that the experiment was not powered to detect effects on this measure (as discussed above), participants in the hybrid ($d = 0.36$) and expert ($d = 0.37$) conditions each expressed marginally more positive attitudes towards women in STEM post-intervention than did participants in the control condition. While the narrative condition did not differ from control ($d = 0.00$), the hybrid and expert conditions both differed significantly from the narrative condition (but did not significantly differ from each other). Likely as a function of the ceiling effect on this measure, no conditions reflected significant improvement from baseline (i.e., because attitudes towards women in STEM were already quite high, there was insufficient room for them to improve on this scale).

There was no significant interaction between condition and time, although there was significant decay in the expert interview condition[6]. No VIDS conditions

_____

6 However, these results together indicate that the decay effect was not significantly different from the mere passage of time observed in the control condition (the reference group). Thus, significant reductions from baseline (as indicated by "t" superscripts in Table 7) are not particularly noteworthy in the absence of a significant time by condition interaction in Table 6.

significantly differed from control at follow-up, although as noted above, the experiment was not powered to detect pairwise differences for this measure with restricted range. However, participants in the hybrid condition reported significantly more positive attitudes towards women in STEM one week later than did participants in the narrative condition. The expert interview and narrative conditions did not differ from control (or each other) at follow-up. As at post-intervention, results for all conditions at follow-up did not significantly improve from baseline (again likely due the ceiling effect on this measure).

**Gender bias.** We then examined the effect of condition on gender bias itself. As expected, faculty in the narrative ($d = 0.45$), hybrid ($d = 0.74$), and expert interview ($d = 0.93$) conditions each expressed significantly lower gender bias post-intervention than did faculty in the control condition. The expert interview condition also produced less gender bias than did the narrative condition, but the hybrid condition did not differ significantly from either of the others. Further, academic scientists in all VIDS conditions expressed a significant decrease in gender bias post-intervention relative to baseline, whereas those in the control condition did not significantly change from baseline.

Moreover, there was no significant decay over time (i.e., no significant interactions between condition and time), and participants in the hybrid ($d = 0.60$), and expert interview ($d = 0.73$) conditions still expressed significantly less (and the narrative condition marginally less, $d = 0.42$) gender bias one week later than did participants in the control condition. None of the VIDS conditions significantly differed from each other at follow-up. However, all VIDS conditions remained significantly lower than baseline (while the control condition did not differ from baseline). As predicted, these results

suggest that VIDS successfully reduced faculty members' gender biases over time, whereas the control condition did not.

**Empathy.** We then turned to results for empathy. As in Experiment 1, participants in the narrative ($d = 1.44$), hybrid ($d = 1.52$), and expert ($d = 1.25$) conditions each expressed significantly greater empathy post-intervention than did participants assigned to the control condition; however, the intervention videos did not differ significantly from each other. Additionally, faculty in the narrative and hybrid conditions expressed a significant increase in empathy relative to baseline, while those in the control condition reported a significant *decrease* in empathy (again consistent with attenuation effects) and those in the expert interview condition reported no change from baseline.

Unlike the other outcomes, there was some significant decay over time (as indicated by significant interactions between condition and time). These emotional decay effects are perhaps not surprising given that emotions are frequently conceptualized as affective states that, by definition, fluctuate in response to temporal stimuli (e.g., Eaton & Funder, 2001; Frijda, Mesquita, Sonnemans, & Van Goozen, 1991). Despite the decay, participants in the narrative ($d = 0.71$), hybrid ($d = 0.62$), and expert ($d = 0.54$) conditions still reported significantly more empathy one week later than did participants in the control condition. As at post-intervention, the conditions did not differ significantly from each other; nor were they significantly improved from baseline. In other words, VIDS activated participants' empathy even at follow-up, relative to participants in the control condition.

**Anger.** We then examined direct effects on anger. Participants in the narrative ($d = 1.45$), hybrid ($d = 1.37$), and expert ($d = 1.04$) conditions each expressed significantly

greater anger post-intervention than did participants in the control condition. The narrative condition produced significantly more anger than the expert interview condition, but the hybrid did not significantly differ from either. Further, participants in the VIDS conditions expressed a significant increase in anger post-intervention relative to baseline, while those in the control condition reported a significant *decrease* in anger, likely due to attenuation.

Although there was again significant decay (as indicated by interactions between conditions and time), participants in the hybrid ($d = 0.49$) and expert ($d = 0.68$) conditions (but not the narrative condition, $d = 0.40$) still expressed significantly more anger one week later than did participants in the control condition. The VIDS conditions did not differ significantly from each other, and faculty in the VIDS conditions did not express significantly more anger than at baseline.

**Behavioral intentions.** Finally, we tested the impact of condition on the new measure of faculty participants' intentions to engage in behaviors to promote gender parity in STEM. Participants in the hybrid ($d = 0.50$) and expert ($d = 0.58$) conditions each expressed significantly stronger intentions to engage in parity-promoting behaviors post-intervention than did participants assigned to the control condition. The hybrid and expert conditions did not significantly differ from each other, nor did the narrative condition differ significantly from the control condition ($d = 0.33$). Participants in the control condition reported significantly *fewer* behavioral intentions to combat gender bias than at baseline, while those in the VIDS conditions did not significantly differ from baseline.

Of interest, a slightly different pattern emerged one week later. At follow-up, participants in the narrative condition expressed significantly stronger behavioral intentions relative to those in the control condition ($d = 0.51$), but participants in the hybrid and expert interview conditions no longer significantly differed from control (or each other). And yet, there was no significant condition by time interaction, suggesting that despite the decline in statistical significance in the contrasts between the control condition and the hybrid and expert interview conditions at follow-up, there was no evidence that participants in these VIDS conditions experienced levels of decay beyond what was observed in the control condition. As observed post-intervention, participants in the control condition reported significantly *less* behavioral intentions to combat gender bias than at baseline, while those in the VIDS conditions did not significantly differ from baseline. Thus, although each video led to significantly greater behavioral intentions relative to the control condition, these results may suggest that the narrative videos may lead to stronger long-term intentions, whereas the expert interview videos may lead to stronger immediate intentions[7].

**Mediation analyses.** There was not sufficient existing literature to warrant a-priori predictions regarding mediational processes. However, after further reflecting upon the links between action-oriented emotions and subsequent behavioral tendencies observed in previous research (Batson et al., 2002; Carver & Harmon-Jones,

---

7 As in Study 1, we found no moderating effect of gender on any outcome, $p$s > ,08, nor did results change when adjusting for gender. No other demographic variables consistently moderated condition effects, nor did baseline levels of gender bias. This again suggests that our intervention is equally effective for men and women, for individuals who were both higher and lower on initial bias, and regardless of any other measured aspect of faculty members' identity or experience (e.g., time in academia, STEM field, rank).

2009; Gordijn et al., 2006) per the recommendation of an anonymous reviewer of a previous version of this manuscript, we sought to more fully explore the potential impact of the action-oriented emotions.  We estimated two path models using MPlus version 7.4 in which we requested 95% confidence intervals of the indirect effects using 1,000 bootstrapped samples (Hayes, 2013; Kline, 2011). We modeled behavioral intentions at both post-intervention and follow-up, allowing them to correlate, predicted by condition (dummy coded, with the control video as the reference category) via post-intervention empathy (Model 1) or anger (Model 2). Consistent with the estimate of the direct effects, we modeled change from baseline. In other words, the indirect effects reflect the degree to which changes in the mediating process post-intervention (relative to baseline) led to changes in behavioral intentions at post-intervention and at follow-up. Baseline measurement of behavioral intentions and the mediator were adjusted for in each model.

　　　　We first examined the indirect effect of condition on bias outcomes via empathy (Figure 3). Change in empathy post-intervention (relative to baseline) significantly mediated the direct effect of all three VIDS conditions on changes in behavioral intentions, compared to the control video condition, both at post-intervention and at follow-up (unstandardized 95% CI post-intervention: narrative [.10; .33]), hybrid [.11; . 35]), expert interview [.07; .29]; 95% CI follow-up: narrative [.06; .28], hybrid [.07; .29], expert interview [.05; .23]). Similarly, change in anger (Figure 4) post-intervention (relative to baseline) significantly mediated the direct effect of all three VIDS conditions on changes in behavioral intentions, compared to the control video condition, both at post-intervention and at follow-up (95% CI post-intervention: narrative [.01; .22]), hybrid

[.02; .20]), expert interview [.01; .15]; 95% CI follow-up: narrative [.01; .18], hybrid

[.01; .17], expert interview [.01; .13]).[8] Together, these findings suggest that VIDS

successfully increased feelings of empathy and anger about gender bias in STEM, which

then led to increased short-term and long-term intentions to take action.

**Discussion**

Results from Experiment 2 provided more support for the efficacy

of VIDS, and particularly speak to its effectiveness among STEM

professionals. All VIDS conditions significantly improved awareness of

bias, reduced modern sexism, and heightened the action-oriented

emotions of empathy and anger more so than the control condition,

both immediately post-intervention and at follow-up. The hybrid and

expert interview conditions also marginally significantly improved

attitudes towards women in STEM compared to the control condition at

both post-intervention and follow-up, despite another ceiling effect

constraining results for this measure. Novel to Experiment 2, the

hybrid and expert interview conditions resulted in greater intentions to

engage in behaviors promoting gender parity than did the control

condition post-intervention, while the narrative condition outperformed

---

8 We also estimated a dual mediation model in which the mediational pathways for empathy and anger were modeled simultaneously (as well as the residual correlation between the two mediators and the baseline covariates of each). In this model, empathy remained a significant mediator of VIDS on behavioral intentions at both post-intervention and follow-up, but anger did not. This suggests that, after accounting for the shared variance in emotional response to VIDS, empathy may have been somewhat more effective in inspiring intentions to reduce gender bias in STEM. However, anger and empathy were correlated post-intervention, $r(171) = .56$, $p < .001$, and we hesitate to draw strong conclusions about the relative impact of empathy and anger on inspiring action from a posthoc analysis of a single study. Future research should continue to examine the emotions that are most likely to lead to action in the context of gender bias.

the control at follow-up. Finally, we found that VIDS-induced increases in anger and empathy led to intentions to promote gender parity.

As in Experiment 1, effect sizes relative to the video control condition were generally medium to very large (post-intervention maximum $d$ = 1.52, average $d$ = .83; follow up maximum $d$ = 1.13, average $d$ = .54). Further, our findings were largely robust to decay effects; significant decay emerged only in the case of the action-oriented emotions (consistent with past work on the context-sensitivity of emotional states; e.g. Eaton & Funder, 2001; Frijda et al., 1991). In sum, results underscore the broader pattern of VIDS' effectiveness across multiple measures of bias and related constructs.

Of note, participants in the control condition reported a significant *decrease* in awareness of bias, empathy, anger, and behavioral intentions at post-intervention and follow-up relative to baseline. This pattern likely reflects established attenuation effects, whereby reduced social desirability pressures often result in less inflated self-reporting over time (Knowles et al., 1996). In other words, self-reported awareness of gender bias, empathy, anger, and intentions to engage in behaviors that promote gender parity appear to naturally weaken over time. It is thus critical to note that the many positive results obtained here highlight VIDS' ability to produce *improved* outcomes over and above this tendency for these outcomes

to *worsen* over time (at least as reported by participants), furthering bolstering the efficacy of VIDS.

Replicating and extending Experiment 1, results broadly supported the utility of VIDS among the relevant specialized STEM population. However, consistent with Experiment 1, some interesting differences between VIDS conditions emerged. The results for behavioral intentions were particularly interesting, in that the expert interview was more effective than the narrative post-intervention, while this pattern reversed at follow-up. Although we caution against drawing strong conclusions from this finding without further replication, examining the short- and long-term impacts of each VIDS condition on intentions to combat gender bias is an intriguing area for future research. More broadly, the expert interview condition tended to outperform the narrative condition in changing attitudes and intentions (i.e., awareness of gender bias, attitudes towards women in STEM, modern sexism, and behavioral intentions). However, the narrative condition tended to result in superior outcomes in eliciting emotions. Thus, these results generally suggest that practitioners should make evidence-based decisions about which VIDS condition (or combination) would most effectively target the relevant outcomes of interest to specific populations.

**General Discussion**

Results from two experiments and multiple assessment time points suggested that VIDS functioned as successful interventions. Relative to control conditions, VIDS heightened awareness of STEM gender bias (a critical manifestation of bias literacy; Carnes et al., 2012; Pietri et al., 2017; Zawadzki et al., 2012), promoted attitudes towards women in STEM, reduced gender bias, engaged the action-oriented emotions of empathy and anger, and increased intentions to engage in behaviors that increase STEM gender parity. Of importance, results were observed with both adults from the general population and STEM professionals, and largely persisted over time. Further, in contrast to some past interventions (e.g., Jackson et al., 2014), VIDS were equally effective for both men and women. Additionally, because we utilized filmed media rather than live trainings, our brief online interventions are consistent, inexpensive, easily scalable for implementation, and bolster literature on the power of media to promote positive social change over time (Green & Brock, 2000; Paluck, 2009). VIDS offer a powerful and affordable tool for institutions balancing the importance of employing validated diversity interventions with the realities of budgetary and time constraints. The videos, along with suggested discussion questions, are freely available for download via the VIDS website (https://academics.skidmore.edu/blogs/vids/).

Effect sizes for most outcomes were large or very large, revealing the power and promise of VIDS. Indeed, across both Experiments, 13 of 45 (29%) significant effect sizes were classified as small in size ($d$ = 0.20 - 0.49), while 16 (35%) were classified as medium ($d$ = 0.50 - 0.79), and 16 (35%) were classified as large or very large ($d$ > .80), with an average effect size of $d$ = 0.75. The most obvious exception to this general trend was attitudes towards women in STEM, which had three small effects and six effects smaller than $d$ = 0.20 when compared to the video control condition.  However, it is important to note that in Experiment 1, each VIDS condition did improve attitudes towards women in STEM relative to the non-intervention control. Further, as noted above, this measure was constrained by a ceiling effect (Experiment 1 grand $M$ = 4.41 out of 5.00, $SD$ = 0.58; Experiment 2 baseline $M$ = 4.70 out of 5.00, $SD$ = 0.36), significantly limiting possible results. In other words, because most participants already expressed fairly positive attitudes towards women in STEM as assessed by this measure, it was not possible to adequately investigate the extent to which VIDS might improve these attitudes. As a result, future research should develop and utilize a more sensitive measure of this construct in order to accurately assess its responsiveness to VIDS.

We developed multiple versions of VIDS in order to examine the effectiveness of different communication styles drawn from the

existing literature (Green & Brock, 2000; Petty & Cacioppo, 1986), and in the hopes of enabling appropriate targeting across audiences and contexts. Because there was insufficient existing evidence to justify predicting condition differences, we simply hypothesized that VIDS would outperform the video control condition. While evidence supported this hypothesis by revealing the utility of each VIDS condition, some results suggested that different versions of VIDS might impact certain outcomes more effectively. Specifically, an interesting trend emerged, such that the expert interviews tended to be more effective than the narratives—though not dramatically so—in targeting attitudes and behavioral intentions. In contrast, the narratives appeared to be particularly effective in engaging action-oriented emotions. This pattern is consistent with some prior work revealing that narratives are uniquely effective in engaging empathy (Johnson et al., 2013) and are highly emotionally evocative (Prentice & Gerrig, 1999).

It is perhaps more surprising that the expert interviews—which were interesting and entertaining, but relied on a more straightforward presentation of facts—were so effective in targeting attitudes and behavioral intentions. Further, it is interesting to note that the expert interviews appeared to be a particularly successful approach for the STEM faculty participants in Experiment 2. In keeping with their training to analyze evidence, this community may have been

particularly motivated to carefully scrutinize the information contained in each video. As a result, the clear discussion of research may have been particularly persuasive for this audience (Cacioppo, Kao, Petty, & Rodriguez, 1986). However, it is possible that the narrative videos may generate additional positive outcomes that were not measured here, such as improvements in bias-reducing behaviors (i.e., actively seeking out information about gender bias). Thus, future research should systematically investigate the extent to which the VIDS conditions are differentially effective in targeting various manifestations of gender bias in STEM, and among different types of participant populations. For example, additional research could fruitfully assess the effectiveness of the videos for undergraduate, secondary, and elementary students, both with and without expertise in STEM.

Additionally, it is worth noting that the hybrid condition may provide a useful compromise between the strengths of the narrative and expert interview conditions. On the one hand, the hybrid condition frequently fell between the other experimental conditions. However, it outperformed the other conditions at promoting positive attitudes towards women in STEM at follow-up in Experiment 2. More broadly, it often matched the expert interviews' efficacy at improving attitudes (e.g., gender bias in Experiment 1, awareness of bias and behavioral intentions post-intervention in Experiment 2) and the narrative's ability to engage action-oriented emotions (e.g., anger at follow-up in Experiment 2). As such, it may balance the strengths of the other two conditions, and thus

reflect a "best of both worlds" approach for diversity practitioners eager to maximize benefits across a wide range of outcomes.

The current research moved beyond preliminary work validating VIDS (Pietri et al., 2017) in at least five critical ways. First, the current research allowed for a more robust causal test of VIDS' effectiveness in that it added non-intervention control and hybrid conditions. Results from Experiment 1 indicated that VIDS generally tended to improve outcomes relative to both the non-intervention and video control conditions, suggesting that the effectiveness of VIDS is not attributable solely to the passage of time, simply watching any video, or mere exposure to women scientists. Second, the current research exposed participants to six five-minute VIDS, rather than one, as in prior work. Of importance, because the average VIDS effect size on awareness of bias in prior research (Pietri et al., 2017) was $d = 0.27$ and the smallest effect in the comparable Experiment 1 in the current research was $d = 0.65$ (and the largest was $d = 0.82$), increasing the number of videos viewed was associated with more than a doubling in effect size. These results suggest that there may be an additive effect associated with amount of VIDS exposure. Future research should seek to determine the optimal amount of exposure, in order to maximize both efficiency and outcome magnitude.

Third, the current research was the first to explore the effectiveness of VIDS with the critically relevant population of STEM

professionals. Positive results from Experiment 1 (and preliminary research; Pietri et al., 2017) with adults from the general population are heartening; meaningfully impacting STEM gender bias is likely not possible without shifting prevalent cultural stereotypes about women's science incompetence, which are expressed by those working outside of STEM as well as within it (Nosek et al., 2002; Smyth & Nosek, 2015). Thus, it is critical that VIDS operate effectively with non-STEM professionals. However, those working in STEM are clearly in the best position to interrupt (or, alternately, to perpetuate) the gender status quo in their own communities. As a result, VIDS would not be a promising intervention if it did not work effectively with STEM professionals. Thus, results from the current work provide critical novel evidence supporting the utility of VIDS.

Fourth, the longitudinal design of Experiment 2 allowed us to conduct a first test of VIDS effectiveness over time. Results revealed that VIDS' impact on awareness of STEM gender bias, attitudes towards women in STEM, gender bias, and behavioral intentions did not significantly decay over time. Given the brief nature of the intervention, these results are heartening, and add to the growing body of evidence (e.g., Mazziotta et al., 2011; Ortiz & Harwood, 2007; Paluck, 2009) suggesting that even short evidence-based media interventions can have lasting positive effects. Additionally, the significant decay effects that were observed for empathy and anger

are not particularly surprising, given the established transient nature of emotions (e.g., Eaton & Funder, 2001; Frijda et al., 1991). Put another way, because emotions have long been conceptualized as *states* that fluctuate in response to current stimuli, it would be highly unexpected (and perhaps even concerning!) if VIDS had produced prolonged disruptions in participants' generalized anger and empathy over time.

Fifth, the current work greatly expanded the outcomes investigated. The initial work focused on the development and validation of VIDS, and primarily assessed their impact on bias literacy. The current research revealed that VIDS not only increased awareness of bias (a key indicator of bias literacy; Carnes et al., 2012; Zawadzki et al., 2012), but also promoted positive attitudes towards women in STEM, reduced STEM gender bias, engaged the action-oriented emotions of anger and empathy, and increased intentions to engage in behaviors combatting gender bias. We assessed self-reported attitudes and behavioral intentions, because they are often strongly linked to subsequent actions (Ajzen, 1991; Carnes et al., 2015; Cundiff et al., 2014; Kollmuss & Agymen, 2002; Moss-Racusin et al., 2016). Future research should probe the intervention's effectiveness with additional outcome variables, such as implicit measures and direct measures of behavior. Indeed, while some recent work reveals that evidence-based interventions can produce long-term reductions in undergraduates' implicit racial biases (Devine, Forscher, Austin, & Cox, 2012), other

work has shown mixed effects of interventions on implicit racial bias (e.g., Lai et al., 2014) and no effects on implicit gender bias (Carnes et al., 2015), suggesting that implicit biases may be quite difficult to change. Thus, systematically exploring the extent to which VIDS and other interventions may effectively target implicit STEM gender biases would be a particularly useful topic for future work.

A concern was that our results might be constrained by demand characteristics, or pressure experienced by participants to respond in certain socially desirable, "politically correct" ways. However, several factors mitigate this concern. Although all three sets of videos improved outcomes, as discussed above they did so to different degrees. These systematic differences would be unlikely if results simply reflected demand characteristics, which would likely be equally strong across VIDS conditions. Additionally, the effects for attitudes and intentions were not significantly weaker at follow-up in Experiment 2. It is unlikely that demand characteristics experienced at the time of intervention would persist robustly over time and in a different measurement context. The attenuation effects provide some support for this argument—as evidenced in the video control condition in Experiment 2, social desirability effects typically *decline* over time, suggesting that the experimental conditions produced effects that counteracted this natural tendency to provide less politically correct responses after repeated measurement.

Also, it is worth noting that even if demand effects contributed to the results, this need not necessarily fatally undermine the utility of the interventions. That is, if the videos lead participants to believe that reduced bias reflects the socially desirable response and motivated them to express it, then the videos may still promote positive outcomes. A large body of evidence suggests that biased behavior is the result of both the activation and expression of stereotypes (e.g., Bodenhausen & Macrae, 1998). As a result, effective interventions need not necessarily prevent the very *activation* of participants' biases, but could instead operate successfully by preventing participants from *acting upon and expressing* biases (Monteith, Sherman, & Devine, 1998). In this way, even if our positive results reflect some amount of socially desirable responding, it is promising that our interventions motivated and enabled participants to respond in a less gender-biased manner.

**Limitations and Future Directions**

The current findings highlight additional fruitful avenues for future research. As in other related experiments (Carnes et al., 2015; Jackson et al., 2014), participation in the current research was voluntary, and was thus subject to concerns related to selection biases. That is, because our samples were self-selected to a certain degree, there may be limits to the extent to which the current results generalize to other groups. These concerns may be mitigated by the fact that we successfully recruited a large, relatively diverse sample in Experiment 1, and in Experiment 2, the demographics of the underlying NASI

faculty population largely matched those who elected to participate. Additionally, we were careful to mask the true focus of the experiments, and instead, the cover story informed participants that our research was about their impressions of and ability to remember different videos. This should reduce concerns that those who elected to participate were highly interested in or motivated to address gender bias. Nonetheless, future research should utilize randomly selected samples drawn from both the underlying populations of adults from the general population and STEM professionals in order to determine the extent to which our results are generalizable.

Relatedly, the majority of participants in Experiment 2 were from the biological sciences. Although STEM discipline did not significantly moderate the current results, we did not recruit enough participants from other disciplines to afford sufficient statistical power to thoroughly examine the potential impact of STEM discipline. This is particularly important to consider because women are better represented (e.g., AAUW, 2010; Sonnert, Fox, & Adkins, 2007) and may report encountering less gender bias (Robnett, 2016) in the biological sciences relative to more math-intensive STEM fields (e.g., mathematics, engineering, physics). As a result, Experiment 2 may reflect a relatively liberal test of VIDS' effectiveness with STEM faculty, given that those in the biological sciences may be more receptive than those in more math-intensive fields. Indeed, the ceiling effect in positive attitudes towards women in STEM in Experiment 2 partially supports this possibility.

However, several factors mitigate these concerns. First, the ceiling effects were also observed with general population participants in Experiment 1, suggesting that they may be more reflective of limitations associated with this particular measure than with

biological science faculty's uniquely positive attitudes. Additionally, the lack of ceiling (or floor) effects on the remaining five outcome measures contradict the idea that biological science faculty were generally unusually highly supportive of women in STEM. More importantly, prior experimental evidence of gender bias among STEM professionals has not revealed moderation by STEM field (e.g., Moss-Racusin et al., 2012). Additionally, women only report experiencing less gender bias in biological science (relative to mor math-intensive) fields at the undergraduate level; these differences were not significant at the graduate or high school levels (Robnett, 2016). Still, additional work should examine the extent to which the intervention's effectiveness varies by participants' STEM field, and in particular, if VIDS is equally effective with faculty from math-intensive fields.

Also, because the videos convey the seriousness of gender bias, future work should evaluate whether they result in unintended consequences (such as making obstacles to success in STEM salient for women or increasing susceptibility to stereotype threat), and if so, how best to combat these obstacles. However, some existing evidence mitigates this concern; an intervention program designed to promote interest in science that included a discussion of gender discrimination increased adolescent girls' science self-efficacy and belief in the value of science, whereas the identical intervention that omitted discussion of gender bias had no impact on girls' outcomes (Weisgram & Bigler, 2007). Of importance, the discrimination intervention did not result in unintended negative consequences relative to the standard intervention on any measures assessed by the authors. However, future work could further investigate the possible

negative consequences associated with VIDS and other interventions using a wider range of outcomes.

Despite the non-significant overall effects of time, some individual condition differences were no longer significant at follow-up. Future research should thus consider ways to bolster the long-term efficacy of the intervention, such as reminder emails, brief refresher videos, or a short follow-up course. Further, although the current results were promising, it was only possible to conduct the follow-up measurement after one week (due to the need to complete all measurements before participants arrived at the NASI). Although other tests of bias interventions have included even shorter follow-ups (e.g., 24 hours; Dasgupta & Greenwald, 2001), the extent to which VIDS continues to impact outcomes over longer time periods should be further assessed.

Because the video content differed across conditions, it was not feasible to ask participants to report their action-oriented emotions in response to a specific target. Thus, because it was important to be able to compare participants' emotional responses across conditions, we simply asked participants to report the extent to which they felt each emotion and did not specify or constrain the emotional target. For this reason, the decay effects associated with anger and empathy are perhaps not surprising. Future research should determine whether VIDS also increases action-oriented emotions in response to specific targets,

such as characters portrayed in VIDS, women in STEM more broadly, or the concept of gender bias itself.

Future research should also investigate optimal methods of utilizing VIDS on a large scale. The current research cannot determine whether the videos would be most effective if viewed online individually (as tested here), or whether group viewings accompanied by structured discussions (either in-person or online) could increase efficacy for certain audiences. It is possible that our interventions, when paired with such a structured discussion, would produce even greater reductions in bias-related outcomes. Future work should also examine whether the videos operate effectively in non-STEM contexts. Although the videos focus on gender bias in STEM, they illustrate broad principles from the underlying literature. Moreover, they were effective with adults from the general population, and thus may also work to reduce bias among participants from other specialized non-STEM workplaces and institutions.

A further area to pursue in this research sphere is to determine whether similar intervention videos could reduce bias against other stigmatized groups, such as people of color or people with physical disabilities. Indeed, given the efficacy of VIDS demonstrated here, results suggest that the development of additional media-based interventions may be a fruitful approach to combating bias against— and ultimately promoting the full participation of—different groups across various institution types. Thus, our research serves as a roadmap for interdisciplinary collaboration, and a robust foundation

upon which to base research testing the utility of visual media in addressing diversity issues more broadly.

**Conclusions**

Recruiting and retaining the most talented individuals—regardless of their demographic background—is in the best interest of scientific progress and national competitiveness (Moss-Racusin et al., 2012; 2014; Williams & Ceci, 2015). Gender biases that systematically deter women from STEM thus undermine not only the careers of individual women, but the meritocratic spirit of discovery that enhances the lives of all those who benefit from scientific achievements. The STEM community has argued that it is time to remove lingering obstacles to women's full participation (e.g., Al-Gazali et al., 2013; Ferguson & Ghorayshi, 2015; Handelsman et al., 2005; Nature, 2012; 2015; Raymond, 2013; Shen, 2013); however, effective gender bias interventions have not been readily available. The current work provides a timely, novel demonstration of VIDS' efficacy in targeting STEM gender bias. Our hope is that VIDS will foster a continued scientific approach to the development, testing, and widespread implementation of interventions that effectively promote STEM diversity.

## References

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In J. Kuhl & J. Beckmann (Eds.), *Action-control: From cognition to behavior* (pp. 1l-39). Heidelberg: Springer.

Ajzen, I. (1987). Attitudes, traits, and actions: Dispositional prediction of behavior in personality and social psychology. *Advances in Experimental Social Psychology, 20,* l-63. New York: Academic Press.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50,* 179-211.

Al-Gazali, L., Valian, V., Barres, B., Wu, L., Andrei E. Y., Handelsman, J., … & Husu, L.  (2013). Scientists of the world speak up for equality. *Nature, 495,* 35-38.

Allport, F. H. (1954). The structuring of events: outline of a general theory with applications to psychology. *Psychological Review, 61,* 281-303.

American Association of University Women. (2010). *Why so few? Women in science, technology, engineering, and mathematics.* Washington, D.C: AAUW.

American Association of University Women. (2015). *Solving the equation: The variables for women's success in engineering and computing.* Washington, D.C: AAUW.

Bamberg, S., Ajzen, I., & Schmidt, P. (2003). Choice of travel mode in the Theory of Planned Behavior: The roles of past behavior,

habit, and reasoned action. *Basic and Applied Social Psychology, 25,* 175-187.

Batson, C. D., Chang, J., Orr, R., & Rowland, J. (2002). Empathy, attitudes, and action: Can feeling for a member of a stigmatized group motivate one to help the group? *Personality and Social Psychology Bulletin, 28,* 1654-1666.

Batson, C. D., Early, S., & Salvarani, G. (1997). Perspective taking: Imagining how another feels versus imaging how you would feel. *Personality and Social Psychology Bulletin*, *23*, 751-758.

Behrend, T. S., Sharek D. J., Meade A. W., & Wiebe, E. G. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods, 43,* 800-813.

Bodenhausen, G. V., & Macrae, C. N. (1998). Stereotype activation and inhibition. In R. S. Wyer (Ed.), *Stereotype activation and inhibition: Advances in social cognition* (pp. 1-52). Mahwah, NJ: Erlbaum.

Buhrmester, M., Kwang T., & Gosling S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science, 6,* 3-5.

Cacioppo, J. T., Kao, C. F., Petty, R. E., & Rodriguez, R. (1986). Central and peripheral routes to persuasion: An individual difference perspective. *Journal of Personality and Social Psychology, 15,* 1032-1043.

Capaldi, D., & Paterson G. R. (1987). An approach to the problem of

    recruitment and retention rates for longitudinal research.

    *Behavioral Assessment, 9,* 169-177.

Carnes, M., Devine, P. G., Isaac, C., Manwell, L. B., Ford, C. E., Byars-

    Winston, A., … Sheridan, J. T. (2012). Promoting institutional

    change through bias literacy. *Journal of Diversity in Higher*

    *Education, 5,* 63-77.

Carnes, M., Devine, P. G., Manwell, B. L., Byars-Winston, A., Fine E.,

    Ford, C. E., … Sheridan, J. (2015). The effect of an intervention to

    break the gender bias habit for faculty at one institution: A

    cluster randomized, controlled trial. *Academic Medicine, 90,* 221-

    230.

Carver, C. S., & Harmon-Jones, E. (2009). Anger is an approach-related

    affect: Evidence and implications. *Psychological Bulletin*, *135*,

    183-204.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.*

    Hillside, NJ: Lawrence Erlbaum Associates.

Cundiff, J. L., Zawadzki, M. J., Danube, C. L., & Shields, S. A.

    (2014). Using experiential learning to increase the recognition of

    everyday sexism as harmful: The WAGES intervention. *Journal of*

    *Social Issues, 70,* 703-721.

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of

    automatic attitudes: Combating automatic prejudice with images

of admired and disliked individuals. *Journal of Personality and Social Psychology, 81,* 800-814.

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology, 48,* 1267-1278.

Dobbin, F., Schrage, D., & Kalev, A. (2015). Rage against the iron cage: The varied effects of bureaucratic personnel reforms on diversity. *American Sociological Review*, *80,* 1014-1044.

Eaton, L. G., & Funder, D. C. (2001). Emotional experience in daily life: Valence, variability, and rate of change. *Emotion*, *1*, 413-421.

Ferguson, C., & Ghorayshi, A. (2015). Read this letter from scientists accusing top publisher of sexism. *BuzzFeedNews*. Retrieved from https://www.buzzfeed.com/catferguson/science-journal-reinforces-dangerous-stereotypes?utm_term=.gyq718Ke#.kyOOJLgv

Foschi, M. (1996). Double standards in the evaluation of men and women. *Social Psychology Quarterly*, *59*, 237-254.

Frijda, N. H., Mesquita, B., Sonnemans, J., & Van Goozen, S. (1991). The duration of affective phenomena or emotions, sentiments and passions. *International Review of Studies on Emotions, 1,* 186-225.

Glick, P., & Fiske, S. T. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, *70*, 491-512.

Glick, P., & Fiske, S. T. (1999). The Ambivalence toward Men Inventory: Differentiating hostile and benevolent beliefs about men. *Psychology of Women Quarterly*, *23*, 519-536.

Glick, P., & Fiske, S. T. (2001). An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American Psychologist*, *56*, 109-118.

Godin, G., & Kok, G. (1996). The theory of planned behavior: A review of its applications to health-related behaviors. *American Journal of Health Promotion, 11,* 87-98.

Gordijn, E. H., Yzerbyt, V., Wigboldus, D., & Dumont, M. (2006). Emotional reactions to harmful intergroup behavior. *European Journal of Social Psychology*, *36*, 15-30.

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology, 79,* 701-721.

Green, M. C., Garst, J., Brock, T. C., & Chung, S. (2006). Fact versus fiction labeling: Persuasion parity despite heightened scrutiny of fact. *Media Psychology*, *8*, 267-285.

Handelsman, J., Cantor, N., Carnes, M., Denton, D., Fine, E., …
Sheridan, J. (2005). More women in science. *Science, 309*, 1190-1191.

Handelsman, J., Miller, S., & Pfund, C. (2007). *Scientific Teaching.* New York: Freeman.

Handley, I. M., Brown, E. R., Moss-Racusin, C. A., & Smith, J. L. (2015). The quality of evidence revealing subtle gender biases in science is in the eye of the beholder. *Proceedings of the National Academy of Sciences, 43,* 13201-13206.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* New York: Guilford Press.

Heilman, M. E., Wallen, A. S., Fuchs, D., & Tamkins, M. M. (2004). Penalties for success: Reactions to women who succeed at male gender-typed tasks. *Journal of Applied Psychology*, *89*, 416-427.

Jackson, S. M., Hillard, A. L., & Schneider, T. R. (2014). Using implicit bias training to improve attitudes toward women in STEM. *Social Psychology of Education, 17*, 419-438.

Johnson, D. R., Jasper, D. M., Griffin, S., & Huffman, B. L. (2013). Reading narrative fiction reduces Arab-Muslim prejudice and offers a safe haven from intergroup anxiety. *Social Cognition*, *31*, 578-598.

Kline, R. B. (2011). *Principals and practice of structural equation modeling (3$^{rd}$ ed).* Guilford Press, New York.

Knobloch-Westerwick, S., Glynn, C. J., & Huge, M. (2013). The Matilda effect in science communication: An experiment on gender bias in publication quality perceptions and collaboration interest. *Science Communication*, *35*, 603-625.

Knowles, E. S., Coker, M. C., Scott, R. A., & Cook, D. A. (1996). Measurement-induced improvements in anxiety: Mean shifts with repeated assessment. *Journal of Personality and Social Psychology, 71*, 352-363.

Kollmuss, A., & Agyeman, J. (2002). Mind the gap: why do people act environmentally and what are the barriers to pro-environmental behavior? *Environmental Education Research*, *8*, 239-260.

Lai, C., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. -R. L., Joy-Gaba, J. A., … Nosek, B. A. (2014). Reducing implicit racial preferences: 1. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General, 142*, 1765-1785.

Lane, S. P., & Hennes, E. P. (in press). Power struggles: Estimating sample size for multilevel relationships research. *Journal of Social and Personal Relationships*.

Lariviere, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature*, *504*, 211-213.

Legault, L., Gutsell, J. N., & Inzlicht, M. (2011). Ironic effects of antiprejudice messages: How motivational interventions can reduce (but also increase) prejudice. *Psychological Science, 22*, 1472-1477.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*, 1-23.

Mazziotta, A., Mummendey, A., & Wright, S. C. (2011). Vicarious intergroup contact effects: Applying social-cognitive theory to intergroup contact research. *Group Processes and Intergroup Relations, 14,* 255-274.

McArdle, J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60,* 577-605.

Milkman, K. L., Akinola, M., & Chugh, D. (2012). Temporal distance and discrimination: An audit study in academia. *Psychological Science*, *23*, 710-717.

Milkman, K. L., Akinola, M., & Chugh, D. (2015). What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. *Journal of Applied Psychology, 100,* 1678-1712.

Monteith, M. J., Sherman, J. W., & Devine, P. G. (1998). Suppression as a stereotype control strategy. *Personality and Social Psychology Review, 2*, 63-82.

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, 16474-16479.

Moss-Racusin, C. A., Molenda, A. K., & Cramer, C. R. (2015). Can evidence impact attitudes? Public reactions to evidence of gender bias in STEM fields. *Psychology of Women Quarterly, 39,* 194-209.

Moss-Racusin, C. A., van der Toorn, J., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2014). Scientific diversity interventions. *Science, 343*, 615-616.

Moss-Racusin, C. A., van der Toorn, J., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2016). A "scientific diversity" intervention to reduce gender bias in a sample of life scientists. *CBE-Life Sciences Education*, *15*, 29.

National Science Foundation, National Center for Science and Engineering Statistics. (2017). *Women, minorities, and persons with disabilities in science and engineering: 2017.* Special Report NSF 17-310. Arlington, VA. Retrieved from www.nsf.gov/statistics/wmpd/

Nature. (2012). Nature's sexism. *Nature, 491,* 495.

Nature. (2015). Sexism has no place in science. *Nature, 522,* 255.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, And Practice*, *6*, 101-115.

Okimoto, T. G., & Brescoll, V. L. (2010). The price of power: Power

seeking and backlash against female politicians. *Personality and

Social Psychology Bulletin*, *36*, 923-936.

Ortiz, M., & Harwood, J. (2007). A social cognitive theory approach to

the effects of mediated intergroup contact on intergroup

attitudes. *Journal of Broadcasting and Electronic Media, 41,* 615-

631.

Paluck, E. L. (2006). Diversity training and intergroup contact: A call to

action research. *Journal of Social Issues, 62,* 577-595.

Paluck, E. L. (2009). Reducing intergroup prejudice and conflict using

the media: A field

experiment in Rwanda. *Journal of Personality and Social Psychology,

96,* 574-587.

Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A

review and assessment of research and practice. *Annual Review

of Psychology, 60,* 339-367.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments

on Amazon Mechanical Turk. *Judgment and Decision Making, 5,*

411-419.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact

theory. *Journal of Personality and Social Psychology*, *90*, 751-783.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, *19,* 123-205*.*

Pietri, E. S., Moss-Racusin, C. A., Dovidio, J. F., Guha, D., Roussos, G., Brescoll, V. L., & Handelsman, J. (2017). Using video to increase gender bias literacy toward women in science. *Psychology of Women Quarterly, 41,* 175-196.

Prentice, D. A., & Gerrig, R. J. (1999). Exploring the boundary between fiction and reality. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 529-546). New York: Guilford Press.

Prochaska, J. M., Mauriello, L. M., Sherman, K. J., Harlow, L., Silver, B., & Trubatch, J. (2006). Assessing readiness for advancing women scientists using the transtheoretical model. *Sex Roles, 54,* 869-880.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*, Thousand Oaks, CA: Sage.

Raymond, J. (2013) Sexist attitudes: Most of us are biased. *Nature, 495,* 33-34.

Reuben, E., Sapienza, P., Zingales, L. (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences, 111,* 4403-4408.

Robnett, R. D. (2016). Gender bias in STEM fields: Variation in
    prevalence and links to STEM self-concept. *Psychology of Women
    Quarterly, 40,* 65-79.

Rudman, L. A., & Glick, P. (1999). Feminized management and backlash
    toward agentic women: The hidden costs to women of a kinder,
    gentler image of middle managers. *Journal of Personality and
    Social Psychology, 77,* 1004-1010.

Rudman, L. A., Moss-Racusin, C. A., Glick, P., & Phelan, J. E. (2012).
    Reactions to vanguards: Advances in backlash theory. *Advances
    in Experimental Social Psychology, 45,* 167-227.

Rudman, L. A., & Phelan, J. E. (2008). Backlash effects for disconfirming
    gender stereotypes in organizations. *Research in Organizational
    Behavior, 28,* 61-79.

Sheltzer, J. M., & Smith, J. C. (2014). Elite male faculty in the life
    sciences employ fewer women.  *Proceedings of the National
    Academy of Sciences, 111*, 10107-10112.

Shen, H. (2013). Mind the gender gap. *Nature,* 495: 22-24.

Shields, S. A., Zawadzki, M. J., & Johnson, R. N. (2011). The impact of
    the Workshop Activity for Gender Equity Simulation in the
    Academy (WAGES–Academic) in demonstrating cumulative
    effects of gender bias. *Journal of Diversity in Higher Education*, *4*,
    120-129.

Shrout, P. E., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clavél, F. D., Iida, M., Gleason, M. E. J., Xu, J. H., & Bolger, N. (submitted). Initial elevation bias in subjective reports.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359-1366.

Slater, M. D., Rouner, D., Murphy, K., Beauvais, F., van Leuven, J., & Domenech-Rodriguez, M. M. (1996). Adolescent counterarguing of TV beer advertisements: Evidence for effectiveness of alcohol education and critical viewing discussions. *Journal of Drug Education*, *26*, 143-158.

Smyth, F. L., & Nosek, B. A. (2015). On the gender–science stereotypes held by scientists: Explicit accord with gender-ratios, implicit accord with scientific identity. *Frontiers in Psychology*, *6*, 415.

Sonnert, G., Fox, M. F., & Adkins, K. (2007). Undergraduate women in science and engineering: Effects of faculty, fields, and institutions over time. *Social Science Quarterly, 88,* 1333-1356.

Steele, J., James, J. B., & Barnett, R. C. (2002). Learning in a man's world: Examining the perceptions of undergraduate women in male-dominated academic areas. *Psychology of Women Quarterly, 26,* 46-50.

Steinpreis, R. E., Anders, K. A., & Ritzke, D. (1999). The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles*, *41*, 509-528.

Swim, J. K., Aikin, K. J., Hall, W. S., & Hunter, B. A. (1995). Sexism and racism: Old-fashioned and modern prejudices. *Journal of Personality and Social Psychology*, *68*, 199-214.

Weisgram, E. S., & Bigler, R. S. (2007). Effects of learning about gender discrimination on adolescent girls' attitudes toward and interest in science. *Psychology of Women Quarterly, 31*, 262-269.

Wenneras, C., & Wold, A. (1997). Nepotism and sexism in peer-review. *Nature, 387*, 341-343.

White House, Office of the Press Secretary, The. (2015). *Remarks by the President at White House Science Fair*. Retrieved from https://www.whitehouse.gov/the-press-office/2015/03/23/remarks-president-white-house-science-fair

Williams, W. M., & Ceci, S. J. (2015). National hiring experiments reveal 2:1 faculty preference for women on STEM tenure track. *Proceedings of the National Academy of Sciences*, *112*, 5360-5365.

Williams, J. C., Phillips, K. W., & Hall, E. V. (2016). Tools for change: Boosting the retention of women in the STEM pipeline. *Journal of Research in Gender Studies, 1,* 11-75.

Wood, W. B., & Handelsman, J. (2004). Meeting report: The 2004 National Academies Summer Institute on Undergraduate Education in Biology. *Cell Biology Education, 3,* 215-217.

Zawadzki, M. J., Danube, C. L., & Shields, S. A. (2012). How to talk about gender inequity in the workplace: Using WAGES as an experiential learning tool to reduce reactance and promote self-efficacy. *Sex Roles*, 67, 605-616.

Zawadzki, M. J., Shields, S. A., Danube, C. L., & Swim, J. K. (2014). Reducing the endorsement of sexism using experiential learning: The workshop activity for gender equity simulation (WAGES). *Psychology of Women Quarterly, 38*, 75-92.

**Table 1: Descriptive and Reliability Statistics (Experiments 1 and 2).**

| Variable | Experiment 1 | | | Experiment 2 Baseline | | | Post-Intervention | | | Follow-Up | | | $R_{IF}$ | $R_C$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | α | *M* | *SD* | α | *M* | *SD* | α | *M* | *SD* | α | | |
| Awareness of Gender Bias in STEM | 3.69 | 0.75 | .89 | 3.59 | 0.65 | .87 | 3.78 | 0.65 | .86 | 3.75 | 0.62 | .90 | .89 | .66 |
| Attitudes toward Women in STEM | 4.41 | 0.58 | .84 | 4.70 | 0.36 | .74 | 4.67 | 0.38 | .79 | 4.60 | 0.44 | .80 | .81 | .63 |
| Gender Bias | 2.26 | 0.76 | .90 | 2.11 | 0.53 | .83 | 1.97 | 0.53 | .84 | 2.03 | 0.55 | .86 | .89 | .39 |
| Behavioral Intentions | | | | 3.85 | 0.55 | .89 | 3.87 | 0.55 | .92 | 3.81 | 0.54 | .91 | .94 | .66 |
| Empathy | 3.03 | 1.23 | .88 | 3.11 | 1.01 | .86 | 2.74 | 1.01 | .85 | 2.74 | 1.01 | .85 | .82 | .83 |
| Anger | 1.97 | 1.18 | .92 | 1.52 | 0.73 | .83 | 2.31 | 1.26 | .91 | 1.51 | 0.82 | .91 | .82 | .86 |

*Note. M* indicates grand mean, *SD* indicates standard deviation, α indicates the Cronbach's alpha estimate of reliability, $R_{IF}$ indicates the generalized reliability, and $R_C$ indicates the reliability of change.

**Table 2: Bivariate Correlation Matrix (Experiments 1 and 2).**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Awareness of Bias | | .37*** | -.66*** | .40*** | .35*** | - | - | - | - | - | - | -.15** | .05 | -.25** |
| 2. Attitudes toward Women | .25*** | | -.45*** | .28*** | .11* | - | - | - | - | - | - | -.12* | .07 | -.20*** |
| 3. Gender Bias | -.67*** | -.49*** | | -.34** | -.26** | - | - | - | - | - | - | .33*** | -.04 | .25*** |
| 4. Empathy | .39*** | .17* | -.41*** | | .57*** | - | - | - | - | - | - | -.13* | .11* | -.22*** |
| 5. Anger | .27*** | .10 | -.29*** | .35*** | | - | - | - | - | - | - | -.18** | .05 | -.08 |
| 6. Behavioral Intentions | .53*** | .39*** | -.44*** | .40*** | .09 | | - | - | - | - | - | - | - | - |
| 7. R01 Institution | -.08 | .01 | -.02 | -.01 | .02 | -.01 | | - | - | - | - | - | - | - |
| 8. Pretenure (vs. tenured) | -.08 | -.07 | .06 | .03 | -.02 | -.02 | .02 | | - | - | - | - | - | - |
| 9. Nontenure (vs. tenured) | .05 | -.07 | -.01 | .00 | .03 | -.11 | .08 | -.52*** | | - | - | - | - | - |
| 10. % Women Faculty | -.03 | -.15† | .09 | .07 | -.15* | .04 | -.43*** | .02 | -.02 | | - | - | - | - |
| 11. % Women in Lab | .18† | .21* | -.20* | -.09 | -.05 | .25** | -.36*** | -.13 | -.08 | .11 | | - | - | - |
| 12. Conservatism | -.26*** | -.21** | .41*** | -.19* | -.06 | -.28*** | -.11 | .06 | .02 | .22** | -.11 | | .17*** | -.01 |

| | | | | | | | | -.47** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 13. Age | .02 | -.01 | -.11 | .15* | .01 | .08 | -.07 | * | -.14$^\dagger$ | .00 | .03 | -.08 | -.06 |
| 14. Gender (W = 0, M = 1) | -.22** | -.12 | .11 | -.11 | -.14$^\dagger$ | -.05 | .14$^\dagger$ | .12 | -.16* | -.16* | .05 | .06 | .14$^\dagger$ |

Note: $^\dagger p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.  Experiment 1 values are presented above the diagonal; Experiment 2 (baseline) values are presented below the diagonal.

**Table 3: Means and Standard Errors for Outcome Variables by Condition (Experiment 1).**

| Variable | *Means (Standard Errors)* | | | | |
| --- | --- | --- | --- | --- | --- |
| | *Non-Intervention Control* (A) (N = 100) | *Video Control* (B) (N = 94) | *Narrative* (C) (N = 101) | *Hybrid* (D) (N = 81) | *Expert* (E) (N = 74) |
| Awareness of Gender Bias in STEM | 3.50 (0.07)$^{cde}$ | 3.35 (0.07)$^{cde}$ | 3.83 (0.07)$^{ab}$ | 3.91 (0.08)$^{ab}$ | 3.96 (0.08)$^{ab}$ |
| Attitudes toward Women in STEM | 4.26 (0.06)$^{cde}$ | 4.41 (0.06) | 4.43 (0.06)$^{a}$ | 4.50 (0.06)$^{a}$ | 4.50 (0.07)$^{a}$ |
| Gender Bias | 2.43 (0.08)$^{de}$ | 2.38 (0.08)$^{de}$ | 2.22 (0.07) | 2.15 (0.08)$^{ab}$ | 2.07 (0.09)$^{ab}$ |
| Empathy | 2.41 (0.12)$^{cde}$ | 2.31 (0.13)$^{cde}$ | 3.70 (0.11)$^{abe}$ | 3.47 (0.13)$^{ab}$ | 3.28 (0.14)$^{abc}$ |
| Anger | 1.39 (0.11)$^{cde}$ | 1.19 (0.11)$^{cde}$ | 2.87 (0.11)$^{abde}$ | 2.53 (0.13)$^{abce}$ | 1.96 (0.12)$^{abcd}$ |

*Note.* Least square means estimates. Superscripts indicate which conditions are significantly different ($p < .05$). For instance, for awareness of bias, the non-intervention control differs significantly from the narrative (condition C), hybrid (condition D), and expert (condition E) conditions.

**Table 4: List of Effect Sizes (Experiments 1 and 2).**

| | | *Effect Sizes (d; vs. video control condition)* | | | | | |
|---|---|---|---|---|---|---|---|
| *Variable* | *Time* | *Experiment 1* | | | *Experiment 2* | | |
| | | *Narrative* | *Hybrid* | *Expert* | *Narrative* | *Hybrid* | *Expert* |
| Awareness of Gender Bias in STEM | Post-Intervention | 0.65 | 0.75 | 0.82 | 0.55 | 0.96 | 1.16 |
| | Follow-Up | - | - | - | 0.64 | 0.98 | 1.13 |
| Attitudes toward Women in STEM | Post-Intervention | 0.03 | 0.15 | 0.15 | 0.00 | 0.36 | 0.37 |
| | Follow-Up | - | - | - | 0.00 | 0.36 | 0.14 |
| Gender Bias | Post-Intervention | 0.21 | 0.31 | 0.41 | 0.45 | 0.74 | 0.93 |
| | Follow-Up | - | - | - | 0.42 | 0.60 | 0.73 |
| Empathy | Post-Intervention | 1.13 | 0.94 | 0.79 | 1.44 | 1.52 | 1.25 |
| | Follow-Up | - | - | - | 0.71 | 0.62 | 0.54 |
| Anger | Post-Intervention | 1.43 | 1.14 | 0.65 | 1.45 | 1.37 | 1.04 |
| | Follow-Up | - | - | - | 0.40 | 0.49 | 0.68 |
| Behavioral Intentions | Post-Intervention | - | - | - | 0.33 | 0.50 | 0.58 |
| | Follow-Up | - | - | - | 0.51 | 0.36 | 0.36 |

*Note.* Effect sizes were estimated by subtracting the condition mean from the video control mean and dividing by the omnibus standard deviation (Experiment 1) or the standard deviation of the change score at that time point (Experiment 2). Effect sizes (*d*) indicate the number of standard deviations separating the experimental condition from the video control condition. $d = 0.2$ is considered a small effect size, $d = 0.5$ is considered a medium effect size, and $d = 0.8$ is considered a large effect size (Cohen, 1988).

| Item Number | Item Wording |
|---|---|
| 1 | I will provide more career development and research support for my female students |
| 2 | I will make sure I professionally mentor female students in the sciences |
| 3 | I will provide more emotional support my female students |
| 4 | I will take special care to personally mentor female students in the sciences |

| 5 | I will create an environment that ensures both female and male students feel welcome in my classroom |
|---|---|
| 6 | I will create an environment that ensures both female and male students feel welcome in my discipline |
| 7 | I will help recruit and hire more female research assistants |
| 8 | I will help to recruit more female students |
| 9 | I will put pressure on the university to create policies to recruit and hire more female students |
| 10 | I will put pressure on the university to create and support women in science mentoring programs |
| 11 | I will put pressure on the university to create and support retention initiatives for women in science |
| 12 | I plan to learn about and find resources to increase women's representation in the sciences |
| 13 | I would benefit from approaches and trainings aimed at increasing women's representation in the sciences |

**Table 5: Behavioral Intentions Scale (Experiment 2).**

*Note.* Responses were indicated on a scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*).

**Table 6: Multilevel Regression Models for Change in Outcome Variables Relative to Baseline (Experiment 2).**

| | Awareness of Gender Bias in STEM | | Attitudes toward Women in STEM | | Gender Bias | | Empathy | | Anger | | Behavioral Intentions | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *b* | *SE* | *b* | *SE* | *b* | *SE* | *b* | *SE* | *b* | *SE* | *b* | *SE* |
| Intercept | -0.18* | 0.08 | -0.10† | 0.05 | 0.04 | 0.06 | -1.57*** | 0.16 | -0.46** | 0.17 | -0.11* | 0.06 |
| Baseline | -0.40*** | 0.05 | -0.44*** | 0.07 | -0.22*** | 0.04 | -0.54*** | 0.06 | -0.49*** | 0.08 | -0.24*** | 0.04 |
| Time | -0.01 | 0.07 | -0.05 | 0.05 | -0.01 | 0.06 | 0.79*** | 0.17 | 0.13 | 0.17 | -0.02 | 0.05 |
| Narrative | 0.30** | 0.10 | -0.02 | 0.07 | -0.16* | 0.08 | 1.99*** | 0.22 | 1.83*** | 0.24 | 0.12 | 0.08 |
| Hybrid | 0.52*** | 0.10 | 0.12† | 0.07 | -0.27*** | 0.08 | 2.10*** | 0.21 | 1.73*** | 0.23 | 0.18* | 0.08 |
| Expert | 0.63*** | 0.11 | 0.13† | 0.08 | -0.34*** | 0.08 | 1.73*** | 0.22 | 1.31*** | 0.24 | 0.21* | 0.08 |
| Narrative*Time | 0.04 | 0.09 | -0.02 | 0.07 | 0.03 | 0.08 | -1.34*** | 0.23 | -1.51*** | 0.24 | 0.06 | 0.07 |
| Hybrid*Time | 0.00 | 0.09 | 0.02 | 0.07 | 0.08 | 0.08 | -1.54*** | 0.23 | -1.33*** | 0.23 | -0.05 | 0.06 |
| Expert*Time | -0.03 | 0.09 | -0.07 | 0.07 | 0.12 | 0.08 | -1.23*** | 0.24 | -0.76** | 0.24 | -0.08 | 0.07 |
| $S^2_{Intercept}$ | 0.11 | 0.02 | 0.08 | 0.01 | 0.04 | 0.01 | 0.26 | 0.07 | 0.27 | 0.07 | 0.07 | 0.01 |
| $S^2_{ResidualTime1}$ | 0.08 | 0.02 | 0.02 | 0.01 | 0.06 | 0.01 | 0.54 | 0.09 | 0.69 | 0.10 | 0.04 | 0.01 |
| $S^2_{ResidualTime2}$ | 0.06 | 0.01 | 0.06 | 0.01 | 0.04 | 0.01 | 0.37 | 0.08 | 0.24 | 0.06 | 0.03 | 0.01 |

*Note.* Condition was dummy coded with the video control as the reference condition. Time was coded with post-intervention = 0 and follow-up = 1. All models estimated a random person intercept and a variance components residual covariance matrix. The unstandardized regression coefficient is indicated by *b*, and the standard error of the estimate is indicated by *SE*. †$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$

**Table 7: Means and Standard Errors for Outcome Variables by Condition (Experiment 2).**

| Variable | Time Point | Mean Change from Baseline (Standard Errors) | | | |
|---|---|---|---|---|---|
| | | Control (A) (N = 37) | Narrative (B) (N = 38) | Hybrid (C) (N = 41) | Expert (D) (N = 36) |
| Awareness of Gender Bias in STEM | Post-Intervention | -0.18 (0.08)*bcd | 0.12 (0.07)acd | 0.34 (0.07)*ab | 0.44 (0.07)*ab |
| | One Week Follow-Up | -0.19 (0.07)*bcd | 0.15 (0.07)*ad | 0.33 (0.06)*a | 0.41 (0.07)*ab |
| Attitudes toward Women in STEM | Post-Intervention | -0.11 (0.05) | -0.13 (0.05)*cd | 0.02 (0.05)b | 0.02 (0.05)b |
| | One Week Follow-Up | -0.15 (0.06)* | -0.19 (0.06)*c | -0.01 (0.06)b | -0.10 (0.07)t |
| Gender Bias | Post-Intervention | 0.04 (0.06)bcd | -0.13 (0.05)*ad | -0.23 (0.05)*a | -0.30 (0.05)*ab |
| | One Week Follow-Up | 0.03 (0.05)cd | -0.10 (0.05)* | -0.16 (0.05)*a | -0.19 (0.05)*a |
| Empathy | Post-Intervention | -1.57 (.16)*bcd | 0.42 (0.15)*a | 0.53 (0.14)*a | 0.15 (0.15)a |
| | One Week Follow-Up | -0.78 (0.14)*bcdt | -0.14 (0.13)at | -0.22 (0.13)at | -0.29 (0.14)*at |
| Anger | Post-Intervention | -0.45 (0.17)*bcd | 1.38 (0.17)*ad | 1.27 (0.16)*a | 0.86 (0.17)*ab |
| | One Week Follow-Up | -0.32 (0.12)*cd | 0.00 (0.12)t | 0.08 (0.11)at | 0.23 (0.13)at |
| Behavioral Intentions | Post-Intervention | -0.12 (0.06)*cd | 0.00 (0.05) | 0.06 (0.06) a | 0.09 (0.05)a |
| | One Week Follow-Up | -0.13 (0.05)*b | 0.05 (0.05)a | -0.01 (0.06) | -0.01 (0.05)t |

*Note.* Least square mean estimates. Superscripts indicate which conditions are significantly different ($p < .05$). For instance, for awareness of bias at the post-intervention time point, the control differs significantly from the narrative (condition B), hybrid (condition C), and expert (condition D) conditions. *t* superscripts at follow-up indicate significant decline within condition from post-intervention. * indicates significant change from baseline.

**Figure 1: Bias Outcomes among the General Population (Experiment 1)**
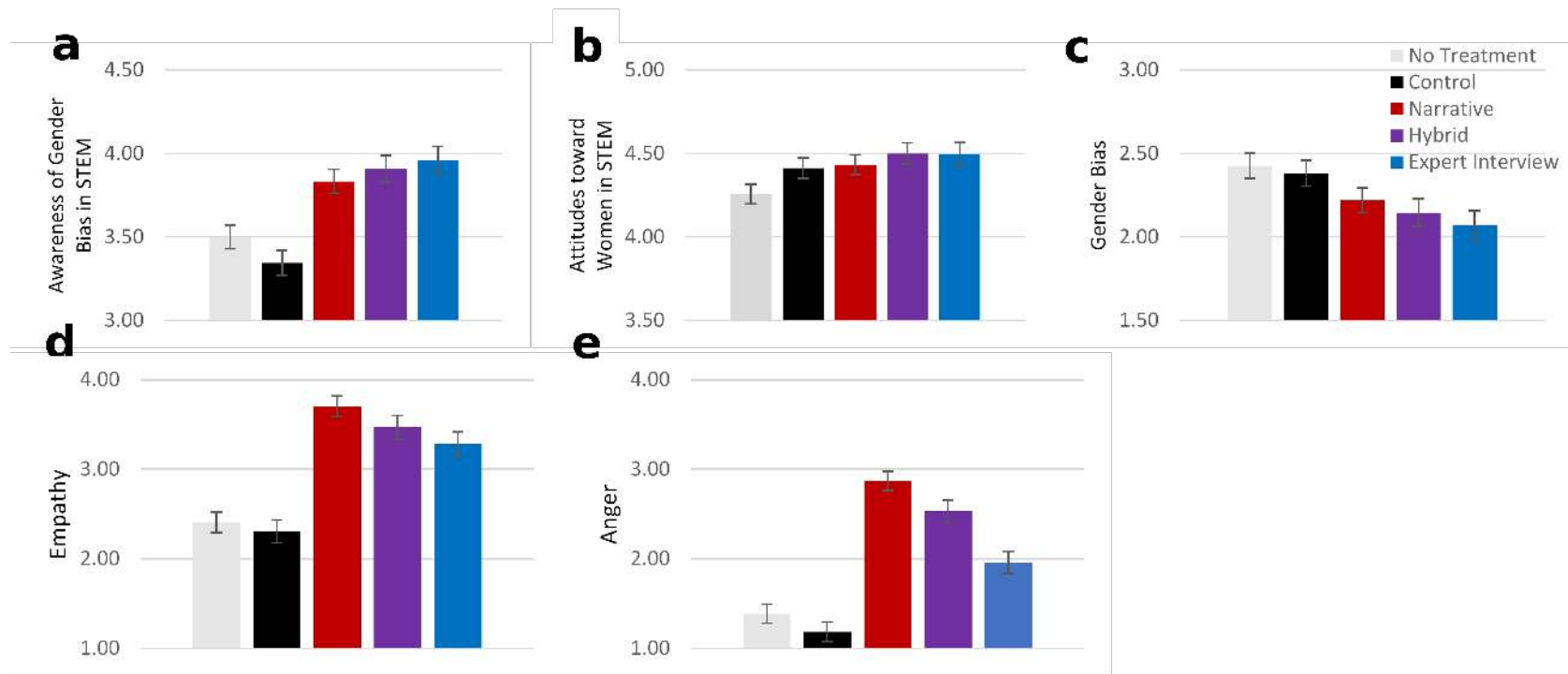


*Note.* Post-intervention awareness of gender bias (a), attitudes toward women in STEM (b), gender bias (c), empathy (d), and anger (e), by condition. Error bars indicate standard errors. "No treatment" refers to the non-intervention control condition, while "control" refers to the video control condition.

**Figure 2: Change in Bias Outcomes among STEM Faculty (Experiment 2)**
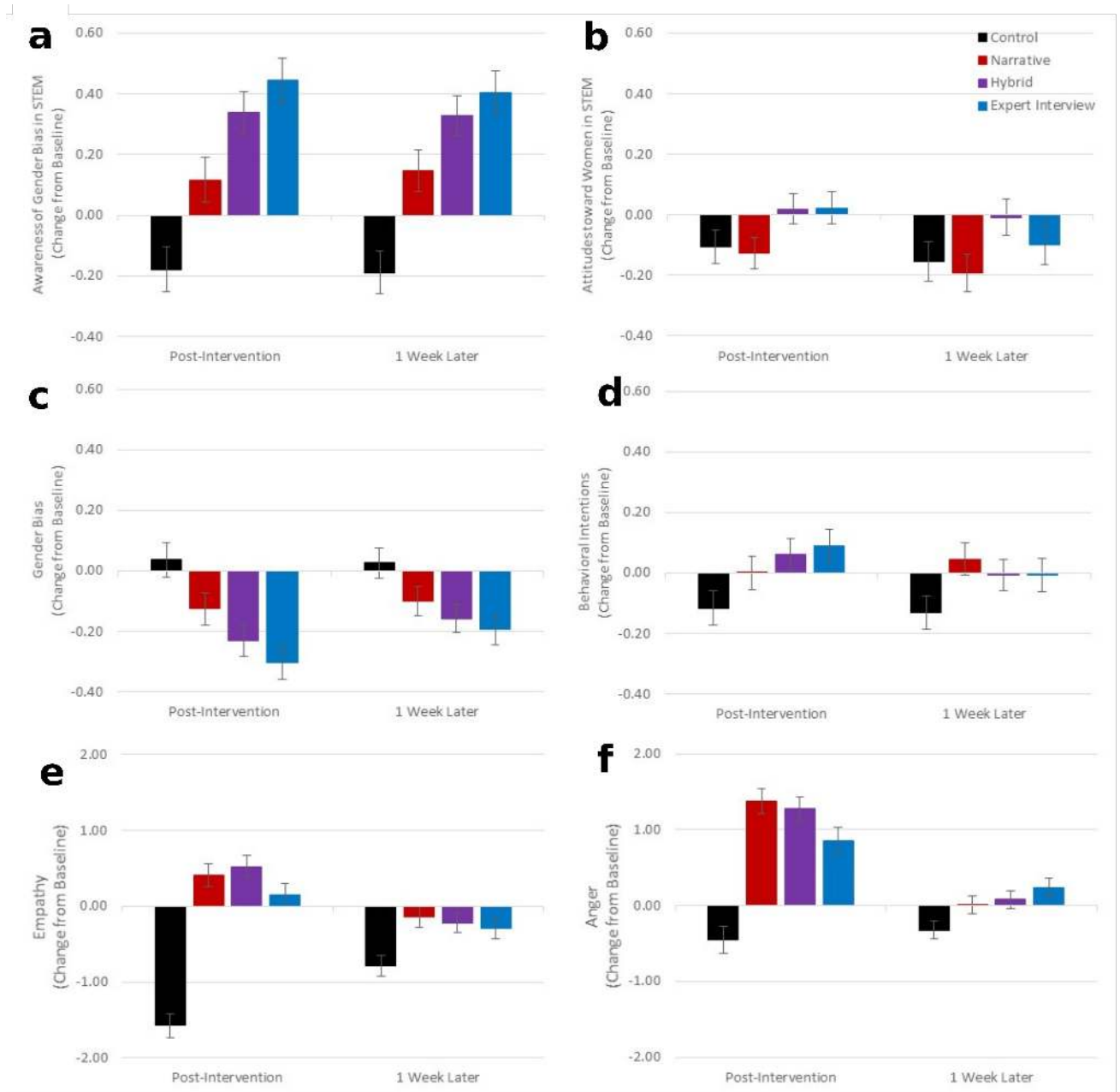


*Note.* Change in awareness of gender bias (a), attitudes toward women in STEM (b), gender bias (c), behavioral intentions (d), empathy (e), and anger (f), relative to baseline by condition. Worsening outcomes over time in the video control condition reflect established attenuation effects (i.e., a response bias toward inflated self-report at Time 1). Error bars indicate standard errors.

**Figure 3: Indirect Effects of Empathy on Behavioral Intentions (Experiment 2)**



*Note:* Coefficient estimates are unstandardized. Bolded paths indicate significant primary results. Solid lines indicate additional significant effects. Dotted lines indicate nonsignificant effects. **$p < .01$, ***$p < .001$

**Figure 4: Indirect Effects of Anger on Behavioral Intentions (Experiment 2)**



*Note:* Coefficient estimates are unstandardized. Bolded paths indicate significant primary results. Solid lines indicate additional significant effects. Dotted lines indicate nonsignificant effects. *p < .05, **p < .01, ***p < .001