Organizational Behavior and Human Decision Processes 152 (2019) 105–121

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

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

Updating impressions: The differential effects of new performance information on evaluations of women and men☆

Madeline E. Heilman[a,*], Francesca Manzi[a], Suzette Caleo[b]

[a] Department of Psychology, New York University, United States
[b] E.J. Ourso College of Business, Louisiana State University, United States

ARTICLE INFO

Keywords:
Lack of fit
Gender stereotypes
Gender bias
Competence perceptions
Performance evaluation

ABSTRACT

In three experimental studies we investigated whether changes in performance would have different consequences on the competence perceptions and performance evaluations of women and men whose earlier performance had been unmistakably successful or unsuccessful. We reasoned that the ambiguity created by new performance information that was inconsistent with previous performance information would facilitate stereotype-based gender bias. The results provided support for this idea. Whereas no differences emerged between reactions to men and women when performance remained the same, differences emerged when performance changed. Moreover, regardless of the nature of the change in performance, in male gender-typed domains women were evaluated more negatively than men: an improvement in performance had a less beneficial effect for women than for men (Study 1) and a decline in performance had a more detrimental effect for women than for men (Study 2). These effects were shown to be moderated by the gender-type of the field. Women were evaluated more negatively than men whether performance improved or declined only when the field was male gender-typed; when the field was female gender-typed, men were evaluated more negatively than women (Study 3). These findings are consistent with the idea that gender stereotypes and the performance expectations they produce can influence responses to new information about men's and women's performance.

1. Introduction

Gender stereotypes have been identified as possible impediments to women's and men's progress in traditionally gendered occupations, and research has focused on the nature of these stereotypes and the role they play in biasing evaluations of men's and women's performance (Heilman, 2012). An important and often overlooked limitation of this work lies in its focus on one-shot performance evaluation – participants are typically asked to evaluate men and women after receiving information of their performance at one point in time. In most settings, however, performance is ongoing and reviews occur repeatedly, giving evaluators the opportunity to revise their past judgments as new information becomes available (Heslin, Latham, & VandeWalle, 2005). Understanding how subsequent information about performance affects competence perceptions and performance evaluations is critical to expanding our knowledge about the reach of gender stereotypes and the processes underlying their influence.

The studies reported here extend earlier work by exploring how the performance expectations produced by gender stereotypes shape responsiveness to new information about performance. We propose that fluctuations in performance give rise to ambiguity, thereby fueling the influence of stereotype-based expectations. Specifically, we investigate how negative expectations about women's performance in traditionally male domains affect the rigidity or elasticity of competence perceptions, and the degree to which views of their competence are relinquished and views of their incompetence are retained when their performance changes over time. In addition, we address the question of whether stereotype-based expectations about men in traditionally female domains have analogous consequences.

1.1. Gender stereotypes and lack of fit perceptions

Evidence suggests that despite changes in workforce participation and educational attainment, stereotypic beliefs about men and women persist (Haines et al., 2016). Gender stereotypes depict women as relationship-focused, intuitive, emotionally sensitive, and concerned about others (communal), but not as task-focused, analytic, assertive, or independent (agentic), attributes that are ascribed to men. These beliefs

☆ This work was supported by the National Science Foundation ADVANCE Research Challenge Grant [grant number HRD-0820202].

* Corresponding author at: 6 Washington Place, New York University, Department of Psychology, New York, NY 10003, United States.
E-mail address: madeline.heilman@nyu.edu (M.E. Heilman).

https://doi.org/10.1016/j.obhdp.2019.03.010
Received 28 June 2016; Received in revised form 13 March 2019; Accepted 25 March 2019
Available online 24 May 2019

about what women and men are like are shared across cultures (Williams & Best, 1990), and can be found in work, social, and domestic settings (Eagly & Wood, 2011; Heilman, Block, & Martell, 1995; Schein, 2001). Moreover, they are pervasive in their effects. Because gender is such a salient feature and is readily noticed and remembered (Fiske, Haslam, & Fiske, 1991), it serves as a powerful cue for stereotypic thinking (Blair & Banaji, 1996). Even without people's awareness of their impact, gender stereotypes tend to dominate in impression formation, functioning as heuristics that guide the processing of information (Banaji, Hardin, & Rothman, 1993; Macrae, Milne, & Bodenhausen, 1994).

Gender stereotypes create problems due to a perceived "lack of fit" between what women and men are thought to be like and the presumed requirements of certain tasks or roles (Heilman, 1983, 2001, 2012). Research suggests that this is particularly problematic for women seeking careers in traditionally male fields – fields that not only are numerically dominated by men but also are considered to require "male" attributes for success (Cejka & Eagly, 1999; Gaucher, Friesen, & Kay, 2011; Johnson, Murphy, Zewdie, & Reichard, 2008; Koenig, Eagly, Mitchell, & Ristikari, 2011; Smyth & Nosek, 2015). These male gender-typed fields are various, including science and technology, finance, and upper level management, and are thought to necessitate an achievement-oriented aggressiveness, emotional toughness, and a cognitive style that is antithetical to the stereotyped view of what women are like. Though examined less frequently, research also shows that men can be affected by gender stereotypes in traditionally female fields (Davison & Burke, 2000; Eagly, Karau, & Makhijani, 1995; Ko, Kotrba, & Roebuck, 2015; Paustian-Underdahl, Walker, & Woehr, 2014). These so-called pink-collared fields, including health care, social services, and education, are thought to require empathy, interpersonal acuity, and a relationship focus that is antithetical to the stereotyped view of what men are like (Cejka & Eagly, 1999).

The perceived incongruity between stereotypic attributes and gender-typed task requirements leads to the conclusion that women and men are not equipped to handle the work in fields that are thought to be gender-inconsistent. It also produces the expectation that if they were to engage in such pursuits they would be unlikely to succeed. This is the essence of the Lack of Fit Model (Heilman, 1983, 2001, 2012), which further specifies that the negative performance expectations deriving from lack of fit perceptions create a predisposition to see people as incompetent – a predisposition that forms the basis of gender bias in many traditionally gender-typed settings.

1.2. *Stereotype-based expectations and performance evaluation*

Stereotype-based expectations about performance in gender-typed fields not only are powerful, but also are tenacious. Expectations perpetuate themselves through cognitive distortion and affect the way in which information is attended to, processed, and remembered (Fyock & Stangor, 1994; Kunda, Sinclair, & Griffin, 1997; Pittinsky, Shih, & Ambady, 2000; Robbins & DeNisi, 1993; Scott & Brown, 2006). The acceptance of expectation-disconfirming information necessitates a restructuring of beliefs, and the easiest response is to reject it. Thus, expectations often create self-fulfilling prophecies that allow evaluators to see precisely what they expect to see. Consequently, stereotype-based performance expectations produced by lack of fit perceptions are likely to have considerable effects on how women and men are regarded and their performance evaluated.

An important body of research attests to the consequences of negative performance expectations for gender bias in performance evaluation. Researchers have repeatedly shown the difficulty women face in being viewed as competent in traditionally male pursuits (e.g., Biernat, Fuegen, & Kobrynowicz, 2010; Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012). For example, women tend to receive lower performance ratings than men in positions thought of as male in gender-type (e.g., Davison & Burke, 2000; Lyness & Heilman, 2006). Similarly, research has shown that when men are successful in female gender-typed positions, they are judged to be more ineffective than equally performing women (e.g., Heilman & Wallen, 2010; Ko et al., 2015). But negative performance expectations also are apt to have more subtle and less immediate effects on performance evaluation. Because they create a predisposition toward perceptions of incompetence and encourage the dismissing of disconfirming evidence, they are likely to influence responsiveness to new information about performance.

Previous investigations of gender bias in performance evaluation have tended to document reactions to women and men when given one-time information about performance. Though these studies provide valuable information about the ways in which stereotypes affect evaluations, they fail to reflect the context of most performance evaluation settings. Indeed, performance is a dynamic process where fluctuation – large or small – is typically the norm (Cappelli & Conyon, 2016; Zyphur, Chaturvedi, & Arvey, 2008). Under these conditions, performance evaluations occur repeatedly, and judgments are likely to vary in conjunction with changing performance. To fully understand the effect of stereotype-based expectations on performance evaluation, it is necessary to examine how and when perceptions of competence change (or do not change) in light of new performance information.

1.3. *The role of ambiguity*

Central to our ideas is a consideration of ambiguity and its role in facilitating gender bias. By necessitating inference, and thus leaving room for multiple interpretations, ambiguity enables evaluators to distort their judgments in line with expectations, serving to maintain and potentially reinforce them. Indeed, it is when ambiguity is high that stereotype-based expectations have been found to exert their strongest effect, with greater disparities emerging in the evaluation of men and women (Koch, D'Mello, & Sackett, 2015). Conversely, when ambiguity is low, stereotype-based expectations have been shown to have little effect. Specifically, bias in performance ratings has not been found when a target's credentials are either very strong or very weak, leaving little room for debate about performance quality (Biernat & Vescio, 2002; Dovidio & Gaertner, 2000; Hodson, Dovidio, & Gaertner, 2002), when the measures used to assess performance are objective rather than subjective, limiting alternative interpretations (Haslam, Ryan, Kulich, Trojanowski, & Atkins, 2010; Heilman & Haynes, 2008; Jirjahn & Stephan, 2004), or when performance quality has been validated (Heilman, Wallen, Fuchs, & Tamkins, 2004). Under such conditions, the unequivocal nature of the information prevents stereotype-based expectations from influencing evaluation.

However, evaluators are rarely presented with such clear-cut information, and this realization has produced many attempts to reduce or eliminate ambiguity in the evaluation process. These strategies, which focus on the amassing and reporting of performance information, aim to limit the opportunity for inference that enables stereotype-based expectations and facilitates gender bias (Heilman & Caleo, 2018). But eliminating ambiguity from the evaluation process is not easily accomplished. Research attests to the various ways in which ambiguity can prevail in evaluative contexts, emerging when information is impoverished, vague, or irrelevant, and when evaluative criteria are idiosyncratic, poorly defined, or not amenable to standardized measurement (Heilman & Haynes, 2008). These findings highlight the importance of providing clear and unequivocal information about performance quality, and well-delineated evaluative criteria, if gender bias is to be averted.

Yet, even when ambiguity is expunged from performance information, it can prevail when ongoing performance is evaluated over time. Lack of clarity and objectivity are not the only sources of ambiguity in performance evaluation; another key source of ambiguity is lack of consistency (Hodson et al., 2002). Information imparted at one time period, no matter how clear and unequivocal, can create ambiguity if it conflicts with information imparted at another time period. Thus, unless sequential performance information is constant, it is likely to produce ambiguity and be facilitative of gender bias. In short, the emergence of new information that is inconsistent with the previous instance creates conditions that enable bias by leaving room for inference, allowing stereotype-based expectations to reshape evaluations over time.

How are evaluations likely to be affected? Given the extraordinary pull of expectations, it is likely that when ambiguity is high because performance information over time is inconsistent, information that is aligned with expectations will be more determinative of evaluations than information that is not. This idea is consistent with research indicating that it takes only a minor indication of expectation-consistent information for an expectation to be fulfilled (Brescoll, Dawson, & Uhlmann, 2010; Dovidio & Gaertner, 2000). It also is consistent with research demonstrating that existing expectations can affect the way managers adapt to new information about their employees' performance, profoundly affecting subsequent ratings (Heslin et al., 2005; Manzoni & Barsoux, 1998). Thus, even when gender bias is not evident in the beginning of an evaluation process, gender-based expectations may inadvertently seep into subsequent evaluations when performance changes.

1.4. Anticipated evaluative reactions to changes in performance

We propose that evaluations that are based on initial information that is a poor fit with stereotype-based expectations are unstable and will be revised in favor of new information that is a better fit when the opportunity arises for revision of initial views. We further propose that evaluations that are based on initial information that is a good fit with stereotype-based expectations will be resistant to change when the opportunity for revision arises. We initially test these ideas within male-dominated fields, focusing our hypotheses on women.

We anticipate that when women have been successful in male gender-typed fields, perceptions of them as competent are likely to be fragile and more easily relinquished in the face of subsequent negative performance information than those of equivalently performing men. Additionally, we anticipate that when their performance in male gender-typed fields is initially unsuccessful, views of women as incompetent are likely to be more recalcitrant and less amenable to revision when confronted with additional positive performance information than will views of men. We therefore expect women to be at a disadvantage relative to men in a male gender-typed field when sequential performance is inconsistent – regardless of whether it declines or improves – with competence perceptions inexorably pulled towards consistency with stereotype-based negative performance expectations, and performance evaluations following suit.

Because the lack of fit model holds that negative performance expectations arise from a mismatch between gender stereotypic conceptions and the requirements thought necessary for success in certain tasks or roles, these expectations should not arise for women in areas that are not male in gender-type. However, they should arise for men in areas that are female gender-typed. When men's performance is inconsistent with stereotype-based expectations, it, too, should produce competence perceptions that move toward stereotype-consistent perceptions when there is the opportunity to revise original perceptions, and when it is consistent with stereotype-based expectations it should produce competence perceptions that are resistant to change. Thus, when a field is female in gender-type, we anticipate men to be disadvantaged more than women when their performance worsens over time and to benefit less than women when their performance improves. As with women in male gender-typed domains, the expected tendency is for competence perceptions and performance evaluations to move toward conformity with stereotype-based expectations.

1.5. Overview of research

In the following three studies, we sought to determine if and when additional performance information has different consequences on the competence perceptions and performance evaluations of women and men whose earlier performance has been clearly successful or unsuccessful. We used an academic context to test our ideas. Participants evaluated a male or female student whose performance had either changed or stayed the same over two separate and independent units of a college course in a gender-typed field. In order to isolate the unique effects of ambiguity produced by sequential assessments, we chose a type of performance information (grades) that, on its own, would be unambiguous in its implications about performance quality, and therefore unlikely to be susceptible to distortion from stereotype-driven expectations.

We expected differences in the extent to which people would revise their initial judgments when the additional information introduced ambiguity by conflicting with the initial performance information. Because lack of success in male gender-typed fields is gender-consistent for women, we expected improved performance in a course in a male gender-typed field to have a *less beneficial effect* on the competence perceptions and performance evaluations of previously unsuccessful women than men (Study 1). Additionally, because success in male gender-typed fields is gender-inconsistent for women, we expected worsened performance to have *a more detrimental effect* on the competence perceptions and performance evaluations of previously successful women than men (Study 2). In neither of these two studies did we expect differences in competence perceptions and performance evaluations of male and female targets when the performance information was constant over time and ambiguity was not created by the additional information. In a third study, we tested our assertion about the role of perceived lack of fit and the performance expectations it produces by varying the gender-type of the academic field of study and also examining the effects of inconsistent performance information on evaluations of men. When provided with subsequent performance information that was inconsistent with the initial performance information, we expected more negative responses to women than to men only when the field of study was male gender-typed, and we expected more negative responses to men than to women when the field of study was female gender-typed (Study 3).

2. Study 1

In Study 1, we provided research participants with sequential performance information. We sought to test the idea that improved performance in a male gender-typed endeavor will have a generally beneficial effect on competence perceptions and performance evaluations, but its effect will be less beneficial for previously unsuccessful women than for previously unsuccessful men. Moreover, because information constancy does not produce ambiguity, we expected only negligible differences in reactions to men and women when performance was consistently poor over the two time periods. We therefore hypothesized:

Hypothesis 1.1. *All targets will be viewed as more competent and their performance evaluated more positively when their previously unsuccessful performance has improved than when it remains at the same level.*

Hypothesis 1.2. *There will be a difference in reactions to male and female targets when performance has improved, but not when it has stayed the same. Specifically, when their performance has improved, female targets will be viewed as less competent and their performance evaluated less positively than male targets.*

2.1. Method

2.1.1. Participants and design

203[1] undergraduates (126 female, 76 male, 1 undetermined; mean age 19.62) from a large northeastern university were recruited through the student subject pool. Three additional participants completed the study but were excluded from analysis after incorrectly responding to a manipulation check. The study was a 2 × 2 between-subjects design, with gender of target (male or female) and performance change (no-change or improvement) as the two independent variables. No other conditions were run. Participants were randomly assigned to one of the four experimental conditions.

2.1.2. Procedure

The study was said to be about student performance and evaluation. Participants were told that we were interested in examining how the format and structure of an academic course can affect impressions of students. They were informed that they would be reviewing performance information about a student who had taken a semester-long course divided into two units.

Participants read a brief paragraph of background information about a college student taking an introductory computer science class to fulfill a science requirement. The student was reported to have grown up in the southwest, to be in the junior year of college, and to enjoy reading and sports. The full description provided to participants is included in Appendix A. Computer science was chosen based on previous research suggesting that the field continues to be strongly male-dominated (Cheryan, Ziegler, Montoya & Jiang, 2017). Indeed, a recent report from the National Center for Educational Statistics estimates that men earn over 80% of undergraduate degrees in computer science (NCES, 2015). Importantly, computer science is also perceived by students and the general population to be highly male in gender-type (see Cheryan, Master & Meltzoff, 2015).

Participants then received the syllabus for the class (*Programming I*). In designing the syllabus, it was important to make clear that the course was divided into two independent course units and that performance in the two course units would be assessed separately, each having its own project and exam. We did so to ensure that we could adequately manipulate performance change without implying that performance in the second unit was contingent on performance in the first unit and, therefore given different weight. We also wanted to avoid potential attributions of attenuated or amplified motivation for the change in performance from Unit 1 to Unit 2, and to prevent participants from making the assumption that the course instructor could have used information from Unit 1 when assigning Unit 2 grades. Appendix B provides a full description of the course syllabus.

The target's project and exam grades for each course unit were presented separately. After reviewing the project and exam grades for Unit 1, participants were asked to indicate how well the target had performed in that unit, and then received the project and exam grades for Unit 2. After reviewing the information about the target's course performance in Unit 2, participants responded to a brief questionnaire, giving their reactions to the student, a final grade for the course and, to keep consistent with our cover story, their opinions of the course structure. At this point, participants were thanked for their participation and debriefed.

2.1.3. Experimental manipulations

Gender of target. The gender of the student was varied by the names (Patricia or Thomas Olsen) on the stimulus materials and gender-relevant pronouns used in the background information. In addition, to strengthen our manipulation and ensure that gender was attended to, there was a portrait photo attached to the background information paragraph depicting either a White male or White female college-aged student.

Both names and photos were selected with the intent of minimizing confounding factors. In choosing the target names, we first referred to existing research, finding the names Patricia and Thomas to be rated comparably in attractiveness, age, competence, and race connotation (Kasof, 1993). Photos were selected from the Radboud Faces Database (www.rafd.nl), which contains photos of 20 White male and 19 White female models. Using data from Langner et al. (2010), we selected a set of male and female photos that not only were matched on emotional expression and gaze, but also rated comparably on attractiveness, valence, and clarity of facial expression. As a final check, we ran a pre-test to ensure that the photographs were also matched in perceived intelligence ($M_{female} = 5.30$, $SD_{female} = 1.11$; $M_{male} = 5.20$, $SD_{male} = 1.21$), $t(28) = 0.32$, $p = .76$, age ($M_{female} = 25.27$, $SD_{female} = 2.66$; $M_{male} = 25.13$, $SD_{male} = 4.12$), $t(28) = 0.11$, $p = .92$, and attractiveness ($M_{female} = 5.40$, $SD_{female} = 1.12$; $M_{male} = 5.60$, $SD_{male} = 0.82$), $t(28) = 0.58$, $p = .58$.

Performance change. We operationalized performance change through the project and exam grades presented for each course unit. The grades for Unit 1 (B-, C+) were the same in all conditions; they were selected on the basis of pre-testing and were uniformly viewed by potential participants in our sample to be "unsuccessful". The grades for Unit 2 varied according to the condition. In the no-change condition, Unit 2 grades were the same as Unit 1 grades, but reversed in order (C+, B-). In the improvement condition, Unit 2 grades were markedly better than Unit 1 grades (A, A-).

2.1.4. Measures

There were two dependent variables: competence perceptions and performance evaluations. Competence perceptions were measured with a perceived competence scale and performance evaluations were measured by final grade assignments. All of the questionnaire items comprising the measures for each dependent variable, the manipulation and stimulus checks, and the demographic questions are presented in Appendix C.

Perceived competence. The perceived competence scale was created by combining ratings on three 7-point bipolar adjective scales (incompetent-competent, not smart-smart, ineffective-effective) and a rating of how competent the student was in the field of study (1 = not at all competent, 7 = very competent) ($\alpha = 0.90$).

Grade assignment. Participants were asked to assign a final grade for the student they reviewed using grades ranging from A to F, including pluses (+) and minuses (-). The unit grades that we provided in each performance condition precluded the assignment of a final grade based on sheer averaging – there was no grading option that corresponded to the true average of the grades in the two units. The true grade average fell between B- and C+ in the no-change conditions and between B and B+ in the improvement conditions. Consequently, in asking for a final grade we forced participants to make a choice – they had to either "round up" and assign a grade above the true average, or "round down" and assign a grade below the true average. Our interest was in the frequency with which the assigned grade was above or below the true average between Unit 1 and Unit 2 grades.

Manipulation and stimulus checks. To check on the target gender manipulation, we asked participants to write the name of the student reviewed. To check the performance manipulation, participants were asked to report whether the student's performance had: (a) improved throughout the course, (b) stayed the same throughout the course, or (c) declined throughout the course. In addition, participants' ratings of

[1] An *a priori* power analysis based on the average effect size from similar previous studies ($\eta_p^2 = 0.087$) indicated that 96 participants were needed to obtain 85% power for detecting the expected medium effect at $\alpha = 0.05$. At the reviewers' and editor's request, we increased the original sample size to approximately 50 participants for each of the four experimental conditions. All of the significant effects reported were also significant at the 0.05 level prior to the additional data collection.

the computer science course on a male-female scale (1 = male-oriented, 7 = female-oriented) provided a check on the male gender-typing of the field, and participants' reports of how the target performed in the first unit (1 = poor, 7 = excellent) provided a check on whether the original performance was seen as unsuccessful, as intended, by participants in all conditions.

2.2. Results

2.2.1. Preliminary analyses

Results indicated that our target gender and performance manipulations were successful. All participants correctly reported the name of the student and, with only three exceptions (excluded from further analyses[2]), all participants correctly depicted the target's performance as having improved or as having stayed the same throughout the course. Moreover, participants' ratings of the computer science course confirmed that they viewed it as male gender-typed ($M = 3.19$, $SD = 0.92$) and as significantly different from the midpoint (4) of the "male–female" scale, $t(202) = 12.57$, $p < .001$. Lastly, ratings confirmed that participants in all conditions saw the target's initial performance as unsuccessful (overall $M = 3.23$, $SD = 0.96$) and significantly lower than the midpoint (4) of the 7-point "poor-excellent" scale, $t(117) = 9.60$, $p < .001$.

Analyses testing for differences between male and female participants indicated no main effects or interactions involving participant gender on any of the manipulation checks or dependent measures. Their data were therefore combined for all subsequent analyses.

2.2.2. Dependent measures

We conducted a 2 × 2 analysis of variance (ANOVA) on the perceived competence measure and used follow-up pairwise comparisons to directly test our hypotheses. We used chi-square tests to analyze the distribution of the final grades that participants assigned to the targets.

Perceived competence. ANOVA of the perceived competence ratings yielded a significant main effect of performance change, $F(1, 199) = 175.98$, $p < .001$, $\eta_p^2 = 0.469$, indicating that all targets were seen as more competent when their performance improved ($M = 5.15$, $SD = 0.73$) than when it did not ($M = 3.72$, $SD = 0.83$). There also was a main effect of target gender, $F(1, 199) = 4.49$, $p = .04$, $\eta_p^2 = 0.022$, whereby men ($M = 4.57$, $SD = 1.12$) were rated as more competent than women ($M = 4.33$, $SD = 0.98$). As expected, the interaction between performance change and target gender also was significant, $F(1, 199) = 4.72$, $p = .03$, $\eta_p^2 = 0.023$. Pairwise comparisons indicated that although there was no significant difference between competence ratings of male targets ($M = 3.71$, $SD = 0.84$) and female targets ($M = 3.73$, $SD = 0.83$) when performance had not changed, $t(199) = 0.04$, $p = .97$, there was a significant difference in ratings of male and female targets when performance had improved. As predicted, with an improvement in performance the male target ($M = 5.38$, $SD = 0.66$) was rated as significantly more competent than the female target ($M = 4.92$, $SD = 0.72$), $t(199) = 3.08$, $p = .002$, $d = 0.44$. Fig. 1 presents the mean competence ratings for male and female targets in the no-change and improvement conditions.

Grade assignment. Chi-square tests indicated that, as anticipated, higher grades were assigned more often in the improvement conditions than in the no-change conditions, χ^2 (1, N = 202) = 186.38, $p < .001$. To test our hypotheses about the differences in grades assigned to male and female targets, we divided the grades in each performance change condition into groupings of below and above the true average of the grades across the two course units: either "C+ or lower" or "B− or higher" in the no-change conditions, and either "B or lower" or "B+ or higher" in the improvement conditions. Chi-square tests indicated support for the hypothesized differences in grades assigned to male and female targets for each of the performance change conditions. As shown in Table 1, results indicated no significant difference in the distribution of final grades below and above the true average assigned to male and female targets in the no-change condition, χ^2 (1, N = 99) = 0.43, $p = .51$. However, when performance had improved, there was a significant difference in the distribution of the grades as a result of target gender; although their performance was identical, male targets were assigned the higher grade significantly more frequently than were female targets, χ^2(1, N = 104) = 10.09, $p < .001$.[3]

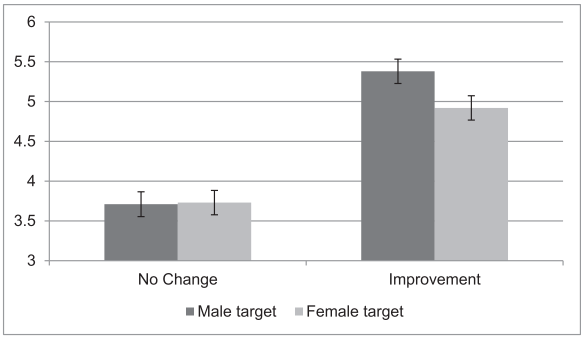


Fig. 1. Study 1: Perceived competence of male and female targets in the no-change and improvement conditions.

2.3. Discussion

Results from Study 1 demonstrated that additional performance information had differing effects for men and women only when the new information was inconsistent with the initial information, supporting the idea that it is ambiguity and the latitude it creates for stereotype-based expectations that produces these differences. As predicted, improvement in performance had a generally positive impact, but its impact was less beneficial for women than for men both in competence perceptions and grades assigned. This result is consistent with the idea that expectations about women influence receptivity to new performance information, and that initial information about performance that is consistent with these expectations is more resistant to interpretative revision.

3. Study 2

In Study 2, we tested our ideas about the effects of changes in performance when the change is negative – when performance begins successfully and then declines. We were interested in examining whether women would be more detrimentally affected than men when their performance in a male gender-typed field has worsened. In this case, we expected that when initial performance is inconsistent with expectations (that is, a woman being clearly successful in a male gender-typed field), the original evaluation would be more amenable to revision. Thus, we expected women's competence to be judged more harshly and their performance to be evaluated more negatively than men's when their performance has similarly declined. As in Study 1, participants were exposed to sequential information about male and female targets' performance in a computer science course and were asked to rate their competence and assign a final grade. However, in this study, performance in Unit 1 was always highly successful and performance in Unit 2 either stayed the same as Unit 1 or declined. We expected a decline in

[2] Analyses including all 206 participants did not change the results reported below. Results for these analyses are reported in Appendix D.

[3] We also analyzed the data treating grade assignment as a continuous measure. These analyses revealed the same pattern of results. The grades assigned to female and male targets did not significantly differ in the no-change condition, $t(97) = 0.66$, $p = .51$, but did differ in the improvement condition, with male targets receiving higher grades than female targets, $t(102) = 2.85$, $p = .005$.

Table 1
Study 1: Frequencies (and percentages) of grades above and below the true average of Unit 1 and Unit 2 grades in the no-change and improvement conditions.

	No change (n = 99)		Improvement (n = 104)	
	Below the average	Above the average	Below the average	Above the average
Male target	10 (20.4%)	39 (79.6%)	14 (26.9%)	38 (73.1%)
Female target	13 (26.0%)	37 (74.0%)	30 (57.7%)	22 (42.3%)

performance to have a generally negative effect on competence perceptions and performance evaluations but expected that worsened performance would be more detrimental for women than for men. Moreover, we again expected only negligible differences in reactions to men and women when performance was consistent – in this case consistently excellent – over the two time periods.

Hypothesis 2.1. *All targets will be viewed as more incompetent and their performance evaluated more negatively when their previously successful performance has declined than when it remains at the same level.*

Hypothesis 2.2. *There will be a difference in reactions to male and female targets when performance has declined, but not when it has stayed the same. Specifically, when their performance has declined, female targets will be viewed as more incompetent and their performance evaluated more negatively than male targets.*

3.1. Method

3.1.1. Participants and design

199[4] undergraduate students (129 female, 69 male, 1 undetermined; mean age 19.88) from a large northeastern university were recruited from the subject pool. Six additional participants completed the study but were excluded from analysis after incorrectly responding to a manipulation check. The study used a 2 × 2 between-subjects design, with gender of target (male or female) and performance change (no-change or decline) as the two independent variables. No other conditions were run. Participants were randomly assigned to one of the four experimental conditions.

3.1.2. Procedure and experimental manipulations

The procedure was largely the same as that of Study 1. Identical names, background information, and photos were used to manipulate target gender, and participants again received separate information about the student's performance in two independent units of a computer science course, as depicted in Appendices A and B. However, we presented different project and exam grades in order to operationalize consistently good or declining performance. In this study, Unit 1 grades were always highly successful (A, A-), and Unit 2 grades either stayed the same (A-, A) or declined (B-, C+). As before, participants completed a final questionnaire and were debriefed and thanked for their participation.

3.1.3. Measures

The questionnaire was the same as that used in Study 1 (see Appendix C), and the same items were used to compose the perceived competence scale ($\alpha = 0.88$). The range for the final grade also was the same (A to F), and once again we set up our unit grades so that there was no final grade available that was the true average of the grades in the two units; rounding up or rounding down was necessary. The manipulation and stimulus checks also were the same as those used in Study 1.

3.2. Results

3.2.1. Preliminary analyses

All participants wrote down the correct name for the student they were rating, indicating that our target gender manipulation was successful. Our performance change manipulation also had its intended effect. With six exceptions (excluded from subsequent analyses[5]) participants responded consistently with condition, correctly reporting that the student's performance had "declined throughout the course" or "stayed the same throughout the course." Ratings of the computer science course again confirmed that participants viewed it as male gender-typed; the mean of participants' ratings ($M = 3.17$, $SD = 0.99$) differed significantly from the midpoint (4) of the "male-female" scale, $t(197) = 11.74$, $p < .001$. Finally, ratings confirmed that in all conditions the targets' initial performance was seen as similarly successful ($M = 6.53$, $SD = 0.50$) and significantly higher than the midpoint (4) of the "poor-excellent" scale, $t(198) = 71.24$, $p < .001$.

Analyses testing for differences between male and female participants indicated no significant main effects or interactions for any of our manipulation checks or dependent measures. Data from both male and female participants were therefore combined for all analyses.

3.2.2. Dependent measures

We used the same data analysis strategy as in Study 1, conducting a 2 × 2 ANOVA and pairwise comparisons on the perceived competence measure, and chi-square tests to analyze the distribution of the grades assigned to male and female targets.

Perceived competence. A two-way ANOVA of the perceived competence scale ratings yielded a significant main effect of performance change, $F(1, 195) = 233.13$, $p < .001$, $\eta_p^2 = 0.545$, indicating that all targets were rated as less competent when their performance declined ($M = 4.72$, $SD = 0.81$) than when it stayed consistently positive ($M = 6.30$, $SD = 0.70$). Results also indicated a significant main effect of target gender, $F(1, 195) = 7.90$, $p = .005$, $\eta_p^2 = 0.039$, with male targets receiving more positive competence ratings ($M = 5.64$, $SD = 0.95$) than female targets ($M = 5.31$, $SD = 1.22$). These effects were qualified by a significant interaction between performance change and target gender, $F(1, 195) = 10.46$, $p = .001$, $\eta_p^2 = 0.051$. Pairwise comparisons showed that, as predicted, there was no significant difference in the competence ratings of male ($M = 6.28$, $SD = 0.81$) and female targets ($M = 6.32$, $SD = 0.57$) when their performance had not changed, $t(195) = 0.29$, $p = .77$, but the female target ($M = 4.40$, $SD = 0.88$) was rated as significantly less competent than the male target ($M = 5.03$, $SD = 0.61$) when performance had declined, $t(195) = 4.35$, $p < .001$, $d = 0.62$. The mean competence ratings for male and female targets in the no-change and decline conditions are presented in Fig. 2.

Grade assignment. Analyses indicated that, as would be expected, targets received lower grades more frequently when their performance declined than when it stayed successful, χ^2 (1, N = 197) = 197.00, $p < .001$. Again, we divided the grades in the performance conditions into groupings of above and below the true average of the grades in the two units: either "A" or "A- or lower" in the no-change conditions, and either "B+ or higher" or "B or lower" in the decline conditions. Results of chi-square tests indicated support for our hypothesized differences in final grade assignments (see Table 2). They indicated no significant difference in the final grade assignment for male and female targets in the no-change condition, $\chi^2(1, N = 95) = 1.05$, $p = .31$, but a

[4] An *a priori* power analysis was conducted by incorporating the effect size from Study 1 into the previously used average effect size. The analysis indicated that 112 participants were needed to obtain 85% power for detecting the estimated effect size ($\eta_p^2 = 0.077$) at $\alpha = 0.05$. At the request of the editor and reviewers, we increased the original sample size to approximately 50 participants for each of the four experimental conditions. All of the significant effects reported were also significant at the 0.05 level prior to the additional data collection.

[5] Analyses including all 205 participants did not change the results (see Appendix D).

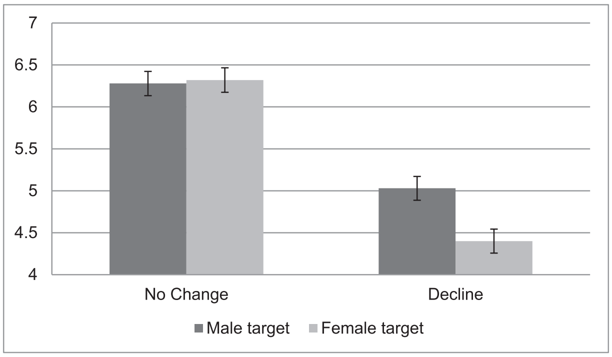


Fig. 2. Study 2: Perceived competence of male and female targets in the no-change and decline conditions.

Table 2
Study 2: Frequencies (and percentages) of grades above and below the true average of Unit 1 and Unit 2 grades in the no-change and decline conditions.

	No change (n = 95)		Decline (n = 103)	
	Below the average	Above the average	Below the average	Above the average
Male target	14 (28.6%)	35 (71.4%)	31 (59.6%)	21 (40.4%)
Female target	9 (19.6%)	37 (80.4%)	43 (84.3%)	8 (15.7%)

significant difference in the distribution of grades as a result of target gender when performance had declined, with female targets assigned the lower grade more frequently than male targets, $\chi^2(1, N = 103) = 7.77$, $p = .005$.[6]

3.3. Discussion

The results once again indicate that only when new information was inconsistent with the initial information provided were there differing reactions to men and women, lending additional support to the importance of ambiguity as a facilitator of gender bias. Furthermore, although a decline in performance in male gender-typed domains negatively affected all targets, it had even more detrimental consequences for women than for men in resulting competence perceptions and assigned course grades. Thus, regardless of whether the new and contradictory information was negative, as in this study, or positive as in the first study, reactions to the changed performance were more adverse for women. These findings are consistent with our assertion that stereotype-based expectations deriving from lack of fit perceptions influence the way updates in performance information are integrated into subsequent impression formation. Study 3 was designed to provide a more direct test of this assertion.

4. Study 3

Studies 1 and 2 showed that in a male gender-typed field, whether performance improved or declined, there were more negative effects on evaluations of women than of men: women were less benefitted than men by an improvement in performance and women were more harmed than men by a decline in performance. In Study 3, we sought to replicate the findings from Studies 1 and 2, and also to demonstrate that the origin of the gender disparity we found lies in expectations about performance that derive from the perceived lack of fit between women's attributes and male gender-typed task requirements. If the greater negativity directed toward women is in fact driven by lack of fit perceptions, then the differential appraisal pattern we have demonstrated should not occur when the field of study is not male in gender-type. Moreover, if perceptions of lack of fit between gendered attributes and gender-inconsistent task requirements are to blame for the differential updating of competence perceptions about women in male gender-typed domains, then an analogous pattern should be observed when evaluators are exposed to a male target whose performance changes in a female gender-typed domain.

In Study 3, we used a different male gender-typed field to ensure that our effects replicate beyond the specific course description and discipline used in the previous studies. We also included a female gender-typed field of study and presented information of either improving or declining performance of a male or a female student. As in Studies 1 and 2, we expected that a change in performance would promote reactions that were more negative for women than for men when the field was male in gender-type. Furthermore, we expected that these reactions would be more negative for men than for women when the field was female in gender-type. We hypothesized:

Hypothesis 3.1. *When their performance changes in a male gender-typed field, female targets will be viewed as more incompetent and their performance evaluated more negatively than male targets; this will occur whether performance has improved or declined.*

Hypothesis 3.2. *When their performance changes in a female gender-typed field, male targets will be viewed as more incompetent and their performance evaluated more negatively than female targets; this will occur whether performance has improved or declined.*

We also included a second measure of perceived competence in this study to demonstrate that our results reflect the actual updating of initial competence impressions when new information about performance is provided, and not merely the residue of initial differences in impressions. The measure, a projection of future success in similar courses, was collected twice – once after information was provided about the first unit's performance and again after information was provided about the second unit's performance. We thereby could measure actual change in competence perceptions and see if the degree of change differed for male and female targets in gender-inconsistent domains. We expected that:

Hypothesis 3.3. *When their performance changes in a male gender-typed field, the change in ratings of likely future success will be more negative for women than men.*

Hypothesis 3.4. *When their performance changes in a female gender-typed field, the change in ratings of likely future success will be more negative for men than for women.*

4.1. Method

4.1.1. Participants and design

491[7] undergraduate students (308 female, 176 male, 7 undetermined; mean age 19.19) were recruited from a large northeastern university. Three additional participants completed the study but were excluded from analysis after incorrectly responding to a manipulation check. The study was a 2 × 2 × 2 between-subjects design, with gender

[6] Again, analyses treating grades as a continuous measure indicated an identical pattern of results. There was no significant difference in the grades assigned in the no-change condition, $t(93) = 1.02$, $p = .31$, but male targets were assigned significantly higher grades than female targets in the decline condition, $t(101) = 3.85$, $p < .001$.

[7] As in previous studies, we conducted an *a priori* power analysis to determine the sample size needed. Because Study 3 included a third independent variable, we conducted this analysis on the basis of a small effect size and 85% power. The analysis indicated that 442 participants were needed to detect an effect size of $\eta_p^2 = 0.02$ at $\alpha = 0.05$. Data from additional male participants were collected after reaching the desired sample size to ensure a more balanced distribution of male and female participants per condition.

of target (male or female), performance change (decline or improvement), and gender-type of field (male or female) as the three independent variables. No other conditions were run. Participants were randomly assigned to one of the eight experimental conditions.

4.1.2. Procedure

The procedure closely paralleled that used in Studies 1 and 2, with a few exceptions. First, although participants again received information about a male or female student's performance in two independent units of a course, we varied the gender-type of the course. Second, we did not include a no-change condition, but varied both improvement and decline in performance. Lastly, we included a new measure designed to examine change in perceptions of competence.

4.1.3. Experimental manipulations

Gender of target and performance change. The manipulation of target gender was identical to that used in the earlier studies (see Appendix A). Indicators of improvement and decline were the same as those used in Study 1 and Study 2, respectively. In the improvement conditions, grades for the project and exam went from unsuccessful (B-, C+) in Unit 1 to successful (A, A-) in Unit 2. The reverse was true in the decline conditions; grades for the project and exam went from successful (A, A-) in Unit 1 to unsuccessful (B-, C+) in Unit 2.

Gender-type of field. We manipulated gender-type of field by varying the discipline in which the course was situated. We chose physics as the male-typed field and early childhood education as the female-typed field because both fields continue to be disproportionately dominated by one gender (Cheryan et al., 2017; Croft, Schmader, & Block, 2015). According to the National Center for Education Statistics, over 80 percent of bachelor's degrees in physics are awarded to men, and over 90 percent of bachelor's degrees in early childhood education are awarded to women (NCES, 2015). Because our sample consisted of college students, we reasoned that these disciplines would be perceived by participants to be strongly gender-typed. Furthermore, in a preliminary study, we verified that there were differences in performance expectations for the female and male student in each of the two courses. Respondents were presented with the stimulus materials from one of our experimental conditions (without performance information) and asked to indicate how they expected Karen/Brian to perform in the course on a seven-point scale ranging from "very poorly" to "very well". Results indicated a significant interaction between discipline and target gender for expected course performance, $F(1, 402) = 8.49$, $p = .004$. Specifically, expectations for Karen ($M = 4.77$, $SD = 1.02$) were lower than expectations for Brian ($M = 5.06$, $SD = 0.85$) in the physics course, $t(402) = 2.15$, $p = .03$, and expectations for Brian ($M = 5.34$, $SD = 1.01$) were lower than expectations for Karen ($M = 5.60$, $SD = 0.94$) in the early childhood education course $t(402) = 1.96$, $p = .05$.

In Study 3, participants were again presented with a course syllabus – used in this study to convey whether the course focused on physics or early childhood education. Both syllabi were designed to closely parallel each other in format and wording and, as before, the syllabi indicated that the course was divided into two independent units. Appendix B includes the two syllabi used in this study.

4.1.4. Measures

We used the same questionnaire as in our earlier studies. The perceived competence scale ($\alpha = 0.80$), the range for the assigned final grade (A to F), and stimulus and manipulation checks for target gender and performance were the same as in the earlier studies (see Appendix C). Participants' ratings of the course on the male-oriented (1) – female-oriented (7) scale were used to check the gender-type of field manipulation.

Change in projected likelihood of success. To measure change in projected likelihood of success, we asked participants to assess how successful they thought the student would be in other physics (early childhood education) courses and did this after performance information about each course unit was presented. Responses were on a 7-point scale, the endpoints of which were "not at all successful (1)" and "very successful (7)". A likelihood of success change score was created by subtracting the rating of likelihood of success after reviewing Unit 1 from the rating of likelihood of success after reviewing Unit 2 (Unit 2 – Unit 1).

4.2. Results

4.2.1. Preliminary analyses

Manipulation checks indicated that both the target gender and performance manipulations were successful. All participants reported the student's correct name. With three exceptions (excluded from subsequent analyses[8]), participants in the improvement condition indicated that "the student's performance improved throughout the course" and participants in the decline condition indicated that "the student's performance declined throughout the course".

Participants' ratings also indicated that the gender-type of field manipulation had its intended effect. ANOVA demonstrated that the physics course ($M = 3.33$, $SD = 0.91$) was seen as significantly more male-oriented than the early childhood education course ($M = 4.70$, $SD = 0.93$), $F(1, 480) = 275.14$, $p < .001$, $\eta_p^2 = 0.364$. Additional analyses of the mean difference between participants' gender-typing ratings and the scale midpoint (4.0) yielded further information about the gender-typing of the two fields of study. The analyses confirmed that the physics course was perceived by participants as male-typed: the mean score of ratings for the physics course on the "male-female" scale was significantly lower than the scale midpoint (4), $t(243) = 11.52$, $p < .001$. The opposite pattern was observed for the early childhood education course, where the mean score of ratings on the "male–female" scale was significantly higher than the scale midpoint (4), $t(243) = 11.80$, $p < .001$.

As in earlier studies, ratings confirmed that targets' initial performance in the improvement conditions was regarded as unsuccessful ($M = 3.29$, $SD = 0.96$) and significantly lower than the midpoint (4) of the 7-point "poor-excellent" scale, $t(241) = 11.53$, $p < .001$, and targets' initial performance in the decline conditions was regarded as successful ($M = 6.47$, $SD = 0.59$) and significantly higher than the midpoint of the 7-point "poor-excellent" scale, $t(248) = 66.31$, $p < .001$.

Analyses including participant gender revealed only one significant effect: an interaction between discipline, target gender, and participant gender for the change in perceived likelihood of success measure, $F(1, 468) = 8.56$, $p = .004$, $\eta_p^2 = 0.018$. Though the pattern of results was the same among male and female participants, men were more extreme in their judgements than women; that is, they evaluated targets more harshly when performance changed in a gender-inconsistent field. The results for the analysis of the change scores did not vary after adjusting for participant gender; therefore, the responses of male and female participants were combined in the results reported below.

4.2.2. Dependent measures

We conducted a 2 × 2 × 2 ANOVA and pairwise comparisons to analyze the perceived competence ratings, and chi-square tests to test our hypotheses about the final grades. A 2 × 2 × 2 ANOVA and pairwise comparisons also were conducted on the likelihood of success change scores.

Perceived competence. We had hypothesized that with a change in performance female targets would be rated as less competent than male targets when the field was male gender-typed but that male targets would be rated as less competent than female targets when the field was female gender-typed, and we expected this to happen regardless of whether their performance had improved or declined. A three-way ANOVA on perceived competence ratings yielded a significant main

[8] Analyses including all 494 participants did not differ from the results reported below (see Appendix D).

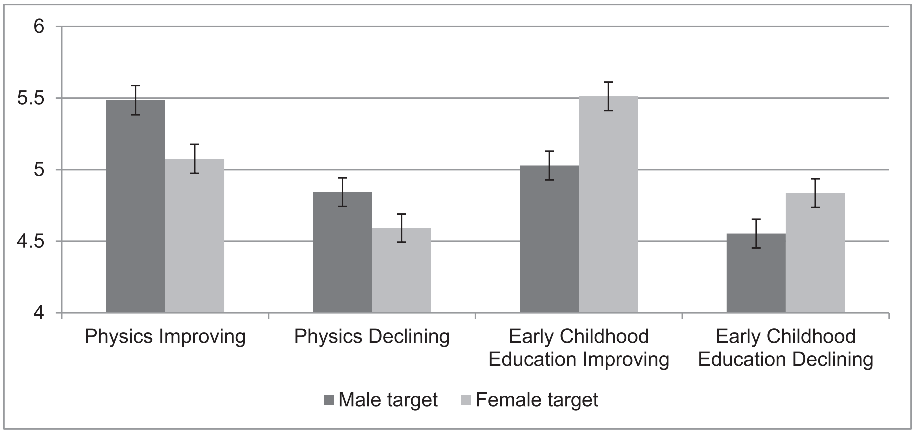


Fig. 3. Study 3: Perceived competence of male and female targets whose performance improved or declined in physics and early childhood education.

effect for performance, $F(1, 483) = 64.55, p < .001, \eta_p^2 = 0.118$, with improving performance ($M = 5.28$, $SD = 0.81$) receiving higher competence ratings than declining performance ($M = 4.71$, $SD = 0.79$). It also revealed a significant interaction between target gender and gender-type of field, $F(1, 483) = 25.32, p < .001, \eta_p^2 = 0.050$.

Pairwise comparisons revealed that, as hypothesized, across performance change conditions female targets were perceived as significantly less competent ($M = 4.83$, $SD = 0.86$) than male targets ($M = 5.16$, $SD = 0.83$) in the physics course, $t(483) = 3.30, p = .001, d = 0.30$; that is, the female target was seen as less competent than the male target when performance had improved, and more incompetent than the male target when performance had declined. Also as predicted, male targets were perceived as significantly less competent ($M = 4.79$, $SD = 0.78$) than female targets ($M = 5.17$, $SD = 0.86$) in the early childhood education course, $t(483) = 3.83$, $p < .001$, $d = 0.35$; this was the case whether performance had improved or declined. Fig. 3 illustrates the pattern of mean competence ratings for male and female targets in all conditions.

Grade assignment. We again broke our final grades into groupings of above and below the true average of the grades in the two course units (in this case "B+ and above" or "B or below" for all conditions). Chi-square tests indicated that, as predicted, for the physics course, female targets were assigned the lower grade more frequently than male targets, whatever the nature of the change in performance, χ^2 (1, N = 245) = 14.21, $p < .001$[9]. That is, when performance had improved, female targets were less often given the grade that was above the true average than were male targets, and when performance had declined, female targets were more often given the grade below the true average than were male targets. This pattern reversed when the field was female in gender-type. In the early childhood education course, male targets were assigned the lower grade more frequently than female targets, regardless of whether performance had improved or declined, χ^2 (1, N = 246) = 10.99, $p = .001$. In this case, male targets received a grade that was above the true average less often than female targets when their performance improved, and male targets received a grade that was below the true average more often than female targets when their performance declined. Table 3 presents the grade frequencies for male and female targets in the physics and education courses.

Change in projected likelihood of success. An ANOVA of the likelihood of success change scores yielded a significant main effect of performance, $F(1, 483) = 1342.71$, $p < .001$, $\eta_p^2 = 0.735$, with improving performance leading to positive change ($M = 1.68$, $SD = 1.17$) and declining performance leading to negative change ($M = -1.91$, $SD = 1.03$). It also indicated a significant interaction between target gender and gender-type of field, $F(1, 483) = 16.68$, $p < .001, \eta_p^2 = .033$.[10]

As expected, follow-up tests indicated that when performance in the physics course changed, female targets' likelihood of success change scores ($M = -0.31$, $SD = 2.09$) were more negative than those of male targets ($M = 0.06$, $SD = 2.17$), $t(483) = 2.58$, $p = .01$, $d = 0.24$, with their ratings increasing less than the males' ratings when performance had improved and decreasing more than the males' ratings when performance had declined. However, with performance change in the early education course, it was male targets' likelihood of success change scores that were more negative ($M_{female} = 0.07$, $SD_{female} = 2.07$; $M_{male} = -0.38$, $SD_{male} = 2.08$), $t(483) = 3.20$, $p = .002$, $d = 0.29$; their ratings increased less than those of female targets when performance had improved and decreased more than those of female targets when performance had declined. Fig. 4 depicts the likelihood of success change scores for male and female targets in each condition.

4.3. Discussion

The results of Study 3 both replicated and extended the findings from the first two studies. Again, we found women to be more negatively affected by changes in performance than were men when the field was male in gender-type: improvement was shown to have a less beneficial effect and decline to have a more adverse effect on women's than men's perceived competence and assigned grades. Furthermore, these results occurred in a different male gender-typed field than the previous studies, lending support to the idea that it is the perceived gender inconsistency of a field, not its particular nature, that gives rise to the differential updating process. These differential effects for women and men reversed when the course was female in gender-type, supporting the idea that they are the result of perceived attribute/task requirement lack of fit, and not a general tendency to perceive women as incompetent or negatively evaluate them. In fact, the findings from Study 3 are consistent with the idea that the stereotype-based expectations that derive from lack of fit perceptions have symmetrical effects – affecting not only the updating of competence perceptions regarding women, but also the updating of competence perceptions regarding men, in gender-inconsistent contexts.

Lastly, and very importantly, the likelihood of success results

[9] Once again, when we treated grade assignments as a continuous measure, analyses were consistent with these results. Female students were assigned significantly lower grades than male students in the physics course, $t(243) = 3.19$, $p = .002$, and male students were assigned significantly lower grades than female students in the early childhood education course, $t(244) = 3.29$, $p = .001$.

[10] An ANCOVA of the likelihood of success scores including participant gender as a covariate also yielded a significant main effect of performance, $F(1, 475) = 1334.97$, $p < .001$ and a significant interaction between discipline and target gender, $F(1, 475) = 16.11$, $p < .001$.

Table 3
Study 3: Frequencies (and percentages) of grades above and below the true average of Unit 1 and Unit 2 grades in the physics and early childhood education courses when performance improved or declined.

	Physics				Early childhood education			
	Improve (n = 119)		Decline (n = 126)		Improve (n = 123)		Decline (n = 123)	
	Below the average	Above the average	Below the average	Above the average	Below the average	Above the average	Below the average	Above the average
Male target	16 (27.1%)	43 (72.9%)	36 (58.1%)	26 (41.9%)	27 (44.3%)	34 (55.7%)	47 (77.0%)	14 (23.0%)
Female target	31 (51.7%)	29 (48.3%)	52 (81.2%)	12 (18.8%)	13 (21.0%)	49 (79.0%)	36 (58.1%)	26 (41.9%)

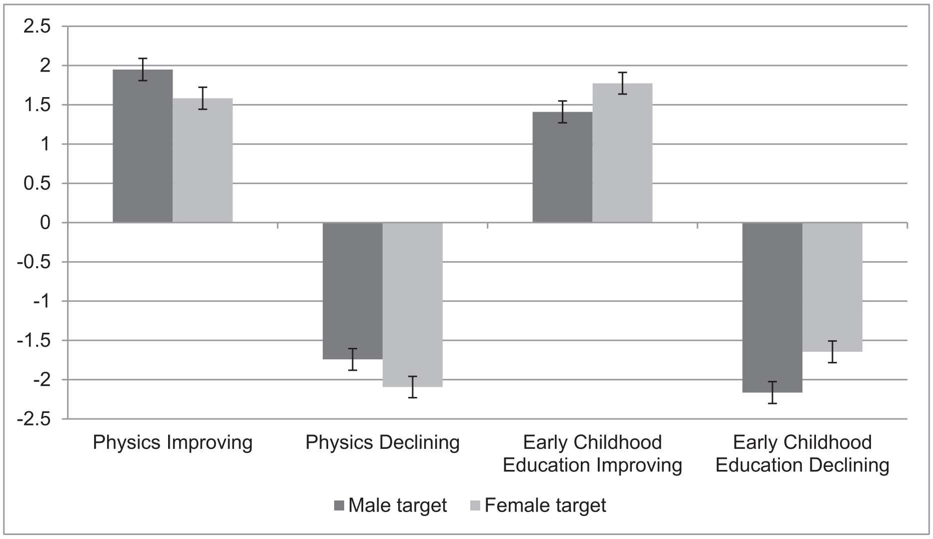


Fig. 4. Study 3: Change in projected likelihood of success of male and female targets whose performance improved or declined in physics and early childhood education.

demonstrated that success projections were altered when new information was presented, supporting the idea that the pattern of competence perceptions we found after evidence of improvement or decline was indicative of a revision in earlier views. The fact that these changes were different for men and women in gender-inconsistent fields lends additional support to the idea that stereotype-based expectations, derived from lack of fit perceptions, lead to a differential updating of perceivers' impressions of male and female targets.

5. General discussion

5.1. Theoretical implications

These results provide evidence that sequential information about performance can differentially affect how men and women are evaluated. In the first two studies, sequential information produced differences in evaluations of men and women when performance changed and was consequently inconsistent, but not when it was consistently excellent or poor. This finding supports the idea that ambiguity, and the room it leaves for inference, facilitates gender bias. It also supports the idea that ambiguity about performance is not only a product of deficiencies in information clarity, but also a result of the inconsistencies that occur when performance is evaluated over time.

The results also provide evidence of the way in which evaluations of men and women were affected by changes in performance. Although in a male gender-typed field improvement positively affected everyone, it had less beneficial effects for women than men. Likewise, though decline in a male gender-typed field negatively affected everyone, it had more detrimental effects for women than men. Importantly, these results occurred only when the performance was in a field that was male gender-typed. In fact, the opposite occurred in a field that was female gender-typed. The results therefore are consistent with the idea that stereotype-based expectations born of lack of fit perceptions can influence the updating of impressions.

Specifically, our data suggest that when ambiguity is created by inconsistent performance information, stereotype-based expectations determine whether the new performance information is accepted or rejected. When women's or men's initial performance was inconsistent with stereotype-based expectations, and therefore unexpected, people seemed quite ready to downwardly revise their impressions in light of new information that was more in line with expectations. However, when their initial performance was consistent with stereotype-based expectations, and therefore was expected, people seemed to resist upwardly revising their original impressions. And this occurred even though the performance information provided was "objective", that is, was quantitative and clear-cut. This suggests that the way in which gender stereotypes exert their effect in evaluative contexts can be far-reaching, yet very subtle. Their influence appears to supersede considerations of primacy and recency (Murdock, 1962), and to override the differential diagnosticity that has been found for positive and negative information (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001). Much like informational anchors, gender stereotypes appear to exert "drag" on final judgments (Epley & Gilovich, 2006).

Our results also provide a compelling demonstration that the problem of gender bias is not necessarily resolved when no gender disparities are apparent. Despite the uniformity of competence impressions in initial success projections in Study 3, the success projections resulting from the second installment of performance information were distorted in line with gender stereotypic expectations. That is, gender bias was not gone, but only dormant, and surfaced when the opportunity for it presented itself. This finding is stark in its implications, suggesting that there are potential downstream consequences of gender stereotype-based expectations that might go unrecognized, and therefore are not attended to, because their effects are not evident in early judgments.

Lastly, it is important to note that in addition to studying perceptions of competence, we sought to study evaluative decisions based on the performance information provided. Accordingly, we obtained performance evaluations in the form of course grade assignments indicating whether men or women were more likely to be given "the benefit of the doubt" when scoring was equivocal. Finding that the pattern of grades assigned closely paralleled that of the reported competence perceptions gives added importance to our findings. It lends credence to the idea that gender stereotypes and the expectations they produce not only promote biased judgments, but also promote discriminatory behavior.

5.2. Limitations and future research

Our findings raise several issues to be addressed in future research. First, additional work is needed to further explore the effects observed here. Although the current results provide support for the idea that the differential updating of impressions is driven by a perceived lack of fit between gender stereotypes and the gender-type of the field, it is possible that other processes might also be involved. For example, changes in performance may give rise to shifts in what is regarded as a sufficient standard of competence (or incompetence), and these standards can differ for men and for women (Kobrynowicz & Biernat, 1997). It is also possible that under some conditions individuals differentially revise their performance judgments to legitimize the current status quo and maintain a system of gender inequality (Kay et al., 2009). Indeed, past research suggests that discrimination might emerge most strongly under conditions of ambiguity because it provides evaluators with a plausible justification for their negative ratings of stereotyped group members (Hodson et al., 2002; Norton, Vandello, & Darley, 2004; Uhlmann & Cohen, 2005). Future research should aim to clarify our understanding of when and why differences in the updating of impressions of men and women occur, and to identify the individual and contextual factors that regulate their occurrence.

It also is important to demonstrate the effects we found using other gender-typed fields and alternative evaluative contexts, and to determine how our results are affected when the ambiguity due to inconsistency in performance information is mitigated by other information. Moreover, future studies should examine the updating of impressions over greater time periods and multiple iterations, and explore the difference it makes whether the incremental information is a strong or weak indication of improvement or decline. They also should explore the effects of sequential evaluation when the performance information is vaguer and/or more subjective than the information provided in these studies, and may therefore foster gender bias in initial judgments; in these situations gender bias would likely occur despite consistency in sequential evaluation, and inconsistency in sequential evaluation may augment the ambiguity inherent in the information itself and further exacerbate gender bias. Identifying boundary conditions for our effects are critical for eliminating alternative explanations of our results and for further identifying the conditions under which they do and do not occur. Such investigations could shed light on how to prevent this type of bias – solutions that we think might be rooted in counteracting the inconsistency-induced ambiguity that provides such fertile ground for expectations to dominate in impression formation.

Our desire for experimental control led us to use names and photographs of White, relatively young targets, but social identity is far more nuanced and complex. Future endeavors could benefit from adopting an intersectional approach that examines how people update their impressions of men and women of various races, ethnicities, and ages. Consistent with the lack of fit model, we would expect effects to be driven by the stereotypes associated with the group in question and the degree to which they are believed to fit with the requirements of a given field.

Though this research was primarily concerned with the effects of gender stereotypes on competence perceptions, future research might also consider whether performance fluctuations in gender-inconsistent domains affect other stereotype-based evaluative outcomes. For example, extreme improvements in performance may be accepted as an indicator of competence but be interpreted as a sign of excessive competitiveness and ambition for women in male fields – behaviors that violate female gender prescriptions and, in turn, lead to social penalties (Heilman & Okimoto, 2007; Rudman & Glick, 2001). A similar situation can occur for men who improve in female fields; their incremental success may be seen as a violation of male gender prescriptions and result in them being seen as wimpy and not worthy of respect (Heilman & Wallen, 2010).

Additionally, research should be conducted to determine if the differential updating of competence perceptions evidenced in these studies also is reflected in self-evaluations. There are questions of whether women and men view themselves differently when their performance has improved or deteriorated, and whether they are prone to modulate their self-perceptions to conform to stereotype-based expectations. Such research could potentially extend the ideas presented here, allowing a comparison of the effects of stereotyped-based expectations about others to the effects of stereotyped-based expectations about oneself.

Lastly, but very importantly, to fully understand the importance of these results it is essential for this research to be followed up in ongoing performance settings. Additional research is necessary to determine when the process of updating impressions both of women in traditionally male contexts and men in traditionally female contexts coincides with, or differs from, what we have demonstrated in our studies. The amount of information available to evaluators in these settings is typically far more plentiful than that provided to our research participants. It also is likely to be more textured, with relevant information coming from a multitude of sources and being accumulated over a substantial period of time. It is not clear that our results would be replicated with these types of differences in the amount, type, and timing of information and, if they are not, significant moderators of the effects we have found would be discovered – all of which would help refine and elaborate our ideas as well as suggest avenues for intervention.

5.3. Practical implications

A reason beyond that of good science to endeavor to replicate and extend these results and identify their boundary conditions is that they have potentially important implications. Although our effect sizes were relatively small, this should not cause an underestimation of their importance. Research on gender bias has shown that even small effect sizes can add up and have profound effects on careers (Martell, Lane, & Emrich, 1996). Because repeated evaluation is the norm in work settings, and impressions are continually being revised and updated, the small to medium effects demonstrated here are likely to have significant practical consequences.

Our results show that both men and women can be burdened by gender stereotypes and the expectations they produce in the pursuit of traditionally gendered jobs and interests. The negative effects of gender bias on men are often overlooked, and this is an issue that is stirring interest among current researchers (Croft et al., 2015). However, because of the preponderance of male gender-typing of prestigious, lucrative, and powerful positions, it is women striving to advance their careers who are particularly likely to be the victims of the processes documented in this research.

In fact, our findings resonate strongly with the often heard claim made by upwardly striving women that for them, as compared to their male counterparts, unequivocal and continuous evidence of effectiveness is required to earn acclaim, and one misstep is sufficient to derail a promising career. They also are consistent with the idea that women in male gender-typed settings are held to a higher standard than men and imply that gender stereotype-inspired negative expectations continue to beleaguer them even when they have demonstrated their competence or their failures have given way to success. Given the rigidity of gender stereotypes and their stubborn resistance to change (Haines et al., 2016), it seems unlikely that women will escape these problems anytime soon. On the contrary, it appears that gender stereotypes are still with us, and that they can exert influence in ways that can be elusive and hidden, requiring continued vigilance on the part of evaluators if gender equity in evaluation is truly to be achieved.

Appendix A. Target Background Information and Gender Manipulation for All Studies

Female target

STUDENT NAME: Patricia Olsen
ASSIGNED COURSE: [Programming I/ Principles of Physics /Early Childhood Education]
Patricia Olsen grew up in a suburb outside of Phoenix. She is currently a junior at [redacted] University and is taking a [computer science/physics/education] class to fulfill [a science/an interdisciplinary] requirement. She hopes to get an internship next summer to provide her with hands-on experience in the work world. In her spare time, Patricia enjoys reading and playing tennis.

Male target

STUDENT NAME: Thomas Olsen
ASSIGNED COURSE: [Programming I/ Principles of Physics /Early Childhood Education]
Thomas Olsen grew up in a suburb outside of Phoenix. He is currently a junior at [redacted] University and is taking a [computer science/physics/education] class to fulfill [a science/an interdisciplinary] requirement. He hopes to get an internship next summer to provide him with hands-on experience in the work world. In his spare time, Thomas enjoys reading and playing tennis.

Appendix B. Course Syllabi

Male-Typed Field, Studies 1 and 2

Department of Computer Science

Course: Programming I

Course design: Two independent units

Overview of the course: This seminar examines programming techniques and methods. The objectives of the course are to inform students of the issues involved in developing large-scale software systems, teach how programming style conventions and language restrictions can ease software development, and provide students with insight into interface design.

Course structure: The course is divided into **two independent units**. In each course unit, students will learn how to design and implement a simulation in several stages. The different techniques required to develop each of these programs will be covered during the respective course unit.

- During the first unit, students will be working with C++.
- During the second unit, students will be working with Java.

Course requirements: Performance in the two course units will be assessed independently. For each course unit there will be an in-class exam and an individual project.

Note: The exam for the second unit of the course will not include material covered in the first unit of the course.

Male-Typed Field, Study 3
Department of Physics

Course: Principles of Physics

Course design: Two independent units

Overview of the course: This seminar covers the basic principles of physics. The objectives of the course are to introduce students to the concepts involved in momentum and energy conservation, learn how to apply principles of oscillation and wave propagation, and provide students with a general understanding of acceleration and uniform circular motion.

Course structure: The course is divided into **two independent units**. In each course unit, students will review and discuss physical principles and their application to different states of matter. Specific problem solving techniques and formulations will be covered during the respective course unit.

- During the first unit, students will examine physical laws and their applications for motion and gravitation
- During the second unit, students will examine physical laws and their applications for fluid mechanics and kinetics

Course requirements: Performance in the two course units will be assessed independently. For each course unit there will be an in-class exam and an individual project.

Note: The exam for the second unit of the course will not include material covered in the first unit of the course.

Female-Typed Field, Study 3

Department of Education and Human Development

Course: Early Childhood Education

Course design: Two independent units

Overview of the course: This seminar covers the basic principles of early childhood education. The objectives of the course are to introduce students to the socio-cultural foundations of the education of children from birth to age 5. Students will learn the techniques involved in guiding a child’s social and emotional development, and provide hands on experience with infants, toddlers and young children.

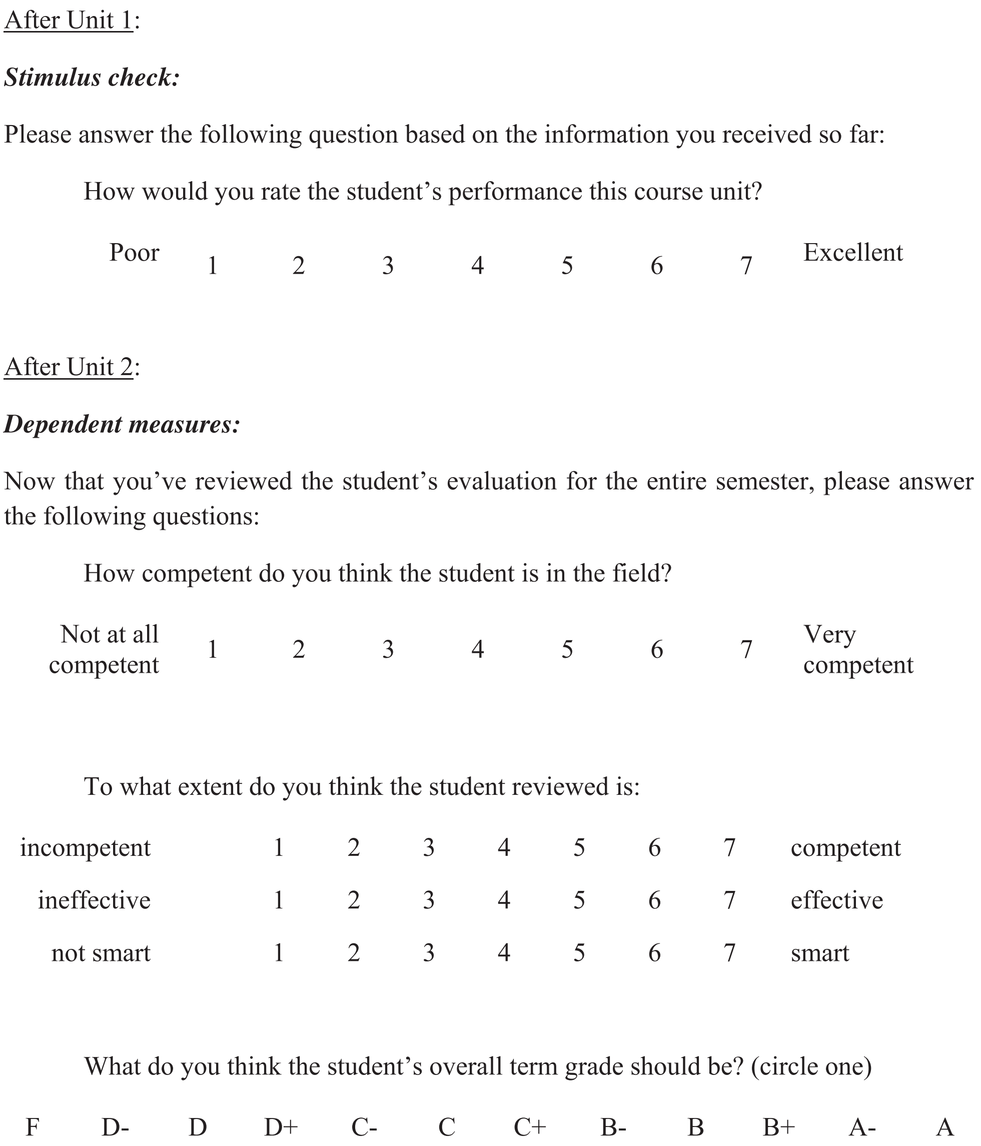
Course structure: The course is divided into **two independent units**. In each course unit, students will review and discuss educational principles and apply them in actual educational settings. Specific teaching and evaluation techniques will be covered during the respective course unit.

- During the first unit, students will focus on educational practices specific to infants and toddlers (age 0 to 3)
- During the second unit, students will focus on educational practices specific to young children (age 3 to 5)

Course requirements: Performance in the two course units will be assessed independently. For each course unit there will be an in-class exam and an individual project.

Note: The exam for the second unit of the course will not include material covered in the first unit of the course.

Appendix C. Manipulation and Stimulus Checks, Dependent Measures, and Demographic Questions for All Studies



After Unit 1:

Stimulus check:

Please answer the following question based on the information you received so far:

How would you rate the student's performance this course unit?

Poor	1	2	3	4	5	6	7	Excellent

After Unit 2:

Dependent measures:

Now that you've reviewed the student's evaluation for the entire semester, please answer the following questions:

How competent do you think the student is in the field?

Not at all competent	1	2	3	4	5	6	7	Very competent

To what extent do you think the student reviewed is:

incompetent	1	2	3	4	5	6	7	competent
ineffective	1	2	3	4	5	6	7	effective
not smart	1	2	3	4	5	6	7	smart

What do you think the student's overall term grade should be? (circle one)

F	D-	D	D+	C-	C	C+	B-	B	B+	A-	A

Manipulation checks:

Name of student reviewed: ___________________________________

Which of the following best describes the student's performance in the course? (circle one)

a. The student's performance improved throughout the course

b. The student's performance stayed the same throughout the course

c. The student's performance declined throughout the course

Stimulus check:

Please provide your thoughts on the specific course that you reviewed:

What was the name of the course? ___________________________

To what extent do you think the course reviewed was:

male-oriented	1	2	3	4	5	6	7	female-oriented

Demographic questions:

For clerical purposes, please provide the following information:

Age: ________

Sex: ____ Male ____ Female

Race/ethnicity: __________________________

Major: _________________________________

Additional Dependent Measure For Study 3
(asked after Unit 1 and after Unit 2)

How successful do you think this student would be in similar courses?

Not at all successful	1	2	3	4	5	6	7	Very successful

Appendix D. Analyses Without Exclusions

Study 1

Perceived competence
Main effect of performance change: $F(1, 202) = 165.87$, $p < .001$, $\eta_p^2 = 0.451$.
Main effect of target gender: $F(1, 202) = 4.17$, $p = .04$, $\eta_p^2 = 0.020$.
Interaction between performance change and target gender: $F(1, 202) = 4.94$, $p = .03$, $\eta_p^2 = 0.024$.

Pairwise comparisons:
Difference between male and female targets in no-change condition: $t(202) = 0.13$, $p = .90$,
Difference between male and female targets in improve condition: $t(202) = 3.04$, $p = .003$, $d = 0.43$.

Grade assignment
Difference in the distribution of final grades below and above the true average assigned to male and female targets in the no-change condition, χ^2 (1, N = 102) = 0.22, $p = .64$.

Difference in the distribution of final grades below and above the true average assigned to male and female targets in the improve condition, χ^2 (1, N = 104) = 10.09, p = .001.

Study 2

Perceived competence

Main effect of performance change: $F(1, 201) = 232.33$, $p < .001$, $\eta_p^2 = 0.536$.
Main effect of target gender: $F(1, 201) = 9.97$, $p = .002$, $\eta_p^2 = 0.047$.
Interaction between performance change and target gender: $F(1, 201) = 9.00$, $p = .003$, $\eta_p^2 = 0.043$.

Pairwise comparisons:
Difference between male and female targets in no-change condition: $t(201) = 0.11$, $p = .91$,
Difference between male and female targets in decline condition: $t(201) = 4.38$, $p < .001$, $d = 0.62$.

Grade assignment

Difference in the distribution of final grades below and above the true average assigned to male and female targets in the no-change condition, χ^2 (1, N = 100) = 0.88, p = .35.

Difference in the distribution of final grades below and above the true average assigned to male and female targets in the improve condition, χ^2 (1, N = 104) = 8.08, p = .004.

Study 3

Perceived competence

Main effect for performance change: $F(1, 486) = 64.51$, $p < .001$, $\eta_p^2 = 0.117$.
Interaction between target gender and gender-type of field: $F(1, 486) = 25.52$, $p < .001$, $\eta_p^2 = 0.050$.

Pairwise comparisons:
Difference between male and female targets in the physics course: $t(486) = 3.23$, $p = .001$, $d = 0.29$.
Difference between male and female targets in the early childhood education course: $t(486) = 3.93$, $p < .001$, $d = 0.36$.

Grade assignment

Difference in the distribution of final grades below and above the true average assigned to male and female targets in the physics course: χ^2 (1, N = 246) = 13.63, $p < .001$.

Difference in the distribution of final grades below and above the true average assigned to male and female targets in the early childhood education course: χ^2 (1, N = 248) = 10.08, p = .001.

Change in projected likelihood of success

Main effect of performance change: $F(1, 486) = 1{,}344.05$, $p < .001$, $\eta_p^2 = 0.734$.
Interaction between target gender and gender-type of field: $F(1, 486) = 17.71$, $p < .001$, $\eta_p^2 = 0.035$.

Pairwise comparisons:
Difference between male and female targets in the physics course: $t(486) = 2.70$, $p = .007$, $d = 0.25$,
Difference between male and female targets in the early childhood education course: $t(486) = 3.26$, $p = .001$, $d = 0.30$.

References

Banaji, M. R., Hardin, C., & Rothman, A. J. (1993). Implicit stereotyping in person judgment. *Journal of Personality and Social Psychology, 65*, 272–281.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*, 323–370.

Biernat, M., Fuegen, K., & Kobrynowicz, D. (2010). Shifting standards and the inference of incompetence: Effects of formal and informal evaluation tools. *Personality and Social Psychology Bulletin, 36*, 855–868.

Biernat, M., & Vescio, T. K. (2002). She swings, she hits, she's great, she's benched: Implications of gender-based shifting standards for judgment and behavior. *Personality and Social Psychology Bulletin, 28*, 66–77.

Blair, I. V., & Banaji, M. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology, 70*, 1142–1163.

Brescoll, V. L., Dawson, E., & Uhlmann, E. L. (2010). Hard-won and easily lost: The fragile status of leaders in gender-stereotype-incongruent occupations. *Psychological Science, 21*, 1640–1642.

Cappelli, P., & Canyon, M. J. (2016). What to do performance appraisals do? *ILR Review, 71*(1), 88–116.

Cejka, M. A., & Eagly, A. H. (1999). Gender-stereotypic images of occupations correspond to the gender segregation of employment. *Personality and Social Psychology Bulletin, 25*, 413–423.

Cheryan, S., Master, A., & Meltzoff, A. N. (2015). Cultural stereotypes as gatekeepers: increasing girls' interest in computer science and engineering by diversifying stereotypes. *Frontiers in Psychology, 6*.

Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2017). Why are some STEM fields more gender balanced than others? *Psychological Bulletin, 143*, 1–35.

Croft, A., Schmader, T., & Block, K. (2015). An underexamined inequality: Cultural and psychological barriers to men's engagement with communal roles. *Personality and Social Psychology Review, 19*, 343–370.

Davison, H. K., & Burke, M. J. (2000). Sex discrimination in simulated employment contexts: A meta-analytic investigation. *Journal of Vocational Behavior, 56*, 225–248.

Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science, 11*, 315–319.

Eagly, A. H., Karau, S. J., & Makhijani, M. G. (1995). Gender and the effectiveness of leaders: A meta-analysis. *Psychological Bulletin, 117*, 125–145.

Eagly, A. H., & Wood, W. (2011). Social role theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.). *Handbook of theories in social psychology* (pp. 458–476). Los Angeles, CA: SAGE.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science, 17*, 311–318.

Fiske, A. P., Haslam, N., & Fiske, S. T. (1991). Confusing one person with another: What errors reveal about the elementary forms of social relations. *Journal of Personality and Social Psychology, 60*, 656–674.

Fyock, J., & Stangor, C. (1994). The role of memory biases in stereotype maintenance. *British Journal of Social Psychology, 33*, 331–343.

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology, 101*, 109–128.

Haines, E. L., Deaux, K., ... Lofaro, N. (2016). The times they are a-changing… or are they not? A comparison of components of gender stereotypes, 1983 to 2014. *Sex Roles, 40*, 353–363.

Haslam, S. A., Ryan, M. K., Kulich, C., Trojanowski, G., & Atkins, C. (2010). Investing with prejudice: The relationship between women's presence on company boards and objective and subjective measures of company performance. *British Journal of Management, 21*, 484–497.

Heilman, M. E. (1983). Sex bias in work settings: The lack of fit model. *Research in Organizational Behavior, 5*, 269–298.

Heilman, M. E. (2001). Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of Social Issues, 57*, 657–674.

Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Research in Organizational Behavior, 32*, 113–135.

Heilman, M. E., Block, C. J., & Martell, R. F. (1995). Sex stereotypes: Do they influence perceptions of managers? *Journal of Social Behavior and Personality, 10*, 237–252.

Heilman, M. E., & Caleo, S. (2018). Combatting gender bias: A lack of fit framework. *Group Processes and Intergroup Relations, 21*, 725–744.

Heilman, M. E., & Haynes, M. C. (2008). Subjectivity in the appraisal process: A facilitator of gender bias in work settings. In E. Borgida, & S. T. Fiske (Eds.). *Beyond common knowledge: Psychological science in court*. Mahwah, NJ: Larry Erlbaum Associates.

Heilman, M. E., & Okimoto, T. G. (2007). Why are women penalized for success at male tasks? The implied communality deficit. *Journal of Applied Psychology, 92*, 81.

Heilman, M. E., & Wallen, A. S. (2010). Wimpy and undeserving of respect: Penalties for men's gender-inconsistent success. *Journal of Experimental Social Psychology, 46*, 664–667.

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

Heslin, P. A., Latham, G. P., & VandeWalle, D. (2005). The effect of implicit person theory on performance appraisals. *Journal of Applied Psychology, 90*, 842–856.

Hodson, G., Dovidio, J. F., & Gaertner, S. L. (2002). Processes in racial discrimination: Differential weighting of conflicting information. *Personality and Social Psychology Bulletin, 28*, 460–471.

Jirjahn, U., & Stephan, G. (2004). Gender, piece rates and wages: Evidence from matched employer-employee data. *Cambridge Journal of Economics, 28*, 683–704.

Johnson, S. K., Murphy, S. E., Zewdie, S., & Reichard, R. J. (2008). The strong, sensitive type: Evidence for gender-specific leadership prototypes. *Organizational Behavior and Human Decision Processes, 106*, 39–60.

Kasof, J. (1993). Sex bias in the naming of stimulus persons. *Psychological Bulletin, 113*, 140–163.

Kay, A., Gaucher, D., Peach, J. M., Laurin, K., Friesen, J., Zanna, M. P., & Spencer, S. J. (2009). Inequality, discrimination, and the power of the status quo: Direct evidence for a motivation to see the way things are as they should be. *Journal of Personality and Social Psychology, 97*, 421–434.

Ko, I., Kotrba, L., & Roebuck, A. (2015). Leaders as males? The role of industry gender composition. *Sex Roles, 72*, 294–307.

Kobrynowicz, D., & Biernat, M. (1997). Decoding subjective evaluations: How stereotypes provide shifting standards. *Journal of Experimental Social Psychology, 33*, 579–601.

Koch, A. J., D'Mello, S. D., & Sackett, P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of Applied Psychology, 100*, 128–161.

Koenig, A. M., Eagly, A. H., Mitchell, A. A., & Ristikari, T. (2011). Are leader stereotypes masculine? A meta-analysis of three research paradigms. *Psychological Bulletin, 137*, 616–642.

Kunda, Z., Sinclair, L., & Griffin, D. (1997). Equal ratings but separate meanings: Stereotypes and the construal of traits. *Journal of Personality and Social Psychology, 72*, 720–734.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition & Emotion, 24*, 1377–1388.

Lyness, K. S., & Heilman, M. E. (2006). When fit is fundamental: Performance evaluations and promotions of upper-level female and male managers. *Journal of Applied Psychology, 91*, 777–785.

Macrae, C. N., Milne, A. B., & Bodenhausen, G. V. (1994). Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *Journal of Personality and Social Psychology, 66*, 37–47.

Manzoni, J. F., & Barsoux, J. L. (1998, March–April). How bosses create their own poor performers: The set-up-to-fail syndrome. Harvard Business Review, 101–113.

Martell, R. F., Lane, D. M., & Emrich, C. (1996). Male-female differences: A computer simulation. *American Psychologist, 51*, 157–158.

Moss-Racusin, C., 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.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64*, 482–488.

National Center for Education Statistics (2015). Bachelor's, master's, and doctor's degrees conferred by postsecondary institutions, by sex of student and discipline division: 2014–15. Retrieved from https://nces.ed.gov/programs/digest/d16/tables/dt16_318.30.asp?current = yes.

Norton, M. I., Vandello, J. A., & Darley, J. M. (2004). Casuistry and social category bias. *Journal of Personality and Social Psychology, 87*, 817.

Paustian-Underdahl, S. C., Walker, L. S., & Woehr, D. J. (2014). Gender and perceptions of leadership effectiveness: A meta-analysis of contextual moderators. *Journal of Applied Psychology, 99*, 1129–1145.

Pittinsky, T. L., Shih, M., & Ambady, N. (2000). Will a category cue affect you? Category cues, positive stereotypes, and reviewer recall for applicants. *Social Psychology of Education, 4*, 53–65.

Robbins, T. L., & DeNisi, A. S. (1993). Moderators of sex bias in the performance appraisal process: A cognitive analysis. *Journal of Management, 19*, 113–126.

Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues, 57*, 743–762.

Schein, V. E. (2001). A global look at psychological barriers to women's progress in management. *Journal of Social Issues, 57*, 675–688.

Scott, K. A., & Brown, D. J. (2006). Female first, leader second? Gender bias in the encoding of leadership behavior. *Organizational Behavior and Human Decision Processes, 101*, 230–242.

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.

Uhlmann, E. L., & Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science, 16*, 474–480.

Williams, J., & Best, D. (1990), Measuring Sex Stereotypes: A Multination Study (revised edition), Beverly Hills, CA, USA: Sage Publications.

Zyphur, M. J., Chaturvedi, S., & Arvey, R. D. (2008). Job performance over time is a function of latent trajectories and previous performance. *Journal of Applied Psychology, 93*, 217.