

# The Gender Minority Gaps in Confidence and Self-Evaluation\*

Billur Aksoy<sup>†</sup> Christine L. Exley<sup>‡</sup> Judd B. Kessler<sup>§</sup>

June 5, 2025

## Abstract

An increasing share of the population identifies their gender in more ways than just as men or women. Analyzing data on 1,500 adults and 10,000 students, we find gender minority gaps. Relative to equally performing men, gender minorities are less confident and provide less favorable self-evaluations on a math and science test. These gaps are robust and as large as gender gaps between men and women. But, unlike gender gaps between men and women, gender minority gaps are unexpected. People are unsure about the confidence and other traits of gender minorities, a clear barrier for addressing gender minority gaps.

**Keywords:** Gender Identity, LGBTQ+ Identity, Confidence, Self-Evaluation

**JEL codes:** C91, D91, J16

---

\*This paper was supported by Character Lab and facilitated through the Character Lab Research Network, a consortium of schools across the country working collaboratively with scientists to advance scientific insights that help kids thrive. Harvard Business School provided generous financial support. We would like to thank Christopher Carpenter and Dario Sansone, and various seminar and conference participants for their helpful comments and feedback. This paper includes data from two pre-registered studies: see details in Footnote 9 for the Adult Study and Footnote 21 for the Predictions Study. As explained in Section 3.1 and Footnote 13, the authors had access to the data for the Student Study prior to the idea for this paper, so that data was not pre-registered.

<sup>†</sup>Department of Economics, Rensselaer Polytechnic Institute, Email: aksoyb3@rpi.edu.

<sup>‡</sup>Department of Economics, University of Michigan, Email: exley@umich.edu.

<sup>§</sup>The Wharton School, University of Pennsylvania and NBER, Email: judd.kessler@wharton.upenn.edu.

# 1 Introduction

A sizable share of the population identifies as part of a gender minority group. Examples include individuals who identify as transgender, non-binary, or genderqueer; such identities can overlap and evolve over time. Among adults in the United States, it is estimated that around 1–2% identify as part of a gender minority group (Jones, 2022; Brown, 2022). Moreover, there is a growing share of the population in this category, with an estimate of about 5% among U.S. adults under 30 (Brown, 2022). Understanding the traits and beliefs of gender minority groups is clearly important.<sup>1</sup>

Inspired by the rich line of prior work on gender differences between men and women in various traits including confidence (Barber and Odean, 2001; Niederle and Vesterlund, 2007) and self-evaluations (Exley and Kessler, 2022), we initiate a line of research exploring the confidence and self-evaluations of gender minorities.<sup>2</sup> We specifically consider *gender diverse individuals* who identify in some way other than “male” or “female” when asked about their gender in our survey. We investigate whether there are differences in confidence, measured by beliefs about absolute performance, and differences in subjective self-evaluations about performance between gender diverse individuals and those who identify as either male or female.

One challenge with conducting research on gender minority groups is that data on gender identity is often recorded as binary or is missing in administrative records. The lack of data—and thus research—on gender identity in the United States is likely to worsen given the current political climate, particularly in light of Executive Order 14168, issued on January 20, 2025.<sup>3</sup> This order mandates that federal agencies recognize only biological sex, disregarding gender identity in all official matters. As a result, data on gender identity has already been

---

<sup>1</sup>Gender identity is currently understood as a person’s internal sense or individual experience of their gender, which may or may not align with their sex assigned at birth. It is important to note that gender identity is distinct from sexual identity, which pertains to a person’s emotional and/or sexual attraction to individuals of a certain gender or genders. Sexual minorities include, but are not limited to, those who are gay, lesbian, or bisexual. In this paper, given our desire to study a gender minority group, we focus on gender identity and not sexual identity.

<sup>2</sup>Prior work on gender has focused on gender gaps between men and women with an eye toward explaining gaps between those genders in pay, representation in certain fields, and roles in corporate and political leadership (Bertrand, Goldin and Katz, 2010; Blau and Kahn, 2017; Grossman et al., 2019; Bütikofer, Løken and Willén, 2022). To explain these differences, researchers have leveraged observational data—to consider factors such as occupational selection and institutional and policy features—and have measured various traits in experiments, identifying gender differences between men and women in relation to traits such as negotiation (Babcock and Laschever, 2003), competitiveness (Niederle and Vesterlund, 2007), risk taking (Eckel and Grossman, 2008), the contribution of ideas (Coffman, 2014), and image concerns (Bursztyn, Fujiwara and Pallais, 2017) (as well as confidence and self-evaluation, as referenced in the main text).

<sup>3</sup><https://www.whitehouse.gov/presidential-actions/2025/01/defending-women-from-gender-ideology-extremism-and-restoring-biological-truth-to-the-federal-government>

removed from various federal datasets.<sup>4</sup> We overcome this challenge by collecting new data that allow people to self-identify their gender as part of our studies.

Another challenge is that it is often hard to recruit a sufficient number of gender minorities, particularly among older populations. We overcome this challenge in two ways. For our studies with adults, we recruit an online sample of 1,494 adults with a pre-registered protocol that overweights individuals whose prior answers on Prolific suggest they might be gender minorities; we identify 330 people in these studies as gender diverse. For our student study, we recruit a large sample of young individuals, which allows us to analyze data from 10,807 students in grades 6–12; we identify 180 students as gender diverse.

Our two main studies proceed in six stages. First, following the prior literature—which has identified particularly large gender gaps between men and women in confidence and self-evaluations in stereotypically male-typed domains such as math and science (Lundeberg, Fox and Punčohař, 1994; Niederle and Vesterlund, 2007; Coffman, 2014; Bordalo et al., 2019; Coffman, Collis and Kulkarni, 2019; Exley and Kessler, 2022) (for related reviews, see also Niederle and Vesterlund, 2011; Blau and Kahn, 2017; Hernandez-Arenaz and Iriberry, 2019; Niederle, 2016), participants complete a math and science test. Second, we elicit participants’ *beliefs* about their absolute performance by asking them to guess how many questions they got right on the test, which serves as a measure of *confidence*. Third, we elicit participants’ *uninformed self-evaluations* with four questions that ask them to provide subjective evaluations of their performance on the test, such as by indicating their level of agreement with the statement “I performed well on the test.” Fourth, we inform participants about how many questions they actually got right on the test. Fifth, we elicit participants’ *informed self-evaluations* with the same four questions asking them to provide subjective evaluations of their performance. Finally, we ask participants to complete a survey that gathers demographic information, including on gender identity, which allows us to classify participants as either male, female, or gender diverse.

In our *Adult Study*, we observe large *gender minority gaps*. When compared to equally-performing men, gender diverse individuals display lower confidence: they believe they got fewer questions right on the math and science test. When compared to equally-performing men, gender diverse individuals also provide less favorable self-evaluations about their performance on the math and science test. They indicate less agreement with the statement that they “performed well” on the test, they report being less inclined to apply for a job that involves the math and science topics covered on the test, and they believe they would be less

---

<sup>4</sup>Additionally, the executive order prohibits the use of federal funds for research on gender identity, which will likely hinder future studies in this area. These recent developments in the United States further underscore the importance of the research undertaken in our paper.

likely to succeed in such a job. These differences in self-evaluations persist when participants are informed about how many questions they actually got right on the test. In addition, the gender minority gaps are just as large as the gender gaps between male and female adults.

In the *Student Study*, we again observe large gender minority gaps. When compared to equally-performing male students, gender diverse students report lower confidence and provide less favorable self-evaluations. These differences in self-evaluations persist when students are informed about how many questions they actually got right on the test. In addition, we find that gender diverse students assess their performance *more negatively* than equally-performing women, so the gender minority gaps among students are *larger* than the gender gaps between male and female students that we observe in our data, adding to prior work that explores gender differences among children and adolescents (see for example [Dreber, von Essen and Ranehill, 2014](#)).<sup>5</sup>

Before seeing the results in this paper, it is unclear whether one should have expected gender minority gaps. One might have expected gaps in confidence and self-evaluation between gender diverse people and equally performing men because gender diverse people are part of a marginalized group, and marginalized groups often display lower confidence than majority groups.<sup>6</sup> On the other hand, self-identifying as gender diverse means rejecting society’s imposed gender identity classification and perhaps subjecting oneself to additional discrimination, so gender diverse individuals could be even more confident and self-assured than other groups. But, akin the discussion in [Schaerer et al. \(2023\)](#) on the importance of their results on the extent to which differences between men and women in hiring processes are accurately expected, it is important to understand beliefs about gender minority gaps to efficiently counter them.

Thus, since prior literature provides little insight into beliefs about the confidence of gender minorities—and how those beliefs compare to beliefs about men and women—we ran the *Predictions Study*. In this study, a new group of participants assigned to the role of “predictors” are asked to make incentivized guesses about gender diverse, male, and female participants from our Adult Study.

In one set of questions, predictors are asked whether a participant is overconfident, underconfident, or accurate about their performance on the math and science test; or, alterna-

---

<sup>5</sup>The gender gaps between male and female students in this data are largely *not* novel. As explained in Section 3.1, [Exley and Kessler \(2022\)](#) also analyze this youth data; however, that paper relies on administrative data that only has a binary classification of gender (i.e., only male or female), whereas this paper relies on self-reported gender that allows students to identify as gender diverse. The novel analyses in this paper relates to the examination of gender diverse students.

<sup>6</sup>For recent evidence supporting lower confidence among non-binary individuals, see also [Coffman, Coffman and Ericson \(2024\)](#). For evidence in the context of sexual minorities, rather than gender minorities, see [Aksoy and Chadd \(2025\)](#).

tively, predictors can indicate that they are unsure about a participant’s confidence. Most predictors expect men to be overconfident and only 12% of predictors indicate that they are unsure of men’s confidence. Most predictors expect women to be underconfident and only 11% of predictors indicate that they are unsure of women’s confidence. By contrast, predictors indicate substantial uncertainty when asked about gender diverse participants, with 16% predicting overconfidence, 31% predicting accuracy, 30% predicting underconfidence and 24% directly indicating that they are unsure.

Results from the Predictions Study highlight a key difference between the well documented gender gaps in self-evaluations and confidence between men and women and the novel gender minority gaps that we document here. The former gender gaps have come to be expected (Exley and Nielsen, 2024), while the gender minority gaps we document here are largely *unexpected*.

The unexpected nature of the gender minority gaps is important to highlight for several reasons. First, it may contribute to little attention toward interventions or policies that may combat these gender minority gaps in confidence and self-evaluations, even though the existing (albeit small) body of work on gender minorities indicates that, compared to the general population, gender minorities have significantly worse economic outcomes (Badgett, Carpenter and Sansone, 2021; Carpenter, Eppink and Gonzales, 2020; Carpenter, Lee and Nettuno, 2022); have worse educational outcomes (Meyer et al., 2017; Downing and Przedworski, 2018; Sansone, 2019); and are more likely to be unemployed, be in low-income households, and be uninsured (Badgett, Carpenter and Sansone, 2021). Unlike the widespread push for women to “lean-in,” and initiatives that seek to encourage more confidence in women, there is no similar focus on gender minorities.<sup>7</sup>

Second, the unexpected nature of the gender minority gaps in confidence and self-evaluations reflects one potentially unifying feature about gender minority gaps: individuals may be broadly *unsure or uncertain* about the traits of gender minorities. Indeed, in addition to documenting that predictors are unsure about the confidence levels of gender diverse participants, predictors also report being unsure about the risk-taking behavior, competitiveness, and generosity of gender diverse participants. This finding may be particularly valuable for future work to explore since—as shown in Coffman, Coffman and Ericson (2024), which investigates a wide range of contexts and traits—the direction and size of gender minority gaps indeed vary across contexts and traits.

---

<sup>7</sup>For academic and related policy discussions on leaning in, see Exley, Niederle and Vesterlund (2020). See Demiral and Mollerstrom (2024) for the negative consequences of signaling excessive confidence.

## 2 The Adult Study

### 2.1 The Design of the Adult Study

The *Adult Study* follows the design of [Exley and Kessler \(2022\)](#) and proceeds in six stages. Participants earn a fixed payment of \$4 and have an opportunity to earn bonus payment. Additional design details, including screenshots, can be found in [Appendix E](#).

In the first stage, participants answer 20 test questions from the Armed Services Vocational Aptitude Battery and are told they will receive 5 cents for each correct answer on the test if the first part of the study is chosen to determine bonus payments (otherwise they receive 25 cents as a bonus payment). Each question appears on a separate page, and participants have 30 seconds to answer each question (see [Appendix Figure E.2](#) for an example question).

In the second stage, we collect each participant’s belief about their absolute performance by asking how many questions out of 20 they thought they answered correctly. This gives us a measure of their confidence in their absolute performance.

In the third stage, we elicit each participant’s uninformed self-evaluations by asking a free response question about their performance and four quantitative self-evaluation questions. Like [Exley and Kessler \(2022\)](#), we focus on the quantitative answers to the self-evaluation questions. In the *performance-bucket* question, participants are asked to indicate how well they think they performed on the test by choosing from the following list of seven adjectives: terrible, very poor, poor, neutral, good, very good, and exceptional. In the remaining three self-evaluation questions, participants are asked to indicate their agreement—on a scale from 0 (entirely disagree) to 100 (entirely agree)—with various statements. In the *performance self-evaluation* question, participants are asked to indicate their agreement with “I performed well on the test.” In the *willingness* question, participants are asked to indicate their agreement with “I would apply for a job that required me to perform well on the test I took in Part 1.” In the *success* question, participants are asked to indicate their agreement with “I would succeed in a job that required me to perform well on the test I took in Part 1.”

In the fourth stage, we inform participants of how many questions they got right on the test and then require them to correctly report back that number. By informing participants about their absolute performance, we mechanically close any gap in beliefs about absolute performance once we condition on participants having the same score, which we do in our regression analysis.

In the fifth stage, we elicit participants’ informed self-evaluations by asking the same set of questions they were asked before they received information about their performance.

In the sixth stage, we ask a demographic survey and adopt the gender question proposed by Miller and Willson (2022) and recommended as following best practices in 2022 for the collection of self-reported sexual orientation and gender identity data on Federal statistical surveys. Specifically, our gender question allows participants to choose all applicable options from the following: “Male,” “Female,” or “Transgender, non-binary, or another gender (see Figure E.9).”<sup>8</sup>

## 2.2 Gender Identity in the Adult Study

A total of 746 people participated in the *Adult Study* on Prolific during June and July of 2023. Since gender minorities constitute a relatively small share of the adult population in the U.S., we implemented a pre-registered stratified protocol to recruit a relatively large number of gender minorities from the Prolific platform.<sup>9</sup> In response to our gender survey question, 41.0% (n=306) selected only “Male,” 36.3% (n=271) selected only “Female,” and the remaining 22.7% (n=169) selected “Transgender, non-binary, or another gender” or multiple options, which leads us to classify them as gender diverse.<sup>10</sup>

## 2.3 The Gender Minority Gap in Confidence among Adults

Gender diverse participants got an average of 12.51 questions correct out of 20. This performance is better than male participants who got an average of 11.43 questions correct. Both of these performances are better than the performance of female participants, who got an average of 10.25 questions correct. Despite outperforming men, however, gender diverse participants report lower confidence in their performance: gender diverse participants believe they answered 9.56 questions correctly, while male participants believe they answered 10.30 questions correctly.

To examine whether there is a gender minority gap in confidence between gender diverse participants and *equally-performing* male participants, we run regressions that control for performance. Table 1 presents regression results related to a participant’s performance (i.e., the number of questions they got right on the test) and reported confidence (i.e., their belief

---

<sup>8</sup>Following the June 2022 Executive Order 14075 on “Advancing Equality for Lesbian, Gay, Bisexual, Transgender, Queer, and Intersex Individuals,” in January 2023 the Office of the Chief Statistician of the United States developed the “Recommendations on Best Practices for the Collection of Sexual Orientation and Gender Identity Data on Federal Statistical Surveys” report to provide recommendations for Federal agencies on the current best practices for the collection of self-reported sexual orientation and gender identity data on Federal statistical surveys. The gender question we use in our study is highlighted in this report as an example gender question.

<sup>9</sup>This recruitment procedure was pre-registered on AsPredicted (#136119) which can be accessed here: [https://aspredicted.org/2FW\\_Z5H](https://aspredicted.org/2FW_Z5H). Further details about our recruitment procedure is discussed in the Online Appendix E.1.

<sup>10</sup>Specifically, 122 participants only selected “Transgender, non-binary, or another gender,” 27 participants selected “Transgender, non-binary, or another gender” and “Male,” 19 participants selected “Transgender, non-binary, or another gender” and “Female,” and 1 participant selected “Male” and “Female.”



about the number of questions they got right on the test). *Gender Diverse* and *Female* are indicators for gender diverse participants and female participants, respectively, while male participants are the excluded category in these regressions. Thus, the coefficient estimates on *Gender Diverse* compares gender diverse participants to male participants, and the coefficient estimate on *Female* compares female participants to male participants. At the bottom of the table, the coefficient estimate on *Difference* reports the difference between gender diverse and female participants along with its corresponding p-value for a two-sided t-test of the difference in these coefficient estimates.

Table 1: In the *Adult Study*, Participants' Performance (i.e., score on test) and Reported Confidence (i.e., believed scored on test)

	Performance	Reported Confidence	Reported Confidence– Performance	1(Reported Confidence < Performance)
	(1)	(2)	(3)	(4)
Gender Diverse	1.08*** (0.30)	-1.45*** (0.30)	-1.82*** (0.31)	0.20*** (0.04)
Female	-1.18*** (0.27)	-1.75*** (0.30)	-1.54*** (0.29)	0.12*** (0.04)
Male Average	11.43	10.30	-1.14	0.60
Gender Diverse – Female (= Gender Minority Gap – Gender Gap)				
Difference	2.26	0.31	-0.28	0.08
p-value	<0.01	0.34	0.38	0.06
Performance FEs	No	Yes	No	No
N	746	746	746	746

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . SEs are robust and clustered at participant level. This table presents data from the *Adult Study*. Results are from OLS regressions of the dependent variable noted in the column. *Performance* is the number of questions that the participant answered correctly on the math and science test. *Reported Confidence* is the number of questions the participant believe that they answered correctly out of the 20 questions on the test. *Reported Confidence–Performance* is a participant's belief minus their actual performance. *1(Reported Confidence < Performance)* is a binary variable that takes the value of 1 if the participant was under-confident in their performance (i.e. their believed performance was worse than their actual performance) and otherwise zero. *Gender Diverse* is an indicator for the participant selecting "Transgender, non-binary, or another gender" or multiple options, when asked about their gender. *Female* is an indicator for the participant selecting only female when asked about their gender. *Male Average* is the average of the dependent value for participants selecting only male when asked about their gender. *Difference* is the difference between the *Female* and *Gender Diverse* coefficient estimates and *p-value* presents the corresponding p-value for a two-sided t-test of these two coefficient estimates. Performance FEs are dummies for each possible number of questions a participant got right out of the 20 questions on the test. Performance FEs are omitted from the analysis in Column (1) because the dependent variable is performance and from the analysis in Column (3) because the dependent variable is participant's belief minus actual performance.



Column (1) of Table 1 confirms that gender diverse participants perform significantly better than male and female participants, while Columns (2)–(4) show outcomes related to the gender minority gaps in confidence (as well as the gender gaps between men and women in confidence).

Column (2) of Table 1 includes performance fixed effects, i.e., indicators for each possible score that a participant could have received on the test, to allow us to compare equally-performing participants. The coefficient estimate on *Gender Diverse* reveals the gender minority gap in confidence: gender diverse participants believe they answered 1.45 fewer questions correctly than equally-performing male participants. The coefficient estimate on *Female* reveals the gender gap in confidence: female participants believe they answered 1.75 fewer questions correctly than equally performing male participants. The coefficient estimate on *Difference* and corresponding p-value then confirm that gender minority gap is approximately as large as the gender gap between men and women.

Column (3) presents similar results with a slightly different specification. Rather than including performance fixed effects as in Column (2), in Column (3) the dependent variable is adjusted to be the difference between a participant’s reported confidence and performance (i.e., the number of questions they report they got right minus the number of questions they actually got right). When this dependent variable is negative it suggests underconfidence; when it is positive it suggests overconfidence. The “Male Average” of  $-1.14$  suggests that on average men are underconfident in this setting. Nevertheless, the coefficient estimate on *Gender Diverse* confirms that there is still a gender minority gap in confidence: gender diverse participants are 1.82 questions *more underconfident* than male participants. The coefficient estimate on *Female* confirms there is also a gender gap in confidence measured this way: female participants are 1.54 questions more underconfident than male participants. The coefficient estimate on *Difference* and corresponding p-value confirms that the gender minority gap in confidence is approximately as large as the gender gap between men and women.

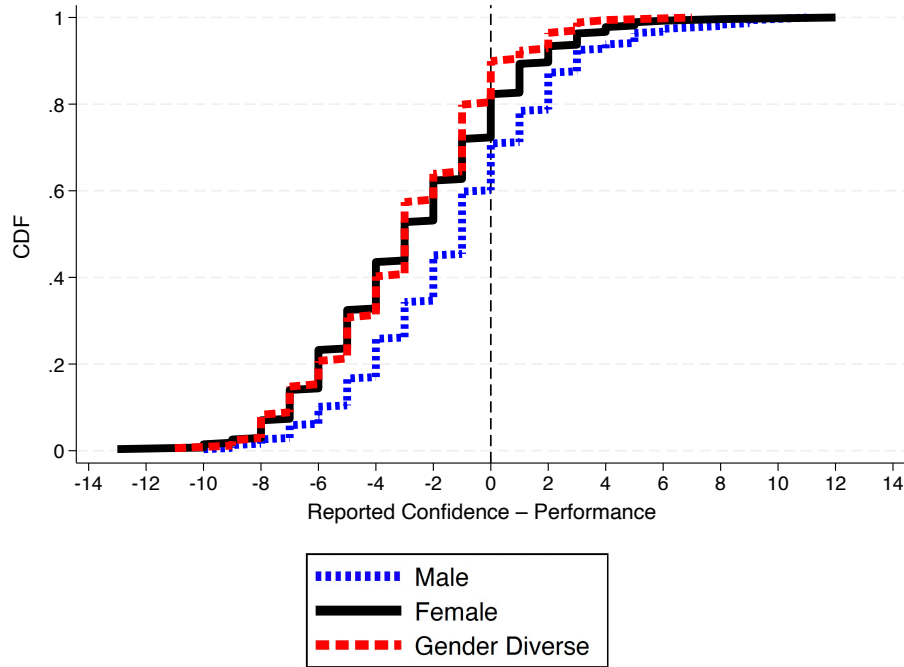
Column (4) presents similar results when instead considering a binary measure of a participant’s confidence, specifically, whether a participant is underconfident (i.e., whether their reported confidence falls below their actual performance). The coefficient estimate on *Gender Diverse* reveals the gender minority gap in underconfidence: gender diverse participants are 20 percentage points more likely to be underconfident than male participants. The coefficient estimate on *Female* reveals the gender gap in underconfidence: female participants are 12 percentage points more likely to be underconfident than male participants. The coefficient estimate on *Difference* and corresponding p-value further shows that the gender minority gap in underconfidence is somewhat larger than the gender gap in underconfidence

( $p = 0.06$ ).<sup>11</sup>

Finally, Figure 1 shows the CDFs of the differences between reported confidence and performance for the three groups and confirms that nearly the entire distribution is shifted to the left for the gender diverse participants (and female participants) relative to male participants.

**Result 1** (Gender Minority Gap in Confidence among Adults) Gender diverse adults report that they got fewer questions right, on average, than equally-performing male adults. Gender diverse adults are also more likely to underestimate the number of questions they got right and less likely to overestimate the number of questions they got right than male participants. These gender minority gaps in confidence are approximately as large as the gender gaps in confidence between men and women.

Figure 1: In the *Adult Study*, Reported Confidence–Performance Distributions



Graph shows CDFs for *Reported Confidence–Performance*, the number of questions a participant believes they answered correctly minus the number of questions a participant answered correctly. Positive responses suggest overconfidence while negative numbers suggest underconfidence.

## 2.4 The Gender Minority Gaps in Self-Evaluations Among Adults

Table 2 presents the regression results of participants’ self-evaluations. Each regression includes performance fixed effects. Thus, the coefficient estimates on *Gender Diverse* com-

<sup>11</sup>Appendix Table A.1 shows that gender diverse participants are less likely to be overconfident (see Column 1) and that there is no differences in the likelihood of them being accurate (see Column 2).

Table 2: In the *Adult Study*, Uninformed and Informed Self-Evaluations

	Performance Self- Evaluation (1)	Performance- Bucket (2)	Willingness (3)	Success (4)
<b>Panel A: Uninformed Self-Evaluations</b>				
Gender Diverse	-9.11*** (2.15)	-0.47*** (0.11)	-15.53*** (2.64)	-13.81*** (2.70)
Female	-9.82*** (1.98)	-0.50*** (0.10)	-15.64*** (2.35)	-13.98*** (2.40)
Male Average	49.55	3.97	44.04	48.05
Gender Diverse – Female (= Gender Minority Gap – Gender Gap)				
Difference	0.71	0.03	0.12	0.17
p-value	0.76	0.81	0.97	0.95
<b>Panel B: Informed Self-Evaluations</b>				
Gender Diverse	-2.99 (1.82)	-0.18** (0.09)	-14.92*** (2.53)	-12.01*** (2.54)
Female	-3.86** (1.60)	-0.17** (0.08)	-11.49*** (2.12)	-9.75*** (2.17)
Male Average	51.84	4.18	46.06	48.97
Gender Diverse – Female (= Gender Minority Gap – Gender Gap)				
Difference	0.87	-0.01	-3.43	-2.25
p-value	0.66	0.95	0.20	0.41
Performance FEs	Yes	Yes	Yes	Yes
N	746	746	746	746

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . SEs are robust and clustered at the participant level. This table presents data from the *Adult Study*. Results are from OLS regressions of a participant’s response to the uninformed (elicited before the participant learns their test performance) (Panel A) and informed (Panel B) self-evaluation noted in the column. *Performance Self-Evaluation* is the responses to the question in which participants are asked to indicate their agreement with the statement, “I performed well on the test.” *Performance-bucket* is the responses to the question in which participants are asked to indicate how well they think they performed on the test by choosing from the following list of seven adjectives: terrible, very poor, poor, neutral, good, very good, and exceptional. *Willingness* is the responses to the question in which participants are asked to indicate their agreement—on a scale from 0 (entirely disagree) to 100 (entirely agree)—with the statement, “I would apply for a job that required me to perform well on the test I took in Part 1.” *Success* is the responses to the question in which participants are asked to indicate their agreement—on a scale from 0 (entirely disagree) to 100 (entirely agree)—with the statement, “I would succeed in a job that required me to perform well on the test I took in Part 1.” See Table 1 for definitions of the independent variables, *Difference*, *p-value*, and Performance FEs.

pares gender diverse participants to equally-performing male participants, and the coefficient estimate on *Female* compares female participants to equally-performing male participants.

At the bottom of the table, the coefficient estimate on *Difference* reports the difference between gender diverse and equally-performing female participants.

Panel A of Table 2 presents participants’ uninformed self-evaluations. Column (1) presents results for the *performance self-evaluation* question that asked participants to indicate their agreement on a scale from 0 (entirely disagree) to 100 (entirely agree) with having “performed well on the test.” We find that gender diverse participants provide self-evaluation that are 9.11 points (18.3%) significantly lower on average than those provided by male participants. This gender minority gap is approximately as large as the 9.82-point gender gap observed between equally performing male and female participants.

Column (2) of Panel A of Table 2 presents results for the *performance-bucket* question that asked participants to indicate how well they think they performed on the test on a seven-point Likert scale. The average response provided by gender diverse participants is 0.47 points (11.8%) lower than the average response of equally performing male participants. This gender minority gap is approximately as large as the 0.50-point gender gap observed between equally performing male and female participants.

Column (3) of Panel A of Table 2 presents results for the *willingness* question that asked participants to indicate their agreement on a scale from 0 (entirely disagree) to 100 (entirely agree) with “I would apply for a job that required me to perform well on the test I took in Part 1.” The average response provided by gender diverse participants is 15.53 points (35.3%) lower than the average response of equally performing male participants. This gender minority gap is approximately as large as the 15.64-point gender gap observed between equally performing male and female participants.

Column (4) of Panel A of Table 2 presents results for the *success* question that asked participants to indicate their agreement on a scale from 0 (entirely disagree) to 100 (entirely agree) with “I would succeed in a job that required me to perform well on the test I took.” The average response provided by gender diverse participants is 13.81 points (28.7%) lower than the average response of equally performing male participants. This gender minority gap is approximately as large as the 13.98-point gender gap observed between equally performing male and female participants.

Appendix Figure A.1 shows CDFs of the responses to each of the four uninformed self-evaluation questions. Differences in the distributions of responses may be harder to interpret, however, because—unlike the regressions—they do not account for underlying differences in performance between the groups.

**Result 2** (Gender Minority Gaps in Uninformed Self-Evaluations Among Adults)  
Gender diverse adults provide worse self-evaluations of their performance on a math and science test than equally performing male adults on all four self-evaluation questions.

The gender minority gaps in self-evaluations are just as large as the gender gaps in self-evaluations between men and women.

Since gender diverse participants believe they answered fewer questions correctly on the test, one may wonder whether the gender minority gaps in self-evaluations persist even when we compare participants who answered the same number of questions correctly on the test and *know* how many questions they answered correctly on the test. To investigate this, we tell participants exactly how many questions they answered correctly on the test (and then have them report this number back to us to confirm they actually saw it). We then ask them the same four self-evaluation questions to elicit informed self-evaluations.

The informed self-evaluation results are presented in Panel B of Table 2. Even after the participants are told how many questions they answered right, we again observe gender minority gaps, albeit to a smaller degree in two of the four self-evaluation questions (compare the results across Panel A and Panel B).

As shown in Columns (1)–(4) of Panel B, gender diverse participants provide lower self-evaluations than equally-performing male participants. These gender minority gaps are statistically significant in three out of the four informed self-evaluation questions (see Columns (2)–(4)) and marginally significant in the remaining self-evaluation question ( $p = 0.10$  in Column (1)). As with uniformed self-evaluations, the gender minority gap in informed self-evaluations is approximately as large as the gender gap in informed self-evaluations.

Appendix Figure A.2 shows CDFs of the responses to each of the four informed self-evaluation questions. As mentioned earlier, differences in the distributions of responses may be harder to interpret, however, because—unlike the regressions—they do not account for underlying differences in performance between the groups.

**Result 3** (Gender Minority Gaps in Informed Self-Evaluations Among Adults) Even after they are informed of how many questions they got correct, gender diverse adults provide worse self-evaluations of their performance on a math and science test than equally-performing male adults on three out of four self-evaluation questions. The gender minority gaps in informed self-evaluations are just as large as the gender gaps in informed self-evaluations between men and women.

### 3 The Student Study

One may wonder whether gender minority gaps arise for younger individuals, particularly given their higher likelihood of identifying as gender diverse (Brown, 2022). Results from the *Student Study* show that the gender minority gaps also arise with a younger population.

### 3.1 The Design of the Student Study

The *Student Study* was conducted in the fall semester of 2020 with the partnership of the Character Lab Research Network (CLRN), which helped us recruit 10,807 students in grades 6–12 from a large school district in the United States. The students agreed to participate in a short study during the school day.<sup>12</sup> Some of the student data we analyze here was also analyzed in Exley and Kessler (2022). While that paper primarily leverages adult data to document gender gaps in self-evaluations between men and women (e.g., while varying the presence of incentives to self-promote), Section V of that paper explores gender gaps among middle school and high school students.<sup>13</sup> The Exley and Kessler (2022) analysis of the student data, however, exploits administrative data identifying every student as either male or female. In this paper, we instead explore students’ self-reported gender to generate new results on gender diverse individuals. Furthermore, for this paper, we use supplementary data on academic performance from our student sample during the academic quarter in which they participated in our study and the next seven quarters. This supplementary data allows us to document a correlation between our confidence and self-evaluation measures and student GPAs, as shown in Appendix Section B.1.<sup>14</sup>

Like the *Adult Study*, the *Student Study* also had six stages, but it had some minor differences. First, students were asked to answer 10 (instead of 20) math and science questions from the Armed Services Vocational Aptitude Battery. We requested that students try their best when answering, but there were no financial incentives in the study.<sup>15</sup> Second, we

---

<sup>12</sup>The following text from the CLRN website explains the data collection process in more detail: “This investigation was part of a larger data collection effort that included a variety of studies designed by scientists affiliated with Character Lab Research Network (CLRN)... This study was conducted on school computers during class time in participating schools over the course of a two- to three-week testing window. On a predetermined testing day, a teacher proctor at each school administered the CLRN research activities to students. To introduce the study, teachers read a script that explained to students that all research activities were part of an educational research initiative at their school, that participation was voluntary and they were not being graded, and that teachers would not see their answers. Teachers also instructed students to focus on their own computers and (if relevant) not to look at classmates’ screens. Upon logging into the CLRN platform, all students first viewed an assent screen that reiterated this information and, in addition, explained that parents would not see their responses and that their names and any other unique identifying information would not be shared with researchers. Students who agreed to participate were then directed to the survey.” This text was copied and pasted from the CLRN website. Website: <https://clrn.characterlab.org/resources/publishing-and-promotion#how-should-i-describe-character-lab-research-network-in-my-manuscript-s-methods-section> (accessed: October 13, 2023).

<sup>13</sup>Prior to having the idea of this paper, two of the authors on this paper already had access to the student data from their prior work in Exley and Kessler (2022). Thus, we did not pre-register the Student Study in this paper.

<sup>14</sup>This additional data collection also allowed us to validate more survey responses—which was done by matching unique identifiers in our data with the unique identifiers in the CLRN’s data—resulting in a slightly larger sample size than Exley and Kessler (2022).

<sup>15</sup>That the gender gaps among adults in Exley and Kessler (2022) are roughly identical with and without incentives to self-promote helps to mitigate potential concerns about the lack of monetary incentives in the

elicited each student’s belief about their absolute performance and their informed and uninformed self-evaluations similar to the *Adult Study* with two exceptions. In the *willingness* question, students were asked to indicate their agreement with “If given an option, I would choose to take a class that involves topics like those covered on the test.” In the *success* question, students were asked to indicate their agreement with “I would succeed in a class that involves topics like those covered on the test.” Third, again similar to the *Adult Study*, we asked students to complete a short follow-up survey at the end to gather demographic information, including a question about their gender where they select “male,” “female,” or “other.” If they selected other, they could choose to provide free response text about their gender identity. As explained in greater detail below (see Section 3.2), we use these responses to classify students by gender and to identify the students who are gender diverse. Figure E.18 shows specifically how we ask students to self-report their gender.

## 3.2 Gender Identity in the Student Study

A total of 10,807 students completed the *Student Study* in Fall 2020. Of these students, 48% selected male ( $n=5,187$ ), 50% selected female ( $n=5,412$ ), and 2% selected other ( $n=208$ ) when asked about their gender.<sup>16</sup> Out of the 208 students who selected other, we exclude 28 students who provided offensive responses in the corresponding free response text box. We classify the remaining 180 students as *gender diverse* (since they selected other as their gender identity and did not provide an offensive free response answer).<sup>17</sup> However, as detailed in Section 3.4.1, our results are robust to alternative classifications.

## 3.3 Gender Minority Gaps in Confidence in the Student Study

Gender diverse students answered an average of 5.87 questions correctly out of 10. This performance is statistically indistinguishable from the average performance of male students who answered an average of 5.90 questions correctly. Both of these performances are better than the average performance of female students, who answered an average of 5.44 questions correctly. Despite performing similarly to male students, gender diverse students believe they only answered an average of 5.21 questions correctly while male students believe they answered 6.65 questions correctly. To examine whether there is a gender minority gap in confidence between gender diverse students and *equally-performing* male students, we again turn to regression analyses that allow for such comparisons.

---

Student Study.

<sup>16</sup>The proportion of students who selected other is similar across students aged 11–18, where we have good data coverage. (We also have data on 11 ten-year-olds and 5 nineteen-year-olds; none of these 16 students selected other.)

<sup>17</sup>Some transgender students might not have chosen their gender identity as “other” and so would not be included in our definition of gender diverse. Since we do not have data on sex assigned at birth, we cannot identify if such individuals are present in our student data.



Table 3 follows the structure of Table 1 except for also controlling for year in school fixed effects (i.e., indicators for being in 6th grade, 7th grade, etc.) and school fixed effects (i.e., indicators for each school in the data). The regression results confirm that gender diverse students perform similarly to male students and better than female students.

Column (1) of Table 3 confirms that gender diverse students perform similarly to male students, and Columns (2)–(4) show that gender diverse students, nonetheless, have lower confidence. When compared to equally-performing male students, Column (2) shows that gender diverse students think they answered 1.41 (out of 10) fewer questions correctly. Relative to the actual number of questions they answered correctly, Column (3) shows that while male students overestimated their performance by 0.74 questions on average, gender diverse students were 1.41 questions more pessimistic about their performance relative to men and underestimated their performance on average. Column (4) shows that this pattern is also evident on the extensive margin: compared to the 29% of male students who are underconfident, gender diverse students are 20 percentage points more likely to be underconfident (and Appendix Table B.4 further shows that gender diverse students are 19% less likely to be overconfident). In addition, the gender minority gaps are consistently as large (or larger) than the gender gaps in confidence between men and women. See also Figure 2 to see these differences in confidence graphically.

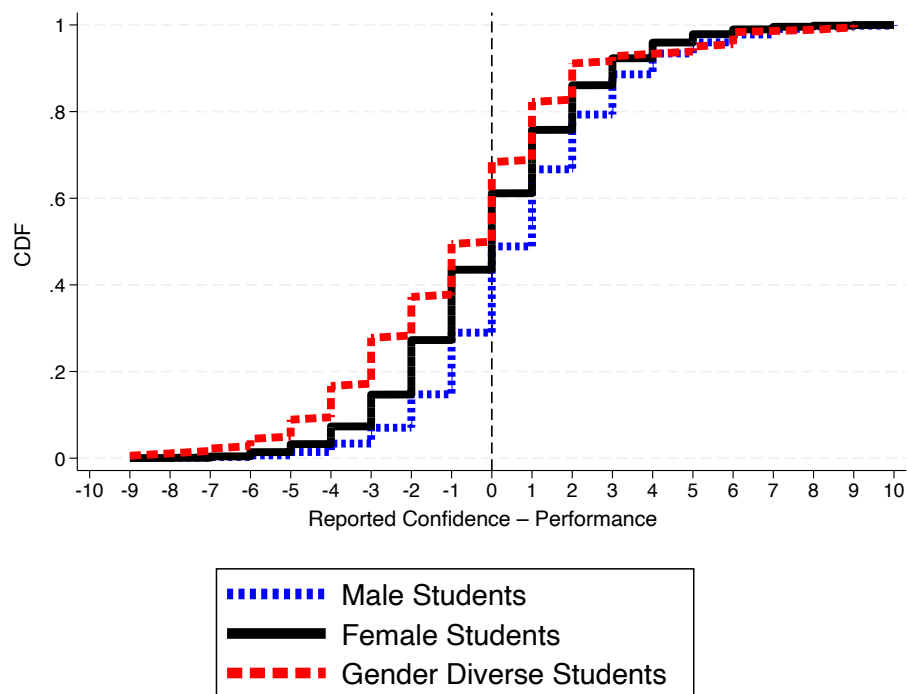
**Result 4** (Gender Minority Gap in Confidence Among Students) Gender diverse students report that they got fewer questions right, on average, than equally-performing male students. Gender diverse students are also more likely to underestimate the number of questions they got right and less likely to overestimate the number of questions they got right than male students. These gender minority gaps in confidence are as large (or larger) than the gender gaps in confidence between male and female students.

Table 3: In the *Student Study*, Performance and Beliefs

	Performance	Reported Confidence	Reported Confidence– Performance	1(Reported Confidence < Performance)
	(1)	(2)	(3)	(4)
Gender Diverse	-0.06 (0.16)	-1.41*** (0.21)	-1.41*** (0.22)	0.20*** (0.04)
Female	-0.46*** (0.04)	-1.03*** (0.04)	-0.80*** (0.05)	0.15*** (0.01)
Male Average	5.90	6.65	0.74	0.29
Gender Diverse – Female (= Gender Minority Gap – Gender Gap)				
Difference	0.40	-0.39	-0.61	0.05
p-value	0.01	0.06	0.01	0.12
Year in School FEs	Yes	Yes	Yes	Yes
School FEs	Yes	Yes	Yes	Yes
Performance FEs	No	Yes	No	No
N	10,779	10,779	10,779	10,779

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . SEs are robust and clustered at the participant level. This table presents data from the *Student Study*. Results are from OLS regressions of the dependent variable noted in the column. This follows a similar structure as Table 1 which provides definitions of the dependent variables, *Difference*, *p-value*, and Performance FEs. The only difference between Table 1 and this table is that this table presents data from the *Student Study* where the math and science test had 10 questions instead of 20. *Gender Diverse* is an indicator for the participant selecting other when asked about their gender and identifies the “gender minority gap.” *Female* is an indicator for the participant selecting female when asked about their gender and identifies the “gender gap.” Year in School FE and School FEs are dummies for each participant’s year in school (e.g., 6th grade, 7th grade, etc.) and school, respectively. The analysis excludes the 28 participants who selected other and provided an offensive response when asked about their gender and presents results from the remaining 10,779 participants.

Figure 2: In the *Student Study*, Reported Confidence–Performance Distributions



Graph shows CDFs for *Reported Confidence–Performance*, the number of questions a participant believes they answered correctly minus the number of questions a participant answered correctly. Positive responses suggest overconfidence while negative numbers suggest underconfidence.

### 3.4 Gender Minority Gaps in Self-Evaluations in the Student Study

Table 4 presents the regression results of students’ self-evaluations. Following the structure of Table 2, each regression includes performance fixed effects and thus allows us to compare equally-performing students. Like Table 3, Table 4 also controls for year in school fixed effects and school fixed effects.

Panel A of Table 4 presents participants’ uninformed self-evaluations. Relative to male students, Columns (1)–(4) show that gender diverse students provide self-evaluations that, on average, are: 17.46 points (26.3%) lower on the 0–100 *performance self-evaluation* question, 0.75 points (16.0%) lower on 7-point Likert *performance-bucket* question, 9.62 points (17.0%) lower on the 0–100 *willingness* question, and 16.09 points (23.5%) lower on the 0–100 *success* question. All of these gender minority gaps are statistically significant and they are all significantly larger than the corresponding gender gaps between male and female students.

**Result 5** (Gender Minority Gaps in Uninformed Self-Evaluations Among Students)

Gender diverse students provide worse self-evaluations of their performance on a math and science test than equally performing male students on all four self-evaluation questions. The gender minority gaps in self-evaluations are larger than the gender gaps in self-evaluations between male and female students.

Since gender diverse students believe they answered fewer questions correctly on the test, one may again wonder whether the gender minority gaps in self-evaluations persist even when we compare students who answered the same number of questions correctly on the test and *know* how many questions they answered correctly on the test. As also observed in the *Adult Study*, the answer is clearly yes.

The results about informed self-evaluations are presented in Panel B of Table 4. Even after the students are told how many questions they answered right, we again observe significant and substantial gender minority gaps in response to all four self-evaluation questions. In addition, as with uninformed self-evaluations, these gender minority gaps are larger than the gender gaps between male and female students.<sup>18</sup>

**Result 6** (Gender Minority Gaps in Informed Self-Evaluations Among Students)

Even after students are informed of how many questions they got correct, gender diverse students provide worse self-evaluations of their performance on a math and science test than equally-performing male students in all four self-evaluation questions. The gender minority gaps in informed self-evaluations are larger than the gender gaps in

---

<sup>18</sup>See Appendix Figures B.1 and B.2 for CDFs of the responses to each of the four uninformed and informed self-evaluation questions. But, we note that differences in the distributions of responses may be harder to interpret because—unlike the regressions—they do not account for underlying differences in performance between the groups.

Table 4: In the *Student Study*, Uninformed and Informed Self-Evaluations

	Performance Self- Evaluation (1)	Performance- Bucket (2)	Willingness (3)	Success (4)
<b>Panel A: Uninformed Self-Evaluations</b>				
Gender Diverse	-17.46*** (2.13)	-0.75*** (0.11)	-9.62*** (2.46)	-16.09*** (2.38)
Female	-10.97*** (0.45)	-0.52*** (0.02)	-4.27*** (0.58)	-7.48*** (0.54)
Male Average	66.42	4.70	56.52	68.34
Gender Diverse – Female (= Gender Minority Gap – Gender Gap)				
Difference	-6.49	-0.23	-5.35	-8.60
p-value	<0.01	0.04	0.03	<0.01
<b>Panel B: Informed Self-Evaluations</b>				
Gender Diverse	-13.61*** (2.23)	-0.54*** (0.12)	-11.94*** (2.52)	-17.34*** (2.46)
Female	-6.44*** (0.52)	-0.26*** (0.03)	-2.94*** (0.60)	-5.34*** (0.59)
Male Average	45.84	3.60	51.27	57.52
Gender Diverse – Female (= Gender Minority Gap – Gender Gap)				
Difference	-7.17	-0.28	-9.01	-11.99
p-value	<0.01	0.02	<0.01	<0.01
Year in School FEs	Yes	Yes	Yes	Yes
School FEs	Yes	Yes	Yes	Yes
Performance FEs	Yes	Yes	Yes	Yes
N	10,779	10,779	10,779	10,779

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . SEs are robust and clustered at the participant level. This table presents data from the *Student Study*. Results are from OLS regressions of a student’s response to the uninformed (elicited before the student learns their test performance) (Panel A) and informed (Panel B) self-evaluation questions noted in the column. See Table 2 for definitions of the dependent variables with the exception of two dependent variables where the questions were worded slightly differently for the *Student Study*. *Willingness* and *Success* were different. *Willingness* is the responses to the question in which students are asked to indicate their agreement—on a scale from 0 (entirely disagree) to 100 (entirely agree)—with the statement, “If given an option, I would choose to take a class that involves topics like those covered on the test.” *Success* is the responses to the question in which students are asked to indicate their agreement—on a scale from 0 (entirely disagree) to 100 (entirely agree)—with the statement, “I would succeed in a class that involves topics like those covered on the test.” Also see Table 3 for definitions of the independent variables, *Difference*, *p-value*, and FEs. Our analysis excludes the 28 students who selected other and provided an offensive response when asked about their gender and presents the remaining 10,779 students.

informed self-evaluations between male and female students.

### 3.4.1 Robustness in the Student Study

The gender minority gaps in confidence and in self-evaluation that we identify are robust to different ways of classifying students as gender diverse. Appendix Table B.5 describes four sets of robustness checks that we ran with our data, which we summarize here.

In the first set of robustness tests, we define gender diverse as anyone who selected “other” (i.e., we include the 28 students with offensive responses in the text box as gender diverse).

In the second set of robustness tests, we classify the 74 students who selected other and provided details on the nature of their gender identity in the corresponding free response text box as *explicitly gender diverse*.<sup>19</sup> We then only keep these explicitly gender diverse students and drop everyone else who selected other (i.e., we drop students who provided offensive responses, those who left the text box blank, and those who did not provide an informative response about their gender identity).

In the third and fourth set of robustness tests, we rely on gender data collected by the Character Lab Research Network (CLRN) in a demographics survey that was run before our study. Using that survey, we classify participants as male, as female, as those who selected “Other” when asked about their gender, and as those who selected “Prefer not to say” when asked about their gender. In the third set of robustness tests, we use the CLRN survey for gender classification, dropping the 535 students who selected “Prefer not to say.” In the fourth set of robustness tests, we primarily use the CLRN survey for gender classification and use responses to our survey question only to classify those who selected “Prefer not to say” (for more details, see Appendix Table B.5).

Appendix Tables B.6–B.8 replicate the analysis conducted in Tables 3 and 4, showing results for each of the four sets of robustness tests. The results identified in Sections 3.3 and 3.4 are highly robust. All 40 of the differences in confidence and self evaluations that we estimate between students who we classify as gender diverse and students we classify as male are statistically significant at  $p < 0.05$  (with 39 significant at  $p < 0.01$ ). Across all specifications, these gender minority gaps are large and, in most cases, also larger than the corresponding gender gaps we see when we compare responses of male and female students.<sup>20</sup>

---

<sup>19</sup>Most of these students mentioned their gender being something different than male or female such as non-binary, transgender, agender, demigirl, demiboy, gender fluid, or pangender; others provided their gender pronouns (such as she/they, he/they, they/them); and a few noted that they were still questioning. The remaining 106 students who selected other provided either no response or a response that was not specific enough for us to classify them as explicitly gender diverse. Specifically, 99 of them left the text box empty, 1 wrote “boy,” 1 wrote “kid,” 1 wrote “uhhhhh,” 1 mentioned that they answered this question already, and 3 mentioned that they prefer not to say.

<sup>20</sup>In particular, 35 out of 40 of the differences in confidence and self evaluations that we estimate between students who we classify as gender diverse and students we classify as female are statistically significant at

## 4 The Predictions Study

Sections 2 and 3 document robust evidence of gender minority gaps in confidence and self-evaluations across various measures in both an adult population and a youth population. Given this evidence, one could be tempted to conclude that these gender minority gaps are akin to gender differences between men and women in confidence and self-evaluations.

However, we posited the possibility of a key difference between these gaps. Unlike the well-documented gender differences in self-evaluations and confidence between men and women that are expected (Exley and Nielsen, 2024), we speculated that these novel gender minority gaps might be *unexpected*. The *Predictions Study*, detailed in the following two subsections, investigates people’s beliefs about gender diverse, male, and female participants.

### 4.1 Predictions Study Design

We recruited 600 participants to be “predictors” in our pre-registered Predictions Study.<sup>21</sup> Specifically, we recruited a sex-balanced sample of U.S. Prolific participants who did not participate in other studies discussed in this paper. Predictors are paid \$3 to participate in the study. Additionally, one of the 21 questions they answer is chosen at random and the predictor can earn an additional \$1 if they answer that question correctly. The study proceeded as follows (further instructions and design details can be found in Online Appendix E.4).

First, we provide predictors with information about the *Adult Study* (we refer to this as the “prior study” for the predictors). Then, we elicit incentivized predictions about the performance of participants in the Adult Study and about the reported confidence of participants in the Adult Study. The order of these two types of predictions is randomized at the predictor level.

For the incentivized predictions about performance, predictors are provided with information about the confidence or self-evaluations of a group of participants and are asked to predict their actual test performance. In particular, they are asked two sets of nine questions for a total of 18 questions. The order of these sets is randomized at the predictor level, and the order of questions within each set is also randomized at the predictor level.

In nine of these questions, we ask predictors to consider the group of either female, male, or gender diverse participants who guessed that they answered either 5, 10, or 15 questions correctly on the math and science test (out of 20). We then ask them to predict how many questions, on average, these participants in that group actually answered correctly on the

---

$p < 0.05$  (with 27 significant at  $p < 0.01$ ), indicating that the gender minority gap (between male and gender diverse students) is bigger than the corresponding gender gap (between male and female students).

<sup>21</sup>The study was pre-registered on AsPredicted (#184073) which can be accessed here: <https://aspredicted.org/y97t-22mj.pdf>.



math and science test. Predictors indicate their answer on a slider (see, e.g., Appendix Figure E.22), and their answer is correct if the slider includes the true average.

In the other nine questions, we ask predictors to consider the group of either female, male, or gender diverse participants who assigned their performance either a low rating (between 0 and 33), a medium rating (between 34 and 66), or a high rating (between 67 and 100) in response to the *performance self-evaluation* question (see Table 2 for a description of this self-evaluation question). We then ask them to predict how many questions, on average, participants in that group answered correctly on the math and science test. Predictors indicate their answer on a slider (see, e.g., Appendix Figure E.24), and their answer is correct if the range selected by the slider includes the truth.

For the incentivized predictions about the reported confidence of participants in the Adult Study, we directly ask predictors three questions about the reported confidence of each group of participants: female participants, male participants, and gender diverse participants. To begin, we inform predictors that a participant is overconfident if they overestimated how many questions they got right, accurate if they correctly guessed how many they got right, or underconfident if underestimated how many questions they got right (see Figure E.25). We then ask predictors to guess whether a randomly selected participant—who is known to be female, male, or gender diverse—is either overconfident, accurate, or underconfident. We also give predictors the option to indicate if they are unsure. Predictors are told that if they choose the “I’m unsure” option in the question chosen for payment, they will earn \$1 with a 50% chance. An example decision screen is shown in Figure E.27. The order of the three questions is randomized at the predictor level.

Finally, we also ask 15 unincentivized questions to measure broader beliefs about female, male, and gender diverse individuals in general—rather than in relation to participants in our Adult Study. In these questions, we ask predictors whether female, male, and gender diverse people are likely to be overconfident or underconfident in their performance and abilities in math and science tasks and whether, in general, female, male, and gender diverse people are likely to take risks, be competitive, and be generous. The order of these questions is also randomized at the predictor level.

## 4.2 Predictions Study Results

Table 5 presents the average predicted performance when predictors are asked about male participants (see Column 1), gender diverse participants (see Column 2), and female participants (see Column 3).

The first three rows of Table 5 show the average predicted performance (i.e., the number of questions predictors think participants got right) for participants who *self-report* that they got 5, 10 or 15 questions right. The next three rows show the average predicted performance

for participants who reported low (0–33), medium (34–66), or high (67–100) *self-evaluations*. As one may expect, predictors expect higher performance among participants with higher reported confidence and higher self-evaluations.<sup>22</sup> These results also reveal that predictors—for a given level of reported confidence or self-evaluation—expect very similar performances from male participants (Column 1) and gender diverse participants (Column 2), on average.<sup>23</sup>

Since we have shown above that gender diverse participants have worse confidence and self-evaluations than equally-performing male participants, this pattern of average beliefs—i.e., believing men and gender diverse people perform equally well given the same confidence and self-evaluations—provides evidence that predictors fail to anticipate gender minority gaps. Appendix Table C.1 indeed confirms that the predicted performance does not statistically significantly differ when predictors are asked about gender diverse versus male participants with the same reported confidence or same self-evaluations (see the close-to-zero and insignificant coefficient estimate on *Predicted Performance of Gender Diverse Participant*).<sup>24</sup> Correspondingly, Appendix Table C.4 shows that this pattern of average beliefs results in predictors underestimating the true performance of gender diverse participants by 1.64–2.31 questions but only underestimating the performance of male participants by a much more modest 0.23–0.77 questions.

Extending beyond predictions that directly elicit the performance of participants, additional results—specifically predictions that directly ask about the confidence of participants—provide novel insight. Figure 3 shows how often predictors indicate that participants are overconfident, accurate, underconfident—or that they are “unsure”—when predictors are asked about male participants (see Panel A), gender diverse participants (see Panel B), or female participants (see Panel C). Clear differences in predictions about all three groups of participants now emerge.

Panel A of Figure 3 shows that the vast majority of predictors (72%) indicate that men are overconfident, which is about 22 times greater than the percentage indicating that men are underconfident (3.3%) and about 6 times greater than the percentage indicating that

---

<sup>22</sup>As one also may mechanically expect given the available room for underestimating or overestimating one’s performance (or given pull to center effects), predictors also expect that participants with lower (higher) self-reports are more likely to underestimate (overestimate) their performance.

<sup>23</sup>After being provided with such performance signals, predictors do not expect the same performances between female participants (Column 3) and male participants (Column 1). Rather, predictors expect that female participants have a better performance than male participants. While one could be tempted to conclude that this presents evidence in favor of predictors “accurately accounting for” the gender gap in confidence, we note that—absent knowing the full distribution of prior beliefs—it is difficult to calculate what predictors’ beliefs should be if they accurately accounted for the gender minority gaps. In addition, as with the findings in Exley and Nielsen (2024), it could also be that predictors expect gender gaps but do not accurately account for them. To gain insight into whether gaps are expected, we directly elicit predictors’ expectations in another set of predictions explained next.

<sup>24</sup>See also Appendix Tables C.2 and C.3 for these results by specific self-reports and self-evaluations.

Table 5: In the *Predictions Study*, Average Predicted Performances

	Predictions about:		
	Male Participants	Gender Diverse Participants	Female Participants
<i>Average predicted performance of participants who</i>			
report that they got 5 questions right	7.64	7.79	8.15
report that they got 10 questions right	10.52	10.68	11.22
report that they got 15 questions right	13.03	13.06	13.84
assign low self-evaluations of 0 to 33	7.95	7.94	8.29
assign medium self-evaluations of 34 to 66	11.34	11.33	11.84
assign high self-evaluations of 67 to 100	14.24	14.08	14.81
N	600	600	600

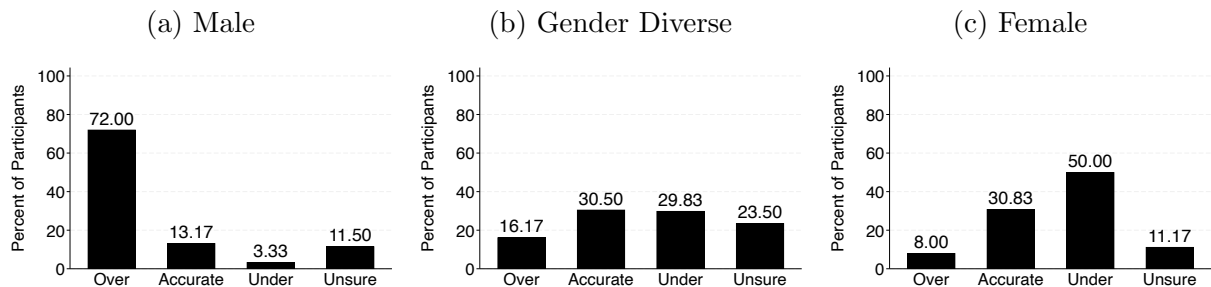
This table provides average responses to each of the incentivized questions used in the Predictions Study. Predictors are asked to consider the group of [female]/[male]/[gender diverse] prior participants who guessed that they answered [5]/[10]/[15] questions (out of 20) correctly on the math and science test, then, asked to predict how many questions, on average, they think these previous participants answered correctly on the math and science test. The first three rows report average responses to each of these incentivized questions. Predictors are also asked to consider the group of [female]/[male]/[gender diverse] prior participants who assigned their performance [a low rating of 0 to 33]/[a medium rating of 34 to 66]/[a high rating of 67 to 100] in response to the *performance self-evaluation* question (see Table 2 for a description of this self-evaluation question), then, asked to predict how many questions, on average, they think these prior participants answered correctly on the math and science test. The bottom three rows report average responses to each of these incentivized questions.

men are accurate (13.2%). By contrast, Panel B shows a much less skewed distribution for predictions about gender diverse participants relative to predictions about male participants. Only a minority of predictors indicate that gender diverse participants are underconfident (16.2%), and predictors are equally split between indicating that gender diverse participants are accurate (30.5%) and underconfident (29.8%). In addition, a substantial and statistically significantly ( $p < 0.01$ ) larger percentage of predictors directly indicate that they are unsure about the confidence of gender diverse participants (23.5%) relative to the percentage indicating they are unsure about male participants (11.5%) and female participants (11.2%, as shown in Panel C).<sup>25</sup>

That predictors are unsure about the confidence of gender diverse people is also evident from our unincentivized questions that measure broader beliefs. As shown in Appendix Figure C.1, predictors are much less sure about whether gender diverse people are overconfident

<sup>25</sup> $p < 0.01$  results from a linear probability model of being unsure on an indicator for predictions about male participants and an indicator for predictions about female participants, with SEs clustered at the prediction level.

Figure 3: Predictions about Reported Confidence



Participants are asked to guess whether [male]/[gender diverse]/[female] participants are overconfident, accurate, underconfident or that they are unsure (coin flip). Graphs show distributions of responses for men, gender diverse, and female prior participants.

or underconfident in general. This contrasts with predictors being very likely to report that men are overconfident (and not underconfident) and that women are underconfident (and not overconfident) in general.

In addition, akin to our finding that predictors are much more likely to say they are “unsure” (to receive a lottery payoff rather than payoff that depends on whether their prediction is correct) when asked about gender diverse participants, Appendix Figure C.1 reveals that one-third of predictors respond “maybe” when asked about whether gender diverse people are overconfident or underconfident in general (while “maybe” is never selected more than 10% of the time when predictors are asked about men or women).

Similar results follow when we ask about three other traits of gender diverse people, men, and women. Anywhere from 29–37% of predictors answer “maybe” and indicate they are unsure about the behavior of gender diverse people when they are asked about whether they are likely to take risks (Appendix Figure C.2), to be competitive (Appendix Figure C.3), or to be generous (Appendix Figure C.4). By contrast, the percent of predictors selecting “maybe” is anywhere from 2–18% for men and 8–12% for women—with predictors almost uniformly expecting men to be take risks, men to be competitive, and women to be prosocial. This last finding echoes the robust believed gender differences in social preferences documented in Exley et al. (2024).<sup>26</sup>

**Result 7** (Gender Minority Gaps in Confidence and Self-Evaluations are Unexpected) The gender minority gaps in confidence and self-evaluations are largely unexpected. Moreover, predictors are significantly more likely to express uncertainty about the confidence and other traits of gender minorities while they have more certain beliefs about male and female individuals.

<sup>26</sup>For a meta-analysis on gender differences in generosity, see Bilén, Dreber and Johannesson (2021).

## 5 Discussion

In this paper, we document gender minority gaps in confidence and self-evaluation. In the *Adult Study*, we document gender minority gaps in confidence and self-evaluations on a math and science test. Gender diverse adults believe they answered fewer questions correctly and provide less favorable self-evaluations—even after being informed of exactly how many questions they answered correctly—than equally-performing men. These gender minority gaps are sizable and are just as large as the gender gaps between men and women. In the *Student Study*, we again document gender minority gaps in confidence and self-evaluations. These gender minority gaps are even larger than the gender gaps between men and women.

Additional results reveal that the gender minority gaps are unexpected. While men are typically expected to be overconfident and women are typically expected to be underconfident, there is no clear expectation about the confidence of gender diverse people. People frequently admit to being unsure about the confidence of gender diverse people, highlighting a clear challenge to countering the gender minority gaps we document.

The results in our paper open up many important avenues for future work. One direction is to explore various settings to see when gender minority gaps are more or less likely to arise. For instance, motivated by the literature that shows that gender differences between men and women can be domain specific (Günther et al., 2010; Shurchkov, 2012; Coffman, 2014; Dreber, von Essen and Ranehill, 2014; Bordalo et al., 2019; Boschini et al., 2019; Coffman, Collis and Kulkarni, 2019; Coffman, Flikkema and Shurchkov, 2019; Atwater and Saygin, 2020), we randomized a set of adult participants to take a verbal test rather than a math and science test (see Appendix D for details and analysis of this *Adult (Verbal) Study*). When we switch from a math and science domain, which is typically male-stereotyped, to a verbal test, which is typically considered less male-stereotyped, the gender minority gaps shrink dramatically and almost all go away. We hope future work, and ultimately a body of literature, how different features of a setting affect gender minority gaps.

Future work may also explore the norms, beliefs, and stereotypes related to gender diverse individuals, particularly given the connection between gender norms and many important outcomes (Bertrand, Kamenica and Pan, 2015; Dhar, Jain and Jayachandran, 2022; Field et al., 2021; Pande and Roy, 2021; Jayachandran et al., 2023) as well as prior work on inaccurate gender beliefs and misperceptions of gender norms (Bordalo et al., 2019; Bursztyn, González and Yanagizawa-Drott, 2020; Coffman, Exley and Niederle, 2021; Bohren et al., 2023; Bursztyn, Cappelen and Tungodden, 2023).

Results from our *Predictors Study* already highlight that, particularly relative to beliefs about men and women, individuals express substantial *uncertainty* in their predictions about gender diverse people. This uncertainty arises when participants are asked about traits such

as confidence (as is the focus in this paper) as well as when asked about other important traits and behaviors such as risk taking, competitiveness, and generosity. Indeed, we speculate that this uncertainty may be a defining feature of beliefs about gender minorities and could contribute to policy roadblocks, discrimination, inattention, and a lack of understanding. We hope future work considers these possibilities among other important avenues, which, in addition to directly informing the gender literature, may yield broader insights into under-represented and historically disadvantaged minority groups.

We also hope future work explores diversity among gender minorities. For example, future work may seek to separately study those who identify as transgender men, transgender women, non-binary individuals, genderqueer individuals, or gender non-conforming individuals. Future work may also aim to study other minority groups, such as those related to sexual orientation (e.g., see [Coffman, Coffman and Ericson, 2017](#); [Buser, Geijtenbeek and Plug, 2018](#); [Lewis et al., 2017](#); [Aksoy, Carpenter and Sansone, 2025](#); [Aksoy and Chadd, 2025](#)), and the impact of intersectionality, such as the impact of being a gender minority and a sexual minority.<sup>27</sup>

As we look toward future work, it is worth noting that even if individual studies or experiments may be under-powered to examine gender minorities, allowing for more inclusive measures of gender in surveys (e.g., options beyond binary gender options) can be potentially quite useful in generating data for subsequent meta-analyses to provide insights. Future work may also include additional measures of gender, such as the continuous gender identity measures in [Brenøe et al. \(2022\)](#) and [Piasenti and Süer \(2024\)](#). For a discussion of current best practices for inclusive gender identity (and sexual orientation) questions, see [Aksoy et al. \(2024\)](#). The need for this work will only increase as policies, such as the aforementioned executive order, limit insights that other data sources may provide.

---

<sup>27</sup>Considering intersectionality, [Aksoy, Chadd and Koh \(2023\)](#) find that women, relative to men, are more likely to hide their LGBTQ+ affinity due to anticipated discrimination. Many of these lines of future work would contribute to a growing field of LGBTQ+ economics (for a literature review, see [Badgett, Carpenter and Sansone, 2021](#); [Badgett et al., 2023](#))

## References

- Aksoy, Billur, and Ian Chadd.** 2025. “Competitiveness at the Intersection of Gender and Sexual Orientation.” *Journal of Economic Behavior & Organization*, 233: 106987.
- Aksoy, Billur, Christopher S. Carpenter, and Dario Sansone.** 2025. “Understanding Labor Market Discrimination Against Transgender People: Evidence from a Double List Experiment and a Survey.” *Management Science*, 71(1): 659–677.
- Aksoy, Billur, Ian Chadd, and Boon Han Koh.** 2023. “Sexual Identity, Gender, and Anticipated Discrimination in Prosocial Behavior.” *European Economic Review*, 154: 104427.
- Aksoy, Billur, Ian Chadd, Brit Grosskopf, and Boon Han Koh.** 2024. “Sexual Orientation and Gender Identity.” *Available at SSRN*.
- Atwater, Ann, and Perihan O. Saygin.** 2020. “Gender Differences in Leaving Questions Blank on High-Stakes Standardized Test.” *Working Paper*.
- Babcock, Linda, and Sara Laschever.** 2003. *Women don’t ask: negotiation and the gender divide*. Princeton, NJ:Princeton University Press.
- Badgett, M.V. Lee, Christopher S. Carpenter, and Dario Sansone.** 2021. “LGBTQ economics.” *Journal of Economic Perspectives*, 35(2): 141–70.
- Badgett, M.V. Lee, Christopher S. Carpenter, Maxine J. Lee, and Dario Sansone.** 2023. “A Review of the Economics of Sexual Orientation and Gender Identity.” *Journal of Economic Literature*, Forthcoming.
- Barber, Brad M., and Terrance Odean.** 2001. “Boys will be boys: Gender, overconfidence, and common stock investment.” *The quarterly journal of economics*, 116(1): 261–292.
- Bertrand, Marianne, Claudia Goldin, and Lawrence F Katz.** 2010. “Dynamics of the gender gap for young professionals in the financial and corporate sectors.” *American economic journal: applied economics*, 2(3): 228–255.
- Bertrand, Marianne, Emir Kamenica, and Jessica Pan.** 2015. “Gender identity and relative income within households.” *The Quarterly Journal of Economics*, 130(2): 571–614.
- Bilén, David, Anna Dreber, and Magnus Johannesson.** 2021. “Are women more generous than men? A meta-analysis.” *Journal of the Economic Science Association*, 7(1): 1–18.
- Blau, Francine D., and Lawrence M. Kahn.** 2017. “The Gender Wage Gap: Extent, Trends, and Explanations.” *Journal of Economic Literature*, 55(3).



- Bohren, J. Aislinn, Kareem Haggag, Alex Imas, and Devin G. Pope.** 2023. “Inaccurate statistical discrimination: An identification problem.” *Review of Economics and Statistics*, 1–45.
- Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer.** 2019. “Beliefs about Gender.” *American Economic Review*.
- Boschini, Anne, Anna Dreber, Emma Von Essen, Astri Muren, and Eva Ranehill.** 2019. “Gender, risk preferences and willingness to compete in a random sample of the Swedish population.” *Journal of Behavioral and Experimental Economics*, 83: 101467.
- Brenøe, Anne Ardila, Lea Heursen, Eva Ranehill, and Roberto A Weber.** 2022. “Continuous gender identity and economics.” *AEA Papers and Proceedings*, 112: 573–577.
- Brown, Anna.** 2022. “About 5% of young adults in the U.S. say their gender is different from their sex assigned at birth.” *Pew Research Center*.
- Bursztyn, Leonardo, Alessandra L González, and David Yanagizawa-Drott.** 2020. “Misperceived social norms: Women working outside the home in Saudi Arabia.” *American Economic Review*, 110(10): 2997–3029.
- Bursztyn, Leonardo, Alexander W Cappelen, and Bertil Tungodden.** 2023. “How are gender norms perceived?” *National Bureau of Economic Research*.
- Bursztyn, Leonardo, Thomas Fujiwara, and Amanda Pallais.** 2017. “‘Acting Wife’: Marriage Market Incentives and Labor Market Investments.” *The American Economic Review*, 107(11): 3288–3319.
- Buser, Thomas, Lydia Geijtenbeek, and Erik Plug.** 2018. “Sexual orientation, competitiveness and income.” *Journal of Economic Behavior & Organization*, 151: 191–198.
- Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek.** 2014. “Gender, competitiveness, and career choices.” *The quarterly journal of economics*, 129(3): 1409–1447.
- Bütikofer, Aline, Katrine V. Løken, and Alexander Willén.** 2022. “Building Bridges and Widening Gaps.” *Review of Economics and Statistics*.
- Carpenter, Christopher S., Maxine J. Lee, and Laura Nettuno.** 2022. “Economic outcomes for transgender people and other gender minorities in the United States: First estimates from a nationally representative sample.” *Southern Economic Journal*, 89(2): 280–304.
- Carpenter, Christopher S., Samuel T. Eppink, and Gilbert Gonzales.** 2020. “Transgender status, gender identity, and socioeconomic outcomes in the United States.” *ILR Review*, 73(3): 573–599.

- Coffman, Katherine Baldiga.** 2014. “Evidence on Self-Stereotyping and the Contribution of Ideas.” *The Quarterly Journal of Economics*, 129(4): 1625–1660.
- Coffman, Katherine B., Christine L. Exley, and Muriel Niederle.** 2021. “The Role of Beliefs in Driving Gender Discrimination.” *Management Science*, 67(6): 3321–3984.
- Coffman, Katherine B, Lucas C Coffman, and Keith Marzilli Ericson.** 2024. “Non-Binary Gender Economics.” National Bureau of Economic Research.
- Coffman, Katherine B, Lucas C Coffman, and Keith M Marzilli Ericson.** 2017. “The size of the LGBT population and the magnitude of antigay sentiment are substantially underestimated.” *Management Science*, 63(10): 3168–3186.
- Coffman, Katherine, Clio Bryant Flikkema, and Olga Shurchkov.** 2019. “Gender Stereotypes in Deliberation and Team Decisions.” *Harvard Business School Working Paper*.
- Coffman, Katherine, Manuela Collis, and Leena Kulkarni.** 2019. “Stereotypes and Belief Updating.” *Working Paper*.
- Demiral, Elif E, and Johanna Mollerstrom.** 2024. “Signaling confidence.” *Journal of Economic Behavior & Organization*, 226: 106691.
- Dhar, Diva, Tarun Jain, and Seema Jayachandran.** 2022. “Reshaping adolescents’ gender attitudes: Evidence from a school-based experiment in India.” *American economic review*, 112(3): 899–927.
- Downing, Janelle M., and Julia M. Przedworski.** 2018. “Health of transgender adults in the US, 2014–2016.” *American Journal of Preventive Medicine*, 55(3): 336–344.
- Dreber, Anna, Emma von Essen, and Eva Ranehill.** 2014. “Gender and competition in adolescence: task matters.” *Experimental Economics*, 17(1): 154–172.
- Eckel, Catherine C., and Philip J. Grossman.** 2008. *Men, women and risk aversion: experimental evidence*, [in:] *Handbook of Experimental Economics Results*. Vol. 1, Amsterdam, Oxford, North Holland.
- Exley, Christine L., and Judd B. Kessler.** 2022. “The Gender Gap in Self-Promotion.” *Quarterly Journal of Economics*.
- Exley, Christine L., and Kirby Nielsen.** 2024. “The Gender Gap in Confidence: Expected But Not Accounted For.” *American Economic Review*, 14(3): 851–885.
- Exley, Christine L., Muriel Niederle, and Lise Vesterlund.** 2020. “Knowing When to Ask: The Cost of Leaning-in.” *Journal of Political Economy*, 128(3): 816–854.

- Exley, Christine L, Oliver P Hauser, Molly Moore, and John-Henry Pezzuto.** 2024. “Believed gender differences in social preferences.” *The Quarterly Journal of Economics*, 140(1): 403–458.
- Field, Erica, Rohini Pande, Natalia Rigol, Simone Schaner, and Charity Troyer Moore.** 2021. “On her own account: How strengthening women’s financial control impacts labor supply and gender norms.” *American Economic Review*, 111(7): 2342–75.
- Grossman, Philip J., Catherine C. Eckel, Mana Komai, and Wei Zhan.** 2019. “It pays to be a man: Rewards for leaders in a coordination game.” *Journal of Economic Behavior & Organization*, 161: 197–215.
- Günther, Christina, Neslihan Arslan Ekinici, Christiane Schwierén, and Martin Strobel.** 2010. “Women can’t jump?—An experiment on competitive attitudes and stereotype threat.” *Journal of Economic Behavior & Organization*, 75(3): 395–401.
- Hernandez-Arenaz, Iñigo, and Nagore Iriberrí.** 2019. “A review of gender differences in negotiation.” *Oxford Research Encyclopedia of Economics and Finance*.
- Jayachandran, Seema, Lea Nassal, Matthew Notowidigdo, Marie Paul, Heather Sarrons, and Elin Sundberg.** 2023. “Moving to opportunity, together.”
- Jones, Jeffrey M.** 2022. “LGBT Identification in U.S. Ticks Up to 7.1%.”
- Lewis, Daniel C., Andrew R. Flores, Donald P. Haider-Markel, Patrick R. Miller, Barry L. Tadlock, and Jami K. Taylor.** 2017. “Degrees of acceptance: Variation in public attitudes toward segments of the LGBT community.” *Political Research Quarterly*, 70(4): 861–875.
- Lundeberg, Mary A., Paul W. Fox, and Judith Punčcohař.** 1994. “Highly confident but wrong: Gender differences and similarities in confidence judgments.” *Journal of educational psychology*, 86(1).
- Meyer, Ilan H., Taylor N.T. Brown, Jody L. Herman, Sari L. Reisner, and Walter O. Bockting.** 2017. “Demographic characteristics and health status of transgender adults in select US regions: Behavioral Risk Factor Surveillance System, 2014.” *American journal of public health*, 107(4): 582–589.
- Miller, Kristen, and Stephanie Willson.** 2022. “Development and Evaluation of a Single, Non-Binary Gender Question for Population-Based Federal Health Surveys.” Hyattsville, MD: National Center for Health Statistics - QDRL.
- Niederle, Muriel.** 2016. “Gender.” In *Handbook of Experimental Economics*. Vol. 2, , ed. John Kagel and Alvin E. Roth, 481–553. Princeton University Press.

- Niederle, Muriel, and Lise Vesterlund.** 2007. “Do Women shy away from competition? Do men compete too much?” *Quarterly Journal of Economics*, 122(3): 1067–1101.
- Niederle, Muriel, and Lise Vesterlund.** 2011. “Gender and Competition.” *Annual Review of Economics*, 3: 601–630.
- Pande, Rohini, and Helena Roy.** 2021. ““If you compete with us, we shan’t marry you” The (Mary Paley and) Alfred Marshall Lecture.” *Journal of the European Economic Association*, 19(6): 2992–3024.
- Piasenti, Stefano, and Müge Süer.** 2024. “Predictive Power of Biological Sex and Gender Identity on Economic Behavior.” *CRC TRR 190 Rationality and Competition Discussion Paper No. 513*.
- Pope, Devin, and Justin Sydnor.** 2010. “A new perspective on stereotypical gender differences in test scores.” *Journal of Economic Perspectives*, 24(95).
- Reuben, Ernesto, Matthew Wiswall, and Basit Zafar.** 2017. “Preferences and biases in educational choices and labour market expectations: Shrinking the black box of gender.” *The Economic Journal*, 127(604): 2153–2186.
- Sansone, Dario.** 2019. “LGBT students: New evidence on demographics and educational outcomes.” *Economics of Education Review*, 73(101933).
- Schaerer, Michael, Christilene Du Plessis, My Hoang Bao Nguyen, Robbie CM Van Aert, Leo Tiokhin, Daniël Lakens, Elena Giulia Clemente, Thomas Pfeiffer, Anna Dreber, Magnus Johannesson, Cory J. Clark, Gender Audits Forecasting Collaboration, and Eric Luis Uhlmann.** 2023. “On the trajectory of discrimination: A meta-analysis and forecasting survey capturing 44 years of field experiments on gender and hiring decisions.” *Organizational Behavior and Human Decision Processes*, 179: 104280.
- Shurchkov, Olga.** 2012. “Under pressure: gender differences in output quality and quantity under competition and time constraints.” *Journal of the European Economic Association*, 10(5): 1189–1213.

## Appendices (For Online Publication Only)

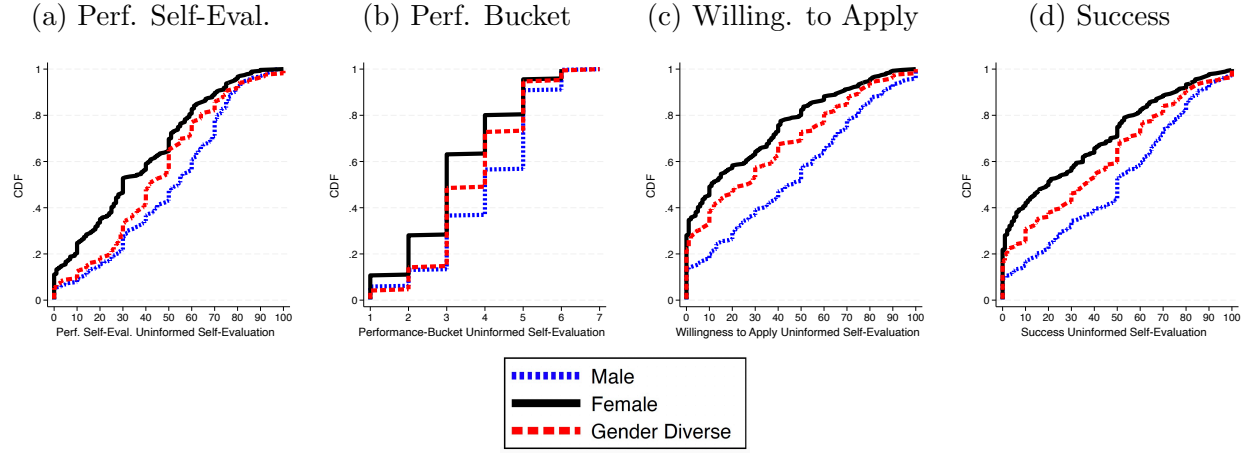
### A Additional Results from the Adult Study

Table A.1: In the *Adult Study*, OLS of Reported Confidence Relative to Truth Outcomes (overconfident, accurate, or overconfident)

	Overconfident: 1 (Reported Confidence > Performance)	Accurate: 1 (Reported Confidence = Performance)	Underconfident: 1 (Reported Confidence < Performance)
Gender Diverse	-0.19*** (0.03)	-0.01 (0.03)	0.20*** (0.04)
Female	-0.11*** (0.03)	-0.01 (0.03)	0.12*** (0.04)
Male Average	0.29	0.11	0.60
GD – F (= Gender Minority Gap – Gender Gap)			
Difference	-0.08	-0.00	0.08
p-value	0.02	0.93	0.06
N	746	746	746

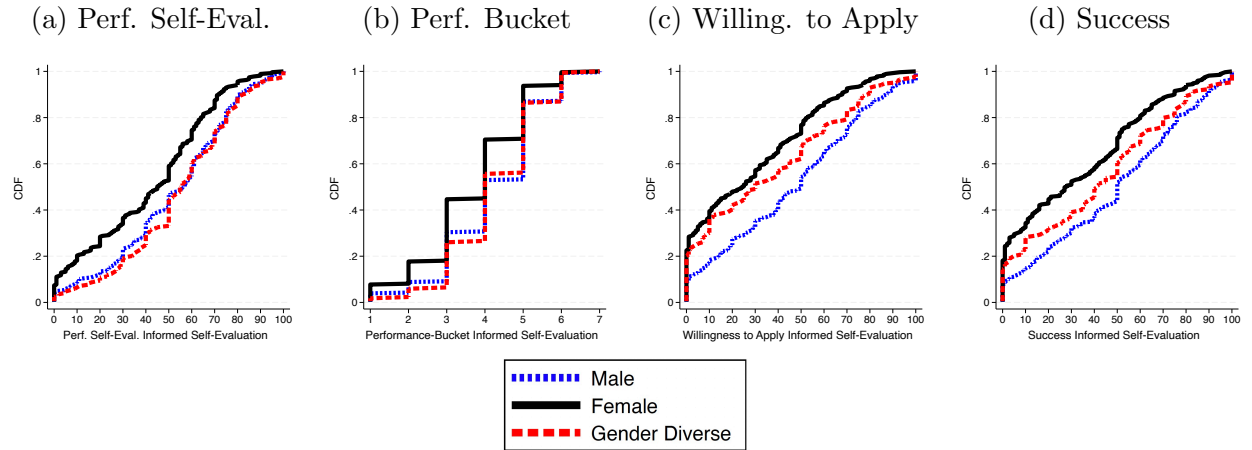
\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . SEs are robust and clustered at the participant level. This table presents data from the *Adult Study*. Results are from OLS regressions of the dependent variable noted in the column. *Overconfident/Accurate/Underconfident* is a binary variable that takes the value of 1 if the participant was overconfident/accurate/underconfident about their performance (i.e. their believed performance was better than/equal to/worse than their actual performance) and otherwise zero. See Table 1 for definitions of the independent variables, *Difference*, and *p-value*.

Figure A.1: In the *Adult Study*, CDFs for Uninformed Self-Evaluations



Graphs show CDFs of responses to the question noted in each panel, elicited before performance information is provided.

Figure A.2: In the *Adult Study*, CDFs for Informed Self-Evaluations



Graphs show CDFs of responses to the question noted in each panel, elicited after performance information is provided.

## B Additional Results from the Student Study

### B.1 Predicting Academic Performance with Confidence and Self-Evaluations from the *Student Study*

Motivated by prior work on how certain behavioral traits that differ by gender can predict educational outcomes (Buser, Niederle and Oosterbeek, 2014; Reuben, Wiswall and Zafar, 2017), we collected additional data on academic performance from our Student Study in the academic quarter that they took our study and the next seven academic quarters. This data reveals that—even after controlling for performance on our test, student gender, year in school, and school—our confidence and self-evaluation measures are highly correlated with student GPA in the quarter of our study and the seven quarters following it. That is, those who are more confident and report more positive self-evaluations in our experiment perform significantly better in school for at least two years after our study. Future work might explore the potential connections between gender minority gaps in confidence and self-evaluations and various educational and labor market outcomes.

Specifically, Table B.1 shows that our measures are highly correlated with academic performance, as measured by a student’s overall GPA within a quarter, both in the quarter of the school year in which our study was run (Q1, shown in the first column) and in each of the next seven quarters, which includes the entire next academic year (Q5–Q8).<sup>28</sup> These regressions control for the student’s performance on our test, the student’s year in school (i.e., 6th grade, 7th grade, etc.), the student’s school, and the student’s gender identity. The regression show that students who are more confident about their absolute performance on the test (Panel A) and who report higher self-evaluations (Panels B–I) have higher GPAs across the quarters.

All eight correlations between confidence and academic performance are statistically significant at  $p < 0.01$  and all 32 correlations between uninformed self-evaluations and academic performance (i.e., the four questions in each of the eight quarters) are statistically significant at  $p < 0.01$ . Comparing the uninformed self-evaluations (Panels B–E) to the informed self-evaluations (Panels F–I), we see some evidence that the predictive power of the self-evaluations are muted when students know how many questions they answered correctly, suggesting that some of the predictive power of the uninformed self-evaluations can be explained by beliefs about absolute performance. That said, the coefficient estimates for the informed self-evaluations are all uniformly positive and 25 out of 32 estimates are still statistically significant with  $p < 0.1$  (of those, 22 have  $p < 0.05$  and 19 have  $p < 0.01$ ), suggesting

---

<sup>28</sup>Not all students in our data have overall GPAs in the administrative data. Additionally, the number of students with GPAs decreases over time (e.g., as students graduate or otherwise leave the school district).



that even informed self-evaluations have predictive power. Appendix Table [B.2](#) follows Table [B.1](#) but shows regressions of Math GPA in each quarter, rather than overall GPA. Results are qualitatively very similar.<sup>29</sup>

**Result 1** (Predicting Academic Performance with Confidence and Self-Evaluations)  
Those who are more confident and those who report more favorable self-evaluations have significantly higher grade point averages both in the academic year that the study was run and in the next academic year.

---

<sup>29</sup>While not the focus of this paper, our data also allow us to directly compare the academic performance of gender diverse students to the academic performance of students who identify as male and who identify as female. Appendix Table [B.3](#) does these comparisons and shows that—considering overall GPA in Panel A or just Math GPA in Panel B—gender diverse students typically perform worse than both male and female students in the academic year our study was run. Looking at the later quarters (i.e., the year after our study was run), gender diverse students continue to underperform students who identify as female but their performance is not statistically distinguishable from students who identify as male. Given the rich literature exploring differences between men and women in test scores and other academic outcomes (e.g., see discussions in [Pope and Sydnor \(2010\)](#) and [Niederle and Vesterlund \(2011\)](#)), an important avenue for future work is to also consider the academic performance of gender diverse students.

Table B.1: Regressions of Overall GPA

	Academic Quarter (Q1–Q8) from 2020–2021 & 2021–2022							
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
<b>Panel A: X = Absolute Belief (0–10)</b>								
<i>X</i>	0.238*** (0.046)	0.261*** (0.047)	0.258*** (0.047)	0.246*** (0.048)	0.180*** (0.046)	0.217*** (0.048)	0.209*** (0.051)	0.175*** (0.053)
<b>Panel B: X = Uninformed Performance Self-Evaluations (0–100)</b>								
<i>X</i>	0.028*** (0.004)	0.033*** (0.004)	0.031*** (0.004)	0.030*** (0.004)	0.023*** (0.004)	0.026*** (0.004)	0.028*** (0.005)	0.028*** (0.005)
<b>Panel C: X = Uninformed Performance-Bucket Self-Evaluations (1–7)</b>								
<i>X</i>	0.458*** (0.085)	0.517*** (0.088)	0.485*** (0.087)	0.487*** (0.088)	0.256*** (0.084)	0.376*** (0.088)	0.474*** (0.097)	0.419*** (0.099)
<b>Panel D: X = Uninformed Willingness Self-Evaluations (0–100)</b>								
<i>X</i>	0.014*** (0.003)	0.019*** (0.003)	0.017*** (0.003)	0.017*** (0.003)	0.013*** (0.003)	0.014*** (0.003)	0.016*** (0.004)	0.017*** (0.004)
<b>Panel E: X = Uninformed Success Self-Evaluations (0–100)</b>								
<i>X</i>	0.037*** (0.004)	0.041*** (0.004)	0.036*** (0.004)	0.035*** (0.004)	0.026*** (0.003)	0.026*** (0.004)	0.029*** (0.004)	0.029*** (0.004)
<b>Panel F: X = Informed Performance Self-Evaluations (0–100)</b>								
<i>X</i>	0.003 (0.004)	0.010*** (0.004)	0.008** (0.004)	0.010*** (0.004)	0.004 (0.003)	0.007* (0.004)	0.011*** (0.004)	0.008* (0.004)
<b>Panel G: X = Informed Performance-Bucket Self-Evaluations (1–7)</b>								
<i>X</i>	0.013 (0.066)	0.120* (0.069)	0.055 (0.069)	0.142** (0.071)	0.020 (0.067)	0.099 (0.071)	0.158** (0.077)	0.077 (0.081)
<b>Panel H: X = Informed Willingness Self-Evaluations (0–100)</b>								
<i>X</i>	0.019*** (0.003)	0.021*** (0.003)	0.020*** (0.003)	0.019*** (0.003)	0.015*** (0.003)	0.017*** (0.003)	0.017*** (0.003)	0.019*** (0.004)
<b>Panel I: X = Informed Success Self-Evaluations (0–100)</b>								
<i>X</i>	0.027*** (0.003)	0.028*** (0.003)	0.026*** (0.003)	0.027*** (0.003)	0.018*** (0.003)	0.020*** (0.003)	0.019*** (0.004)	0.022*** (0.004)
N	10590	10569	10435	9781	7614	7619	7469	7316

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . SEs are robust and clustered at the participant level. Results are from OLS regressions of a student's overall GPA during the academic quarter noted in the column on the confidence or self-evaluation measure listed in the panel. Each regression controls for whether a student identifies as female, male, or other (when asked about their gender) and includes dummies for: each possible number of questions a student got right out of the 10 questions on the test, the student's year in school (i.e., 6th grade, 7th grade, etc.), and for the student's school. The data exclude the 28 students who selected other and provided an offensive response when asked about their gender. Some regressions have smaller sample sizes due to missing values in the administrative data (e.g., because a student's GPA was not recorded in one of the academic quarters).

Table B.2: Regressions of Math GPA

	Academic Quarter (Q1–Q8) from 2020–2021 & 2021–2022							
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
<b>Panel A: X = Absolute Belief (0–10)</b>								
X	0.284*** (0.060)	0.288*** (0.060)	0.252*** (0.059)	0.266*** (0.065)	0.213*** (0.069)	0.232*** (0.069)	0.252*** (0.074)	0.256*** (0.078)
<b>Panel B: X = Uninformed Performance Self-Evaluations (0–100)</b>								
X	0.032*** (0.006)	0.035*** (0.006)	0.030*** (0.006)	0.034*** (0.006)	0.030*** (0.006)	0.029*** (0.006)	0.035*** (0.007)	0.036*** (0.007)
<b>Panel C: X = Uninformed Performance-Bucket Self-Evaluations (1–7)</b>								
X	0.481*** (0.109)	0.488*** (0.110)	0.455*** (0.110)	0.481*** (0.118)	0.387*** (0.125)	0.524*** (0.128)	0.628*** (0.137)	0.717*** (0.145)
<b>Panel D: X = Uninformed Willingness Self-Evaluations (0–100)</b>								
X	0.022*** (0.004)	0.020*** (0.004)	0.017*** (0.004)	0.019*** (0.005)	0.019*** (0.005)	0.018*** (0.005)	0.022*** (0.005)	0.025*** (0.006)
<b>Panel E: X = Uninformed Success Self-Evaluations (0–100)</b>								
X	0.045*** (0.005)	0.045*** (0.005)	0.037*** (0.005)	0.038*** (0.005)	0.037*** (0.005)	0.034*** (0.005)	0.037*** (0.006)	0.038*** (0.006)
<b>Panel F: X = Informed Performance Self-Evaluations (0–100)</b>								
X	0.009** (0.005)	0.012** (0.005)	0.013*** (0.005)	0.010** (0.005)	0.008 (0.005)	0.007 (0.005)	0.012** (0.006)	0.006 (0.006)
<b>Panel G: X = Informed Performance-Bucket Self-Evaluations (1–7)</b>								
X	0.074 (0.087)	0.147 (0.089)	0.081 (0.089)	0.064 (0.098)	0.076 (0.104)	0.164 (0.103)	0.270** (0.111)	0.103 (0.117)
<b>Panel H: X = Informed Willingness Self-Evaluations (0–100)</b>								
X	0.027*** (0.004)	0.024*** (0.004)	0.021*** (0.004)	0.021*** (0.004)	0.021*** (0.005)	0.020*** (0.005)	0.023*** (0.005)	0.023*** (0.005)
<b>Panel I: X = Informed Success Self-Evaluations (0–100)</b>								
X	0.038*** (0.004)	0.033*** (0.004)	0.029*** (0.004)	0.030*** (0.005)	0.026*** (0.005)	0.027*** (0.005)	0.030*** (0.005)	0.029*** (0.005)
N	10348	10272	10212	9577	7393	7383	7246	7075

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . SEs are robust and clustered at the participant level. Results are from OLS regressions of a student's GPA in their math class during the academic quarter noted in the column on the confidence or self-evaluation measure listed in the panel. Each regression controls for whether a student identifies as female, male, or other (when asked about their gender) and includes dummies for: each possible number of questions a student got right out of the 10 questions on the test, the student's year in school (i.e., 6th grade, 7th grade, etc.), and for the student's school. The data exclude the 28 students who selected other and provided an offensive response when asked about their gender. Some regressions have smaller sample sizes due to missing values in the administrative data (e.g., because a student's GPA was not recorded in one of the academic quarters).

Table B.3: Regressions of Overall and Math GPA

	Academic Quarter (Q1–Q8) from 2020–2021 & 2021–2022							
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
<b>Panel A: DV = Overall GPA</b>								
Gender Diverse	-0.61 (0.80)	-1.67** (0.81)	-1.68* (0.90)	-2.49** (0.97)	0.57 (0.88)	0.41 (0.97)	0.43 (0.97)	1.11 (1.09)
Female	3.35*** (0.19)	2.80*** (0.19)	2.74*** (0.19)	2.64*** (0.20)	2.69*** (0.19)	2.93*** (0.20)	3.21*** (0.21)	3.29*** (0.22)
Male Average	83.51	83.10	83.23	83.88	85.42	83.97	83.19	83.54
GD – F <i>Difference</i>	-3.97	-4.47	-4.41	-5.14	-2.12	-2.52	-2.78	-2.18
GD – F <i>p-value</i>	< 0.01	< 0.01	< 0.01	< 0.01	0.02	0.01	< 0.01	0.05
N	10590	10569	10435	9781	7614	7619	7469	7316
<b>Panel B: DV = Math GPA</b>								
Gender Diverse	-1.39 (0.93)	-2.19** (1.05)	-3.07*** (1.06)	-3.06*** (1.10)	-1.00 (1.17)	-0.67 (1.30)	-1.30 (1.31)	1.08 (1.45)
Female	3.36*** (0.25)	2.94*** (0.25)	2.95*** (0.25)	2.52*** (0.27)	2.79*** (0.29)	2.98*** (0.29)	3.73*** (0.31)	3.46*** (0.33)
Male Average	80.77	79.84	79.76	80.91	81.23	80.24	78.74	79.86
GD – F <i>Difference</i>	-4.75	-5.13	-6.01	-5.58	-3.80	-3.65	-5.03	-2.38
GD – F <i>p-value</i>	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	0.10
N	10348	10272	10212	9577	7393	7383	7246	7075

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . SEs are robust and clustered at the participant level. Results are from OLS regressions of a student's GPA in their overall class during the academic quarter noted in the column on the confidence or self-evaluation measure listed in the panel. See Table 3 for definitions of the independent variables. Each regression includes dummies for: each possible number of questions a student got right out of the 10 questions on the test, the student's year in school (i.e., 6th grade, 7th grade, etc.), and for the student's school. The data exclude the 28 students who selected other and provided an offensive response when asked about their gender. Some regressions have smaller sample sizes due to missing values in the administrative data (e.g., because a student's GPA was not recorded in one of the academic quarters).

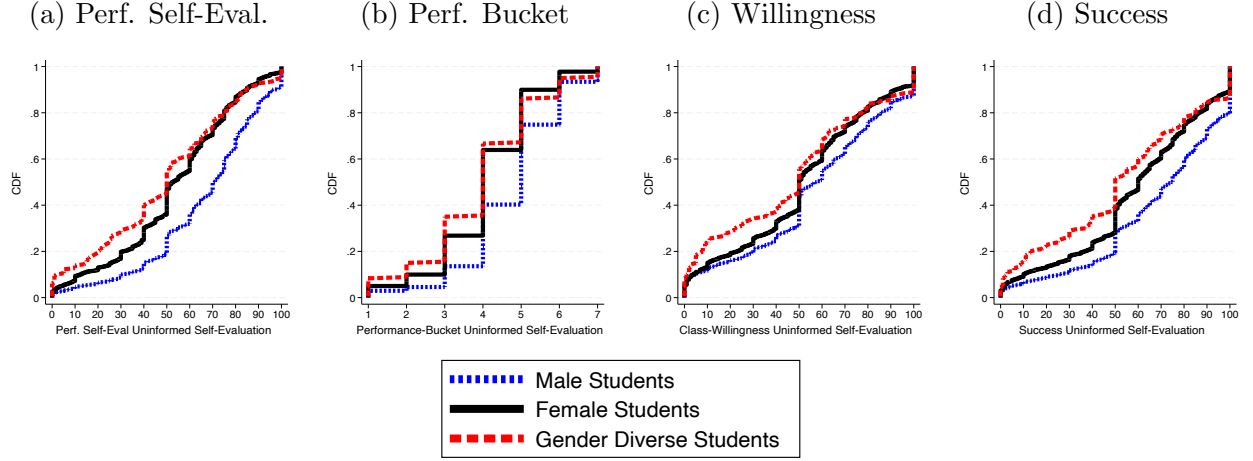
## B.2 Additional Figures and Tables from the Student Study

Table B.4: In the *Student Study*, OLS of Reported Confidence Relative to Truth Outcomes (overconfident, accurate, or overconfident)

	Overconfident: 1 (Reported Confidence > Performance)	Accurate: 1 (Reported Confidence = Performance)	Underconfident: 1 (Reported Confidence < Performance)
Gender Diverse	-0.19*** (0.04)	-0.01 (0.03)	0.20*** (0.04)
Female	-0.12*** (0.01)	-0.02*** (0.01)	0.15*** (0.01)
Male Average	0.51	0.20	0.29
GD – F (= Gender Minority Gap – Gender Gap)			
Difference	-0.07	0.02	0.06
p-value	0.04	0.61	0.14
N	10,779	10,779	10,779

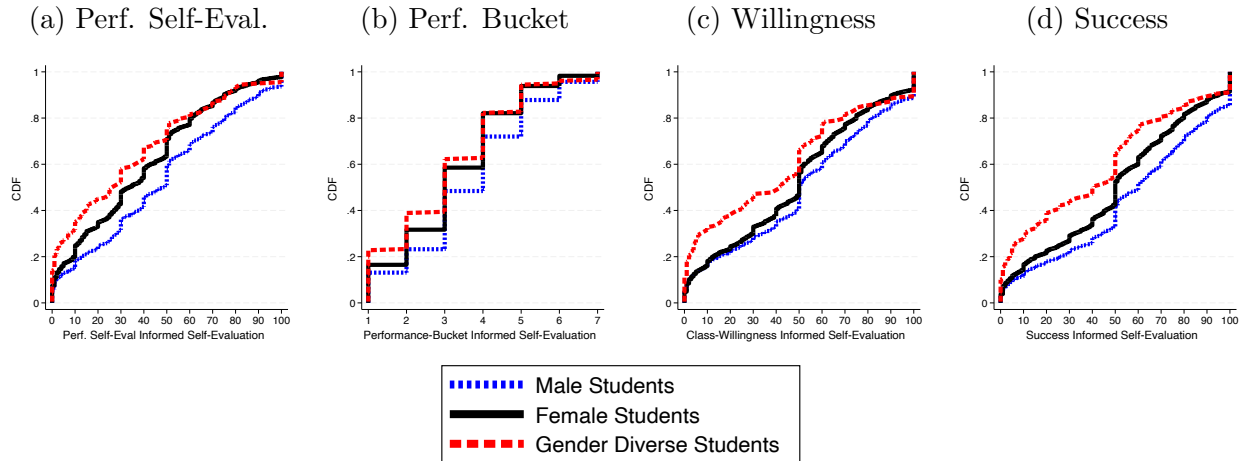
\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . SEs are robust and clustered at the participant level. This table presents data from the *Student Study*. Results are from OLS regressions of a student's level of confidence as noted in the column. *Overconfident/Accurate/Underconfident* is a binary variable that takes the value of 1 if the participant was overconfident/accurate/underconfident about their performance (i.e. their believed performance was better than/equal to/worse than their actual performance) and otherwise zero. See Table 3 for definitions of the independent variables, *Difference*, and *p-value*. Each regression includes dummies for: the student's year in school (i.e., 6th grade, 7th grade, etc.), and for the student's school. The data exclude the 28 students who selected other and provided an offensive response when asked about their gender.

Figure B.1: In the *Student Study*, CDFs for Uninformed Self-Evaluations



Graphs show CDFs of responses to the question noted in each panel, elicited before performance information is provided.

Figure B.2: In the *Student Study*, CDFs for Informed Self-Evaluations



Graphs show CDFs of responses to the question noted in each panel, elicited after performance information is provided.

Table B.5: Sample and Variable Descriptions for Robustness Checks of the *Student Study*

Panel	Notes
Panel A:	These results rely on the gender data from our survey. <i>Female</i> is an indicator for a student selecting “Female.” <i>Gender Diverse</i> is an indicator for a student selecting “Other,” including the 28 students who selected “Other” and provided an offensive response. Panel A thus includes data on all 10,807 students.
Panel B:	These results rely on the gender data from our survey. <i>Female</i> is an indicator for a student selecting “Female.” <i>Explicitly Gender Diverse</i> is an indicator for a student who we classify as explicitly gender diverse. Panel B excludes both the 28 students who provided an offensive response and also excludes the 106 students who selected other but provided either no response or a response that was not specific enough for us to classify them as explicitly gender diverse. Panel B thus includes data on 10,673 students.
Panel C:	These results rely on the gender data from the Character Lab Research Network (CLRN) survey. Panel C excludes the 535 students who selected “Prefer not to say” when asked about their gender. <i>Female</i> is an indicator for female students (50.49% or 5,186) and <i>Gender Diverse</i> is an indicator for a student selecting “Other” when asked about their gender (1.47% or 151). Panel C thus includes data on 10,272 students.
Panel D:	These results rely on the gender data from the Character Lab Research Network (CLRN) survey. Different from Panel C, we do not exclude the 535 students who selected “Prefer not to say” when asked about their gender. Instead, for these 535 students, we replace the missing values with their responses to our survey. Thus, <i>Female</i> is an indicator for a student selecting female gender in the CLRN survey (5,186) or selecting “Prefer not to say” in the CLRN survey but choosing “Female” in our survey (236). <i>Gender Diverse</i> is an indicator for the students selecting “Other” when asked about their gender in the CLRN survey (151) or selecting “Prefer not to say” in the CLRN survey but choosing “Other” in our survey (14). Panel D thus includes data on all 10,807 students.

This table includes information about the variables and each of the samples used in Panels A–D in Tables B.6–B.8.

Table B.6: In the *Student Study*, Performance Beliefs with Alternative Gender Classifications

	Performance	Reported Confidence	Reported Confidence- Performance
<b>Panel A: Our Gender Measure (Full Sample), N=10,807</b>			
Gender Diverse	-0.11 (0.15)	-1.35*** (0.20)	-1.31*** (0.22)
Female	-0.46*** (0.04)	-1.03*** (0.04)	-0.80*** (0.05)
Male Average	5.90	6.65	0.74
Gender Diverse – Female <i>Difference</i>	0.35	-0.32	-0.51
Gender Diverse – Female <i>p-value</i>	0.02	0.11	0.02
<b>Panel B: Our Gender Measure (Restricted Sample), N=10,673</b>			
Explicitly Gender Diverse	0.72*** (0.19)	-1.30*** (0.28)	-1.76*** (0.28)
Female	-0.46*** (0.04)	-1.03*** (0.04)	-0.80*** (0.05)
Male Average	5.90	6.65	0.74
Gender Diverse – Female <i>Difference</i>	1.18	-0.27	-0.96
Gender Diverse – Female <i>p-value</i>	<0.01	0.33	<0.01
<b>Panel C: CLRN Gender Measure, N=10,272</b>			
Gender Diverse	0.17 (0.16)	-1.47*** (0.22)	-1.60*** (0.24)
Female	-0.46*** (0.04)	-1.01*** (0.04)	-0.79*** (0.05)
Male Average	5.94	6.65	0.71
Gender Diverse – Female <i>Difference</i>	0.64	-0.45	-0.81
Gender Diverse – Female <i>p-value</i>	<0.01	0.04	<0.01
<b>Panel D: CLRN Gender Measure (Full Sample), N=10,807</b>			
Gender Diverse	0.09 (0.16)	-1.49*** (0.21)	-1.58*** (0.23)
Female	-0.46*** (0.04)	-1.02*** (0.04)	-0.80*** (0.05)
Male Average	5.90	6.64	0.74
Gender Diverse – Female <i>Difference</i>	0.54	-0.47	-0.78
Gender Diverse – Female <i>p-value</i>	<0.01	0.03	<0.01
Year in School FEs	Yes	Yes	Yes
School FEs	Yes	Yes	Yes
Performance FEs.	No	Yes	No

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . SEs are robust and clustered at the participant level. This table presents data from the *Student Study*. Results are from OLS regressions of the dependent variable noted in the column. See Table 3 and Appendix Table B.5 for more information about samples, dependent variables, and independent variables used in each panel as well as FEs.



Table B.7: In the *Student Study*, Uninformed Self-Evaluations with Alternative Gender Classifications

	Performance Self-Eval.	Performance- Bucket	Willingness	Success
<b>Panel A: Our Gender Measure (Full Sample), N=10,807</b>				
Gender Diverse	-16.39*** (2.10)	-0.75*** (0.11)	-10.51*** (2.37)	-15.80*** (2.31)
Female	-10.97*** (0.45)	-0.52*** (0.02)	-4.27*** (0.58)	-7.48*** (0.54)
Male Average	66.42	4.70	56.52	68.34
Gender Diverse – Female <i>Difference</i>	-5.42	-0.23	-6.24	-8.32
Gender Diverse – Female <i>p-value</i>	0.01	0.04	0.01	<0.01
<b>Panel B: Our Gender Measure (Restricted Sample), N=10,673</b>				
Explicitly Gender Diverse	-19.62*** (2.98)	-0.76*** (0.15)	-8.75** (3.40)	-18.02*** (3.34)
Female	-10.95*** (0.45)	-0.52*** (0.02)	-4.27*** (0.58)	-7.48*** (0.54)
Male Average	66.42	4.70	56.52	68.34
Gender Diverse – Female <i>Difference</i>	-8.67	-0.24	-4.48	-10.54
Gender Diverse – Female <i>p-value</i>	<0.01	0.10	0.19	<0.01
<b>Panel C: CLRN Gender Measure, N=10,272</b>				
Gender Diverse	-18.24*** (2.33)	-0.84*** (0.12)	-9.11*** (2.57)	-15.89*** (2.53)
Female	-10.84*** (0.46)	-0.51*** (0.02)	-4.03*** (0.59)	-7.35*** (0.55)
Male Average	66.50	4.70	56.33	68.36
Gender Diverse – Female <i>Difference</i>	-7.40	-0.33	-5.08	-8.54
Gender Diverse – Female <i>p-value</i>	<0.01	<0.01	0.05	<0.01
<b>Panel D: CLRN Gender Measure (Full Sample), N=10,807</b>				
Gender Diverse	-17.98*** (2.23)	-0.79*** (0.11)	-9.83*** (2.49)	-15.90*** (2.47)
Female	-10.97*** (0.45)	-0.52*** (0.02)	-4.29*** (0.58)	-7.40*** (0.54)
Male Average	66.37	4.70	56.46	68.22
Gender Diverse – Female <i>Difference</i>	-7.01	-0.27	-5.54	-8.50
Gender Diverse – Female <i>p-value</i>	<0.01	0.01	0.03	<0.01
Year in School FEs	Yes	Yes	Yes	Yes
School FEs	Yes	Yes	Yes	Yes
Performance FEs	Yes	Yes	Yes	Yes

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . SEs are robust and clustered at the participant level. This table presents data from the *Student Study*. Results are from OLS regressions of a student's response to the uninformed self-evaluation (elicited before the student learns their test performance) noted in the column. See Table 4 and Appendix Table B.5 for more information about samples, dependent variables, and independent variables used in each panel as well as FEs.

Table B.8: In the *Student Study*, Informed Self-Evaluations with Alternative Gender Classifications

	Performance Self-Eval.	Performance- Bucket	Willingness	Success
<b>Panel A: Our Gender Measure (Full Sample), N=10,807</b>				
Gender Diverse	-11.20*** (2.18)	-0.43*** (0.12)	-11.93*** (2.41)	-15.62*** (2.37)
Female	-6.43*** (0.52)	-0.26*** (0.03)	-2.94*** (0.60)	-5.34*** (0.59)
Male Average	45.84	3.60	51.27	57.52
Gender Diverse – Female <i>Difference</i>	-4.78	-0.17	-8.99	-10.28
Gender Diverse – Female <i>p-value</i>	0.03	0.16	<0.01	<0.01
<b>Panel B: Our Gender Measure (Restricted Sample), N=10,673</b>				
Explicitly Gender Diverse	-17.66*** (3.06)	-0.88*** (0.17)	-11.03*** (3.75)	-18.67*** (3.63)
Female	-6.41*** (0.52)	-0.26*** (0.03)	-2.94*** (0.60)	-5.34*** (0.59)
Male Average	45.84	3.60	51.27	57.52
Gender Diverse – Female <i>Difference</i>	-11.25	-0.62	-8.09	-13.32
Gender Diverse – Female <i>p-value</i>	<0.01	<0.01	0.03	<0.01
<b>Panel C: CLRN Gender Measure, N=10,272</b>				
Gender Diverse	-14.61*** (2.46)	-0.65*** (0.12)	-10.69*** (2.69)	-16.01*** (2.74)
Female	-6.41*** (0.54)	-0.26*** (0.03)	-2.75*** (0.62)	-5.21*** (0.60)
Male Average	45.99	3.61	51.23	57.57
Gender Diverse – Female <i>Difference</i>	-8.20	-0.39	-7.94	-10.80
Gender Diverse – Female <i>p-value</i>	<0.01	<0.01	<0.01	<0.01
<b>Panel D: CLRN Gender Measure (Full Sample), N=10,807</b>				
Gender Diverse	-13.79*** (2.36)	-0.59*** (0.12)	-11.85*** (2.55)	-15.91*** (2.62)
Female	-6.50*** (0.52)	-0.26*** (0.03)	-2.85*** (0.60)	-5.25*** (0.59)
Male Average	45.84	3.60	51.16	57.41
Gender Diverse – Female <i>Difference</i>	-7.29	-0.33	-9.00	-10.66
Gender Diverse – Female <i>p-value</i>	<0.01	<0.01	<0.01	<0.01
Year in School FEs	Yes	Yes	Yes	Yes
School FEs	Yes	Yes	Yes	Yes
Performance FEs	Yes	Yes	Yes	Yes

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . SEs are robust and clustered at the participant level. This table presents data from the *Student Study*. Results are from OLS regressions of a student's response to the uninformed self-evaluation (elicited before the student learns their test performance) noted in the column. See Table 4 and Appendix Table B.5 for more information about samples, dependent variables, and independent variables used in each panel as well as FEs.

## C Additional Results from the Predictions Study

Table C.1: In the *Predictions Study*, OLS of Predicted Performance Given Prior Participant's Reported Confidence and Self-Evaluation

	Predicted Performance Given:	
	Reported Confidence (1)	Self- Evaluation (2)
Gender Diverse Profile	0.11 (0.08)	-0.06 (0.10)
Female Profile	0.67*** (0.09)	0.47*** (0.10)
Constant	10.40*** (0.10)	11.18*** (0.11)
GD – F (= Believed Gender Minority Gap – Believed Gender Gap)		
Difference	-0.56	-0.53
p-value	<0.01	<0.01
N	5400	5400

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . Robust standard errors clustered at the individual level are in parentheses. This table presents data from the *Predictions Study*. Results are from OLS regressions of the following dependent variables: the first column reports predicted performances of [female]/[male]/[gender diverse] prior participants who guessed that they answered [5]/[10]/[15] questions (out of 20) correctly on the math and science test; and the second column reports predicted performances of [female]/[male]/[gender diverse] prior participants who assigned their performance [a low rating of 0 to 33]/[a medium rating of 34 to 66]/[a high rating of 67 to 100] in response to the *performance self-evaluation* question (see Table 2 for a description of this self-evaluation question). *Gender Diverse Profile* is an indicator variable for when the predictors submit their predicted performances for gender diverse prior participants. *Female Profile* is an indicator variable for when the predictors submit their predicted performances for female prior participants. *Difference* is the difference between the Gender Diverse Profile and Female Profile coefficient estimates and *p-value* presents the corresponding p-value for a two-sided t-test of these two coefficient estimates.

Table C.2: In the *Predictions Study*, OLS of Predicted Performance Given Prior Participant's Reported Confidence

	Predicted Performance Given Reported:		
	Confidence of 5 (1)	Confidence of 10 (2)	Confidence of 15 (3)
Female Profile	0.50*** (0.12)	0.71*** (0.12)	0.81*** (0.13)
Gender Diverse Profile	0.15 (0.12)	0.16 (0.11)	0.03 (0.12)
Constant	7.64*** (0.17)	10.52*** (0.11)	13.03*** (0.11)
GD – F (= Believed Gender Minority Gap – Believed Gender Gap)			
Difference	-0.35	-0.54	-0.78
p-value	<0.01	<0.01	<0.01
N	1800	1800	1800

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . Standard errors clustered at the individual level are in parentheses. This table presents data from the *Predictions Study*. Results are from OLS regressions of the following dependent variables: the first column reports predicted performances of [female]/[male]/[gender diverse] prior participants who guessed that they answered 5 questions (out of 20) correctly on the math and science test; the second column reports predicted performances of [female]/[male]/[gender diverse] prior participants who guessed that they answered 10 questions (out of 20) correctly on the math and science test; and the third column reports predicted performances of [female]/[male]/[gender diverse] prior participants who guessed that they answered 15 questions (out of 20) correctly on the math and science test. See C.1 for definitions of independent variables, *Difference*, and *p-value*.

Table C.3: In the *Predictions Study*, OLS of Predicted Performance Given Prior Participant's Self-Evaluations

	Predicted Performance Given Self-Evaluation With:		
	Low Rating of	Medium Rating of	High Rating of
	0 to 33	34 to 66	67 to 100
	(1)	(2)	(3)
Female Profile	0.35** (0.15)	0.49*** (0.13)	0.57*** (0.13)
Gender Diverse Profile	-0.00 (0.15)	-0.02 (0.14)	-0.16 (0.13)
Constant	7.95*** (0.18)	11.34*** (0.13)	14.24*** (0.12)
GD – F (= Believed Gender Minority Gap – Believed Gender Gap)			
Difference	-0.35	-0.51	-0.73
p-value	0.02	<0.01	<0.01

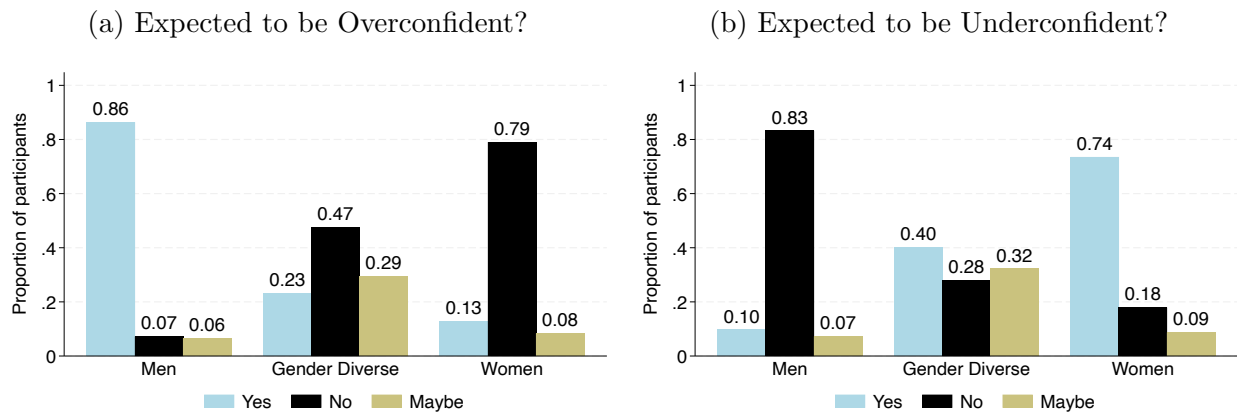
\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . Standard errors clustered at the individual level are in parentheses. This table presents data from the *Predictions Study*. Results are from OLS regressions of the following dependent variables: the first column reports predicted performances of [female]/[male]/[gender diverse] prior participants who assigned their performance a low rating of 0 to 33 in response to the *performance self-evaluation* question (see Table 2 for a description of this self-evaluation question); the second column reports predicted performances of [female]/[male]/[gender diverse] prior participants who assigned their performance a medium rating of 34 to 66 in response to the *performance self-evaluation* question; and the third column reports predicted performances of [female]/[male]/[gender diverse] prior participants who assigned their performance a high rating of 67 to 100 in response to the *performance self-evaluation* question. See C.1 for definitions of independent variables, *Difference*, and *p-value*.

Table C.4: In the *Predictions Study*, Average Predicted Performance – True Performance Given Participants’ Reported Confidence and Self-Evaluations

	Predicted Performance – Truth Given:	
	Reported Confidence	Self-Evaluation
	(1)	(2)
Gender Diverse Profile (n=1800)	-2.31***	-1.64***
Female Profile (n=1800)	0.67***	0.82***
Male Profile (n=1800)	-0.77***	-0.23**

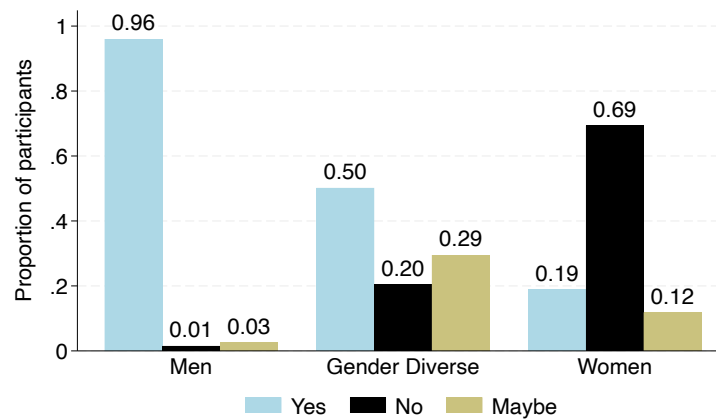
\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . These p-values correspond with estimated differences that have robust standard errors that are clustered at the individual level. This table presents data from the *Predictions Study*. The first column reports the average predicted performance minus the true average performance of gender diverse participants (see the row labeled “Gender Diverse Profile” 1), female participants (see the row labeled “Female Profile”), and male participants (see the row labeled “Male Profile”) who guessed that they answered [5]/[10]/[15] questions (out of 20) correctly on the math and science test. The second column reports the average predicted performance minus the true average performance of these groups of participants who assigned their performance [a low rating of 0 to 33]/[a medium rating of 34 to 66]/[a high rating of 67 to 100] in response to the *performance self-evaluation* question (see Table 2 for a description of this self-evaluation question). The actual average performances for gender diverse, female, and male participants, respectively, used in column 1 are as follows: (i) for those who guessed that they answered 5 questions correctly on the test, the average performances were 10.33, 9.46, and 8.46; (ii) for those who guessed that they answered 10 questions correctly, the average performances were 12.92, 10.19, and 10.68; (iii) for those who guessed that they answered 15 questions correctly, the average performances were 15.21, 11.56, and 14.35. The actual average performances for gender diverse, female, and male participants, respectively, used in column 2 are as follows: (i) for those who assigned their performance a low rating of 0 to 33 in response to the *performance self-evaluation* question, the average performances were 11.09, 9.23, and 9.87; (ii) for those who assigned their performance a medium rating of 34 to 66, the average performances were 12.85, 11.20, and 11.25; (iii) for those who assigned their performance a high rating of 67 to 100, the average performances were 14.34, 12.06, and 13.11.

Figure C.1: In the *Predictions Study*, Expected Overconfidence and Underconfidence in Adult Survey



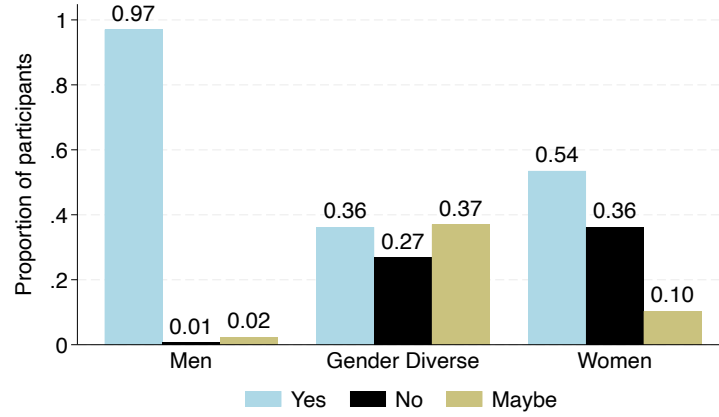
This graph reports the distribution of responses to the following unincentivized survey questions: Do you think that, in general, [female]/[male]/[gender diverse] people are likely to be overconfident (panel a) / underconfident (panel b) in their performance and abilities in math and science tasks?

Figure C.2: Expected Risk Taking in Predictions Study



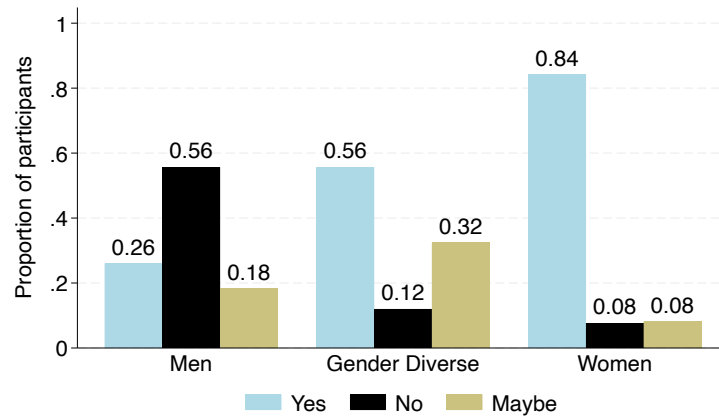
This graph reports the distribution of responses to the following unincentivized survey question: Do you think that, in general, [female]/[male]/[gender diverse] people are likely to take risks?

Figure C.3: Expected Competitiveness in Predictions Study



This graph reports the distribution of responses to the following unincentivized survey question: Do you think that, in general, [female]/[male]/[gender diverse] people are likely to be competitive?

Figure C.4: Expected Generosity in Predictions Study



This graph reports the distribution of responses to the following unincentivized survey question: Do you think that, in general, [female]/[male]/[gender diverse] people are likely to be generous?



## D The Adult (Verbal) Study

We were also interested in exploring how gender differences looked across domains. Since we expected we could recruit enough gender diverse individuals to have the power to do so, when running the Adult Study, we randomized half of our participants to a word knowledge quiz (i.e., a verbal test) rather than a math and science test. This study is very similar to the Adult Study with one difference: the test domain. In this study, we asked participants 20 word knowledge questions from the Armed Services Vocational Aptitude Battery (see Appendix Figure E.11 for an example question). Participants had 15 seconds to answer each question. Everything else is identical to the Adult Study.

A total of 748 participants completed the *Verbal* version run on Prolific in June and July of 2023. On our demographic survey question, 37.6% (n=281) selected only “Male,” 40.9% (n=306) selected only “Female,” and the remaining 21.5% (n=161) selected “Transgender, non-binary, or another gender” or multiple options, which leads us to classify them as gender diverse.<sup>30</sup>

### D.1 Confidence and Self-Evaluations in the Adult (Verbal) Study

Gender diverse participants got an average of 12.45 questions correct out of 20. This performance is better than female participants who got an average of 11.25 questions correct. Both of these performances are better than the performance of male participants, who got an average of 10.58 questions correct. Column (1) of Table D.1 presents regression results of performance and shows that these differences are statistically significant.

Columns (2) and (3) of Table D.1 analyze beliefs about performance. Column (2) examines beliefs while including performance fixed effects (i.e., comparing equally performing male, female, and gender diverse participants). Column (3) examines an individual-level variable of beliefs about performance minus actual performance.

Column (2) shows that gender diverse participants believe they answered 0.10 fewer questions correctly than equally performing male participants. This difference is not statistically significant and small in magnitude, thus revealing no evidence for a gender minority gap. Female participants believe they answered 1.05 fewer questions correctly than equally performing male participants, evidence of a gender gap in confidence on the verbal test.

Column (3) explores the difference between a participant’s belief and their actual performance, calculated for each individual. In this case, we find that male participants are slightly overconfident: they overestimate their performance by 0.50 questions (see *Male Average*).

---

<sup>30</sup>Specifically, 110 participants only selected “Transgender, non-binary, or another gender,” 22 participants selected “Transgender, non-binary, or another gender” and “Male,” 27 participants selected “Transgender, non-binary, or another gender” and “Female,” and 2 participants selected “Male” and “Female.”

On this measure, gender diverse participants are less confident than male participants, evidence of a gender minority gap. (As will become evident below, this is the only case out of our 10 measures of confidence and self-evaluations in which we see a gender minority gap in the *Verbal* version.)<sup>31</sup> We also see that female participants are less confident than male participants with this measure, evidence of a gender gap in confidence in this setting. Figure D.1 shows the CDFs of these differences between beliefs and performance for the three groups.

Table D.1: In the *Adult (Verbal) Study*, Participants’ Performance (i.e., score on the test) and Reported Confidence (i.e., believed score on the test)

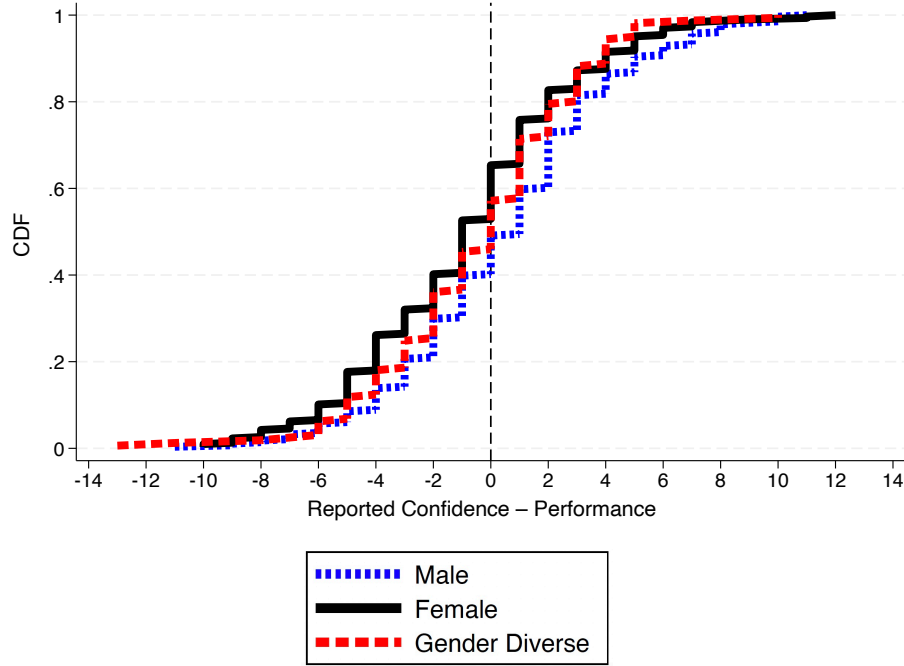
	Performance	Reported Confidence	Reported Confidence– Performance
	(1)	(2)	(3)
Gender Diverse	1.87*** (0.36)	-0.10 (0.33)	-0.85** (0.36)
Female	0.67** (0.31)	-1.05*** (0.29)	-1.32*** (0.32)
Male Average	10.58	11.08	0.50
Gender Diverse – Female (= Gender Minority Gap – Gender Gap) Difference	1.20	0.94	0.47
p-value	<0.01	0.01	0.18
Performance FEs	No	Yes	No
N	748	748	748

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . SEs are robust and clustered at participant level. This table presents data from the *Adult (Verbal) Study*. Results are from OLS regressions of the dependent variable noted in the column. See Table 1 for definitions of the dependent and independent variables, *Difference*, *p-value* and Performance FEs. The only difference between Table 1 and this table is that this table presents data from the *Adult (Verbal) Study*.

**Result 1** (Beliefs about Absolute Performance in the Adult (Verbal) Study) We find limited evidence of a gender minority gap in confidence on a verbal test.

<sup>31</sup>We can also compare the size of the gender minority gaps across versions by comparing the coefficient estimates on *Gender Diverse* in Column (3) of Tables 1 and D.1; the gender minority gap we observe in the *Verbal* study is smaller than the corresponding gender minority gap we see in the *Adult Study* ( $p < 0.01$ ). Indeed, of the 10 comparisons that we can make between gender minority gaps—comparing the coefficient on *Gender Diverse* across the verbal and the main studies (across Tables 1 and D.1 and across Tables 2 and D.2)—the coefficient on *Gender Diverse* is always at least directionally smaller in magnitude in the Adult (Verbal) Study (four comparisons are statistically significantly different at  $p < 0.01$ , four at  $p < 0.05$ , and one at  $p < 0.1$ ; the last comparison has  $p = 0.16$ ). These results emphasize that we see a reduction in gender minority gaps in confidence and self-evaluation going from the math version to the verbal version.

Figure D.1: In the *Adult (Verbal) Study*, Reported Confidence–Performance Distributions



Graph shows CDFs for *Reported Confidence–Performance*, the number of questions a participant believes they answered correctly minus the number of questions a participant answered correctly. Positive responses suggest overconfidence while negative numbers suggest underconfidence.

Panel A of Table D.2 presents regression results on participants’ uninformed self-evaluations about the verbal test and Panel B presents results on participants’ informed self-evaluations (i.e., after they were told how many questions they answered correctly on the verbal test).

We see no evidence of a gender minority gap. Self-evaluations of performance on the verbal test are statistically indistinguishable between gender diverse participants and equally performing male participants. Meanwhile, female participants have worse self-evaluations than equally performing male participants across all four questions (i.e., the coefficient on *Female* is negative and statistically significant in all four columns), evidence of a gender gap.

**Result 2** (Uninformed Self-Evaluations in the Verbal Study) We find no evidence of a gender minority gap in self-evaluation on a verbal test.

Panel B of Table D.2 shows the self-evaluations after participants have been told how many questions they answered correctly on the test. The results are very similar to those from Panel A. We again see no evidence of a gender minority gap between gender diverse participants and male participants. Female participants again have worse self-evaluations than equally performing male participants across all four questions.

**Result 3** (Informed Self-Evaluations in the Verbal Study) Even after individuals are

informed of how many questions they got correct, we still find no evidence of a gender minority gap in self-evaluation on a verbal test.

While we observe little to no evidence for gender minority gaps but more robust evidence for gender gaps in the verbal task, the estimates of the two gaps are often not statistically significantly different from each other. The estimated gender gaps are only larger than the estimated gender minority gaps half of the time (i.e., in column (2) but not (3) of Table D.1 and in columns (1) and (2) but not (3) and (4) of Table D.2). This lack of a significant difference arises in part because the gender gaps in the verbal task are generally smaller than the gender gaps in the math and science task (i.e. the *Adult Study*), suggesting that switching from the math and science test (a male-typed task) to the verbal test (a less male-typed task) mitigates both gender minority gaps and gender gaps.<sup>32</sup>

---

<sup>32</sup>Of the 10 comparisons (two on confidence and eight on self-evaluations) that we can make between gender gaps—comparing the coefficient on *Female* across the *Verbal* and *Main* studies—the gender gap is directionally smaller in the *Verbal* study in 8 of the 10 tests and significantly smaller in 4 of those tests. The gap is never significantly larger in the *Verbal* study. (Exley and Kessler, 2022) also found sizable gender gaps in a math and science task and found that they were absent in a verbal task, consistent with this reduction in the gender gap. That said, further work on the impact of domain on gender gaps between men and women is warranted.

Table D.2: In the *Adult (Verbal) Study*, Informed and Uninformed Self-Evaluations

	Performance Self-Eval. (1)	Performance- Bucket (2)	Willingness (3)	Success (4)
<b>Panel A: Uninformed Self-Evaluations</b>				
Gender Diverse	-0.97 (2.07)	-0.11 (0.11)	-2.98 (2.89)	-4.00 (2.72)
Female	-7.04*** (1.85)	-0.43*** (0.10)	-6.95*** (2.31)	-8.34*** (2.25)
Male Average	58.51	4.48	52.69	57.01
Gender Diverse – Female (= Gender Minority Gap – Gender Gap)				
Difference	6.07	0.32	3.97	4.34
p-value	<0.01	<0.01	0.16	0.10
<b>Panel B: Informed Self-Evaluations</b>				
Gender Diverse	1.60 (1.68)	-0.01 (0.09)	-2.98 (2.59)	-3.65 (2.49)
Female	-4.00** (1.61)	-0.26*** (0.09)	-4.68** (1.99)	-5.78*** (1.98)
Male Average	51.61	4.15	48.00	51.62
Gender Diverse – Female (= Gender Minority Gap – Gender Gap)				
Difference	5.59	0.25	1.70	2.13
p-value	<0.01	<0.01	0.50	0.39
Performance FEs	Yes	Yes	Yes	Yes
N	748	748	748	748

\*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . SEs are robust and clustered at the participant level. This table presents data from the *Adult (Verbal) Study*. Results are from OLS regressions of a participant's response to the uninformed (elicited before the participant learns their test performance) (Panel A) and informed (Panel B) self-evaluation noted in the column. See Table 2 for definitions of the dependent and independent variables, *Difference*, *p-value*, and Performance FEs. The only difference between Table 2 and this table is that this table presents data from the *Adult (Verbal) Study*.

## E Experimental Instructions

### E.1 Experimental Instructions and Protocol for the Adult Study

The instructions for the experiment are displayed in Figure E.1. An example question on the test is displayed in Figure E.2 (note that the timer in the figure indicates the participant has 26 seconds left to answer the question although the timer starts at 30 seconds). After completing the test, participants are asked to complete five additional pages of the study.

First, they are asked about their absolute performance belief (see Figure E.3). Second, they are provided with additional instructions (see Figure E.4) and then asked the self-evaluation questions (see Figure E.5). Third, participants are provided with perfect information on their absolute performance and are required to correctly report back their absolute performance (see Figure E.6). Fourth, they are provided with additional instructions (see Figure E.7) and are asked the self-evaluation questions again (see Figure E.8). Fifth, they are asked for demographic information including their gender identity (see Figure E.9).

Our recruitment procedure was pre-registered on AsPredicted (#136119) which can be accessed here: [https://aspredicted.org/2FW\\_Z5H](https://aspredicted.org/2FW_Z5H). We started by using the “Sex” and “Cisgender and Transgender” screener questions that are set by Prolific (i.e., not by us). The “Sex” question asks: “What is your sex, as recorded on legal/official documents?” with options “Male” and “Female.” The “Cisgender and Transgender” screener question asks: “Does your current gender differ from the one you were assigned at birth?” with answers “Yes,” “No,” and “Rather not say.” Then, we aimed to recruit an equal number of participants who (1) answered “Female” to the “Sex” screener and “No” to the “Cisgender and Transgender” screener, (2) answered “Male” to the “Sex” screener and “No” to the “Cisgender and Transgender” screener, and (3) answered “Yes” to the “Cisgender and Transgender” screener. Since we expected that it would be much more difficult to recruit individuals in the third group—but desired to collect data across all three groups at similar times—we recruited participants in batches on a rolling basis. We first opened recruitment for 100 people in each of the three groups. Once all groups reached 100 completed responses, we opened recruitment for another 100 participants from each group. Our pre-registered recruitment plan was to continue this until we reached 600 people in each group or until we reached a satiation point of any group, whichever came first. The recruitment of the third group reached a satiation point at 500 participants. The first four times we opened recruitment to 100 participants, it took less than a day to collect all responses. The fifth time we opened the study for 100 participants, it took roughly three days to recruit 100 participants who had answered “Yes” to the “Cisgender and Transgender” screener. As a result, we recruited 1,500 participants on Prolific and ended up with 1,494 completed responses. These 1,494

participants were then randomized to either be in the *Adult Study* ( $n = 746$ ) discussed in Section 2 or in the *Adult (Verbal) Study* ( $n = 748$ ) discussed in Section D. We also restricted recruitment to participants who were U.S. nationals who had completed at least 100 prior submissions with at least a 95% approval rate.

Figure E.1: Part 1 Instructions for the test in the Adult Study

**Instructions for Part 1 out of 3:**

In part 1, you will complete a test. On the test, you will be asked to answer up to 20 questions. Each question will test your math and science skills. Specifically, you will be asked about general science, arithmetic reasoning, math knowledge, mechanical comprehension, and assembling objects. Performance on this test is often used as a measure of cognitive ability by academic researchers.

You will be presented with each of the 20 questions on separate pages. You will be given up to 30 seconds to answer each question, although you may push the arrow at the bottom of the page to answer a question before the 30 seconds are up.

If part 1 is randomly selected as the part-that-counts, your additional payment will equal 5 cents times the number of questions you answer correctly on this test.

Figure E.2: Example question on the test in the Adult Study

26

**Question 3 out of 20:**

MECHANICAL COMPREHENSION: Why is it so difficult to hold a beach ball under water?

The ball is full of air, which is much less dense than water.

The ball shrinks under water, making it harder to hold.

The ball expands under water so it rises faster.

The cool water will cool the air in the ball, making it rise.



Figure E.3: Absolute Performance Belief Question in the Adult Study

Congrats! You have now completed part 1 out of 3.

Before pushing the arrow to proceed onto the next part of the study, please answer the following question.

**Out of the 20 questions on the test you took in part 1, how many questions do you think you answered correctly?**



Figure E.4: Additional Instructions in the Adult Study

**Instructions for Part 2 out of 3:**

In part 2, you will be asked several questions -- on the next page -- related to your performance on the test you completed in part 1.

If this part is randomly selected as the part-that-counts, your additional payment will equal 25 cents regardless of how you answer these questions. Thus, we ask that you please answer these questions carefully and honestly.

**Understanding Question:** If this part is randomly selected as the part-that-counts, your additional payment...

will equal 25 cents for sure.

will equal 5 cents times the number of questions you answered correctly on the test in part 1.

will depend on how you answer the questions -- on the next page -- about your performance on the test you took in part 1.



Figure E.5: Self-Evaluation Questions in the Adult Study

Now, please answer the five questions below to complete part 2.

**Please describe how well you think you performed on the test that you took in part 1 and why.**



**Please indicate how well you think you performed on the test you took in part 1.**

Terrible	Very Poor	Poor	Neutral	Good	Very Good	Exceptional
----------	-----------	------	---------	------	-----------	-------------

**On a scale from 0 (entirely disagree) to 100 (entirely agree), please indicate the extent to which you agree with the following statement:**

Entirely Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Neither Disagree Nor Agree	Somewhat Agree	Agree	Strongly Agree	Entirely Agree		
0	10	20	30	40	50	60	70	80	90	100

**I performed well on the test I took in part 1.**



**I would apply for a job that required me to perform well on the test I took in part 1.**



**I would succeed in a job that required me to perform well on the test I took in part 1.**



Figure E.6: Absolute Performance Information in the Adult Study

Congrats! You have now completed part 2 out of 3.

Before pushing the arrow to proceed to the next part in this study, please read the information below on how well you performed on the test in part 1 and answer the corresponding understanding question.

You answered **6 questions correctly out of the 20 questions.**

**Understanding Question:** Out of the 20 questions on the test you took in part 1, how many questions did you answer correctly?

Figure E.7: Additional Instructions in the Adult Study

**Instructions for Part 3 out of 3:**

In part 3, you will be asked several questions -- on the next page -- related to your performance on the test you completed in part 1.

If this part is randomly selected as the part-that-counts, your additional payment will equal 25 cents regardless of how you answer these questions. Thus, we ask that you please answer these questions carefully and honestly.

**Understanding Question:** If this part is randomly selected as the part-that-counts, your additional payment...

will equal 25 cents for sure.

will equal 5 cents times the number of questions you answered correctly on the test in part 1.

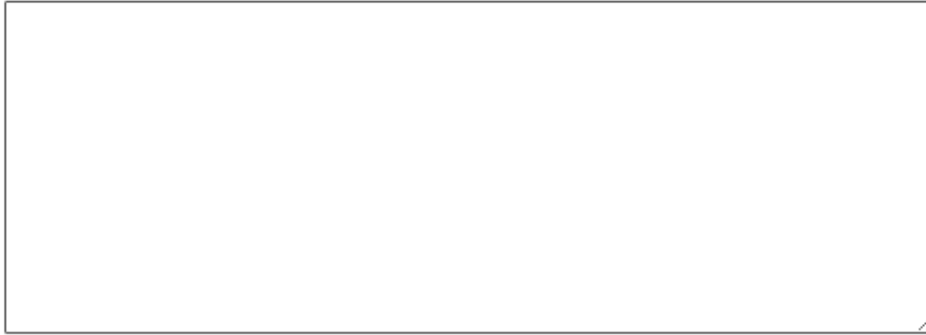
will depend on how you answer the questions -- on the next page -- about your performance on the test you took in part 1.



Figure E.8: Informed Self-Evaluation Questions in the Adult Study

Now, please answer the five questions below to complete part 3.

**Please describe how well you think you performed on the test that you took in part 1 and why.**



**Please indicate how well you think you performed on the test you took in part 1.**

Terrible	Very Poor	Poor	Neutral	Good	Very Good	Exceptional
----------	-----------	------	---------	------	-----------	-------------

**On a scale from 0 (entirely disagree) to 100 (entirely agree), please indicate the extent to which you agree with the following statement:**

Entirely Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Neither Disagree Nor Agree	Somewhat Agree	Agree	Strongly Agree	Entirely Agree		
0	10	20	30	40	50	60	70	80	90	100

**I performed well on the test I took in part 1.**



**I would apply for a job that required me to perform well on the test I took in part 1.**



**I would succeed in a job that required me to perform well on the test I took in part 1.**



Figure E.9: Screenshot of Gender Question in the the Adult Study

Are you: (Mark all that apply)

Male

Female

Transgender, non-binary, or another gender

## E.2 Experimental Instructions for the Adult (Verbal) Study

The Adult (Verbal) Study closely follows the design discussed in Section [E.1](#) with the exceptions discussed in Section [D](#). The instructions for the experiment are displayed in Figure [E.10](#). An example question on the test is displayed in Figure [E.11](#). After completing the test, participants are asked to complete five additional pages of the study which are identical to those described in Appendix Section [E.1](#).

Figure E.10: Part 1 Instructions for the test in the Adult (Verbal) Study

### **Instructions for Part 1 out of 3:**

In part 1, you will complete a test. On the test, you will be asked to answer up to 20 questions. Each question will test your verbal skills. Specifically, you will be asked about word knowledge. Performance on this test is often used as a measure of cognitive ability by academic researchers.

You will be presented with each of the 20 questions on separate pages. You will be given up to 15 seconds to answer each question, although you may push the arrow at the bottom of the page to answer a question before the 15 seconds are up.

If part 1 is randomly selected as the part-that-counts, your additional payment will equal 5 cents times the number of questions you answer correctly on this test.

Figure E.11: Example question on the test in the Adult (Verbal) Study

15

Question 2 out of 20:

WORD KNOWLEDGE: Indolence most nearly means

bliss.

tolerance.

serenity.

laziness.





### E.3 Experimental Instructions for the Student Study

The Student Study closely follows the design discussed in Section E.1 with the exceptions discussed in Section 3.1. Prior to participating in the Student Study, participants must correctly answer a captcha and consent to participate. At the end of the study, participants must complete a short follow-up survey to gather demographic information. Participants are recruited via the Character Lab Research Network and complete this study as part of the curriculum at school. There are no payments associated with this study.

The study begins by informing each participant about the test that they will take. The instructions for the test are displayed in Figure E.12 and an example of a question on the test is displayed in Figure E.13 (note that the timer in that screenshot indicates the participant has 24 seconds left to answer the question although the timer starts at 30 seconds). After completing the test, participants are asked to complete five additional pages of the study.

On the first page, they are asked about their absolute performance belief (see Figure E.14). On the second page, they are asked the self-evaluation questions (see Figure E.15). On the third page, participants are provided with perfect information on their absolute performance and are required to correctly report back their absolute performance (see Figure E.16). On the fourth page, they are asked the self-evaluation questions again (see Figure E.17). On the fifth page, they are asked for demographic information including their gender identity (see Figure E.18).

Figure E.12: Part 1 Instructions for the test in the Student Study

#### **Information about the Test:**

On the test, you will be asked to answer up to 10 questions from the Armed Services Vocational Aptitude Battery (ASVAB). Each question will test your aptitude in one of the following five categories: General Science, Arithmetic Reasoning, Math Knowledge, Mechanical Comprehension, and Assembling Objects. In addition to being used by the military to determine which jobs armed service members are qualified for, performance on the ASVAB is often used as a measure of cognitive ability by academic researchers.

You will be presented with each of the 10 questions on separate pages. You will be given up to 30 seconds to answer each question, although you may push the arrow at the bottom of the page to answer a question before the 30 seconds are up.

**Please try to answer each question as best as you can.**

Figure E.13: Example question on the test in Student Study

24

---

**Question 2 out of 10:**

---

**MATH KNOWLEDGE:** Which number has the greatest value?

9,299

903 tens

93 hundreds

9 thousands

Figure E.14: Absolute Performance Belief Question in Student Study

**Page 1 out of 5**

Please answer the following question.

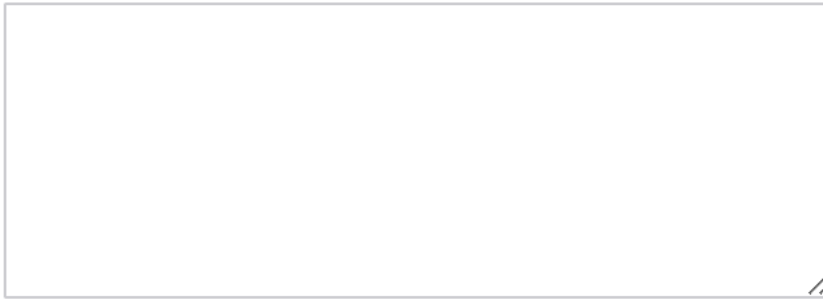
**Out of the 10 questions on the test, how many questions do you think you answered correctly?**

Figure E.15: Self-Evaluation Questions in Student Study

**Page 2 out of 5**

Please answer the following questions.

**Please describe how well you think you performed on the test and why.**



**Please indicate how well you think you performed on the test.**

Terrible	Very Poor	Poor	Neutral	Good	Very Good	Exceptional
----------	-----------	------	---------	------	-----------	-------------

**On a scale from 0 (entirely disagree) to 100 (entirely agree), please indicate the extent to which you agree with the following statement:**

Entirely Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Neither Disagree Nor Agree	Somewhat Agree	Agree	Strongly Agree	Entirely Agree		
0	10	20	30	40	50	60	70	80	90	100

**I performed well on the test.**



**If given an option, I would choose to take a class that involves topics like those covered on the test.**



**I would succeed in a class that involves topics like those covered on the test.**



Figure E.16: Absolute Performance Information in Student Study

**Page 3 out of 5**

On the test, you answered **0 questions correctly out of the 20 questions**. To confirm that you read the prior sentence, please answer the following question.

**Oof the 10 questions on the test you took in part 1, how many questions did you answer correctly?**

Figure E.17: Informed Self-Evaluation Questions in Student Study

**Page 4 out of 5**

Now that you have information on your test performance, please answer the following questions again. Your answers may be the same or different than your previous answers.

**Please describe how well you think you performed on the test and why.**

**Please indicate how well you think you performed on the test.**

Terrible	Very Poor	Poor	Neutral	Good	Very Good	Exceptional
----------	-----------	------	---------	------	-----------	-------------

**On a scale from 0 (entirely disagree) to 100 (entirely agree), please indicate the extent to which you agree with the following statement:**

Entirely Disagree	Strongly Disagree		Disagree	Somewhat Disagree	Neither Disagree Nor Agree	Somewhat Agree	Agree	Strongly Agree	Entirely Agree	
0	10	20	30	40	50	60	70	80	90	100

**I performed well on the test.**

**If given an option, I would choose to take a class that involves topics like those covered on the test.**

**I would succeed in a class that involves topics like those covered on the test.**

Figure E.18: Screenshot of Gender Question in the Student Study

Please select your gender.

Male

Female

Other

## E.4 Experimental Instructions for the Predictions Study

The experiment begins by informing each predictor about the study that they will take as shown in Figure E.19. Next, we provide information about the Prior Study (which is the *Adult Study* of this paper) as well as how gender categories are defined, and then explain the Current Study, as displayed in Figure E.20. Once predictors have an overview understanding of the experiment, they then proceed with the experiment as explained in Section 4.1. Below provides more details on each sets of questions.

We elicit predictors’ beliefs about performance conditional on guessed performance (see Figure E.21). Specifically, we ask predictors to consider the group of [female]/[male]/[gender diverse] prior participants who guessed that they answered [5]/[10]/[15] questions correctly on the math and science test. Then, we ask them to guess how many questions, on average, they think they answered correctly on the math and science test. As shown in Figure E.22, using a slider, predictors select a range and if their chosen range includes the correct answer, they earn 1\$. although the slider indicates the range of 9.5-10.5 in the screenshot, there is no default selection. Thus, predictors must click on the slider before they can move to the next page.

We elicit predictors’ beliefs about performance conditional on the response provided to the following question: “On a scale from 0 (entirely disagree) to 100 (entirely agree), please indicate the extent to which you agree with the following statement: I performed well” (see Figure E.23.) Specifically, we ask predictors to consider the group of [female]/[male]/[gender diverse] participants who assigned their performance [a low rating of 0 to 33]/[a medium rating of 34 to 66]/[a high rating of 67 to 100] in response to the previous question. Then, again using the same sliders, we ask them to guess how many questions, on average, they think they answered correctly on the math and science test. An example decision screen is shown in Figure E.24.

We elicit predictors’ beliefs about [female]/[male]/[gender diverse] participants being overconfident, accurate, or underconfident. We first define what it means for a participant to be overconfident, accurate, or underconfident as shown in Figure E.25, then provide further information about the task (see Figure E.26.) Specifically, we ask predictors to guess whether [female]/[male]/[gender diverse] participants are overconfident, accurate, or underconfident. If their guess is correct, they earn \$1. We also give predictors to option to indicate when they are unsure. If they choose the “I’m unsure” option, they earn \$1 with a 50% chance. An example decision screen is shown in Figure E.27.

Finally, we ask 15 follow-up questions to measure beliefs about female, male, and gender diverse individuals using survey questions (again the order of these questions are randomized at the predictor level). Specifically, we ask predictors the following questions:



- In general, are [female]/[male]/[gender diverse] people likely to be [overconfident]/[underconfident] in their performance and abilities in math and science tasks? [Yes], [No], [I'm not Sure]
- In general, are [female]/[male]/[gender diverse] people likely to [take risks]/[be competitive]/[be generous]? [Yes], [No], [I'm not Sure]

Figure E.19: General Instructions for the Predictions Study

## STUDY INFORMATION

**Study Overview:** To complete this study, you must first answer 21 main questions and then answer 15 follow-up questions and complete a short follow-up questionnaire.

**Payment:** For completing this study, you are guaranteed to receive \$3 within 24 hours. In addition, one of the 21 main questions in this study will be chosen as the question-that-counts, and you will receive the amount of money you will earn in that question as a bonus payment.

**Understanding Question:** Which of the following statements is true?

For completing this study, I will receive \$3 for sure, and I do NOT have a chance of receiving a higher amount.

For completing this study, I will receive at least \$3. I will also receive the amount I'll earn in the question-that-counts.

For completing this study, I will receive at least \$3. The total amount I receive depends on my decisions in all parts in this study.

Figure E.20: Prior and Current Study Instructions in the Predictions Study

## THE PRIOR STUDY

In a prior study, we recruited participants from Prolific to answer math and science questions. They were paid 5 cents for each correct answer. The number of questions they answered correctly is their **test score**.

These participants were then asked a series of questions regarding their test scores and their views on their performance.

Participants were also asked about their gender. We define the gender categories as follows:

- **Male participants** are those who selected **Male**.
- **Female participants** are those who selected **Female**.
- **Gender diverse participants** are those who selected **Transgender, non-binary, or another gender** or who chose multiple options.

## THE CURRENT STUDY

In this current study, we will randomly choose participants or groups of participants from this prior study and ask you a series of questions about these participants.

Figure E.21: Beliefs about Performance (1) Instructions in the Predictions Study

In each of next 9 questions, you will be told how a group of participants answered the following question when asked about their performance on the math and science test:

How many questions did you answer correctly?

Then, in each question, we will ask you to guess how many questions (out of 20) participants answered correctly on the math and science test. Your guess will be provided by choosing a range of numbers on a slider. If the range of numbers you choose includes the right answer, you will earn \$1 in that question.

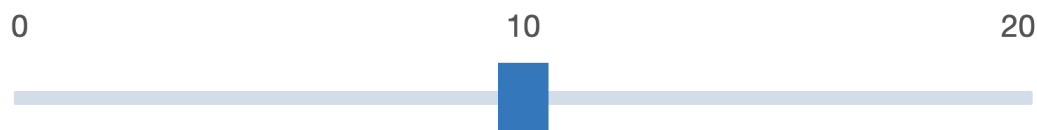
Figure E.22: Beliefs about Performance (1) Decision Screen in the Predictions Study

The continue arrow will enable after you move the slider.

## QUESTION 2 OUT OF 21

Consider the group of **gender diverse** participants who **guessed** that they answered **15** questions correctly on the math and science test.

**On average, how many questions do you think they answered correctly on the math and science test?**



Your guess for these **gender diverse** participants:

9.5 — 10.5

Figure E.23: Belief about Performance (2) in the Predictions Study

In each of next 9 questions, you will be told how a group of participants answered the following question when asked about their performance on the math and science test:

On a scale from 0 (entirely disagree) to 100 (entirely agree), please indicate the extent to which you agree with the following statement: **I performed well.**

Then, in each question, we will ask you to guess how many questions (out of 20) participants answered correctly on the math and science test. Your guess will be provided by choosing a range of numbers on a slider. If the range of numbers you choose includes the right answer, you will earn \$1 in that question.

Figure E.24: Beliefs about Performance (2) Decision Screen in the Predictions Study

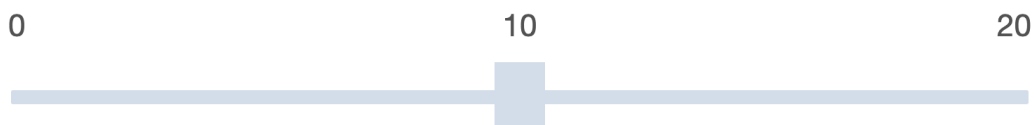
## QUESTION 11 OUT OF 21

Participants were asked the following question:

On a scale from 0 (entirely disagree) to 100 (entirely agree), please indicate the extent to which you agree with the following statement: **I performed well.**

Consider the group of **gender diverse** participants who assigned their performance **a medium rating of 34 to 66** in response to the above question.

**On average, how many questions do you think they answered correctly on the math and science test?**



Your guess for these **gender diverse** participants:

—

Figure E.25: Defining Confidence in the Predictions Study

For the next 3 questions, we categorize individuals based on their perception of their performance on the math and science test as follows:

- **Overconfident:** The number of questions they guessed they answered correctly was larger than the actual number they answered correctly
- **Accurate:** The number of questions they guessed they answered correctly was exactly the same as the actual number they answered correctly
- **Underconfident:** The number of questions they guessed they answered correctly was smaller than the actual number they answered correctly

**Understanding Question: When is a participant accurate?**

If the number of questions they guessed they answered correctly on the math and science test was larger than the actual number they answered correctly

If the number of questions they guessed they answered correctly on the math and science test was exactly the same as the actual number they answered correctly

If the number of questions they guessed they answered correctly on the math and science test was smaller than the actual number they answered correctly

Figure E.26: Beliefs about Confidence Instructions in the Predictions Study

In these 3 questions, we will ask you to choose among four options.

Three of these options involve you guessing that a given participant was accurate, overconfident, or underconfident. If you chose one of these three options, you will earn \$1 if your guess is correct.

The fourth option is to say you are unsure. If you choose the "I'm unsure" option, you will earn \$1 with a 50% chance.

To maximize the chance of earning \$1 in a question, you should choose:

- **The participant is likely overconfident** if you think the chance of the participant in that question being overconfident is more than 50%,
- **The participant is likely accurate** if you think the chance of the participant in that question being accurate is more than 50%,
- **The participant is likely underconfident** if you think the chance of the participant in that question being underconfident is more than 50%, or
- **I'm unsure**, otherwise.

Figure E.27: Beliefs about Confidence Decision Screen in the Predictions Study

### QUESTION 21 OUT OF 21

Consider a randomly selected **gender diverse participant**.

**The Gender Diverse Participant is Likely Overconfident**

**The Gender Diverse Participant is Likely Accurate**

**The Gender Diverse Participant is Likely Underconfident**

**I'm Unsure**