

NBER WORKING PAPER SERIES

THE GENDER MINORITY GAPS IN CONFIDENCE AND SELF-EVALUATION

Billur Aksoy
Christine L. Exley
Judd B. Kessler

Working Paper 32061
<http://www.nber.org/papers/w32061>

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2024

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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2024 by Billur Aksoy, Christine L. Exley, and Judd B. Kessler. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

The Gender Minority Gaps in Confidence and Self-Evaluation
Billur Aksoy, Christine L. Exley, and Judd B. Kessler
NBER Working Paper No. 32061
January 2024
JEL No. C91,D91,J16

ABSTRACT

A rich literature explores gender differences between men and women, but an increasing share of the population identifies their gender in some other way. Analyzing data on roughly 10,000 students and 1,500 adults, we find that such gender minorities are less confident and provide less favorable self-evaluations than equally performing men on a math and science test. We find that these "gender minority gaps" are robust, are as large as—or larger than—gender gaps between men and women, and are domain specific. Administrative data reveals that our confidence and self-evaluation measures are highly predictive of academic performance.

Billur Aksoy
Department of Economics
Rensselaer Polytechnic Institute
3404 Russell Sage Laboratory
110 8th Street
Troy, NY 12180
aksoyb3@rpi.edu

Judd B. Kessler
The Wharton School
University of Pennsylvania
320 Vance Hall
Philadelphia, PA 19104
and NBER
judd.kessler@wharton.upenn.edu

Christine L. Exley
Department of Economics
University of Michigan
Lorch Hall 365B
611 Tappan Ave
Ann Arbor, MI 48109
clexley@gmail.com

1 Introduction

A sizable share of the population identifies as part of a gender minority group. Some 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.¹

Inspired by the rich line of prior work on gender differences between men and women in 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.² 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. We overcome this challenge by allowing subjects to self-identify their gender in a survey as part of our study. 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 first study, the *Student Study*, we recruit a large population of young people. We partner with the Character Lab Research Network to recruit 10,807 students in grades 6–12 to complete our study; we identify 180 students (1.7%) as gender diverse. For our second study, the *Adult Study*, we recruit an online sample of 1,494 subjects with a pre-registered

¹Gender 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.

²Prior 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 competitiveness (Niederle and Vesterlund, 2007), negotiation (Babcock and Laschever, 2003), risk taking (Eckel and Grossman, 2008), and the contribution of ideas (Coffman, 2014) (as well as confidence and self-evaluation, as referenced in the main text). In contrast, we know very little about any of these traits among gender minorities.

protocol that overweights individuals whose prior answers on Prolific suggest they might be gender minorities; we identify 330 subjects in this study (22.1%) as gender diverse.³

Our main studies each proceed in six stages. First, subjects complete a math and science test.⁴ Second, we elicit subjects’ *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 subjects’ *uninformed self-evaluations* via 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 subjects about how many questions they actually got right on the test. Fifth, we elicit subjects’ *informed self-evaluations* via the same four questions that ask them to provide subjective evaluations of their performance. Finally, as is common in experimental economics, we ask subjects to complete a survey at the end of our study that gathers demographic information, including on gender identity, which allows us to classify subjects as either male, female, or gender diverse.

In our *Student Study*, gender diverse students perform roughly equivalent to male students on the math and science test. Nonetheless, when we compare gender diverse to equally performing male students, we observe large *gender minority gaps*. Our measure of confidence reveals that gender diverse students believe that they got fewer questions right on the math and science test than equally performing male students. Gender diverse students also provide less favorable self-evaluations about their performance on the math and science test than equally performing male students. They indicate less agreement with the statement that they “performed well” on the test, they report being less inclined to take a class that involves the math and science topics covered on the test, and they believe they would be less likely to succeed in such a class. These differences in self-evaluations persist when students are informed about how many questions they actually got right on the test. These gender minority gaps are robust to different ways that we can classify gender diverse students. The gender minority gaps are also sizable; they are consistently larger than the gender gaps in

³This recruitment procedure was pre-registered on AsPredicted (#136119) which can be accessed here: https://aspredicted.org/2FW_Z5H. For more details of our pre-registration and our recruitment protocol, see Section 2.2 and Footnote 9.

⁴Following 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 Punčohaf, 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)—our primary analysis involves the subjects who are assigned to complete a math and science test. This includes all of the subjects in the *Student Study* and half of the subjects in the *Adult Study*. To allow us to explore gender differences across domains, our *Adult Study* randomizes subjects to either take a math and science test or a verbal test.

confidence and in self-evaluations we find between male and female students.

In our *Adult Study*, gender diverse subjects perform better than male and female subjects. Nonetheless, when we compare gender diverse adults to equally performing male adults, we again observe large gender minority gaps. Gender diverse adults display less confidence (i.e., believe that they got fewer questions right on the test) and provide less favorable self-evaluations about their performance on the test than equally performing male adults. These gender minority gaps are comparably sized to the gender gaps between male and female adults that we see in our data.

The main contribution of this paper is studying the confidence and self-evaluations of gender diverse individuals and identifying gender minority gaps in these traits. Overall, we find that in the math and science domain—across both students and adults—gender diverse individuals are less confident and provide less favorable self-evaluations than equally performing male subjects (despite performing as well or better than male subjects on average). These gender minority gaps are robust and as large, or larger, than the magnitudes of the gender gaps between men and women that we observe.

It is interesting to reflect on whether we should “expect” gender minority gaps. In line with prior evidence that identifies gender gaps between women and men in confidence and self-evaluations in math and science domains, it could be that we should also expect similar gaps between gender diverse individuals and men because gender diverse individuals are part of a marginalized group, and marginalized groups often display lower confidence than majority groups.⁵ 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 than men. In addition, not only is there no prior work on the confidence and self-evaluations of gender diverse individuals, we note that there is very limited research on gender minorities in the economics literature overall. This paucity arises even though gender minorities represent an increasing share of the population and even though the small body of work that does exist finds 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), are more likely to be unemployed and to be in low-income households and less likely to have health insurance coverage (Badgett, Carpenter and Sansone, 2021), and have worse educational outcomes (Meyer et al., 2017; Downing and Przedworski, 2018; Sansone, 2019). We thus hope that this paper will help open the door to a new set of future work on gender that aims to better understand the

⁵While we focus on gender minorities, see Aksoy and Chadd (2023) for a discussion on this within the context of sexual minorities.

beliefs and traits of gender diverse individuals, rather than just men and women.

We see many avenues for such future work, which we discuss in more detail in Section 5. Additional results from our study allow us to push forward on two particularly promising avenues that we mention here.

First, motivated by prior work on how certain behavioral traits (such as competitiveness, confidence, and self-evaluations) that differ by gender can predict educational and labor market outcomes (Buser, Niederle and Oosterbeek, 2014; Chen, Grove and Hussey, 2017; Reuben, Wiswall and Zafar, 2017; Risse, Farrell and Fry, 2018; Kamas and Preston, 2018), we collected additional data on academic performance from our student sample 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 the academic year of our study and the academic year after it. Future work might explore the potential connections between gender minority gaps in confidence and self-evaluations and various educational and labor market outcomes.

Second, motivated by the literature that shows how gender differences between men and women can be domain specific (Günther et al., 2010; Shurchkov, 2012; Coffman, 2014; Coffman, Flikkema and Shurchkov, 2019; Dreber, von Essen and Ranehill, 2014; Bordalo et al., 2019; Coffman, Collis and Kulkarni, 2019; Atwater and Saygin, 2020), we randomize a set of our adult subjects to take a verbal test rather than a math and science test. When we switch from a math and science domain, which is typically considered male-typed, to a verbal test, which is typically considered less male-typed, the gender minority gaps shrink dramatically and almost all go away. Thus, the gender minority gaps we document appear to be domain specific. Future work might investigate why gender minority gaps arise in some settings and not others. Indeed, as an important factor for future work to consider, we note that gender diverse individuals might be less impacted by gender-based norms, which makes it hard to predict how they would behave in male-typed or female-typed domains.

2 Design

We run two studies to explore the confidence and self-evaluations of gender diverse individuals. The first study, the *Student Study*, run through a partnership with the Character Lab Research Network (CLRN), recruited student subjects and is described in Section 2.1. The second study, the *Adult Study*, recruited adult subjects on Prolific, an online labor market platform, and is described in Section 2.2. Additional details, including screenshots of

both studies, can be found in Appendix C.

2.1 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. The student subjects agreed to participate in a short study during the school day.⁶ 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.⁷ 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 provided at the end of our study to generate new results on gender diverse individuals. Furthermore, for this paper, we gathered supplementary data on academic performance from our student sample during the academic quarter in which they participated in our study and the subsequent seven academic quarters. This supplementary data allows us to explore the correlation between our confidence and self-evaluation measures and student GPAs, which we do in Section 4.⁸

The student study had six stages. First, students were asked to answer 10 math and science questions from the Armed Services Vocational Aptitude Battery. Each question appeared on a separate page, and students had 30 seconds to answer each question (see Appendix Figure C.2 for an example question). We requested that students try their best

⁶The 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).

⁷That the gender gaps in Exley and Kessler (2022) are found to be roughly identical with and without incentives to self-promote help to mitigate potential concerns about the lack of monetary incentives in the student study.

⁸This 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).

when answering, but there were no financial incentives in the study (see footnote 7).

Second, we elicited each student’s belief about their absolute performance on the test by asking them how many questions out of 10 they think they got right. This gives us a measure of their confidence in their absolute performance.

Third, we elicited each subject’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, subjects were 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, subjects were asked to indicate their agreement—on a scale from 0 (entirely disagree) to 100 (entirely agree)—with various statements. In the *performance* question, subjects were asked to indicate their agreement with “I performed well on the test.” In the *willingness* question, subjects 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, subjects were asked to indicate their agreement with “I would succeed in a class that involves topics like those covered on the test.”

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

Fifth, we elicited subjects’ informed self-evaluations by asking the same set of questions they were asked before they received information about their performance.

Sixth, we asked subjects to complete a short follow-up survey to gather demographic information, including a question about their gender where participants 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 3.1), we use these responses to classify subjects by gender and to identify the students who are gender diverse. Figure C.7 shows specifically how we ask subjects to self-report their gender.

2.2 Adult Study

To further explore the behavior of gender minorities, in June and July of 2023 we had 1,494 subjects from Prolific complete our adult study. Since gender minorities constitute such a small share of the adult population in the U.S., we implemented a pre-registered stratified recruitment protocol to recruit a relatively large number of gender minorities from the Prolific platform.

Specifically, we used two screener questions that Prolific asks of its users, which allowed

us to identify—and target for recruitment—potential subjects who we believed to be (1) cis-gender females, (2) cis-gender males, and (3) gender diverse individuals. In particular, we used the “Sex” and “Cisgender and Transgender” screener questions on Prolific. 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.” We recruited 500 subjects who answered “Female” to the “Sex” screener and “No” to the “Cisgender and Transgender” screener. Similarly, we recruited 500 subjects who answered “Male” to the “Sex” screener and “No” to the “Cisgender and Transgender” screener. Finally, we recruited 500 subjects who answered “Yes” to the “Cisgender and Transgender” screener.⁹ All participants were recruited to take part in a 20-minute study with a \$4 completion fee and had the possibility of earning a bonus payment.

The adult study also proceeded in six stages. First, adult subjects answered 20 test questions and were told they would receive 5 cents for each correct answer on the test if the first part of the study was chosen to determine bonus payments (otherwise they received 25 cents as a bonus payment). Second, we collected each subject’s belief about their absolute performance by asking how many questions out of 20 they thought they answered correctly. Third, we elicited self-evaluations: the *performance-bucket* and *performance* self-evaluation questions were the same as in the student study, but we changed the *willingness* and *success* questions to better suit our adult population. Specifically, for the *willingness* question, we asked subjects 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, we asked subjects 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.” Fourth, we informed subjects about how many questions

⁹As explained in our pre-registered recruitment plan, our initial goal was to recruit 600 participants in each of the three gender categories. Since there are many more participants who answer “No” than “Yes” to the “Cisgender and Transgender” screener, we expected that it would be much more difficult to recruit individuals in the third group. In order to ensure we were collecting 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 subjects 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 subjects. 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 subjects, it took roughly three days to recruit 100 subjects who had answered “Yes” to the “Cisgender and Transgender” screener. As a result, we recruited 1,500 subjects on Prolific and ended up with 1494 completed responses. In addition, as is typical in online studies, we restricted recruitment to subjects who were U.S. nationals who had completed at least 100 prior submissions with at least a 95% approval rate.

they answered correctly on the test. Fifth, we elicited self-evaluations again. Finally, we asked a demographic survey and adopted the gender question proposed by [Miller and Willson \(2022\)](#), which allows participants to choose all applicable options from the following: “Male,” “Female,” or “Transgender, non-binary, or another gender.”¹⁰ As explained in greater detail below (see Sections 3.2 and 3.3), we use these responses to classify individuals as gender diverse. Figure C.18 shows how we asked adult subjects to self-report their gender.

An additional and important difference between the adult study and the student study is that—because we were interested in exploring how gender differences looked across domains and because we expected we could recruit enough gender diverse individuals to have the power to do so—we randomized subjects to either a math and science quiz (i.e., the *Math* version of our adult study) or to a word knowledge quiz (i.e., the *Verbal* version of our adult study). In the *Math* version, we asked subjects 20 math and science questions from the Armed Services Vocational Aptitude Battery. In the *Verbal* version, we asked subjects 20 word knowledge questions from the Armed Services Vocational Aptitude Battery. Subjects had 30 seconds to answer each question in the *Math* version and 15 seconds in the *Verbal* version. Appendix Figures C.10 and C.11 provide example questions from each study version.

3 Results

We first present results from the student study, in Section 3.1. We then present the results from the adult study separately for those who participated in the *Math* version, in Section 3.2, and for those who participated in the *Verbal* version, in Section 3.3.

In each section, we first show the performance of gender diverse subjects, male subjects, and female subjects. To explore the possibility of gender minority gaps, we compare the confidence and self-evaluations of equally performing gender diverse and male subjects. In these analyses, we always include dummies for each performance level to ensure we are comparing equal performers. In each section, we also look for gender gaps in confidence and self-evaluations by comparing equally performing male and female subjects. Finally, we benchmark the size of any gender minority gap we observe to the size of any gender gap we observe in our data.

¹⁰Following 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.

3.1 Student Study

A total of 10,807 students completed our 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.¹¹ 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).¹²

We compare gender diverse students to those who identify as male and find large gender minority gaps.¹³ We then compare female students to male students and find gender gaps, replicating prior work. We also show tests comparing gender diverse students to female students, revealing that the gender minority gaps (between male and gender diverse students) we observe in the student study are consistently larger than the gender gaps (between male and female students) that we observe in this setting.

3.1.1 Performance and Confidence in the Student Study

We first examine student performance on the math and science test. 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.

Column (1) of Table 1 presents regression results of the performance while controlling for year in school fixed effects (i.e., dummy variables for being in 6th grade, 7th grade, etc.) and school fixed effects (i.e., a dummy variable for each school in the data). The coefficient estimates on *Gender Diverse* compares gender diverse students to male students, the coefficient estimate on *Female* compares female students to male students. At the bottom of the table, we report the coefficient difference between gender diverse and female students along with its corresponding p-value for a two-sided t-test of the difference in coefficient estimates. The regression results confirm that gender diverse students perform similarly to male students and better than female students.

¹¹The 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.)

¹²Some 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 data. As described in Section 2.2, we changed the way we identified gender diverse subjects in our adult study to ensure we could identify transgender individuals as gender diverse in the adult study.

¹³Section 3.1.3 describes robustness tests showing that we replicate our main findings using alternative classifications of gender diverse students.

Table 1: Performance and Beliefs in the Student Study

	Performance	Belief	Belief–Performance
	(1)	(2)	(3)
Gender Diverse	-0.06	-1.41***	-1.41***
	(0.16)	(0.21)	(0.22)
Female	-0.46***	-1.03***	-0.80***
	(0.04)	(0.04)	(0.05)
Male Average	5.90	6.65	0.74
Gender Diverse – Female (= Gender Minority Gap – Gender Gap)			
Difference	0.40	-0.39	-0.61
p-value	0.01	0.06	0.01
Year in School FEs	Yes	Yes	Yes
School FEs	Yes	Yes	Yes
Performance FEs	No	Yes	No
N	10779	10779	10779

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and clustered at the subject level. Results are from OLS regressions of the dependent variable noted in the column. *Performance* is the number of questions that the subject answered correctly on the math and science test. *Belief* is the number of questions the subject believe that they answered correctly out of the 10 questions on the test. *Belief–Performance* is a subject’s belief minus their actual performance. *Gender Diverse* is an indicator for the subject selecting other when asked about their gender and identifies the “gender minority gap.” *Female* is an indicator for the subject selecting female when asked about their gender and identifies the “gender gap.” *Difference* is the difference between the *Female* and *Gender Diverse* coefficient estimates, which is thus equivalent to the difference between the gender minority gap and the gender gap, and *p-value* presents the corresponding p-value for a two-sided t-test of these two coefficient estimates. *Male Average* is the average of the dependent value for subjects selecting male when asked about their gender. Year in School FE and School FEs are dummies for each subject’s year in school (e.g., 6th grade, 7th grade, etc.) and school, respectively. Performance FEs are dummies for each possible number of questions a subject got right out of the 10 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 student’s belief minus actual performance. The analysis excludes the 28 subjects who selected other and provided an offensive response when asked about their gender and presents results from the remaining 10,779 subjects.

Despite performing similarly to male students and better than female students, gender diverse students have the most pessimistic beliefs about their absolute performance. Gender diverse students believe they answered an average of 5.21 questions correctly; female students believe they answered 5.39 questions correctly; and male students believe they answered 6.65 questions correctly.

Columns (2) and (3) of Table 1 examine beliefs while controlling for performance. Column (2) examines beliefs while including performance fixed effects (i.e., comparing equally performing male, female, and gender diverse students). Column (3) examines an individual-

level variable of beliefs about performance minus actual performance.

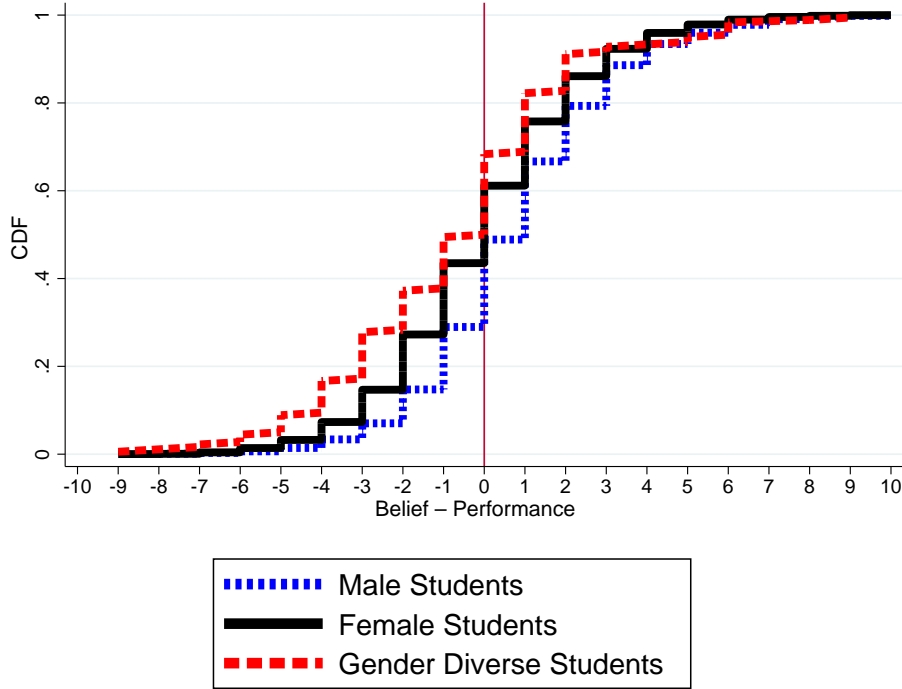
Column (2) shows that, controlling for performance, gender diverse students believe they answered 1.41 fewer questions correctly than equally performing male students, revealing a gender minority gap in confidence. In addition, female students believe they answered 1.03 fewer questions correctly than equally performing male students, which replicates gender gaps in confidence between men and women that are typical of male-typed tasks. The *Difference* and *p-value* rows of the table report -0.39 and $p = 0.06$, highlighting that the gender minority gap in confidence is larger than the gender gap in confidence between male and female students.

Column (3) explores the difference between a student’s belief and the student’s actual performance. These results reveal that, relative to their true performance, gender diverse students on average believe they answered 1.41 fewer questions correctly than male students, which again reveals a gender minority gap in confidence. In addition, relative to their true performance, female students believe they answered 0.80 fewer questions correctly than male students, which again replicates a gender gap in confidence between men and women that are typical of male-typed tasks. The *Difference* and *p-value* rows of the table report -0.61 and $p < 0.01$, highlighting that this gender minority gap in confidence is larger than this gender gap in confidence between male and female students.

The results in Column (3) also make clear the extent to which students (in)correctly estimate their performance. In particular, these results reveal that, on average, male students significantly overestimate their performance by 0.74 questions (see *Male Average*, $p < 0.01$), female students directionally but do not significantly underestimate their performance by 0.06 (0.74–0.80 questions, $p = 0.12$), and gender diverse students significantly underestimate their performance by 0.67 questions (0.74 – 1.41 questions, $p < 0.01$). To complement these average estimates, Figure 1 shows the CDFs of these differences between beliefs and performance for the three groups and confirms that nearly the entire distribution is shifted to the left for the gender diverse students.

Result 1 (Beliefs about Absolute Performance in the Student Study) There are gender minority gaps in confidence. Gender diverse students are less confident than male students.

Figure 1: Belief–Performance Distributions in the Student Study



Graph shows CDFs for *Belief–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.

3.1.2 Self-Evaluations in the Student Study

Next, we turn to self-evaluations that students provided as part of our experiment. Panel A of Table 2 presents regression results on students’ uninformed self-evaluations and Panel B presents results on students’ informed self-evaluations (i.e., after they were told how many questions they answered correctly on the test).

We start by considering students’ uninformed self-evaluations in Panel A of Table 2. Column (1) presents results for the *performance* question that asked students 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 students provide self-evaluations that are 17.46 points lower on average than those provided by male students, corresponding to an average self-evaluation provided by gender diverse students that is 26.3% lower than the average male self-evaluation. This gender minority gap is 6.49 points larger than the gender gap observed between equally performing male and female students, which is 10.97 points.

Column (2) of Panel A of Table 2 presents results for the *performance-bucket* question that asked students to indicate how well they think they performed on the test on a seven-

Table 2: Uninformed and Informed Self-Evaluations in the Student Study

	Performance (1)	Performance- Bucket (2)	Willingness (3)	Success (4)
Panel A: Uninformed Self-Evaluations				
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
Panel B: Informed Self-Evaluations				
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	10779	10779	10779	10779

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and clustered at the subject level. 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 noted in the column. See Table 1 for definitions of the independent variables, *Difference* 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.

point Likert scale. The average response provided by gender diverse students is 0.75 points (16.0%) lower than the average response of equally performing male students. This gender minority gap is 0.23 points larger than the gender gap observed between equally performing male and female students, which is 0.52 points.

Column (3) of Panel A of Table 2 presents results for the *willingness* question that asked students to indicate their agreement on a scale from 0 (entirely disagree) to 100 (entirely agree) with “I would choose to take a class that involves topics like those covered on the

test.” The average response provided by gender diverse students is 9.62 points (17.0%) lower than those provided by male students. This gender minority gap is 5.35 points larger than the gender gap observed between equally performing male and female students, which is 4.27 points.

Column (4) of Panel A of Table 2 presents results for the *success* question that asked students to indicate their agreement on a scale from 0 (entirely disagree) to 100 (entirely agree) with “I would succeed in a class that involves topics like those covered on the test.” The average response provided by gender diverse students is 16.09 points (23.5%). This gender minority gap is 8.60 points larger than the gender gap observed between equally performing male and female students, which is 7.48 points.

Appendix Figure B.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 (Uninformed Self-Evaluations in the Student Study) There are gender minority gaps in self-evaluations. Gender diverse students provide worse self-evaluations of their performance on a math and science test than equally performing male students.

One might wonder whether the gender minority gaps in self-evaluation reflect the fact that gender diverse students also believe they answered fewer questions correctly on the test. To investigate this possibility, we tell students 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 results about informed self-evaluations are presented in Panel B of Table 2. Similar to the uninformed self-evaluations, we again see that gender diverse students’ informed self-evaluations are significantly and substantially lower than their male peers. Looking at Columns (1), (3), and (4) of Panel B of Table 2, where subjects were asked to indicate their agreement with each of the corresponding statements from 0 to 100, we see that the average responses submitted by gender diverse students are 13.61 points (29.7%) lower for the *performance* question, 11.94 points (23.3%) lower for the *willingness* question, and 17.34 points (30.1%) lower for the *success* question relative to equally performing male students. For the seven-point Likert-scale question in Column (2), we similarly see that gender diverse students provide 0.54 points (15.0%) lower self-evaluations compared to their male counterparts. These gender minority gaps are all statistically significantly larger than the corresponding gender gaps between male and female students.¹⁴ Appendix Figure B.4 shows

¹⁴Relative to equally performing male students, female students’ average responses are 6.44 points (14.1%)

CDFs of the responses to each of the four informed self-evaluation questions. As above, however, differences in the distribution of responses may be harder to interpret because they do not account for underlying differences in performance between the groups.

Result 3 (Informed Self-Evaluations in the Student Study) Even after students are informed of how many questions they got correct, the gender minority gaps in self-evaluation persist. Gender diverse students provide worse self-evaluations of their performance on a math and science test than equally performing male students.

3.1.3 Robustness in the Student Study

The gender minority gaps in confidence and in self-evaluation we identify are robust to different ways of identifying students as gender diverse. Appendix Table A.1 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) to show that our results are not sensitive to dropping these students.

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*.¹⁵ 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 subjects 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

lower for the *performance* question, 0.26 points (7.2%) lower for the *performance-bucket* question, 2.94 points (5.7%) lower for the *willingness* question, and 5.34 points (9.3%) lower for the *success* question.

¹⁵Most 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.

“Prefer not to say” (for more details, see Appendix Table A.1).

Appendix Tables A.2–A.4 replicate the analysis conducted in Tables 1 and 2, showing results for each of the four sets of robustness tests. The results identified in Sections 3.1.1 and 3.1.2 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, even larger than the corresponding gender gaps we see when we compare responses of male and female students.¹⁶

3.2 Adult Study, *Math* Version

A total of 746 subjects completed the *Math* version of our Adult Study run on Prolific in June and July of 2023. On our demographic 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.¹⁷ Our data analysis and presentation of these data follow the same structure as the student study results in Section 3.1. We first present results on performance and confidence followed by results on uninformed and informed self-evaluations.

3.2.1 Performance and Confidence in the Adult Study, *Math* Version

Gender diverse subjects got an average of 12.51 questions correct out of 20. This performance is better than male subjects who got an average of 11.43 questions correct. Both of these performances are better than the performance of female subjects, who got an average of 10.25 questions correct. Column (1) of Table 3 presents regression results of performance and shows that these differences are statistically significant.

Despite outperforming men, gender diverse subjects believe they answered fewer questions correctly than male subjects. Gender diverse subjects believe they answered 9.56 questions correctly while male subjects believe they answered 10.30 questions correctly. Female subjects believe they answered 7.58 questions correctly.

Columns (2) and (3) of Table 3 compare these belief measures while controlling for performance. Column (2) examines beliefs while including performance fixed effects (i.e., comparing equally performing male, female, and gender diverse subjects). Column (3) examines an individual-level variable of beliefs about performance minus actual performance.

¹⁶In 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 $p < 0.05$ (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).

¹⁷Specifically, 122 subjects only selected “Transgender, non-binary, or another gender,” 27 subjects selected “Transgender, non-binary, or another gender” and “Male,” 19 subjects selected “Transgender, non-binary, or another gender” and “Female,” and 1 subject selected “Male” and “Female.”

Table 3: Performance and Beliefs in the Adult Study, *Math* Version

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

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and clustered at subject level. Results are from OLS regressions of the dependent variable noted in the column. *Performance* is the number of questions that the subject answered correctly on the math and science test. *Belief* is the number of questions the subject believe that they answered correctly out of the 20 questions on the test. *Belief–Performance* is a subject’s belief minus their actual performance. *Gender Diverse* is an indicator for the subject selecting “Transgender, non-binary, or another gender” or multiple options, when asked about their gender. *Female* is an indicator for the subject selecting only female when asked about their gender. *Male Average* is the average of the dependent value for subjects 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 subject 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 subject’s belief minus actual performance.

Column (2) shows that, controlling for performance, gender diverse subjects believe they answered 1.45 fewer questions correctly than equally performing male subjects, revealing a gender minority gap in confidence. Female subjects believe they answered 1.75 fewer questions correctly than equally performing male subjects, revealing a gender gap in confidence. The gender minority gap and the gender gap are comparably sized and not statistically significantly different ($p = 0.34$).

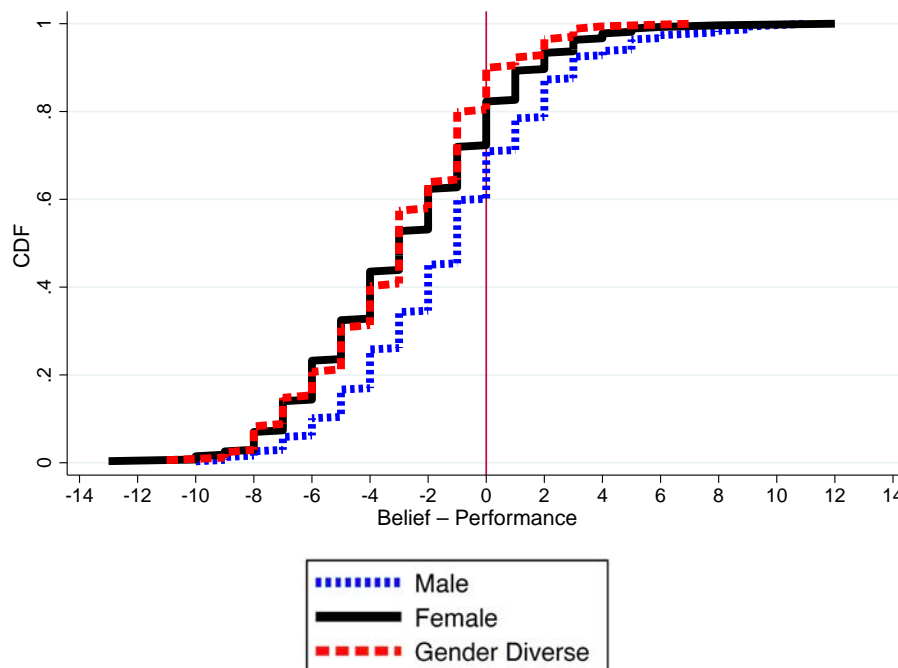
Column (3) explores the difference between a subject’s belief and the subject’s actual performance, calculated for each individual. In this case, we find that male subjects are underconfident: they underestimate their performance by 1.14 questions (see *Male Average*). Gender diverse subjects are less confident than male subjects ($p < 0.01$); on average, gender diverse subjects underestimate their performance by 2.95 questions, again revealing a gender minority gap in confidence. On average, female subjects underestimate their performance by

2.68 questions, again revealing a gender gap in confidence. As before, the gender minority gap and the gender gap are comparably sized and not statistically significantly different ($p = 0.38$).

Figure 2 shows the CDFs of these differences between beliefs and performance for the three groups and confirms that nearly the entire distribution is shifted to the left for the gender diverse subjects (and female subjects) relative to male subjects.

Result 4 (Beliefs about Absolute Performance in the Adult Study, *Math* Version)
There is a gender minority gap in confidence. Gender diverse adults are less confident than male adults.

Figure 2: Belief–Performance Distributions in the Adult Study, *Math* Version



Graph shows CDFs for *Belief–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.

3.2.2 Self-Evaluations in the Adult Study, *Math* Version

Panel A of Table 4 presents regression results on subjects' uninformed self-evaluations. Panel B presents results on subjects' informed self-evaluations (i.e., after they were told how many questions they answered correctly on the math and science test).

The results from Panel A of Table 4 suggest a similar pattern in self-evaluations as seen in beliefs about absolute performance. Gender diverse subjects provide self-evaluations that are less positive than equally performing male subjects. Across all four columns of Panel A, the

Table 4: Uninformed and Informed Self-Evaluations in the Adult Study, *Math* Version

	Performance	Performance- Bucket	Willingness	Success
	(1)	(2)	(3)	(4)
Panel A: Uninformed Self-Evaluations				
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
Panel B: Informed Self-Evaluations				
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 subject level. Results are from OLS regressions of a subject’s response to the uninformed (elicited before the subject learns their test performance) (Panel A) and informed (Panel B) self-evaluation noted in the column. See Table 3 for definitions of the independent variables, *Difference* and Performance FEs.

coefficient on *Gender Diverse* is statistically significantly negative and sizable. In column (1), the coefficient is -9.11 , indicating that, when asked to indicate agreement with the statement “I performed well on the test,” gender diverse subjects report self-evaluations that are 18.4% lower than equally performing male subjects, whose average is 49.55. Similar patterns arise in the other columns, with gender diverse subjects providing self-evaluations that are 11.8% to 35.3% lower than the average self-evaluations of male subjects. The coefficient on *Female* is also statistically significantly negative in all four columns, revealing a gender gap in self-evaluations and replicating prior work. Across all specifications, the gender minority gap and the gender gap are similarly sized and never statistically significantly different (p ranges from 0.76 to 0.95 across the four self-evaluation measures).

Result 5 (Uninformed Self-Evaluations in the Adult Study, *Math* Version) There are gender minority gaps in self-evaluation. Gender diverse adults provide worse self-evaluations of their performance on a math and science test than equally performing male adults.

Panel B of Table 4 shows the self-evaluations after subjects have been told how many questions they answered correctly on the test. Even after adult subjects are told how many questions they answered right, a gender minority gap persists, albeit to a somewhat muted degree. We see the gender minority gap with directionally negative coefficients on *Gender Diverse* in all four columns; results are statistically significant in three out of the four columns and close in the remaining column ($p = 0.10$ in column (1)). Again, we find negative and statistically significant coefficients on *Female*, replicating gender gaps in informed self-evaluations found in prior work. As with uninformed self-evaluations, the gender minority gap and the gender gap are similarly sized in all specifications (p ranges from 0.20 to 0.95 across the four self-evaluation measures).

Comparing results from Panel A to Panel B, the negative coefficients on *Gender Diverse* are smaller for the two self-evaluation questions specifically about prior performance (i.e., the *Performance* and *Performance-Bucket* questions in columns (1) and (2), respectively) than they were in the corresponding columns of Panel A. This latter result is consistent with some of the gender minority gap reflecting differences in beliefs about absolute performance on the test.

Result 6 (Informed Self-Evaluations in the Adult Study, *Math* Version) Even after adults are informed of how many questions they got correct, the gender minority gaps in self-evaluation persist. Gender diverse adults provide significantly worse self-evaluations of their performance on a math and science test than equally performing male adults.

3.3 Adult Study, *Verbal* Version

A total of 748 subjects completed the *Verbal* version of our adult study 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.¹⁸ Again, we first present results on performance and confidence followed by results on uninformed and informed self-evaluations.

¹⁸Specifically, 110 subjects only selected “Transgender, non-binary, or another gender,” 22 subjects selected “Transgender, non-binary, or another gender” and “Male,” 27 subjects selected “Transgender, non-binary, or another gender” and “Female,” and 2 subjects selected “Male” and “Female.”

3.3.1 Performance and Confidence in the Adult Study, *Verbal* Version

Table 5: Performance and Beliefs in the Adult Study, *Verbal* Version

	Performance	Belief	Belief–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	0.95	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 subject level. Results are from OLS regressions of the dependent variable noted in the column. See Table 3 for definitions of the independent variables, *Difference* and Performance FEs. The only difference between Table 3 and this table is that this table presents data from the *Verbal* version.

Gender diverse subjects got an average of 12.45 questions correct out of 20. This performance is better than female subjects who got an average of 11.25 questions correct. Both of these performances are better than the performance of male subjects, who got an average of 10.58 questions correct. Column (1) of Table 5 presents regression results of performance and shows that these differences are statistically significant.

Columns (2) and (3) of Table 5 analyze beliefs about performance. Column (2) examines beliefs while including performance fixed effects (i.e., comparing equally performing male, female, and gender diverse subjects). Column (3) examines an individual-level variable of beliefs about performance minus actual performance.

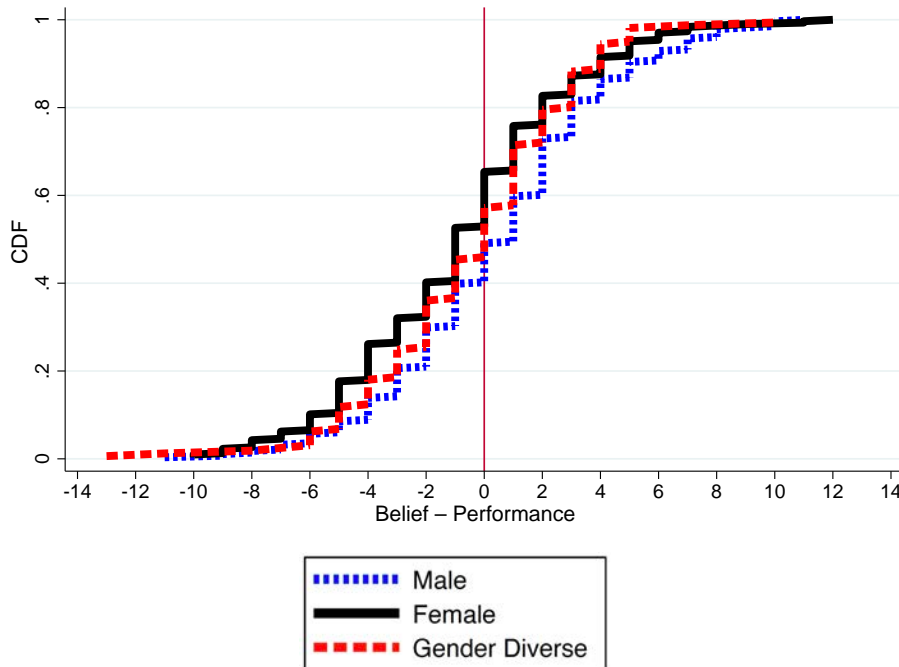
Column (2) shows that gender diverse subjects believe they answered 0.10 fewer questions correctly than equally performing male subjects. This difference is not statistically significant and small in magnitude, thus revealing no evidence for a gender minority gap. Female subjects believe they answered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

Column (3) explores the difference between a subject’s belief and their actual performance, calculated for each individual. In this case, we find that male subjects are slightly overconfident: they overestimate their performance by 0.50 questions (see *Male Average*). On this measure, gender diverse subjects are less confident than male subjects, 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.)¹⁹ We also see that female subjects are less confident than male subjects with this measure, evidence of a gender gap in confidence in this setting. Figure 1 shows the CDFs of these differences between beliefs and performance for the three groups.

Result 7 (Beliefs about Absolute Performance in the Adult Study, *Verbal* Version)
We find limited evidence of a gender minority gap in confidence among adults on a verbal test.

Figure 3: Belief–Performance Distributions in the Adult Study, *Verbal* Version



Graph shows CDFs for *Belief–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.

¹⁹We 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 3 and 5; the gender minority gap we observe in the *Verbal* version is smaller than the corresponding gender minority gap we see in the *Math* version ($p < 0.01$). Indeed, of the 10 comparisons that we can make between gender minority gaps—comparing the coefficient on *Gender Diverse* across the *Verbal* and *Math* versions of the study (across Tables 3 and 5 and across Tables 4 and 6)—the coefficient on *Gender Diverse* is always at least directionally smaller in magnitude in the *Verbal* version (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.

3.3.2 Self-Evaluations in the Adult Study, *Verbal* Version

Panel A of Table 6 presents regression results on subjects’ uninformed self-evaluations about the verbal test and Panel B presents results on subjects’ 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 subjects and equally performing male subjects. Meanwhile, female subjects have worse self-evaluations than equally performing male subjects 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 8 (Uninformed Self-Evaluations in the Adult Study, *Verbal* Version) We find no evidence of a gender minority gap in self-evaluation among adults on a verbal test.

Panel B of Table 6 shows the self-evaluations after subjects 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 subjects and male subjects. Female subjects again have worse self-evaluations than equally performing male subjects across all four questions.

Result 9 (Informed Self-Evaluations in the Adult Study, *Verbal* Version) Even after adults are informed of how many questions they got correct, we still find no evidence of a gender minority gap in self-evaluation among adults 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 5 and in columns (1) and (2) but not (3) and (4) of Table 6). 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, 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.²⁰

²⁰Of 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 *Math* versions of the study—the gender gap is directionally smaller in the *Verbal* version in 8 of the 10 tests and significantly smaller in 4 of those tests. The gap is never significantly larger in the *Verbal* version. (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 6: Informed and Uninformed Self-Evaluations in the Adult Study, *Verbal* Version

	Performance	Performance- Bucket	Willingness	Success
	(1)	(2)	(3)	(4)
Panel A: Uninformed Self-Evaluations				
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
Panel B: Informed Self-Evaluations				
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 subject level. Results are from OLS regressions of a subject’s response to the uninformed (elicited before the subject learns their test performance) (Panel A) and informed (Panel B) self-evaluation noted in the column. See Table 4 for definitions of the independent variables, *Difference* and Performance FEs. The only difference between Table 4 and this table is that this table presents data from the verbal test.

4 Predicting Academic Performance with Confidence and Self-Evaluations

An additional benefit of our partnership with the Character Lab Research Network (CLRN) is that CLRN was able to give us access to administrative data on student performance that we can link to the responses in our student study, which allows us to explore how responses to our confidence and self-evaluation questions correlate with academic performance.

Table 7 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).²¹ 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 that even informed self-evaluations have predictive power. Appendix Table A.5 follows Table 7 but shows regressions of Math GPA in each quarter, rather than overall GPA. Results are qualitatively very similar.²²

Result 10 (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.

²¹Not 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).

²²While 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 A.6 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 7: Regressions of Overall GPA

	Academic Quarter (Q1–Q8) from 2020–2021 & 2021–2022							
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Panel A: X = Absolute Belief (0–10)								
<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)
Panel B: X = Uninformed Performance Self-Evaluations (0–100)								
<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)
Panel C: X = Uninformed Performance-Bucket Self-Evaluations (1–7)								
<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)
Panel D: X = Uninformed Willingness Self-Evaluations (0–100)								
<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)
Panel E: X = Uninformed Success Self-Evaluations (0–100)								
<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)
Panel F: X = Informed Performance Self-Evaluations (0–100)								
<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)
Panel G: X = Informed Performance-Bucket Self-Evaluations (1–7)								
<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)
Panel H: X = Informed Willingness Self-Evaluations (0–100)								
<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)
Panel I: X = Informed Success Self-Evaluations (0–100)								
<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 subject 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).

5 Discussion

We document gender minority gaps in confidence and self-evaluation. In our student study, gender diverse middle and high school students believe they answered fewer questions correctly and provide less favorable self-evaluations about their performance on a math and science test than equally performing male peers. These gender minority gaps in self-evaluation stay large and significant even after we inform students about how many questions they actually answered correctly on the test. In the student study, we also observe gender gaps between men and women in confidence and self-evaluations; the gender minority gaps are even larger than the gender gaps we observe. In our adult study, we again find evidence of gender minority gaps in confidence and self-evaluations on a math and science test, which are a similar magnitude as the gender gaps we observe in that setting. Additional results suggest that the domain of the task can impact gender minority gaps. In a verbal test, we find little-to-no evidence that gender diverse adults were less confident or provided more pessimistic self-evaluations than equally performing male adults. Finally, we find that our measures of confidence and self-evaluation are highly predictive of current and future academic performance among our student sample.

The results in our paper open up many important avenues for future work. For example, future work may investigate the causes and consequences of gender minority gaps in confidence and self-evaluation. In light of previous research on how gender norms between men and women relate to a wide variety of outcomes (Bertrand, Kamenica and Pan, 2015; Dhar, Jain and Jayachandran, 2022; Field et al., 2021; Pande and Roy, 2021; Jayachandran et al., 2023), future work may investigate norms for gender minorities—which may be quite nuanced and certainly warrant more attention themselves—as well as the potential connection between such norms and gender minorities’ confidence and self-evaluation.²³ In addition, gender minority gaps may influence decisions made by gender minorities themselves (e.g., whether they enter a competition, enter a negotiation, apply for a job, etc.) as well as the decisions made by others (e.g., whether they hire them, promote them, etc.).²⁴ A related question is when and whether gender minorities may benefit from—or perhaps alternatively

²³Recent research documents differences in the treatment of gender minorities (Badgett, Carpenter and Sansone, 2021) and the treatment of—as well as the views toward—gender minorities have been found to be worse relative to sexual minorities (Lewis et al., 2017, 2022; Aksoy, Carpenter and Sansone, 2022). Such differences in treatment may contribute to the evolution of differences in traits such as confidence and self-evaluation.

²⁴Indeed, prior work on gender differences between men and women suggests that even if well-intentioned individuals become aware of a gender gap in how individuals evaluate their own performance, they may fail to accurately account for it when reviewing such self-evaluations (Exley and Nielsen, 2023). Such patterns may also arise for gender minority gaps. These patterns could also be exacerbated by underlying discrimination, including the possibility of inaccurate beliefs as also documented in prior literature that focuses on comparing men and women (Coffman, Exley and Niederle, 2021; Bohren et al., 2023).

face backlash from—expressing higher levels of confidence or self-evaluation, which likely also relates to the norms and culture of one’s environment. As shown in prior work on gender-specific backlash ([Riach and Rich, 2002](#); [Bowles, Babcock and Lai, 2007](#); [Rudman and Phelan, 2008](#)) and the potential cost of leaning-in ([Exley, Niederle and Vesterlund, 2020](#)), it need not follow that higher levels of confidence and self-evaluations are optimal.

As suggested by the differences between the math and verbal versions of our adult study, and the differing magnitudes of the gender minority gaps across our student and adult studies in the math and science domain, the extent to which gender minority gaps exist may also depend on the domain and on the population being considered. As the share of gender minorities increases over time, it may naturally follow that the size of gender minority gaps vary across generations. We hope future work investigates these possibilities and important nuances.

Future work might also explore 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 [Buser, Geijtenbeek and Plug, 2018](#); [Aksoy and Chadd, 2023](#)). Relatedly, future work may consider intersectionality more broadly, such as the impact of being a gender minority as well as being a member of other under-represented groups such as those relating to sexual minorities.²⁵

²⁵Considering 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.** 2023. “Competitiveness at the Intersection of Gender and Sexual Orientation.” *Available at SSRN*.
- Aksoy, Billur, Christopher S. Carpenter, and Dario Sansone.** 2022. “Understanding Labor Market Discrimination Against Transgender People: Evidence from a Double List Experiment and a Survey.” *National Bureau of Economic Research Working Paper Series*, , (30483).
- Aksoy, Billur, Ian Chadd, and Boon Han Koh.** 2023. “Sexual Identity, Gender, and Anticipated Discrimination in Prosocial Behavior.” *European Economic Review*, 154: 104427.
- 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, MV, Christopher S Carpenter, and Dario Sansone.** 2021. “LGBTQ economics.” *Journal of Economic Perspectives*, 35(2): 141–70.
- Badgett, MV, 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.
- 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*.

- Bowles, Hannah Riley, Linda Babcock, and Lei Lai.** 2007. “Social incentives for gender differences in the propensity to initiate negotiations: Sometimes it does hurt to ask.” *Organizational Behavior and Human Decision Processes*, 103(1): 84–103.
- 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*.
- 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.
- Chen, Weiwei, Wayne A Grove, and Andrew Hussey.** 2017. “The role of confidence and noncognitive skills for post-baccalaureate academic and labor market outcomes.” *Journal of Economic Behavior & Organization*, 138(10–29).
- 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, 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*.
- 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.** 2023. “The Gender Gap in Confidence: Expected But Not Accounted For.” *Working Paper*.
- 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.
- 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 Schwieren, 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 Iriberry.** 2019. “A review of gender differences in negotiation.” *Oxford Research Encyclopedia of Economics and Finance*.
- Jayachandran, Seema, Lea Nassal, Matthew Notowidigdo, Marie Paul, Heather Sarsons, and Elin Sundberg.** 2023. “Moving to opportunity, together.”
- Jones, Jeffrey M.** 2022. “LGBT Identification in U.S. Ticks Up to 7.1%.”
- Kamas, Linda, and Anne Preston.** 2018. “Competing with confidence: The ticket to labor market success for college-educated women.” *Journal of Economic Behavior & Organization*, 155: 231–252.

- Lewis, Daniel C, Andrew R Flores, Donald P Haider-Markel, Patrick R Miller, and Jami K Taylor.** 2022. "Transitioning opinion? assessing the dynamics of public attitudes toward transgender rights." *Public Opinion Quarterly*, 86(2): 343–368.
- 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 NT 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.
- 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.
- Riach, P. A., and J. Rich.** 2002. "Field Experiments of Discrimination in the Market Place." *The Economic Journal*, 112(483).

- Risse, Leonora, Lisa Farrell, and Tim RL Fry.** 2018. "Personality and pay: do gender gaps in confidence explain gender gaps in wages?" *Oxford Economic Papers*, 70(4): 919–949.
- Rudman, Laurie A, and Julie E Phelan.** 2008. "Backlash effects for disconfirming gender stereotypes in organizations." *Research in organizational behavior*, 28(6-79).
- Sansone, Dario.** 2019. "LGBT students: New evidence on demographics and educational outcomes." *Economics of Education Review*, 73(101933).
- 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 Tables

Table A.1: Sample and Variable Descriptions for Robustness Checks

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 thus 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 [A.2–A.4](#).

Table A.2: Performance Beliefs with Alternative Gender Classifications

	Perf	Belief	Belief-Perf
Panel A: Our Gender Measure (Full Sample), N=10,807			
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
Panel B: Our Gender Measure (Restricted Sample), N=10,673			
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
Panel C: CLRN Gender Measure, N=10,272			
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
Panel D: CLRN Gender Measure (Full Sample), N=10,807			
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 subject level. Results are from OLS regressions of the dependent variable noted in the column. See Table 1 tables notes and A.1 for more information about samples and variables used in each panel and FEs.

Table A.3: Uninformed Self-Evaluations with Alternative Gender Classifications

	Performance	Performance- Bucket	Willingness	Success
Panel A: Our Gender Measure (Full Sample), N=10,807				
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
Panel B: Our Gender Measure (Restricted Sample), N=10,673				
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
Panel C: CLRN Gender Measure, N=10,272				
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
Panel D: CLRN Gender Measure (Full Sample), N=10,807				
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 subject level. 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 1 notes for details on FEs and see Appendix Table A.1 for more information about samples and variables used in each panel.

Table A.4: Informed Self-Evaluations with Alternative Gender Classifications

	Performance	Performance- Bucket	Willingness	Success
Panel A: Our Gender Measure (Full Sample), N=10,807				
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
Panel B: Our Gender Measure (Restricted Sample), N=10,673				
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
Panel C: CLRN Gender Measure, N=10,272				
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
Panel D: CLRN Gender Measure (Full Sample), N=10,807				
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>	-7.30	-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 subject level. 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 1 notes for details on FEs and see Appendix Table A.1 for more information about samples and variables used in each panel.

Table A.5: Regressions of Math GPA

	Academic Quarter (Q1–Q8) from 2020–2021 & 2021–2022							
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Panel A: X = Absolute Belief (0–10)								
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)
Panel B: X = Uninformed Performance Self-Evaluations (0–100)								
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)
Panel C: X = Uninformed Performance-Bucket Self-Evaluations (1–7)								
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)
Panel D: X = Uninformed Willingness Self-Evaluations (0–100)								
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)
Panel E: X = Uninformed Success Self-Evaluations (0–100)								
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)
Panel F: X = Informed Performance Self-Evaluations (0–100)								
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)
Panel G: X = Informed Performance-Bucket Self-Evaluations (1–7)								
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)
Panel H: X = Informed Willingness Self-Evaluations (0–100)								
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)
Panel I: X = Informed Success Self-Evaluations (0–100)								
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 subject 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).

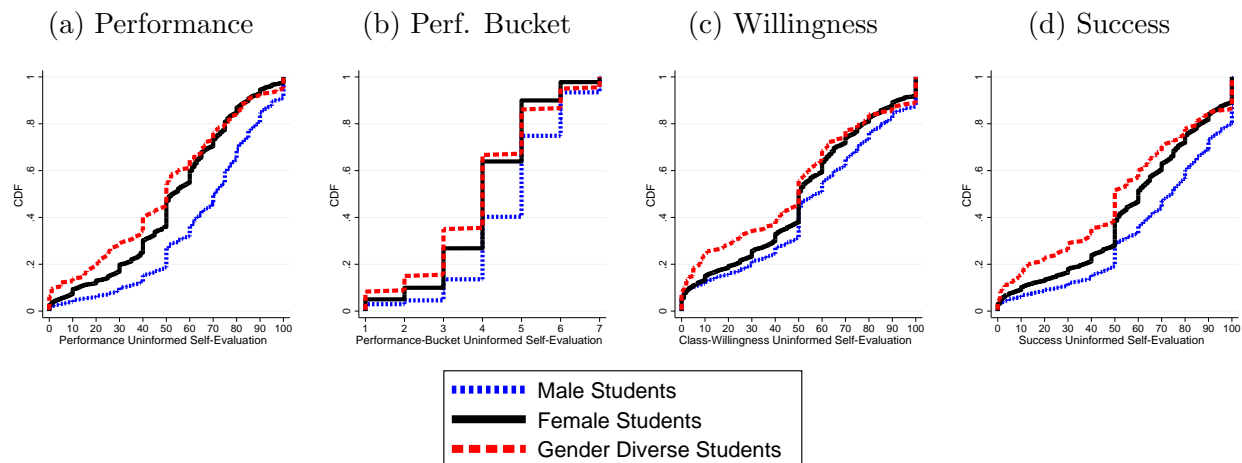
Table A.6: Regressions of Overall and Math GPA

	Academic Quarter (Q1–Q8) from 2020–2021 & 2021–2022							
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Panel A: DV = Overall GPA								
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)
N	10590	10569	10435	9781	7614	7619	7469	7316
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
Panel B: DV = Math GPA								
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)
N	10348	10272	10212	9577	7393	7383	7246	7075
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

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and clustered at the subject 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 1 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).

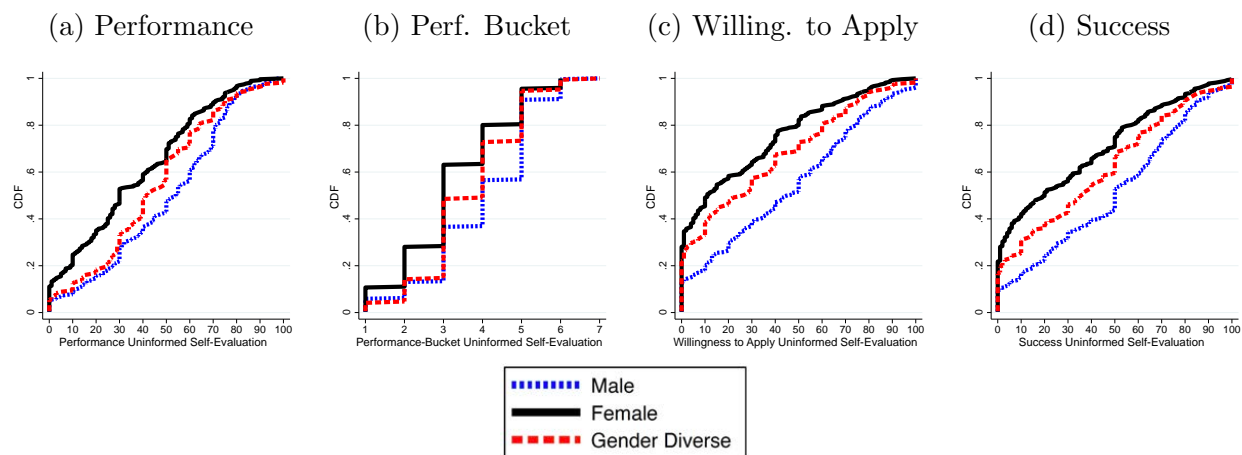
B Additional Figures

Figure B.1: CDFs for Uninformed Self-Evaluations in the Student Study



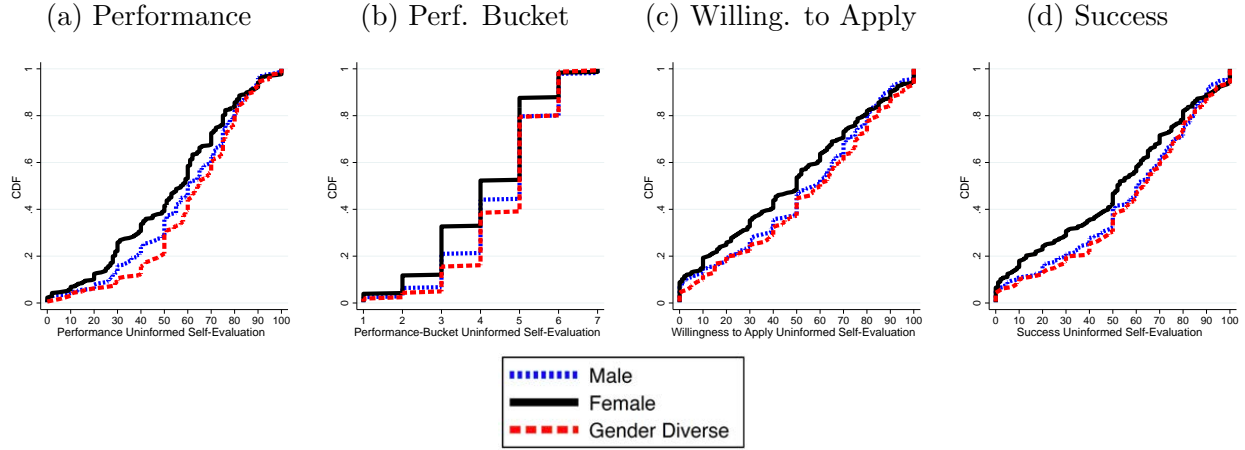
Graphs show CDFs of responses to the question noted in each panel, elicited before performance information is provided.

Figure B.2: CDFs for Uninformed Self-Evaluations in the Adult Study, *Math* Version



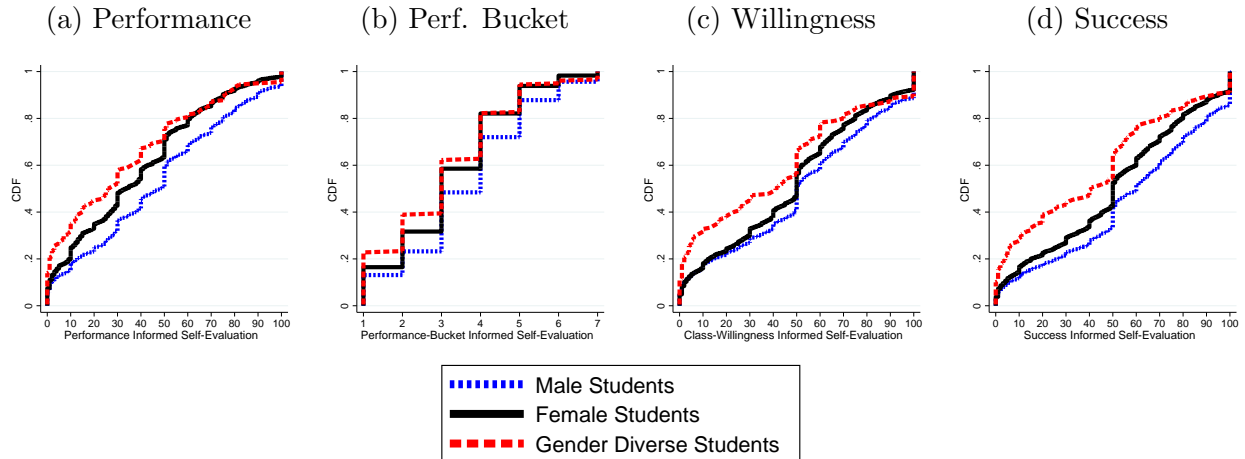
Graphs show CDFs of responses to the question noted in each panel, elicited before performance information is provided.

Figure B.3: CDFs for Uninformed Self-Evaluations in the Adult Study, *Verbal* Version



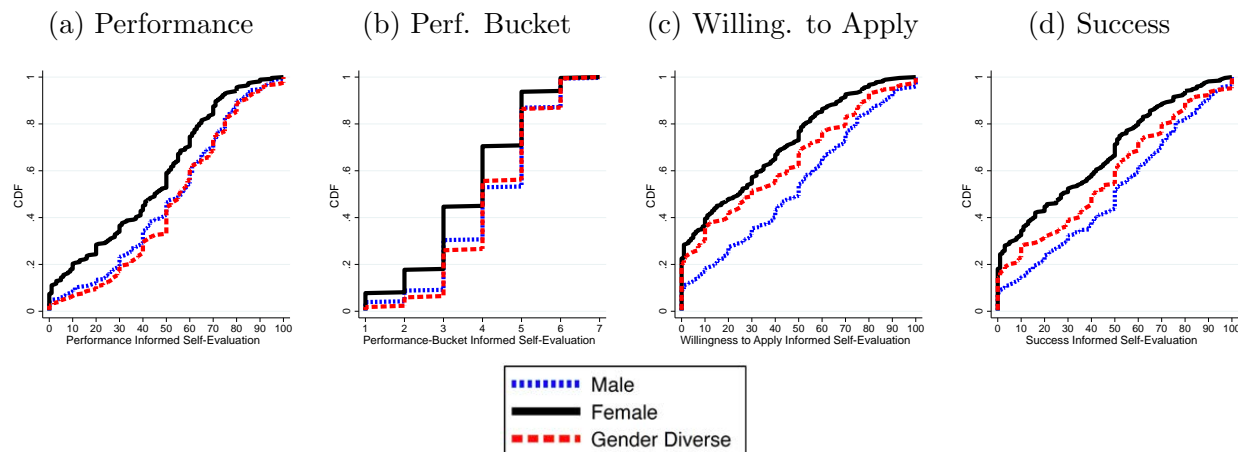
Graphs show CDFs of responses to the question noted in each panel, elicited before performance information is provided.

Figure B.4: CDFs for Informed Self-Evaluations in the Student Study



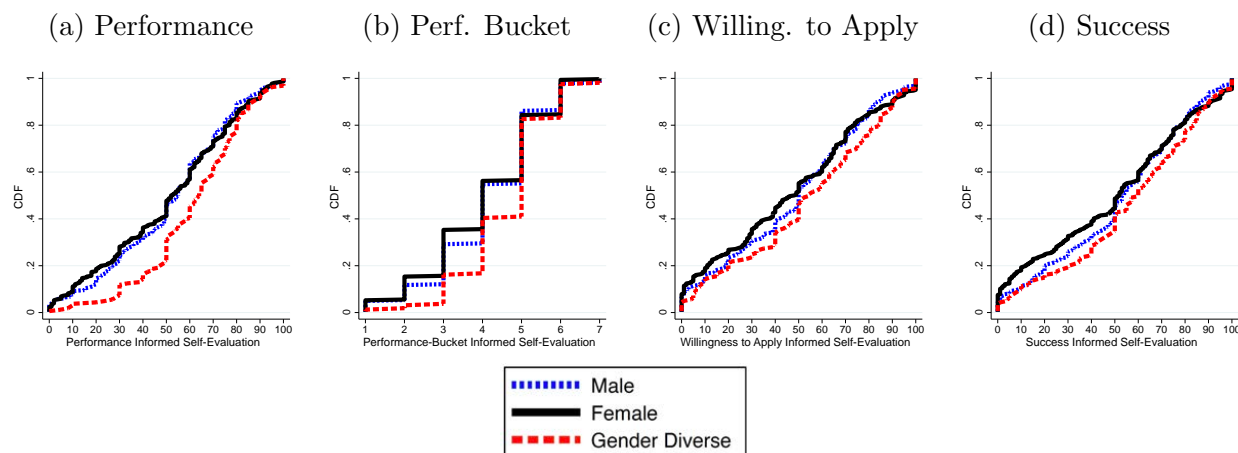
Graphs show CDFs of responses to the question noted in each panel, elicited after performance information is provided.

Figure B.5: CDFs for Informed Self-Evaluations in the Adult Study, *Math* Version



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

Figure B.6: CDFs for Informed Self-Evaluations in the Adult Study, *Verbal* Version



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

C Experimental Instructions

C.1 Experimental Instructions for the Student Study

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 C.1 and an example of a question on the test is displayed in Figure C.2 (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 C.3). On the second page, they are asked the self-evaluation questions (see Figure C.4). 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 C.5). On the fourth page, they are asked the self-evaluation questions again (see Figure C.6). On the fifth page, they are asked for demographic information including their gender identity (see Figure C.7).

Figure C.1: 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 C.2: 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 C.3: 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 C.4: 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 C.5: 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 C.6: Informed Self-Evaluation Questions in Student Study

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 C.7: Screenshot of Gender Question in the Student Study

Please select your gender.

Male

Female

Other

C.2 Experimental Instructions for the Adult Study

The Adult Study closely follows the design discussed in Section C.1 with the exceptions discussed in Section 2. The instructions for the experiment are displayed in Figure C.8 for the *Math* version and Figure C.9 for the *Verbal* version. Examples of questions on the test are displayed in Figures C.10 and C.11 (note that the timer in Figure C.10 indicates the participant has 26 seconds left to answer the question although the timer starts at 30 seconds for the *Math* version and 15 seconds for the *Verbal* version).

After completing the test, they are asked to complete the remainder of the study that follows a similar structure as the Adult Study. First, they are asked about their absolute performance belief (see Figure C.12). Second, they are provided with additional instructions (see Figure C.13) and then asked the self-evaluation questions (see Figure C.14). Third, participants are provided with perfect information on their absolute performance and are required to correctly report back their absolute performance (see Figure C.15). Fourth, they are provided with additional instructions (see Figure C.16) and are asked the self-evaluation questions again (see Figure C.17). Fifth, they are asked for demographic information including their gender identity (see Figure C.18).

Figure C.8: Part 1 Instructions for the test in the Adult Study, *Math* Version

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 C.9: Part 1 Instructions for the test in the Adult Study, *Verbal* Version

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 C.10: Example question on the test in the Adult Study, *Math* Version

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 C.11: Example question on the test in the Adult Study, *Verbal* Version

15

Question 2 out of 20:

WORD KNOWLEDGE: Indolence most nearly means

bliss.

tolerance.

serenity.

laziness.



Figure C.12: 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 C.13: 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 C.14: 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 C.15: 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 C.16: 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.

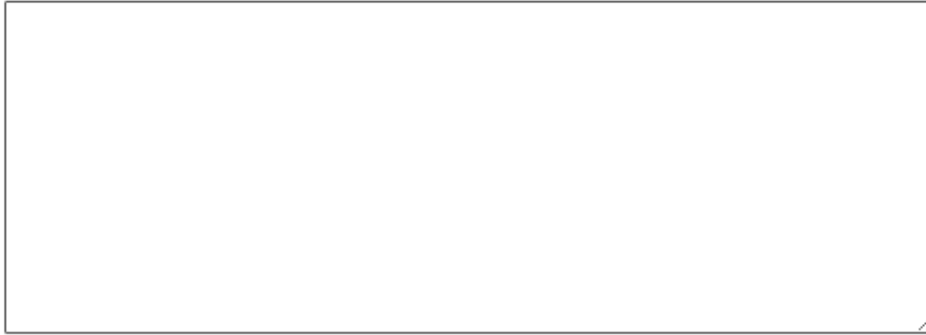
will depend on how you answer the questions -- on the next page -- about your performance on the test you took in part 1.



Figure C.17: 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	Strongly		Somewhat	Neither	Somewhat		Strongly	Entirely		
Disagree	Disagree	Disagree	Disagree	Disagree	Agree	Agree	Agree	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 C.18: Screenshot of Gender Question in the Student Study

Are you: (Mark all that apply)

Male

Female

Transgender, non-binary, or another gender