NBER WORKING PAPER SERIES

THE (UNEXPECTED) GENDER MINORITY GAP IN CONFIDENCE

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, Revised September 2025

This paper was supported by Character Lab and facilitated through the Character Lab Research Network, a consortium of schools across the country working collaboratively with scientists to advance scientific insights that help kids thrive. Harvard Business School provided generous financial support. We would like to thank Christopher Carpenter and Dario Sansone, and various seminar and conference participants for their helpful comments and feedback. This paper includes data from two pre-registered studies: see details in Footnote 16 for the Adult Study and Footnote 9 for the Predictions Study. As explained in Section 4.1 and Footnote 20, the authors had access to the data for the Student Study prior to the idea for this paper, so that data was not pre-registered. This study secured IRB approval from Harvard University and University of Michigan. 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 They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

The (Unexpected) Gender Minority Gap in Confidence
Billur Aksoy, Christine L. Exley, and Judd B. Kessler
NBER Working Paper No. 32061
January 2024, Revised September 2025
JEL No. C91, D91, J16

ABSTRACT

An increasing share of the population identifies their gender in a way other than as a man or woman. Yet, there is little work on beliefs about these gender minorities. We find that subjects, while relatively certain about the traits of men and women, are uncertain about the confidence, competitiveness, risk aversion, and generosity of gender minorities. Despite this uncertainty about gender minorities, however, we find that gender minorities are less confident than equally-performing men, resulting in a gender minority gap that is just as large as the gender gap in confidence between men and women.

Billur Aksoy
Rensselaer Polytechnic Institute
Department of Economics
aksoyb3@rpi.edu

Christine L. Exley
University of Michigan
Department of Economics
clexley@gmail.com

Judd B. Kessler
University of Pennsylvania
The Wharton School
and NBER
judd.kessler@wharton.upenn.edu

1 Introduction

A sizable share of the population identifies as part of a gender minority group. Examples include individuals who identify as transgender, non-binary, or genderqueer; such identities can overlap and evolve over time.[1] 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). Yet, there is little work on gender minorities, and the work that does exist shows that, compared to the general population, gender minorities have significantly worse economic outcomes (Badgett, Carpenter and Sansone, 2021; Carpenter, Eppink and Gonzales, 2020; Carpenter, Lee and Nettuno, 2022); have worse educational outcomes (Meyer et al., 2017; Downing and Przedworski, 2018; Sansone, 2019); and are more likely to be unemployed, be in low-income households, and be uninsured (Badgett, Carpenter and Sansone, 2021).

To help further our understanding of this economically disadvantaged group—and to help encourage the gender literature to more broadly consider gender minorities—this paper investigates the traits of gender minorities and people's *beliefs* about these traits.

Our *Predictions Study* investigates beliefs about five commonly studied behavioral traits: overconfidence, underconfidence, competitiveness, risk taking, and generosity (for related reviews, see, e.g., Croson and Gneezy (2009), Niederle (2016), and Eckel et al. (2021)).[2] Rather than only asking about male and female participants, however, we also ask about *gender diverse participants* who identify in some way other than "male" or "female." In addition, rather than only requiring predictors to make specific guesses about these three groups, we also allow predictors to directly indicate that they are unsure about these behavioral traits. For instance, predictors can say "yes," "no," or "I'm not sure" when asked

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

[2] Prior work on gender has focused on gender gaps between men and women with an eye toward explaining gaps between those genders in pay, representation in certain fields, and roles in corporate and political leadership (Bertrand, Goldin and Katz, 2010; Blau and Kahn, 2017; Grossman et al., 2019; Bütikofer, Løken and Willén, 2022). To explain these differences, researchers have leveraged observational data—to consider factors such as occupational selection and institutional and policy features—and have measured various traits in experiments, identifying gender differences between men and women in traits such as confidence (Barber and Odean, 2001; Niederle and Vesterlund, 2007), self-evaluations (Exley and Kessler, 2022), competitiveness (Niederle and Vesterlund, 2007), risk taking (Eckel and Grossman, 2008), negotiation (Babcock and Laschever, 2003; Hernandez-Arenaz and Iriberri, 2019), the contribution of ideas (Coffman, 2014), and image concerns (Bursztyn, Fujiwara and Pallais, 2017).

whether a gender diverse participant is overconfident. Providing predictors with the option to indicate that they are unsure proves important to understanding their beliefs about gender diverse participants. These results reveal that, while predictors expect a gender gap in confidence between men and women—as also found in Exley and Nielsen (2024)—predictors are much more uncertain about the confidence of gender minorities. Indeed, our first main result is that, relative to the more certain views they hold about men and about women, predictors are between 2 and 10 times more likely to indicate that they are unsure about gender diverse participants when asked about each of the five traits. Similar patterns persist when we incentivize predictions about the confidence and self-evaluations of men, women, and gender minorities who are asked about their performance on a math and science test.

Despite the predictors not expecting a gender minority gap—in part given their uncertainty about gender minorities—we find clear and consistent patterns between gender minorities and other genders in their stated confidence. We find large and robust gender minority gaps. Relative to equally-performing men, gender minorities have significantly worse confidence about their performance on the math and science test. Indeed, this *gender minority gap* is often larger than the *gender gap* in confidence between equally performing men and women that we also document. That is, the gender minority gap is unexpected while the gender gap is expected, despite the former being, if anything, larger.

Additional results document the robustness of the gender minority gap and provide new insights. First, in addition to observing a gender minority gap in confidence—measured by participants' beliefs about their performance on the math and science test—we also observe gender minority gaps in subjective self-evaluations about their performance on that test. Second, we observe that the gender minority gaps persist even after we provide participants with precise information on exactly how many questions they got correct on that test. Third, while our results discussed so far relate to an online sample that we recruited for our *Adult Study*, we also replicate our results with a large sample of middle school and high school youth in our *Student Study*.[3] Fourth, the gender minority gaps we document are robust to a variety of different classifications, including broader subgroups and narrower subgroups of gender diverse participants (e.g., non-binary participants). Fifth, counter to explanations in which gender minorities are broadly lacking confidence, we find that—as also observed in prior literature on how gender differences between men and women can depend on the

[3]While the gender minority gaps we document in our Student Study are novel, as are all the results from the Predictions Study and the Adult Study, the gender gaps between male and female students in our Student Study are largely *not* novel. As explained in Section 4.1, Exley and Kessler (2022) also analyze this youth data; however, that paper relies on administrative data that only has a binary classification of gender (i.e., only male or female), whereas this paper relies on self-reported gender that allows students to identify as gender diverse. The novel analysis of that data in this paper relates to the examination of gender diverse students.

context—the gender minority gaps prove to be context-specific. We do not observe evidence of a significant gender minority gap among our adult population when we instead ask them to complete a test relating to verbal skills.

One challenge with conducting research on gender minority groups is that data on gender identity is often recorded as binary or is missing in administrative records. The lack of data—and thus research—on gender identity in the United States is likely to worsen given the current political climate, particularly in light of Executive Order 14168, issued on January 20, 2025.[4] This order mandates that federal agencies recognize only biological sex, disregarding gender identity in all official matters. As a result, data on gender identity has already been removed from various federal datasets.[5] We overcame this challenge by collecting new data that allowed people to self-identify their gender as part of our studies.

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

A central contribution of our paper relates to exploring—and documenting—gender minority gaps. We view this as an important direction for the gender literature, given the paucity of work on the behaviors and traits of gender minorities, with the notable exception of concurrent work in Coffman, Coffman and Ericson (2024). We are also, to our knowledge, the only paper to document robust gender minority gaps in confidence among a youth population. With many important and open questions that are particularly relevant for this group, as discussed in our Conclusion, we hope future work continues to collect inclusive gender data.[6]

An additional contribution of our paper relates todocumenting the unexpected nature

[4] https://www.whitehouse.gov/presidential-actions/2025/01/defending-women-from-gender-ideology-extremism-and-restoring-biological-truth-to-the-federal-government

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

[6] Indeed, we do not view any of the classic gender findings as having clear implications for this group. For instance, consider the gender minority gap in confidence documented in our paper. On one hand, one might have expected these gaps in confidence between gender diverse people and equally performing men because gender diverse people are part of a marginalized group, and marginalized groups often display lower confidence than majority groups. On the other hand, self-identifying as gender diverse means rejecting societys imposed gender identity classification and perhaps subjecting oneself to additional discrimination, so gender diverse individuals could be even more confident and self-assured than other groups.

of the gender minority gaps as particularly important for a few reasons. First, that we directly explore—and then document—the unexpected nature of the gender minority gaps differentiates us from the limited prior literature on traits of gender minorities. For instance, in concurrent work, Coffman, Coffman and Ericson (2024) show that the direction and size of gaps between non-binary individuals and men and women vary across a wide range of contexts and traits.[7] In prior work, Fornwagner and Städter (2012) find little evidence in support of differences between cisgender and transgender participants in domains relating to competitiveness, risk-taking, and altruism. Unlike our paper, however, this other work neither directly explores nor investigates such uncertainty.

Second, the unexpected nature of the gender minority gaps may contribute to little attention toward interventions or policies to combat gender minority gaps in confidence and self-evaluations. For instance, unlike the widespread push for women to "lean-in," and initiatives that seek to encourage more confidence in women, there is no similar focus on gender minorities even though they are often economically disadvantaged.[8]

Third, the unexpected nature of the gender minority gaps in confidence and self-evaluations reflects one potentially unifying feature about gender minorities: individuals may be broadly *unsure or uncertain* about the traits of gender minorities, which is consistent with predictors also being unsure about the risk-taking, competitiveness, and generosity of gender minorities in our Prediction Study.

2 The Predictions Study

Sections 3 and 4 will document robust evidence of gender minority gaps in confidence and self-evaluations on a math and science test. Given this evidence, one could be tempted to conclude that these gender minority gaps are akin to gender differences between men and women in confidence and self-evaluations.

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

[7] While we choose to focus more broadly on "gender diverse" individuals, we note that we observe similar findings when narrowing in on non-binary individuals, which is very similar to our analyses shown in Panel B or Appendix B.2.

[8] For academic and related policy discussions on leaning in, see Exley, Niederle and Vesterlund (2020). See Demiral and Mollerstrom (2024) for the negative consequences of signaling excessive confidence.

2.1 Predictions Study Design

We recruited 600 participants to be "predictors" in our pre-registered Predictions Study.[9] Specifically, we recruited a sex-balanced sample of U.S. Prolific participants who did not participate in other studies discussed in this paper. Predictors are paid $3 if they complete the study, which involves answering 21 incentivized prediction questions, 15 broad belief questions, and a short follow-up questionnaire. In addition, one of the 21 prediction questions is chosen at random, and the predictor can earn an additional $1 if they answer that question correctly. The study proceeded as follows (further instructions and design details can be found in Online Appendix D.1).

First, we provide predictors with information about the *Adult Study* (we refer to this as the "prior study" for the predictors) which is discussed in Section 3. Then, we elicit incentivized beliefs about the performance of participants in the Adult Study and about the reported confidence of participants in the Adult Study. The order of these two types of beliefs is randomized at the predictor level.

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

In nine of the belief questions about performance, we ask predictors to consider the group of either female, male, or gender diverse participants who guessed that they answered either 5, 10, or 15 questions correctly on the math and science test (out of 20). We then ask them to predict how many questions, on average, these participants in that group actually answered correctly on the math and science test. Predictors indicate their answer on a slider (see, e.g., Appendix Figure D.4), and their answer is correct if the slider includes the true average.

In the other nine belief questions about performance, we ask predictors to consider the group of either female, male, or gender diverse participants who assigned their performance either a low rating (between 0 and 33), a medium rating (between 34 and 66), or a high rating (between 67 and 100) in response to the *performance self-evaluation* question (see Table 2 for a description of this self-evaluation question). We then ask them to predict how many questions, on average, participants in that group answered correctly on the math and science test. Predictors indicate their answer on a slider (see, e.g., Appendix Figure D.6), and their answer is correct if the range selected by the slider includes the truth.

For the belief questions about the reported confidence of participants in the Adult Study,

[9]The study was pre-registered on AsPredicted (#184073) which can be accessed here: https://aspredicted.org/y97t-22mj.pdf.

we directly ask predictors three questions about the reported confidence of each group of participants: female participants, male participants, and gender diverse participants. To begin, we inform predictors that a participant is overconfident if they overestimated how many questions they got right, accurate if they correctly guessed how many they got right, or underconfident if they underestimated how many questions they got right (see Figure D.7). We then ask predictors to guess whether a randomly selected participant—who is known to be female, male, or gender diverse—is either overconfident, accurate, or underconfident. We also give predictors the option to indicate if they are unsure. Predictors are told that if they choose the "I'm unsure" option and if that question is randomly chosen for payment, they will earn $1 with a 50% chance. An example decision screen is shown in Figure D.9. The order of the three questions is randomized at the predictor level.

To measure broad beliefs about five traits for each of the three gender identity groups, we also ask 15 unincentivized questions. Specifically, we ask predictors whether—in general—they believe female, male, and gender diverse people are likely:

1. to be overconfident in their performance and abilities in math and science tasks,
2. to be underconfident in their performance and abilities in math and science tasks,
3. to be competitive,
4. to take risks, and
5. to be generous.

In response to each broad belief questions, predictors can answer: "Yes," "No," or "I'm Not Sure". The order of the broad belief questions is randomized at the predictor level.

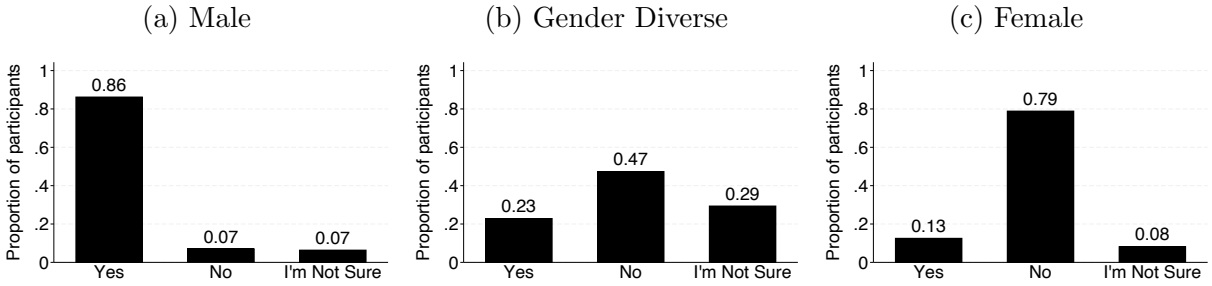
2.2 Predictions Study Results

For both our incentivized belief measures and our broad "in general" belief measures, we observe that predictors are more uncertain about the traits of gender diverse participants than about the traits of men or women.

The broad beliefs demonstrate this finding very clearly. Figure 1 shows that 29% of predictors indicate they are unsure about whether gender diverse are overconfident, which is 3 to 4 times larger than the 7–8% who are unsure about men or women. Figure 2 shows that 32% of predictors indicate they are unsure about whether gender diverse are underconfident, which is 3 to 4 times larger than the 7–9% who are unsure about men or women. Figure 3 shows 37% of predictors indicate they are unsure about whether gender diverse are competitive, which is 3 to 18 times larger than the 2–10% who are unsure about men or

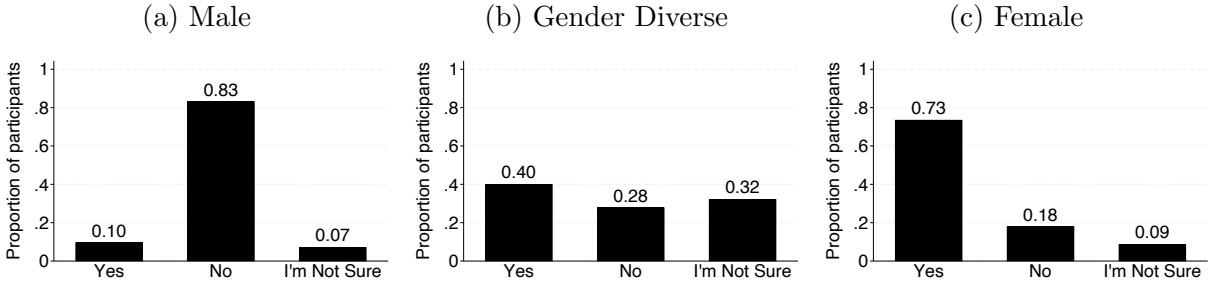
women. Figure 4 shows 29% of predictors indicate they are unsure about whether gender diverse people take risks, which is 2–9 times larger than the 3–12% who are unsure about women or men. Finally, Figure 5 shows 32% of predictors indicate they are unsure about whether gender diverse people are generous, which is 2–10 times larger than the 3–18% who are unsure about women or men. Instead of being unsure about the traits of men, predictors largely believe that men are overconfident, are not underconfident, are competitive, take risks, and are not generous. Instead of being unsure about the traits of women, predictors largely believe that women are not overconfident, are underconfident, do not take risks, are generous, and to a smaller extent, are competitive.[10]

Figure 1: Beliefs about Overconfidence



Participants are asked to guess whether—in general—[male]/[gender diverse]/[female] participants are overconfident in their performance and abilities in math and science tasks. Graphs show distributions of responses.

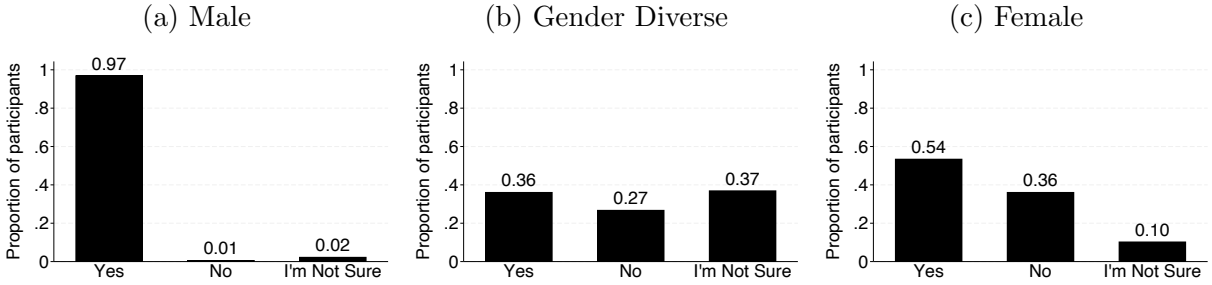
Figure 2: Beliefs about Underconfidence



Participants are asked to guess whether—in general—[male]/[gender diverse]/[female] participants are underconfident in their performance and abilities in math and science tasks. Graphs show distributions of responses.

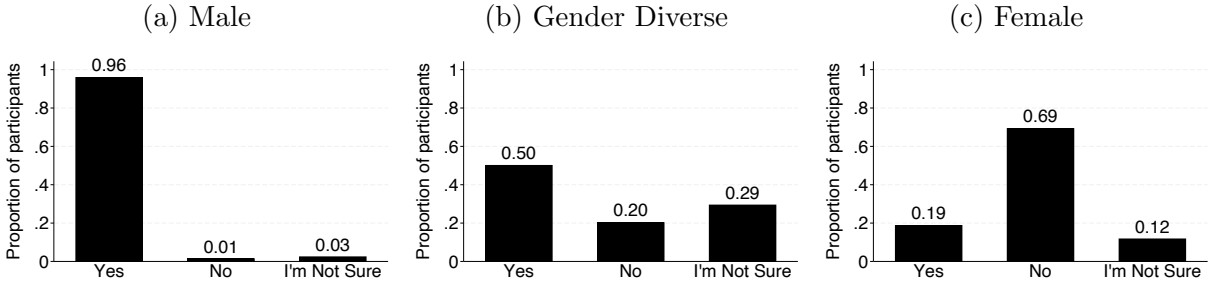
[10]That women are expected to be more generous than men echoes the robust believed gender differences in social preferences documented in Exley et al. (2024). For a meta-analysis on gender differences in generosity, see Bilén, Dreber and Johannesson (2021).

Figure 3: Beliefs about Competitiveness



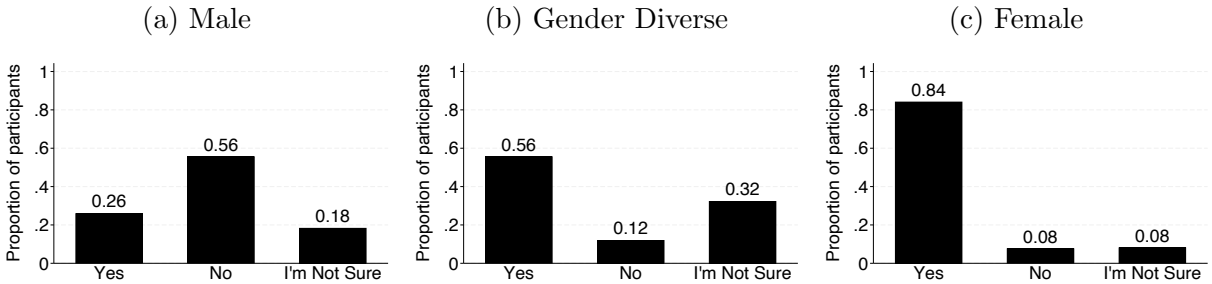
Participants are asked to guess whether—in general—[male]/[gender diverse]/[female] participants are competitive. Graphs show distributions of responses.

Figure 4: Beliefs about Taking Risks



Participants are asked to guess whether–in general—[male]/[gender diverse]/[female] participants are likely to take risks. Graphs show distributions of responses.

Figure 5: Beliefs about Generosity

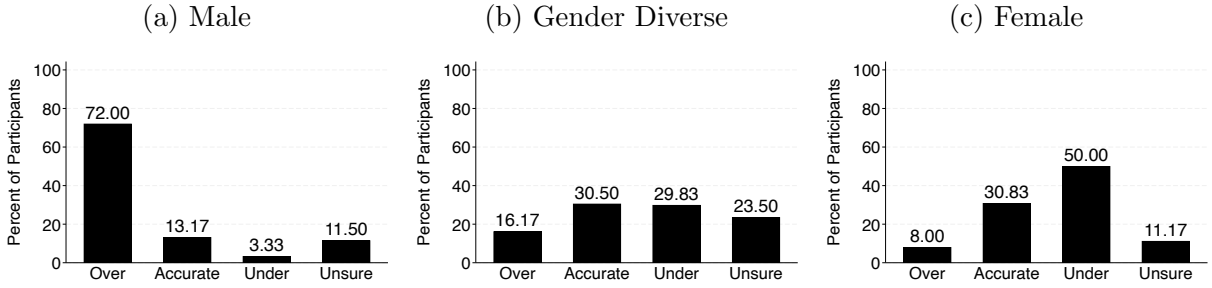


Participants are asked to guess whether–in general—[male]/[gender diverse]/[female] participants are generous. Graphs show distributions of responses.

We find similar results when turning to incentivized belief questions about the reported confidence of participants who take a math and science test. Specifically, Figure 6 shows whether predictors expect participants to be overconfident, accurate, underconfident, or whether predictors are instead "unsure" when asked about the confidence of male participants (see Panel A), gender diverse participants (see Panel B), and female participants (see Panel C). When asked about gender diverse participants, about one quarter of predictors (23.50%) answer "unsure." This is about twice as large as the 11–12% of predictors who answer "unsure" when asked about either male participants or about female participants ($p < 0.01$ for each comparison in a linear probability model with SEs clustered at the prediction level).

Even ignoring the unsure answers, the distribution of answers—across gender identities—also supports the gender minority gap being less expected. For instance, for men, Panel A shows that the rate of expected overconfidence (72%) is 22 times higher than the rate of expected underconfidence (3.3%). For women, Panel C shows that the rate of expected underconfidence (50%) is 6 times larger than the rate of expected overrconfidence (8%). By contrast, for gender minorities, the rate of expected underconfidence (30%) is less than 2 times larger than the rate of expected overconfidence (16%), suggesting more variability in beliefs about the confidence of gender minorities across the population of predictions.

Figure 6: Incentivized Beliefs about Reported Confidence



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

Results from the incentivized belief questions about the performance of prior participants who took a math and science test reinforce the unexpected nature of the gender minority gap in confidence. In particular, Appendix Table A.1 presents the average predicted performance when predictors are asked about male participants (see Column 1), gender diverse participants (see Column 2), and female participants (see Column 3). The first three rows show the average predicted performance (i.e., the number of questions predictors think participants got right) for participants who *self-report* that they got 5, 10 or 15 questions right.

The last three rows show the average predicted performance for participants who reported low (0–33), medium (34–66), or high (67–100) *self-evaluations*. Predictors expect higher performance among participants with higher reported confidence and higher self-evaluations.[11] But, these results also reveal that predictors—for a given level of reported confidence or self-evaluation—expect very similar performances from male participants (Column 1) and gender diverse participants (Column 2), on average.[12] Appendix Table A.2 indeed confirms that the predicted performance does not statistically significantly differ when predictors are asked about gender diverse versus male participants with the same reported confidence or same self-evaluations (see the close-to-zero and insignificant coefficient estimate on *Predicted Performance of Gender Diverse Participant*).[13] That is, evidence from this incentivized belief measure also aligns with predictors not expecting a gender minority gap.

3 The Adult Study

3.1 The Design of the Adult Study

Given these beliefs, we next study the actual confidence (and self-evaluations) of men, women, and gender diverse individuals in a math and science test. The *Adult Study* follows the design of Exley and Kessler (2022) and proceeds in six stages. Participants earn a fixed payment of $4 and have an opportunity to earn a bonus payment. Additional design details, including screenshots, can be found in Appendix D.2.

In the first stage, participants answer 20 math and science questions and are told they will receive 5 cents for each correct answer on the test if the first part of the study is chosen to determine bonus payments (otherwise they receive 25 cents as a bonus payment).[14] Each question appears on a separate page, and participants have 30 seconds to answer each question (see Appendix Figure D.11 for an example question).

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

[11] Predictors also expect that participants with lower (higher) self-reports are more likely to underestimate (overestimate) their performance.

[12] After being provided with such performance signals, predictors do not expect the same performances between female participants (Column 3) and male participants (Column 1). Rather, predictors expect that female participants have a better performance than male participants. While one could be tempted to conclude that this presents evidence in favor of predictors "accurately accounting for" the gender gap in confidence, we note that—absent knowing the full distribution of prior beliefs—it is difficult to calculate what predictors' beliefs should be if they accurately accounted for the gender minority gaps. In addition, as with the findings in Exley and Nielsen (2024), it could also be that predictors expect gender gaps but do not accurately account for them.

[13] See also Appendix Tables A.3 and A.4 for these results by specific self-reports and self-evaluations.

[14] These questions are from the Armed Services Vocational Aptitude Battery, but we make no mention of any relation to the Armed Services. Rather, these questions are accurately described as math and science questions (see Appendix Figure D.11).

us a measure of their confidence in their absolute performance.

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

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

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

In the sixth stage, we ask a demographic survey and adopt the gender question proposed by Miller and Willson (2022) and recommended as following best practices in 2022 for the collection of self-reported sexual orientation and gender identity data on Federal statistical surveys. Specifically, our gender question allows participants to choose all applicable options from the following: "Male," "Female," or "Transgender, non-binary, or another gender (see Figure D.18)."[15]

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

3.2 Gender Identity in the Adult Study

A total of 746 people participated in the *Adult Study* on Prolific during June and July of 2023. Since gender minorities constitute a relatively small share of the adult population in the U.S., we implemented a pre-registered stratified protocol to recruit a relatively large number of gender minorities from the Prolific platform.[16] In response to our gender survey question, 41.0% (n=306) selected only "Male," 36.3% (n=271) selected only "Female," and the remaining 22.7% (n=169) selected "Transgender, non-binary, or another gender" or multiple options, which leads us to classify them as gender diverse.[17]

3.3 The Gender Minority Gap in Confidence among Adults

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

To examine whether there is a gender minority gap in confidence between gender diverse participants and *equally-performing* male participants, we run regressions that control for performance. Table 1 presents regression results related to a participant's performance (i.e., the number of questions they got right on the test) and reported confidence (i.e., their belief about the number of questions they got right on the test). *Gender Diverse* and *Female* are indicators for gender diverse participants and female participants, respectively, while male participants are the excluded category in these regressions. Thus, the coefficient estimates on *Gender Diverse* compares gender diverse participants to male participants, and the coefficient estimate on *Female* compares female participants to male participants. At the bottom of the table, the coefficient estimate on *Difference* reports the difference between gender diverse and female participants along with its corresponding p-value for a two-sided t-test of the difference in these coefficient estimates.

Column (1) of Table 1 confirms that gender diverse participants perform significantly better than male and female participants, while Columns (2)–(4) show outcomes related to

[16]This recruitment procedure was pre-registered on AsPredicted (#136119) which can be accessed here: https://aspredicted.org/2FW_Z5H. We ran the Adult Study after the Student Study to see if our results replicated with an adult sample. Further details about our recruitment procedure is discussed in the Online Appendix D.2.

[17]Specifically, 122 participants only selected "Transgender, non-binary, or another gender," 27 participants selected "Transgender, non-binary, or another gender" and "Male," 19 participants selected "Transgender, non-binary, or another gender" and "Female," and 1 participant selected "Male" and "Female."

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

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

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

the gender minority gaps in confidence (as well as the gender gaps between men and women in confidence).

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

1.75 fewer questions correctly than equally performing male participants. The coefficient estimate on *Difference* and corresponding p-value then confirm that gender minority gap is approximately as large as the gender gap between men and women.

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

Column (4) presents similar results when instead considering a binary measure of a participant's confidence, specifically, whether a participant is underconfident (i.e., whether their reported confidence falls below their actual performance). The coefficient estimate on *Gender Diverse* reveals the gender minority gap in underconfidence: gender diverse participants are 20 percentage points more likely to be underconfident than male participants. The coefficient estimate on *Female* reveals the gender gap in underconfidence: female participants are 12 percentage points more likely to be underconfident than male participants. The coefficient estimate on *Difference* and corresponding p-value further shows that the gender minority gap in underconfidence is somewhat larger than the gender gap in underconfidence ($p = 0.06$).[18]

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

3.4 The Gender Minority Gaps in Self-Evaluations Among Adults

To further investigate the robustness of the gender gap in confidence—broadly construed—Table 2 presents the regression results of participants' subjective self-evaluations. Each

[18] Appendix Table A.5 shows that gender diverse participants are less likely to be overconfident (see Column 1) and that there is no differences in the likelihood of them being accurate (see Column 2).

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

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

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

regression includes performance fixed effects. Thus, the coefficient estimates on *Gender Diverse* compares gender diverse participants to equally-performing male participants, and

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

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

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

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

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

Appendix Figure A.2 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.

Since gender diverse participants believe they answered fewer questions correctly on the test, one may wonder whether the gender minority gaps in self-evaluations persist even when

we compare participants who answered the same number of questions correctly on the test and *know* how many questions they answered correctly on the test. To investigate this, we tell participants exactly how many questions they answered correctly on the test (and then have them report this number back to us to confirm they actually saw it). We then ask them the same four self-evaluation questions to elicit informed self-evaluations.

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

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

Appendix Figure A.3 shows CDFs of the responses to each of the four informed self-evaluation questions. As mentioned earlier, differences in the distributions of responses may be harder to interpret, however, because—unlike the regressions—they do not account for underlying differences in performance between the groups.

Taken together, in the Adult Study, we observe robust evidence—across the reported confidence question and several self-evaluation questions—of gender minority gaps between gender diverse participants and equally-performing men. In addition, these gender minority gaps are just as large as the gender gaps between women and equally-performing men. This is despite the fact that, as previously seen in the Predictions Study, the gender minority gap is largely unexpected while the gender gap is very much expected.

4 The Student Study

As further evidence of the robustness of the gender minority gap, results from the *Student Study* show that the gender minority gaps also arise with a younger population.

4.1 The Design of the Student Study

The *Student Study* was conducted in the fall semester of 2020 with the partnership of the Character Lab Research Network (CLRN), which helped us recruit 10,807 students in grades 6–12 from a large school district in the United States. The students agreed to participate in a short study during the school day.[19] Some of the student data we analyze here was also

[19]The following text from the CLRN website explains the data collection process in more detail: "This investigation was part of a larger data collection effort that included a variety of studies designed by sci-

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.[20] The Exley and Kessler (2022) analysis of the student data, however, exploits administrative data identifying every student as either male or female. In this paper, we instead explore students' self-reported gender to generate new results on gender diverse individuals. Furthermore, for this paper, we use supplementary data on academic performance from our student sample during the academic quarter in which they participated in our study and the next seven quarters. This supplementary data allows us to document a correlation between our confidence and self-evaluation measures and student GPAs, as shown in Appendix Section B.1.[21]

Like the *Adult Study*, the *Student Study* also had six stages, but it had some minor differences. First, students were asked to answer 10 (instead of 20) math and science questions from the Armed Services Vocational Aptitude Battery. We requested that students try their best when answering, but there were no financial incentives in the study.[22] Second, we elicited each student's belief about their absolute performance and their informed and uninformed self-evaluations similar to the *Adult Study* with two exceptions. In the *willingness* question, students were asked to indicate their agreement with "If given an option, I would choose to take a class that involves topics like those covered on the test." In the *success* question, students were asked to indicate their agreement with "I would succeed in a class that involves topics like those covered on the test." Third, again similar to the *Adult Study*,

entists 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).

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

[21]This additional data collection also allowed us to validate more survey responses—which was done by matching unique identifiers in our data with the unique identifiers in the CRLN's data—resulting in a slightly larger sample size than Exley and Kessler (2022).

[22]That the gender gaps among adults in Exley and Kessler (2022) are roughly identical with and without incentives to self-promote helps to mitigate potential concerns about the lack of monetary incentives in the Student Study.

we asked students to complete a short follow-up survey at the end to gather demographic information, including a question about their gender where they select "male," "female," or "other." If they selected other, they could choose to provide free response text about their gender identity. As explained in greater detail below (see Section 4.2), we use these responses to classify students by gender and to identify the students who are gender diverse. Figure D.27 shows specifically how we ask students to self-report their gender.

4.2 Gender Identity in the Student Study

A total of 10,807 students completed the *Student Study* in Fall 2020. Of these students, 48% selected male (n=5,187), 50% selected female (n=5,412), and 2% selected other (n=208) when asked about their gender.[23] 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).[24] However, as detailed in Section 4.4.1, our results are robust to alternative classifications.

4.3 Gender Minority Gaps in Confidence in the Student Study

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

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

Column (1) of Table 3 confirms that gender diverse students perform similarly to male students, and Columns (2)–(4) show that gender diverse students, nonetheless, have lower confidence. When compared to equally-performing male students, Column (2) shows that

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

[24]Some transgender students might not have chosen their gender identity as "other" and so would not be included in our definition of gender diverse. Since we do not have data on sex assigned at birth, we cannot identify if such individuals are present in our student data.

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

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

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

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

4.4 Gender Minority Gaps in Self-Evaluations in the Student Study

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

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

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

The results about informed self-evaluations are presented in Panel B of Table 4. Even after the students are told how many questions they answered right, we again observe significant and substantial gender minority gaps in response to all four self-evaluation questions. In addition, as with uniformed self-evaluations, these gender minority gaps are larger than the gender gaps between male and female students.[25]

[25] See Appendix Figures B.2 and B.3 for CDFs of the responses to each of the four uninformed and informed self-evaluation questions. But, we note that differences in the distributions of responses may be harder to interpret because—unlike the regressions—they do not account for underlying differences in performance between the groups.

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

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

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

4.4.1 Robustness in the Student Study

The gender minority gaps in confidence and in self-evaluation that we identify are robust to different ways of classifying students as gender diverse. Appendix Table B.2 describes four sets of robustness checks that we ran with our data, which we summarize here.

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

In the second set of robustness tests, we classify the 74 students who selected other and provided details on the nature of their gender identity in the corresponding free response text box as *explicitly gender diverse.*[26] We then only keep these explicitly gender diverse students and drop everyone else who selected other (i.e., we drop students who provided offensive responses, those who left the text box blank, and those who did not provide an informative response about their gender identity).

In the third and fourth set of robustness tests, we rely on gender data collected by the Character Lab Research Network (CLRN) in a demographics survey that was run before our study. Using that survey, we classify participants as male, as female, as those who selected "Other" when asked about their gender, and as those who selected "Prefer not to say" when asked about their gender. In the third set of robustness tests, we use the CLRN survey for gender classification, dropping the 535 students who selected "Prefer not to say." In the fourth set of robustness tests, we primarily use the CLRN survey for gender classification and use responses to our survey question only to classify those who selected "Prefer not to say" (for more details, see Appendix Table B.2).

Appendix Tables B.3–B.5 replicate the analysis conducted in Tables 3 and 4, showing results for each of the four sets of robustness tests. The results identified in Sections 4.3 and 4.4 are highly robust. All 40 of the differences in confidence and self evaluations that we estimate between students who we classify as gender diverse and students we classify as male are statistically significant at $p < 0.05$ (with 39 significant at $p < 0.01$). Across all specifications, these gender minority gaps are large and, in most cases, also larger than the corresponding gender gaps we see when we compare responses of male and female students.[27]

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

[27] In particular, 35 out of 40 of the differences in confidence and self evaluations that we estimate between students who we classify as gender diverse and students we classify as female are statistically significant at $p < 0.05$ (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).

5 The Adult (Verbal) Study

The results so far focus on a math and science test—and hence a male-stereotyped domain—and make clear that the gender minority gap in confidence is robust but not expected. This contrasts with the gender gap between men and women, which is both robust and expected.

As observed in prior literature on the gender gap in confidence, however, one may naturally expect evidence for gender minority gaps to depend on the domain studied. Thus, motivated by the literature that shows that gender differences between men and women can be domain specific (Günther et al., 2010; Shurchkov, 2012; Coffman, 2014; Dreber, von Essen and Ranehill, 2014; Bordalo et al., 2019; Boschini et al., 2019; Coffman, Collis and Kulkarni, 2019; Coffman, Flikkema and Shurchkov, 2019; Atwater and Saygin, 2020), we randomized half of our adult participants to the Adult (Verbal). This study is very similar to the Adult Study with one difference: the test domain. In this the Adult (Verbal) Study, we asked participants 20 word knowledge questions rather than a math and science test, and participants had 15 seconds to provide each answer (see Appendix Figure D.20 for an example question). Everything else is identical to the Adult Study.

A total of 748 participants completed the *Verbal* version run on Prolific in June and July of 2023, using the same pre-registered stratified protocol as the Adult Study.[28] 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.[29]

When we switch from a math and science domain to a verbal test, which is typically considered less male-stereotyped, the gender minority gaps shrink dramatically and almost all go away. Following the structure of Tables 1 and 2, Appendix Tables C.1 and C.2 reveal that little to no evidence for a gender minority gap in confidence and in self-evaluations between gender diverse participants and equally-performing men.[30]

Thus, these results show that, as observed via prior literature on the gender gap in confidence, the gender minority gap in confidence is context dependent. We emphasize this

[28]This recruitment procedure was pre-registered on AsPredicted (#136119), which can be accessed here: https://aspredicted.org/2FW_Z5H. As discussed in this pre-registration, participants were randomized into taking either the math and science test or the verbal test.

[29]Specifically, 110 participants only selected "Transgender, non-binary, or another gender," 22 participants selected "Transgender, non-binary, or another gender" and "Male," 27 participants selected "Transgender, non-binary, or another gender" and "Female," and 2 participants selected "Male" and "Female."

[30]In particular, out of the 10 measures of confidence and self-evaluations, we only observe a statistically significant gender minority gap on measure: see Column (3) of Appendix Table C.1. By contrast, we observe some evidence for a gender gap between men and women—albeit a mostly smaller gender gap than what was observed in the math and science test. Appendix Figure C.1 shows the CDFs of these differences between beliefs and performance for the three groups.

point because, we think, one should likely expect most, if not all, gender differences to be context dependent. This is, of course, important to keep in mind for this paper and for the gender literature more broadly.

Our main domain—the math and science domain—is relevant for many educational and career settings, including high-paying occupations in which gender differences in pay and representation are observed. However, we also chose to explore this context to complement prior gender literature that has similarly focused on domains that are more or less male-stereotyped. Ideally, one might have explored a domain considered stereotypical for gender minorities. Given the lack of literature on beliefs about gender minorities, however, there are no such established stereotypical domains. We believe investigating more domains for gender minorities is an important avenue for future work; such work may indeed reveal that—consistent with our finding on uncertainty about the traits of gender minorities—there are *no* stereotypical domains for gender minorities.

6 Discussion

We find that that participants are uncertain about the traits of gender minorities while holding more certain beliefs about men and women. These findings arise in response to broad belief questions and, in the case of confidence, in response to incentivized questions about the reported confidence of gender minorities in a math and science test.

While participants are unsure about the confidence of gender minorities in this domain, we observe a robust gender minority gap in confidence: gender diverse participants report lower confidence and worse self-evaluations than equally performing men. The gender minority gap is as big or bigger than the gender gap in confidence between men and women. But unlike the gender minority gap, the gender gap in confidence is very much expected.

The results in our paper open up many important avenues for future work. For instance, building upon the results in our Predictions Study, future work may investigate whether people's attitudes towards gender minorities are driven by uncertainty about their traits and behaviors across contexts. Future work may examine if such uncertainty contributes to policy roadblocks, discrimination, inattention, and a lack of understanding.[31]

We also hope future work explores diversity among gender minorities. For example, future work may seek to separately study those who identify as transgender men, transgender women, non-binary individuals, genderqueer individuals, or gender non-conforming individuals. Future work may also aim to study other minority groups, such as those related to

[31]Related, see prior work on the connection between gender norms and many important outcomes (Bertrand, Kamenica and Pan, 2015; Dhar, Jain and Jayachandran, 2022; Field et al., 2021; Pande and Roy, 2021; Jayachandran et al., 2023) as well as prior work on inaccurate gender beliefs and misperceptions of gender norms (Bordalo et al., 2019; Bursztyn, González and Yanagizawa-Drott, 2020; Coffman, Exley and Niederle, 2021; Bohren et al., 2023; Bursztyn, Cappelen and Tungodden, 2023).

sexual orientation (e.g., see Coffman, Coffman and Ericson, 2017; Buser, Geijtenbeek and Plug, 2018; Lewis et al., 2017; Aksoy, Carpenter and Sansone, 2025; Aksoy and Chadd, 2025), and the impact of intersectionality, such as the impact of being a gender minority and a sexual minority.[32]

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

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

References

Aksoy, Billur, and Ian Chadd. 2025. "Competitiveness at the Intersection of Gender and Sexual Orientation." *Journal of Economic Behavior & Organization*, 233: 106987.

Aksoy, Billur, Christopher S. Carpenter, and Dario Sansone. 2025. "Understanding Labor Market Discrimination Against Transgender People: Evidence from a Double List Experiment and a Survey." *Management Science*, 71(1): 659–677.

Aksoy, Billur, Ian Chadd, and Boon Han Koh. 2023. "Sexual Identity, Gender, and Anticipated Discrimination in Prosocial Behavior." *European Economic Review*, 154: 104427.

Aksoy, Billur, Ian Chadd, Brit Grosskopf, and Boon Han Koh. 2024. "Sexual Orientation and Gender Identity." *Available at SSRN*.

Atwater, Ann, and Perihan O. Saygin. 2020. "Gender Differences in Leaving Questions Blank on High-Stakes Standardized Test." *Working Paper*.

Babcock, Linda, and Sara Laschever. 2003. *Women don't ask: negotiation and the gender divide.* Princeton, NJ:Princeton University Press.

Badgett, M.V. Lee, Christopher S. Carpenter, and Dario Sansone. 2021. "LGBTQ economics." *Journal of Economic Perspectives*, 35(2): 141–70.

Badgett, M.V. Lee, Christopher S. Carpenter, Maxine J. Lee, and Dario Sansone. 2023. "A Review of the Economics of Sexual Orientation and Gender Identity." *Journal of Economic Literature*, Forthcoming.

Barber, Brad M., and Terrance Odean. 2001. "Boys will be boys: Gender, overconfidence, and common stock investment." *The quarterly journal of economics*, 116(1): 261–292.

Bertrand, Marianne, Claudia Goldin, and Lawrence F Katz. 2010. "Dynamics of the gender gap for young professionals in the financial and corporate sectors." *American economic journal: applied economics*, 2(3): 228–255.

Bertrand, Marianne, Emir Kamenica, and Jessica Pan. 2015. "Gender identity and relative income within households." *The Quarterly Journal of Economics*, 130(2): 571–614.

Bilén, David, Anna Dreber, and Magnus Johannesson. 2021. "Are women more generous than men? A meta-analysis." *Journal of the Economic Science Association*, 7(1): 1–18.

Blau, Francine D., and Lawrence M. Kahn. 2017. "The Gender Wage Gap: Extent, Trends. and Explanations." *Journal of Economic Literature*, 55(3).

Bohren, J. Aislinn, Kareem Haggag, Alex Imas, and Devin G. Pope. 2023. "Inaccurate statistical discrimination: An identification problem." *Review of Economics and Statistics*, 1–45.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2019. "Beliefs about Gender." *American Economic Review*.

Boschini, Anne, Anna Dreber, Emma Von Essen, Astri Muren, and Eva Ranehill. 2019. "Gender, risk preferences and willingness to compete in a random sample of the Swedish population." *Journal of Behavioral and Experimental Economics*, 83: 101467.

Brenøe, Anne Ardila, Lea Heursen, Eva Ranehill, and Roberto A Weber. 2022. "Continuous gender identity and economics." *AEA Papers and Proceedings*, 112: 573–577.

Brown, Anna. 2022. "About 5% of young adults in the U.S. say their gender is different from their sex assigned at birth." *Pew Research Center*.

Bursztyn, Leonardo, Alessandra L González, and David Yanagizawa-Drott. 2020. "Misperceived social norms: Women working outside the home in Saudi Arabia." *American Economic Review*, 110(10): 2997–3029.

Bursztyn, Leonardo, Alexander W Cappelen, and Bertil Tungodden. 2023. "How are gender norms perceived?" *National Bureau of Economic Research*.

Bursztyn, Leonardo, Thomas Fujiwara, and Amanda Pallais. 2017. "'Acting Wife': Marriage Market Incentives and Labor Market Investments." *The American Economic Review*, 107(11): 3288–3319.

Buser, Thomas, Lydia Geijtenbeek, and Erik Plug. 2018. "Sexual orientation, competitiveness and income." *Journal of Economic Behavior & Organization*, 151: 191–198.

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek. 2014. "Gender, competitiveness, and career choices." *The quarterly journal of economics*, 129(3): 1409–1447.

Bütikofer, Aline, Katrine V. Løken, and Alexander Willén. 2022. "Building Bridges and Widening Gaps." *Review of Economics and Statistics*.

Carpenter, Christopher S., Maxine J. Lee, and Laura Nettuno. 2022. "Economic outcomes for transgender people and other gender minorities in the United States: First estimates from a nationally representative sample." *Southern Economic Journal*, 89(2): 280–304.

Carpenter, Christopher S., Samuel T. Eppink, and Gilbert Gonzales. 2020. "Transgender status, gender identity, and socioeconomic outcomes in the United States." *ILR Review*, 73(3): 573–599.

Coffman, Katherine Baldiga. 2014. "Evidence on Self-Stereotyping and the Contribution of Ideas." *The Quarterly Journal of Economics*, 129(4): 1625–1660.

Coffman, Katherine B., Christine L. Exley, and Muriel Niederle. 2021. "The Role of Beliefs in Driving Gender Discrimination." *Management Science*, 67(6): 3321–3984.

Coffman, Katherine B, Lucas C Coffman, and Keith Marzilli Ericson. 2024. "Non-Binary Gender Economics." National Bureau of Economic Research.

Coffman, Katherine B, Lucas C Coffman, and Keith M Marzilli Ericson. 2017. "The size of the LGBT population and the magnitude of antigay sentiment are substantially underestimated." *Management Science*, 63(10): 3168–3186.

Coffman, Katherine, Clio Bryant Flikkema, and Olga Shurchkov. 2019. "Gender Stereotypes in Deliberation and Team Decisions." *Harvard Business School Working Paper*.

Coffman, Katherine, Manuela Collis, and Leena Kulkarni. 2019. "Stereotypes and Belief Updating." *Working Paper*.

Croson, Rachel, and Uri Gneezy. 2009. "Gender Differences in Preferences." *Journal of Economic Literature*, 47(2): 448–474.

Demiral, Elif E, and Johanna Mollerstrom. 2024. "Signaling confidence." *Journal of Economic Behavior & Organization*, 226: 106691.

Dhar, Diva, Tarun Jain, and Seema Jayachandran. 2022. "Reshaping adolescents' gender attitudes: Evidence from a school-based experiment in India." *American economic review*, 112(3): 899–927.

Downing, Janelle M., and Julia M. Przedworski. 2018. "Health of transgender adults in the US, 2014–2016." *American Journal of Preventive Medicine*, 55(3): 336–344.

Dreber, Anna, Emma von Essen, and Eva Ranehill. 2014. "Gender and competition in adolescence: task matters." *Experimental Economics*, 17(1): 154–172.

Eckel, Catherine C., and Philip J. Grossman. 2008. *Men, women and risk aversion: experimental evidence, [in:] Handbook of Experimental Economics Results.* Vol. 1, Amsterdam, Oxford, North Holland.

Eckel, Catherine, Lata Gangadharan, Philip J. Grossman, and Nina Xue. 2021. "A research agenda for experimental economics." Chapter The gender leadership gap: Insights from experiment, 137–162. Edward Elgar Publishing.

Exley, Christine L., and Judd B. Kessler. 2022. "The Gender Gap in Self-Promotion." *Quarterly Journal of Economics*.

Exley, Christine L., and Kirby Nielsen. 2024. "The Gender Gap in Confidence: Expected But Not Accounted For." *American Economic Review*, 14(3): 851–885.

Exley, Christine L., Muriel Niederle, and Lise Vesterlund. 2020. "Knowing When to Ask: The Cost of Leaning-in." *Journal of Political Economy*, 128(3): 816–854.

Exley, Christine L, Oliver P Hauser, Molly Moore, and John-Henry Pezzuto. 2024. "Believed gender differences in social preferences." *The Quarterly Journal of Economics*, 140(1): 403–458.

Field, Erica, Rohini Pande, Natalia Rigol, Simone Schaner, and Charity Troyer Moore. 2021. "On her own account: How strengthening women's financial control impacts labor supply and gender norms." *American Economic Review*, 111(7): 2342–75.

Fornwagner, Helena, Brit Grosskopf Alexander Lauf Vanessa Schöller, and Silvio Städter. 2012. "On the robustness of gender differences in economic behavior." *Scientific Reports*, 12: 21549.

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 Ekinci, 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 Iriberri. 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%."

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.

Meyer, Ilan H., Taylor N.T. Brown, Jody L. Herman, Sari L. Reisner, and Walter O. Bockting. 2017. "Demographic characteristics and health status of transgender adults in select US regions: Behavioral Risk Factor Surveillance System, 2014." *American journal of public health*, 107(4): 582–589.

Miller, Kristen, and Stephanie Willson. 2022. "Development and Evaluation of a Single, Non-Binary Gender Question for Population-Based Federal Health Surveys." Hyattsville, MD: National Center for Health Statistics - QDRL.

Niederle, Muriel. 2016. "Gender." In *Handbook of Experimental Economics*. Vol. 2, , ed. John Kagel and Alvin E. Roth, 481–553. Princeton University Press.

Niederle, Muriel, and Lise Vesterlund. 2007. "Do Women shy away from competition? Do men compete too much?" *Quarterly Journal of Economics*, 122(3): 1067–1101.

Niederle, Muriel, and Lise Vesterlund. 2011. "Gender and Competition." *Annual Review of Economics*, 3: 601–630.

Pande, Rohini, and Helena Roy. 2021. ""If you compete with us, we shan't marry you" The (Mary Paley and) Alfred Marshall Lecture." *Journal of the European Economic Association*, 19(6): 2992–3024.

Piasenti, Stefano, and Müge Süer. 2024. "Predictive Power of Biological Sex and Gender Identity on Economic Behavior." *CRC TRR 190 Rationality and Competition Discussion Paper No. 513*.

Pope, Devin, and Justin Sydnor. 2010. "A new perspective on stereotypical gender differences in test scores." *Journal of Economic Perspectives*, 24(95).

Reuben, Ernesto, Matthew Wiswall, and Basit Zafar. 2017. "Preferences and biases in educational choices and labour market expectations: Shrinking the black box of gender." *The Economic Journal*, 127(604): 2153–2186.

Sansone, Dario. 2019. "LGBT students: New evidence on demographics and educational outcomes." *Economics of Education Review*, 73(101933).

Shurchkov, Olga. 2012. "Under pressure: gender differences in output quality and quantity under competition and time constraints." *Journal of the European Economic Association*, 10(5): 1189–1213.

Appendices (For Online Publication Only)

A Additional Results from the Adult Study and the Predictions Study

Table A.1: In the *Predictions Study*, Average Predicted Performances

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

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

Table A.2: In the In the *Predictions Study*, OLS of Predicted Performance Given Prior Participant's Reported Confidence and Self-Evaluation

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

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

Table A.3: In the *Predictions Study*, OLS of Predicted Performance Given Prior Participant's Reported Confidence

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

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the individual level are in parentheses. This table presents data from the *Predictions Study*. Results are from OLS regressions of the following dependent variables: the first column reports predicted performances of [female]/[male]/[gender diverse] prior participants who guessed that they answered 5 questions (out of 20) correctly on the math and science test; the second column reports predicted performances of [female]/[male]/[gender diverse] prior participants who guessed that they answered 10 questions (out of 20) correctly on the math and science test; and the third column reports predicted performances of [female]/[male]/[gender diverse] prior participants who guessed that they answered 15 questions (out of 20) correctly on the math and science test. See A.2 for definitions of independent variables, *Difference*, and *p-value*.

Table A.4: In the *Predictions Study*, OLS of Predicted Performance Given Prior Participant's Self-Evaluations

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

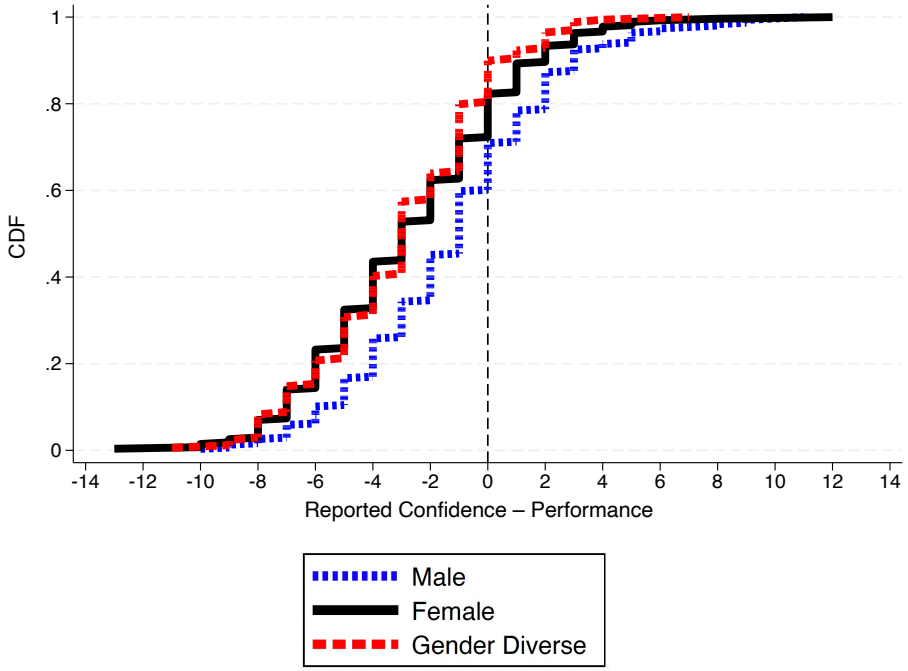
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the individual level are in parentheses. This table presents data from the *Predictions Study*. Results are from OLS regressions of the following dependent variables: the first column reports predicted performances of [female]/[male]/[gender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

Table A.5: In the *Adult Study*, OLS of Reported Confidence Relative to Truth Outcomes (overconfident, accurate, or overconfident)

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

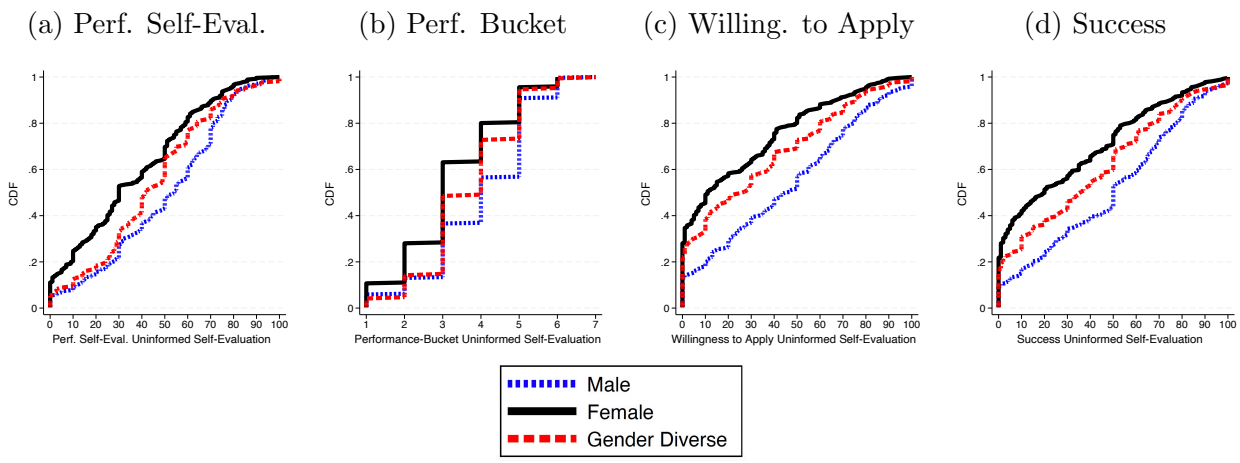
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and clustered at the participant level. This table presents data from the *Adult Study*. Results are from OLS regressions of the dependent variable noted in the column. *Overconfident/Accurate/Underconfident* is a binary variable that takes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

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



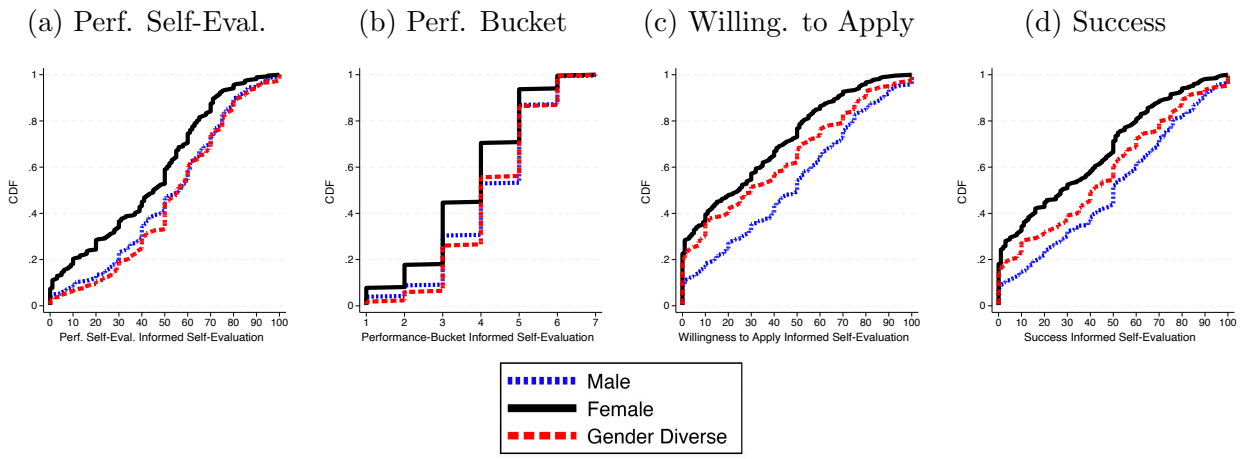
Graph shows CDFs for *Reported Confidence–Performance*, the number of questions a participant believes they answered correctly minus the number of questions a participant answered correctly. Positive responses suggest overconfidence while negative numbers suggest underconfidence.

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



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

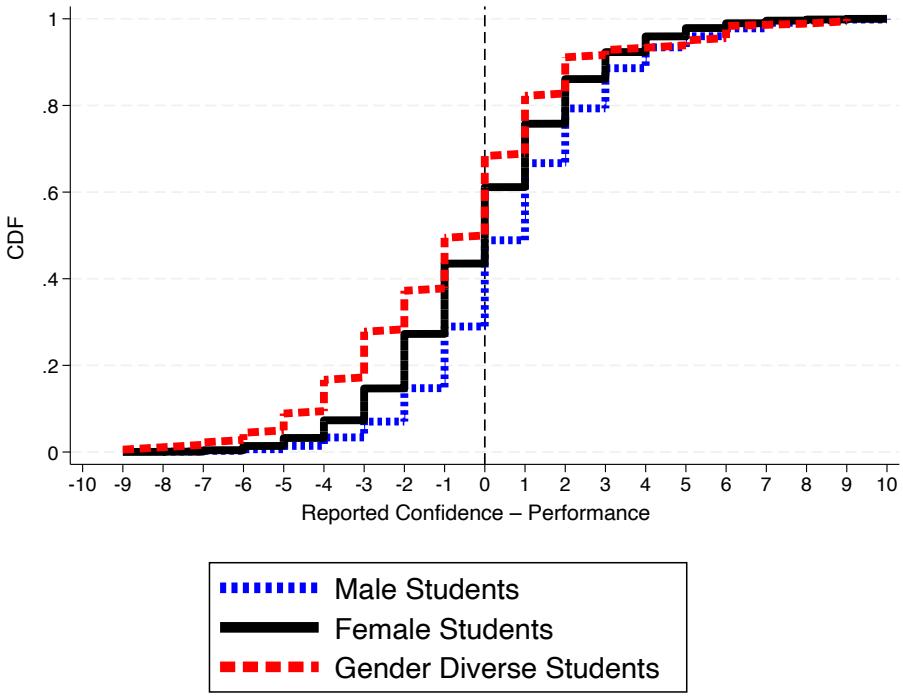
Figure A.3: In the *Adult Study*, CDFs for Informed Self-Evaluations



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

B Additional Results from the Student Study

Figure B.1: In the *Student Study*, Reported Confidence–Performance Distributions



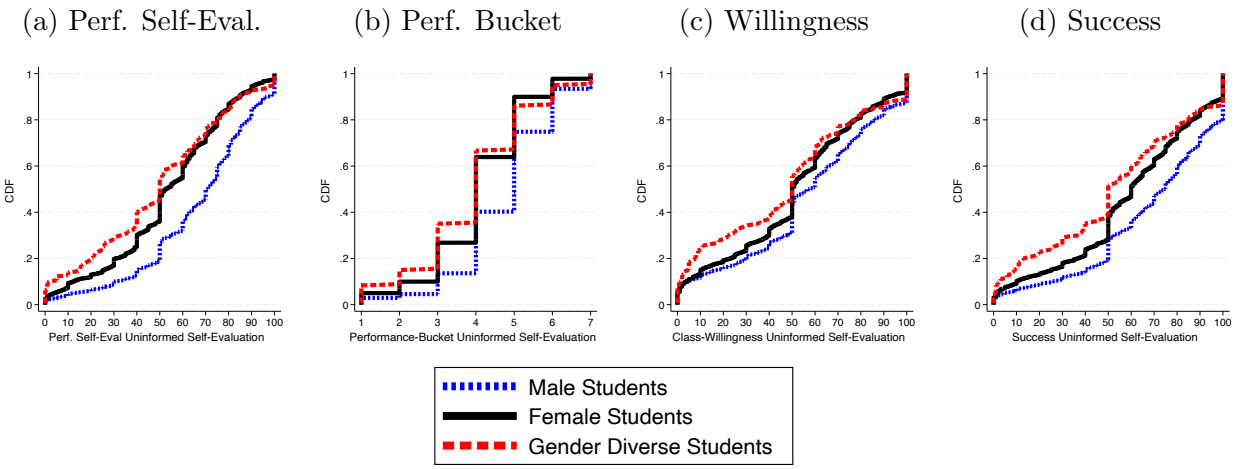
Graph shows CDFs for *Reported Confidence–Performance*, the number of questions a participant believes they answered correctly minus the number of questions a participant answered correctly. Positive responses suggest overconfidence while negative numbers suggest underconfidence.

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

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

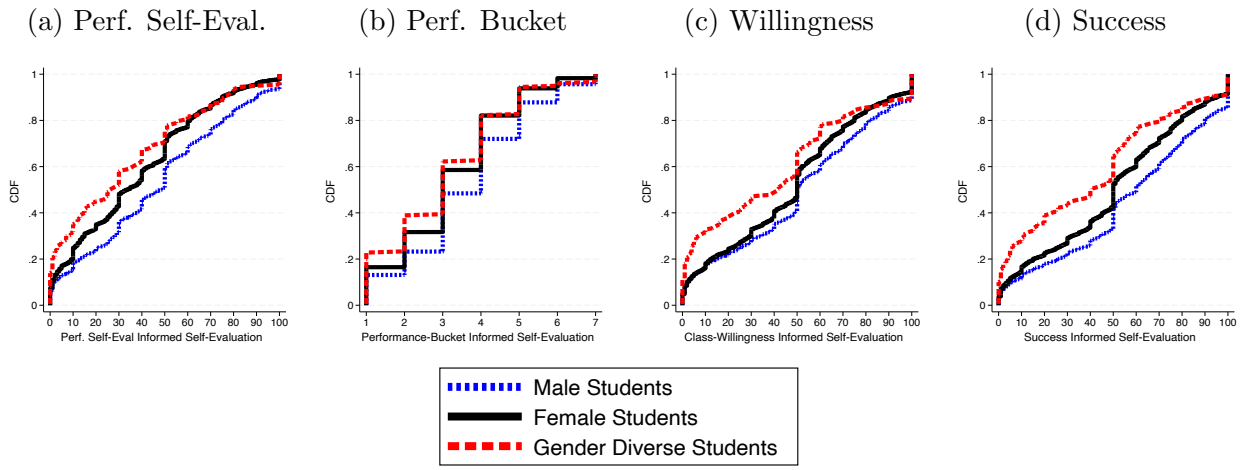
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and clustered at the participant level. This table presents data from the *Student Study*. Results are from OLS regressions of a student's level of confidence as noted in the column. *Overconfident*/*Accurate*/*Underconfident* is a binary variable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

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



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

Figure B.3: In the *Student Study*, CDFs for Informed Self-Evaluations



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

Table B.2: Sample and Variable Descriptions for Robustness Checks of the *Student Study*

Panel	Notes
Panel A:	These results rely on the gender data from our survey. *Female* is an indicator for a student selecting "Female." *Gender Diverse* 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. *Female* is an indicator for a student selecting "Female." *Explicitly Gender Diverse* is an indicator for a student who we classify as explicitly gender diverse. Panel B excludes both the 28 students who provided an offensive response and also excludes the 106 students who selected other but provided either no response or a response that was not specific enough for us to classify them as explicitly gender diverse. Panel B thus includes data on 10,673 students.
Panel C:	These results rely on the gender data from the Character Lab Research Network (CLRN) survey. Panel C excludes the 535 students who selected "Prefer not to say" when asked about their gender. *Female* is an indicator for female students (50.49% or 5,186) and *Gender Diverse* 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, *Female* 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). *Gender Diverse* is an indicator for the students selecting "Other" when asked about their gender in the CLRN survey (151) or selecting "Prefer not to say" in the CLRN survey but choosing "Other" in our survey (14). Panel D thus includes data on all 10,807 students.

This table includes information about the variables and each of the samples used in Panels A–D in Tables B.3–B.5.

Table B.3: In the *Student Study*, Performance Beliefs with Alternative Gender Classifications

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

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and clustered at the participant level. This table presents data from the *Student Study*. Results are from OLS regressions of the dependent variable noted in the column. See Table 3 and Appendix Table B.2 for more information about samples, dependent variables, and independent variables used in each panel as well as FEs.

Table B.4: In the *Student Study*, Uninformed Self-Evaluations with Alternative Gender Classifications

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

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and clustered at the participant level. This table presents data from the *Student Study*. Results are from OLS regressions of a student's response to the uninformed self-evaluation (elicited before the student learns their test performance) noted in the column. See Table 4 and Appendix Table B.2 for more information about samples, dependent variables, and independent variables used in each panel as well as FEs.

Table B.5: In the *Student Study*, Informed Self-Evaluations with Alternative Gender Classifications

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

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust and clustered at the participant level. This table presents data from the *Student Study*. Results are from OLS regressions of a student's response to the uninformed self-evaluation (elicited before the student learns their test performance) noted in the column. See Table 4 and Appendix Table B.2 for more information about samples, dependent variables, and independent variables used in each panel as well as FEs.

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

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

Specifically, Table B.6 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).[33] 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 B.7 follows Table B.6 but shows regressions of Math GPA in each quarter, rather than overall GPA. Results

[33] Not all students in our data have overall GPAs in the administrative data. Additionally, the number of students with GPAs decreases over time (e.g., as students graduate or otherwise leave the school district).

are qualitatively very similar.[34]

Taken together, we thus find that students who are more confident and students 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.

[34]While not the focus of this paper, our data also allow us to directly compare the academic performance of gender diverse students to the academic performance of students who identify as male and who identify as female. Appendix Table B.8 does these comparisons and shows that—considering overall GPA in Panel A or just Math GPA in Panel B—gender diverse students typically perform worse than both male and female students in the academic year our study was run. Looking at the later quarters (i.e., the year after our study was run), gender diverse students continue to underperform students who identify as female but their performance is not statistically distinguishable from students who identify as male. Given the rich literature exploring differences between men and women in test scores and other academic outcomes (e.g., see discussions in Pope and Sydnor (2010) and Niederle and Vesterlund (2011)), an important avenue for future work is to also consider the academic performance of gender diverse students.

Table B.6: 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)								
X	0.238***	0.261***	0.258***	0.246***	0.180***	0.217***	0.209***	0.175***
	(0.046)	(0.047)	(0.047)	(0.048)	(0.046)	(0.048)	(0.051)	(0.053)
Panel B: X = Uninformed Performance Self-Evaluations (0–100)								
X	0.028***	0.033***	0.031***	0.030***	0.023***	0.026***	0.028***	0.028***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)
Panel C: X = Uninformed Performance-Bucket Self-Evaluations (1–7)								
X	0.458***	0.517***	0.485***	0.487***	0.256***	0.376***	0.474***	0.419***
	(0.085)	(0.088)	(0.087)	(0.088)	(0.084)	(0.088)	(0.097)	(0.099)
Panel D: X = Uninformed Willingness Self-Evaluations (0–100)								
X	0.014***	0.019***	0.017***	0.017***	0.013***	0.014***	0.016***	0.017***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)
Panel E: X = Uninformed Success Self-Evaluations (0–100)								
X	0.037***	0.041***	0.036***	0.035***	0.026***	0.026***	0.029***	0.029***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)
Panel F: X = Informed Performance Self-Evaluations (0–100)								
X	0.003	0.010***	0.008**	0.010***	0.004	0.007*	0.011***	0.008*
	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)
Panel G: X = Informed Performance-Bucket Self-Evaluations (1–7)								
X	0.013	0.120*	0.055	0.142**	0.020	0.099	0.158**	0.077
	(0.066)	(0.069)	(0.069)	(0.071)	(0.067)	(0.071)	(0.077)	(0.081)
Panel H: X = Informed Willingness Self-Evaluations (0–100)								
X	0.019***	0.021***	0.020***	0.019***	0.015***	0.017***	0.017***	0.019***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)
Panel I: X = Informed Success Self-Evaluations (0–100)								
X	0.027***	0.028***	0.026***	0.027***	0.018***	0.020***	0.019***	0.022***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)
N	10590	10569	10435	9781	7614	7619	7469	7316

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

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

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

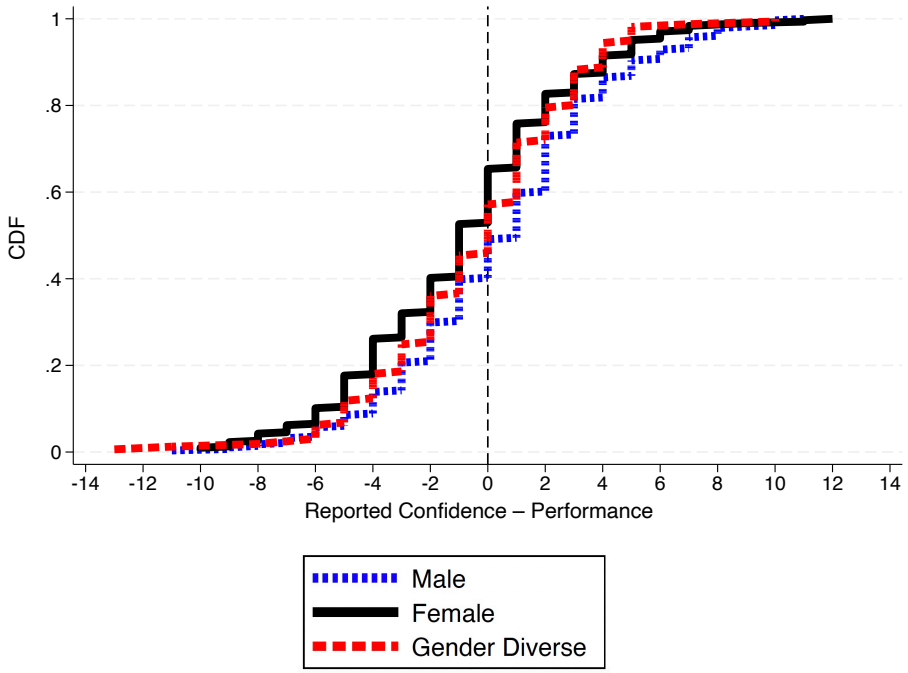
Table B.8: Regressions of Overall and Math GPA

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

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

C Additional Results from the Adult (Verbal) Study

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



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

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

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

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

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

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

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

D Experimental Instructions

D.1 Experimental Instructions for the Predictions Study

The experiment begins by informing each predictor about the study that they will take as shown in Figure D.1. Next, we provide information about the Prior Study (which is the *Adult Study* of this paper) as well as how gender categories are defined, and then explain the Current Study, as displayed in Figure D.2. Once predictors have an overview understanding of the experiment, they then proceed with the experiment as explained in Section 2.1. Below provides more details on each sets of questions.

We elicit predictors' beliefs about performance conditional on guessed performance (see Figure D.3). Specifically, we ask predictors to consider the group of [female]/[male]/[gender diverse] prior participants who guessed that they answered [5]/[10]/[15] questions correctly on the math and science test. Then, we ask them to guess how many questions, on average, they think they answered correctly on the math and science test. As shown in Figure D.4, using a slider, predictors select a range and if their chosen range includes the correct answer, they earn 1$. although the slider indicates the range of 9.5-10.5 in the screenshot, there is no default selection. Thus, predictors must click on the slider before they can move to the next page.

We elicit predictors' beliefs about performance conditional on the response provided to the following question: "On a scale from 0 (entirely disagree) to 100 (entirely agree), please indicate the extent to which you agree with the following statement: I performed well" (see Figure D.5.) Specifically, we ask predictors to consider the group of [female]/[male]/[gender diverse] participants who assigned their performance [a low rating of 0 to 33]/[a medium rating of 34 to 66]/[a high rating of 67 to 100] in response to the previous question. Then, again using the same sliders, we ask them to guess how many questions, on average, they think they answered correctly on the math and science test. An example decision screen is shown in Figure D.6.

We elicit predictors' beliefs about [female]/[male]/[gender diverse] participants being overconfident, accurate, or underconfident. We first define what it means for a participant to be overconfident, accurate, or underconfident as shown in Figure D.7, then provide further information about the task (see Figure D.8.) Specifically, we ask predictors to guess whether [female]/[male]/[gender diverse] participants are overconfident, accurate, or underconfident. If their guess is correct, they earn $1. We also give predictors to option to indicate when they are unsure. If they choose the "I'm unsure" option, they earn $1 with a 50% chance. An example decision screen is shown in Figure D.9.

Finally, we ask 15 follow-up questions to measure beliefs about female, male, and gender

diverse individuals using survey questions (again the order of these questions are randomized at the predictor level). Specifically, we ask predictors the following questions:

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

Figure D.1: General Instructions for the Predictions Study

STUDY INFORMATION

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

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

Understanding Question: Which of the following statements is true?

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

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

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

Figure D.2: Prior and Current Study Instructions in the Predictions Study

THE PRIOR STUDY

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

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

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

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

THE CURRENT STUDY

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

Figure D.3: Beliefs about Performance (1) Instructions in the Predictions Study

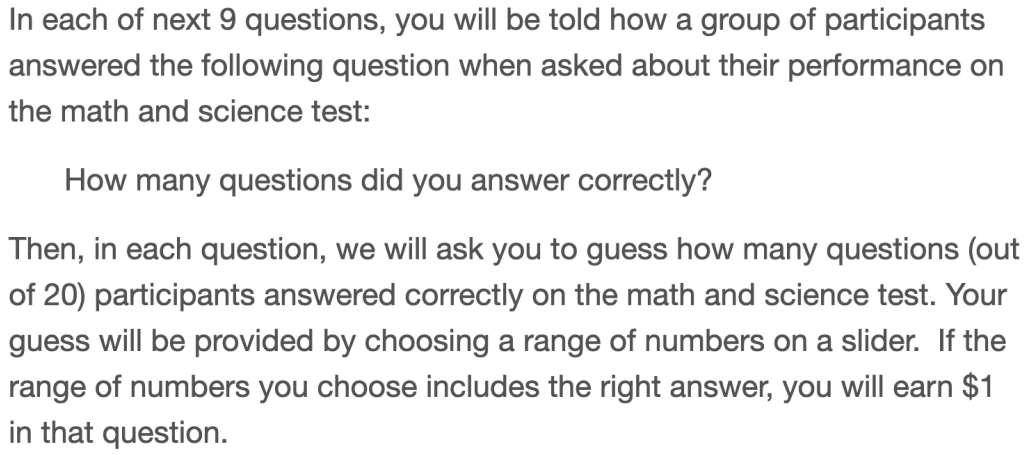


Figure D.4: Beliefs about Performance (1) Decision Screen in the Predictions Study

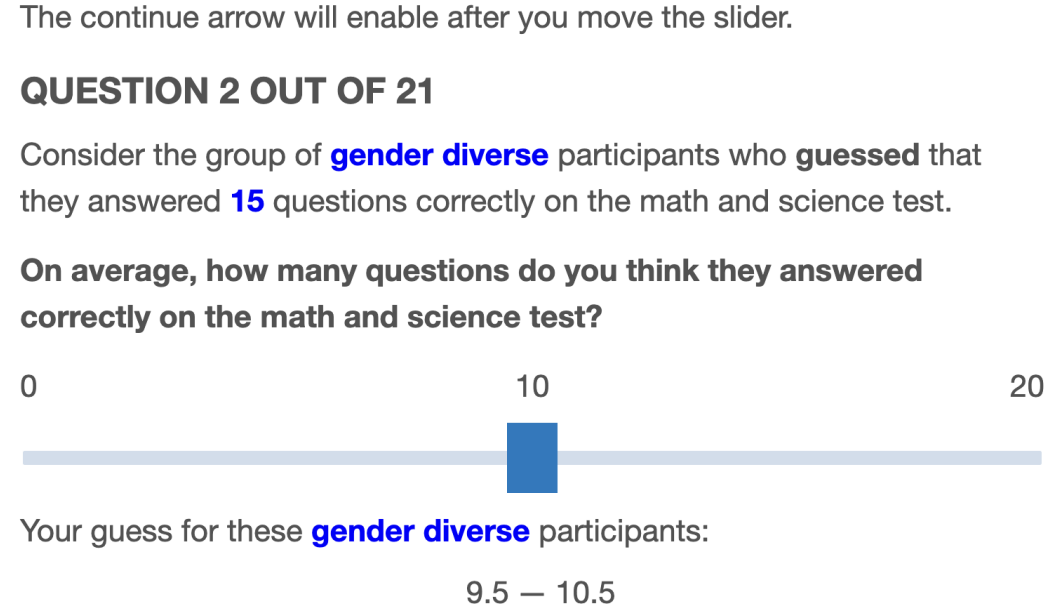


Figure D.5: Belief about Performance (2) in the Predictions Study

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

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

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

Figure D.6: Beliefs about Performance (2) Decision Screen in the Predictions Study

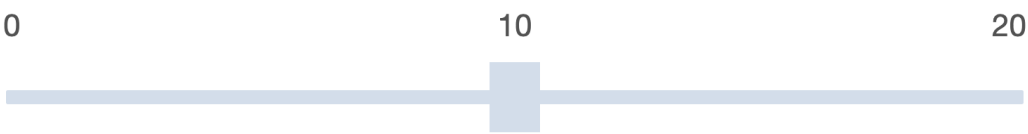
QUESTION 11 OUT OF 21

Participants were asked the following question:

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

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

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



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

—

Figure D.7: Defining Confidence in the Predictions Study

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

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

Understanding Question: When is a participant accurate?

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

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

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

Figure D.8: Beliefs about Confidence Instructions in the Predictions Study

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

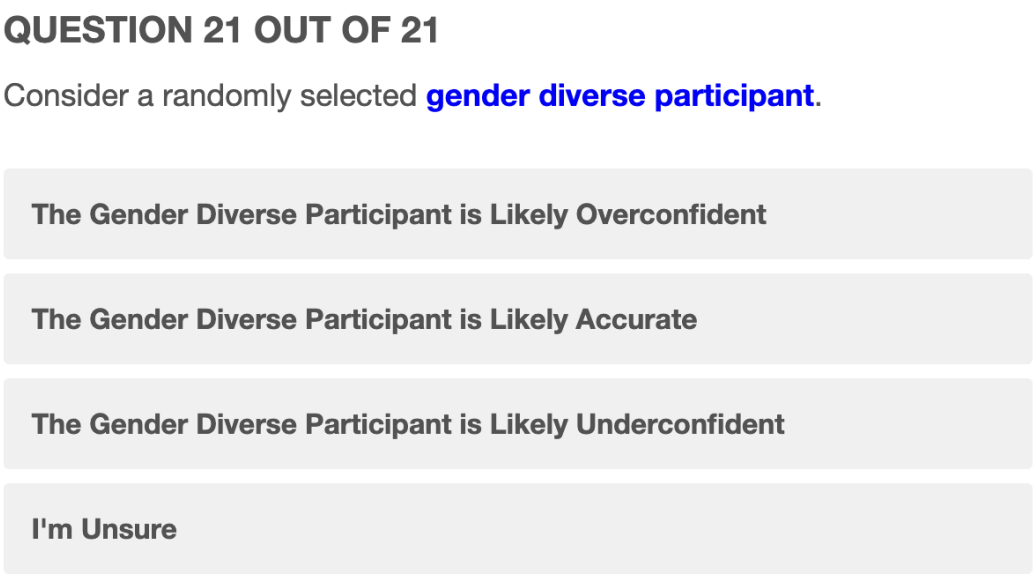
Three of these options involve you guessing that a given participant was accurate, overconfident, or underconfident. If you chose one of these three options, you will earn $1 if your guess is correct.

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

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

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

Figure D.9: Beliefs about Confidence Decision Screen in the Predictions Study



D.2 Experimental Instructions and Protocol for the Adult Study

The instructions for the experiment are displayed in Figure D.10. An example question on the test is displayed in Figure D.11 (note that the timer in the figure indicates the participant has 26 seconds left to answer the question although the timer starts at 30 seconds). After completing the test, participants are asked to complete five additional pages of the study.

First, they are asked about their absolute performance belief (see Figure D.12). Second, they are provided with additional instructions (see Figure D.13) and then asked the self-evaluation questions (see Figure D.14). Third, participants are provided with perfect information on their absolute performance and are required to correctly report back their absolute performance (see Figure D.15). Fourth, they are provided with additional instructions (see Figure D.16) and are asked the self-evaluation questions again (see Figure D.17). Fifth, they are asked for demographic information including their gender identity (see Figure D.18).

Our recruitment procedure was pre-registered on AsPredicted (#136119) which can be accessed here: https://aspredicted.org/2FW_Z5H. We started by using the "Sex" and "Cisgender and Transgender" screener questions that are set by Prolific (i.e., not by us). The "Sex" question asks: "What is your sex, as recorded on legal/official documents?" with options "Male" and "Female." The "Cisgender and Transgender" screener question asks: "Does your current gender differ from the one you were assigned at birth?" with answers "Yes," "No," and "Rather not say." Then, we aimed to recruit an equal number of participants who (1) answered "Female" to the "Sex" screener and "No" to the "Cisgender and Transgender" screener, (2) answered "Male" to the "Sex" screener and "No" to the "Cisgender and Transgender" screener, and (3) answered "Yes" to the "Cisgender and Transgender" screener. Since we expected that it would be much more difficult to recruit individuals in the third group—but desired to collect data across all three groups at similar times—we recruited participants in batches on a rolling basis. We first opened recruitment for 100 people in each of the three groups. Once all groups reached 100 completed responses, we opened recruitment for another 100 participants from each group. Our pre-registered recruitment plan was to continue this until we reached 600 people in each group or until we reached a satiation point of any group, whichever came first. The recruitment of the third group reached a satiation point at 500 participants. The first four times we opened recruitment to 100 participants, it took less than a day to collect all responses. The fifth time we opened the study for 100 participants, it took roughly three days to recruit 100 participants who had answered "Yes" to the "Cisgender and Transgender" screener. As a result, we recruited 1,500 participants on Prolific and ended up with 1,494 completed responses. These 1,494 participants were then randomized to either be in the *Adult Study* (n = 746) discussed in

Section 3 or in the *Adult (Verbal) Study* (n = 748) discussed in Section 5. We also restricted recruitment to participants who were U.S. nationals who had completed at least 100 prior submissions with at least a 95% approval rate.

Figure D.10: Part 1 Instructions for the test in the Adult Study

Instructions for Part 1 out of 3:

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

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

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

Figure D.11: Example question on the test in the Adult Study

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 D.12: Absolute Performance Belief Question in the Adult Study

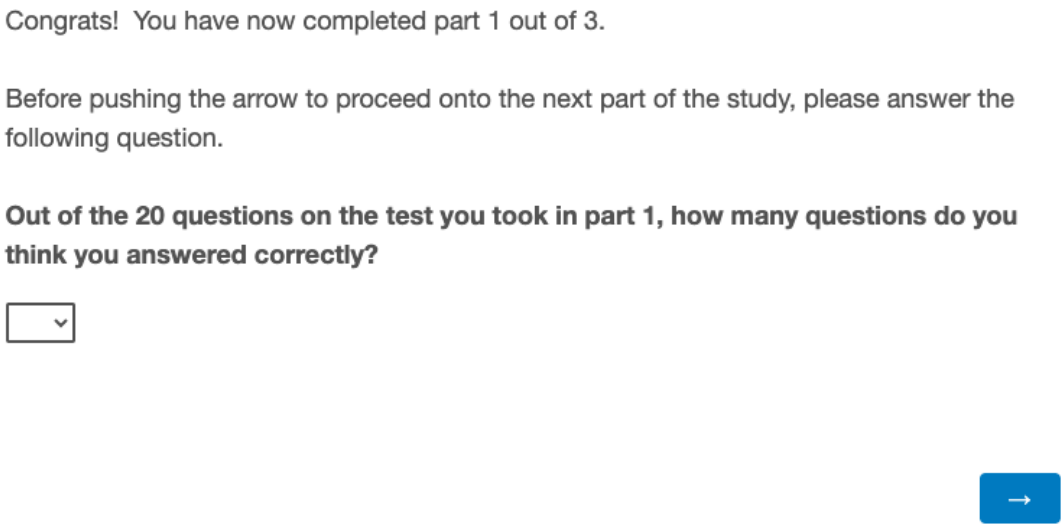


Figure D.13: Additional Instructions in the Adult Study

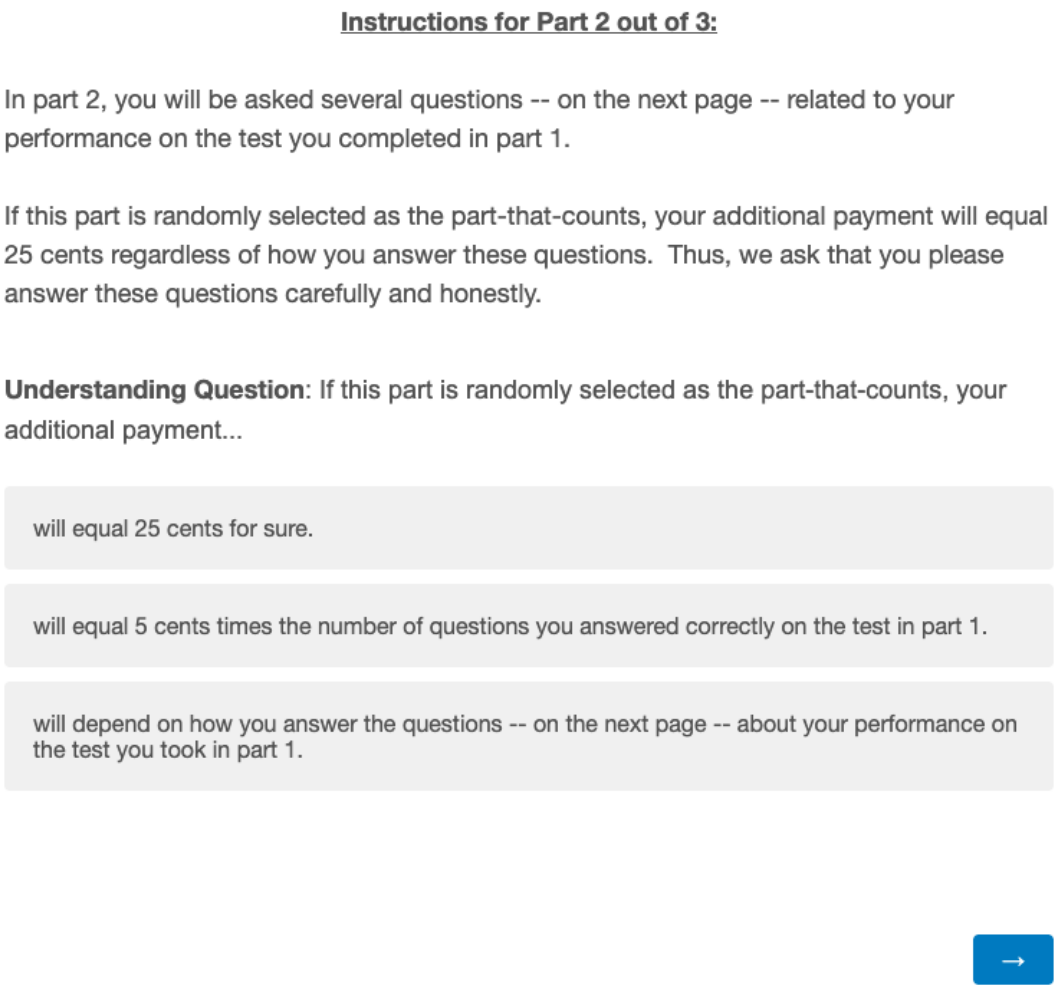


Figure D.14: Self-Evaluation Questions in the Adult Study

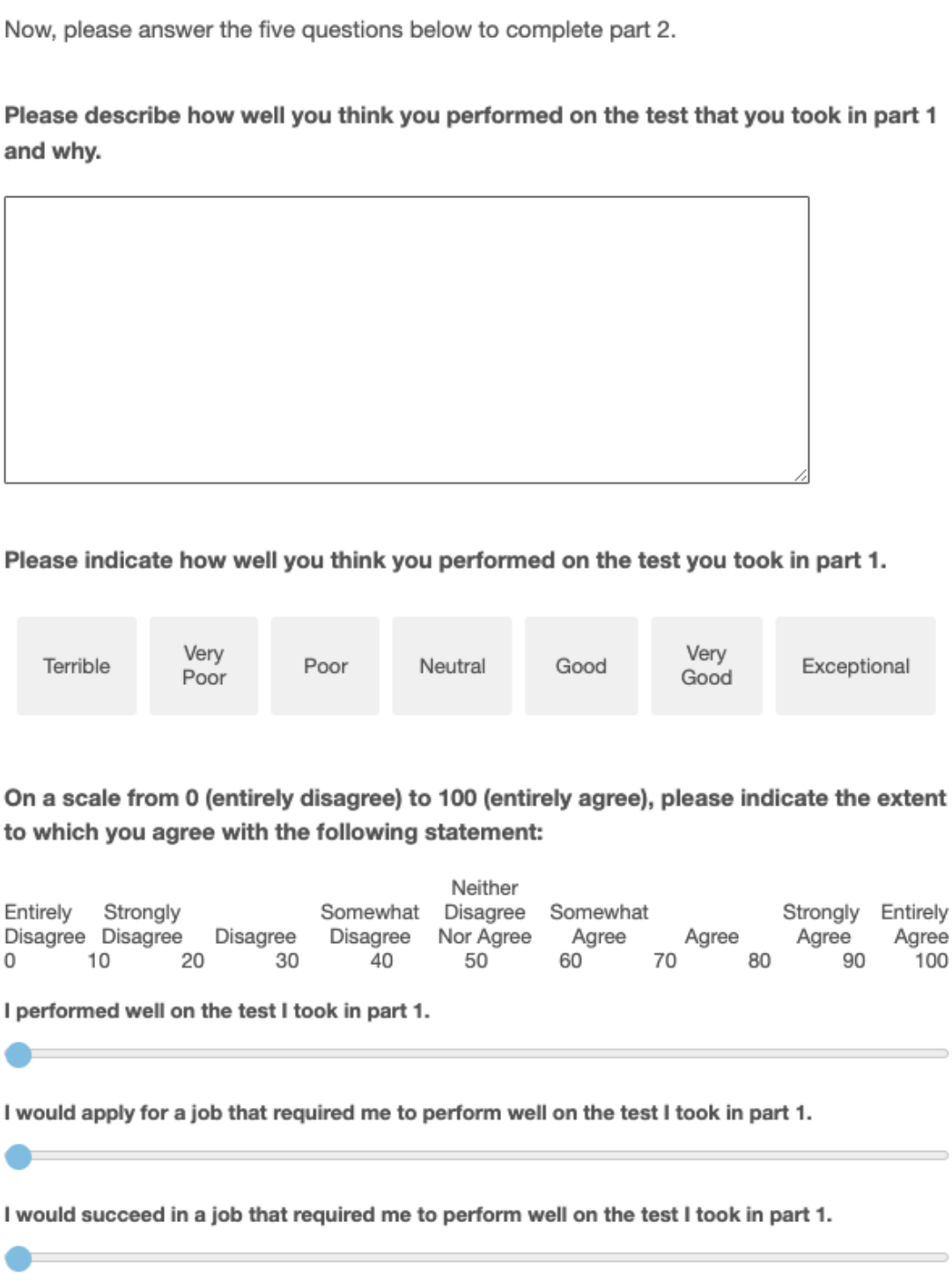


Figure D.15: Absolute Performance Information in the Adult Study

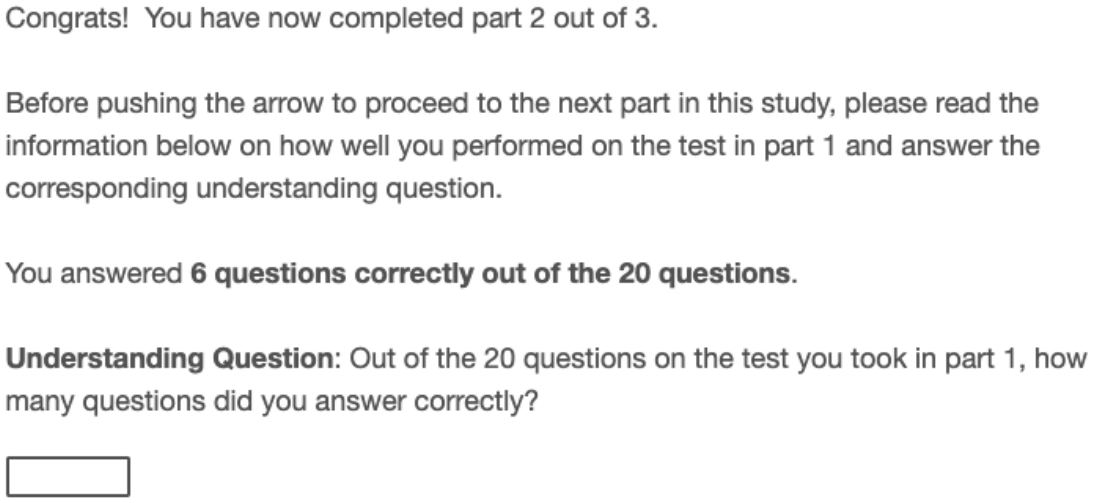


Figure D.16: Additional Instructions in the Adult Study

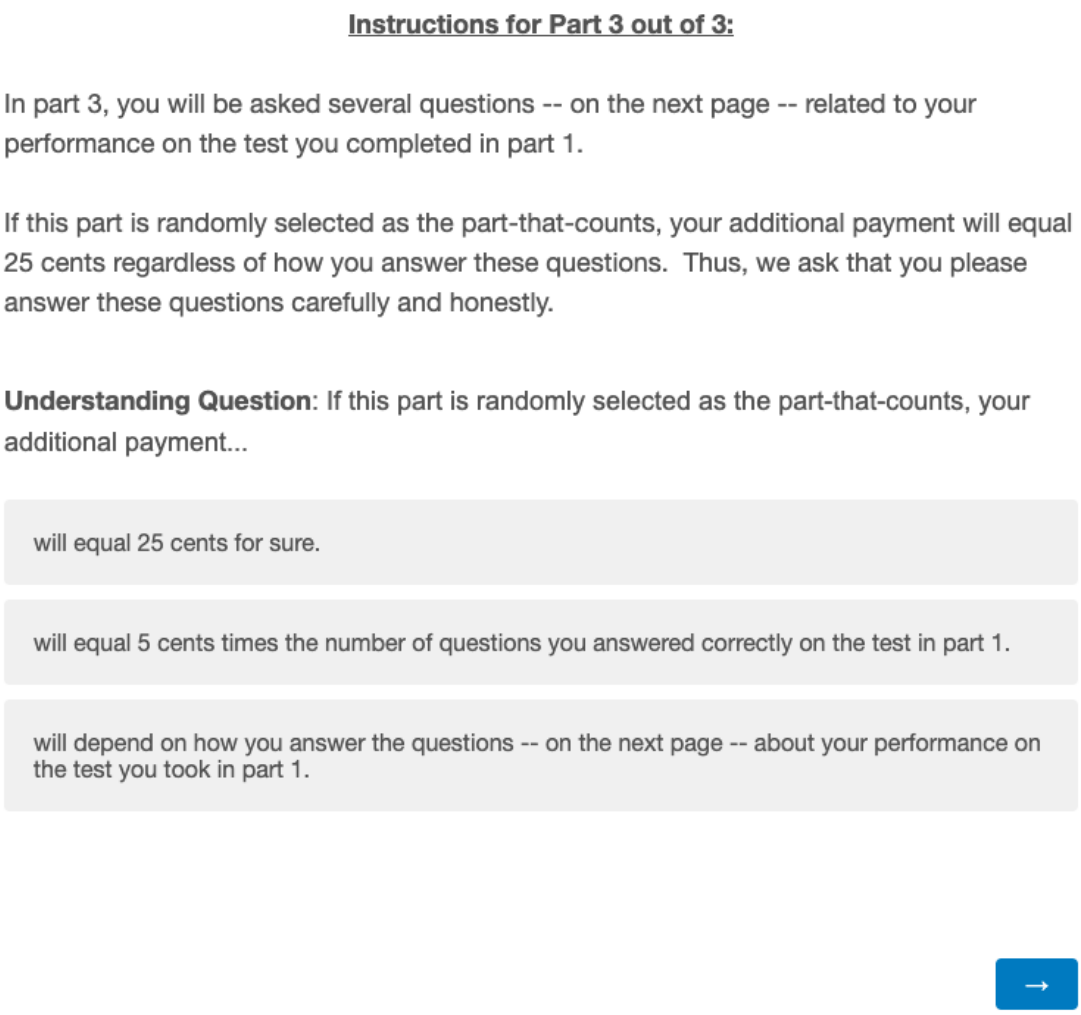


Figure D.17: Informed Self-Evaluation Questions in the Adult Study

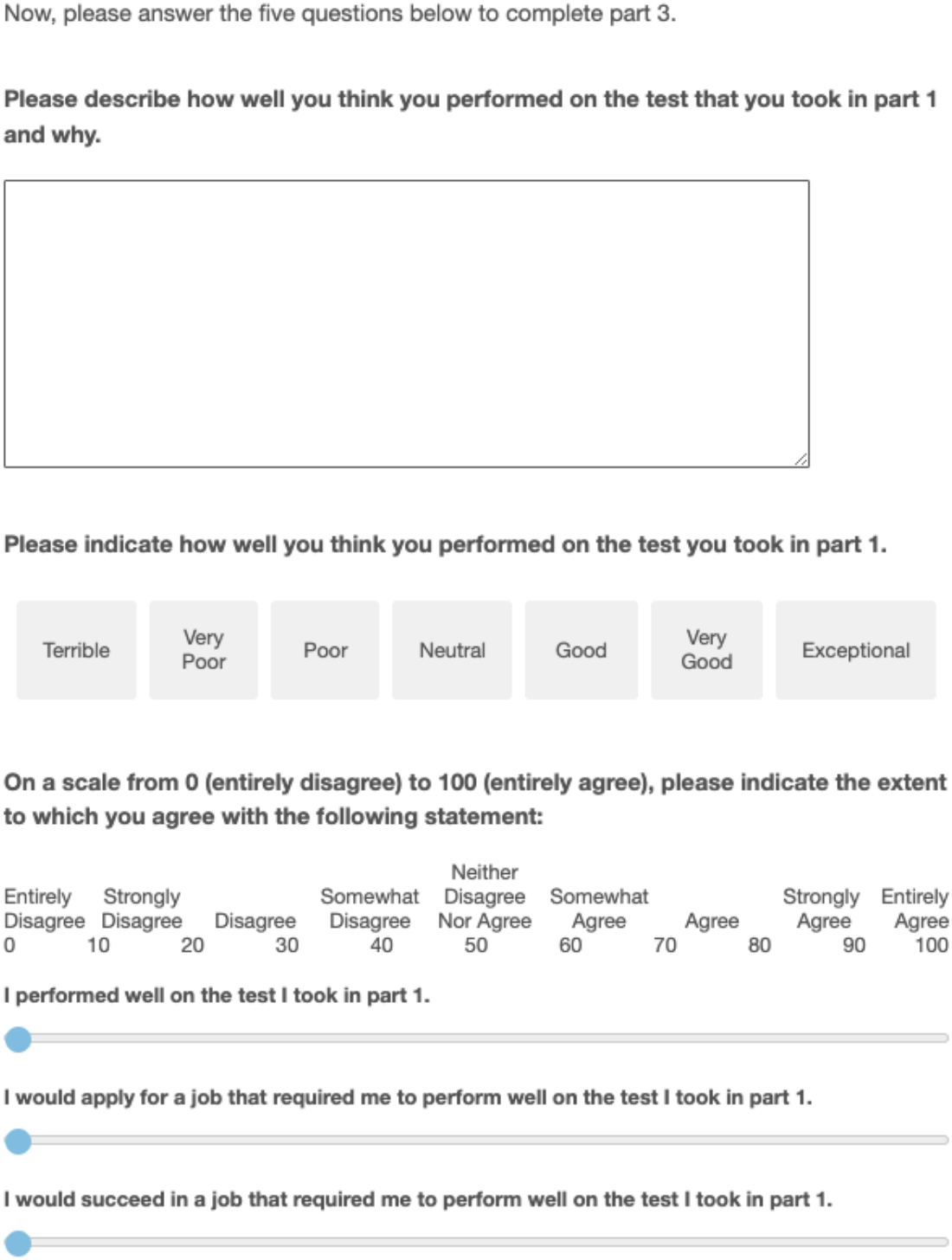


Figure D.18: Screenshot of Gender Question in the the Adult Study

D.3 Experimental Instructions for the Adult (Verbal) Study

The Adult (Verbal) Study closely follows the design discussed in Section D.2 with the exceptions discussed in Section 5. The instructions for the experiment are displayed in Figure D.19. An example question on the test is displayed in Figure D.20. After completing the test, participants are asked to complete five additional pages of the study which are identical to those described in Appendix Section D.2.

Figure D.19: Part 1 Instructions for the test in the Adult (Verbal) Study

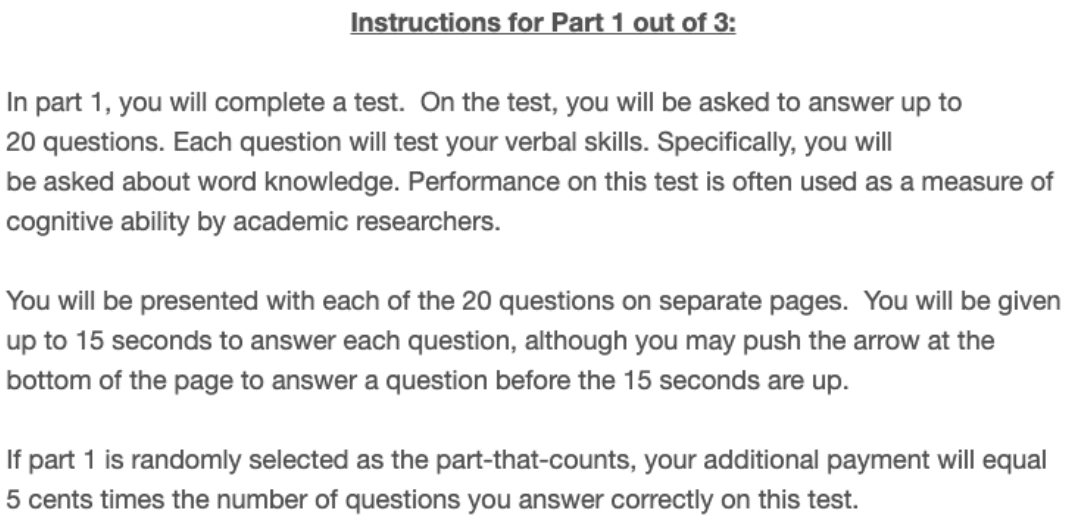
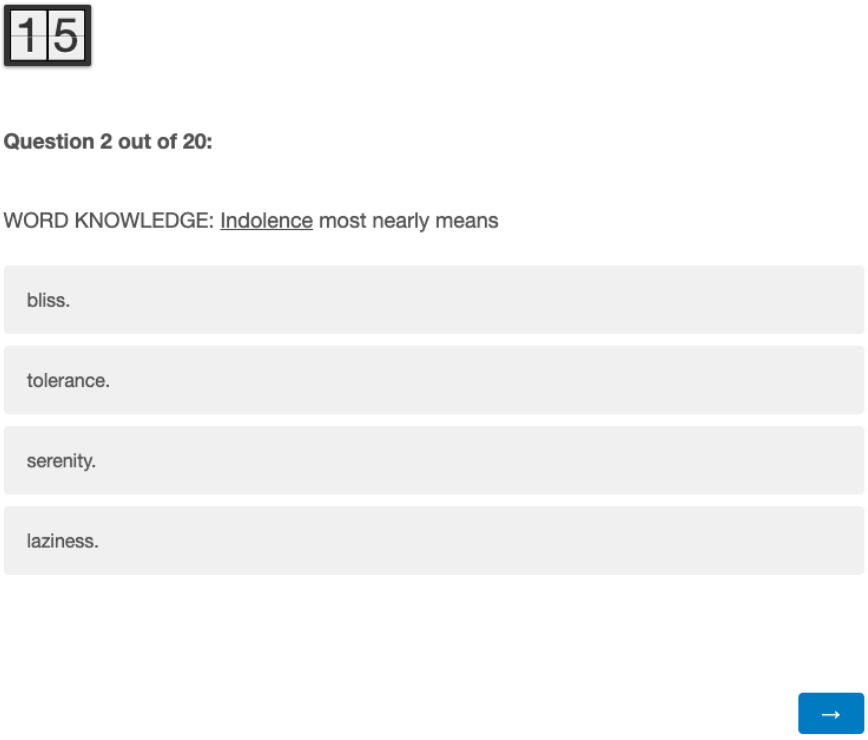


Figure D.20: Example question on the test in the Adult (Verbal) Study



D.4 Experimental Instructions for the Student Study

The Student Study closely follows the design discussed in Section D.2 with the exceptions discussed in Section 4.1. Prior to participating in the Student Study, participants must correctly answer a captcha and consent to participate. At the end of the study, participants must complete a short follow-up survey to gather demographic information. Participants are recruited via the Character Lab Research Network and complete this study as part of the curriculum at school. There are no payments associated with this study.

The study begins by informing each participant about the test that they will take. The instructions for the test are displayed in Figure D.21 and an example of a question on the test is displayed in Figure D.22 (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 D.23). On the second page, they are asked the self-evaluation questions (see Figure D.24). 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 D.25). On the fourth page, they are asked the self-evaluation questions again (see Figure D.26). On the fifth page, they are asked for demographic information including their gender identity (see Figure D.27).

Figure D.21: 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 D.22: Example question on the test in Student Study

Question 2 out of 10:

MATH KNOWLEDGE: Which number has the greatest value?

- 9,299
- 903 tens
- 93 hundreds
- 9 thousands

Figure D.23: Absolute Performance Belief Question in Student Study

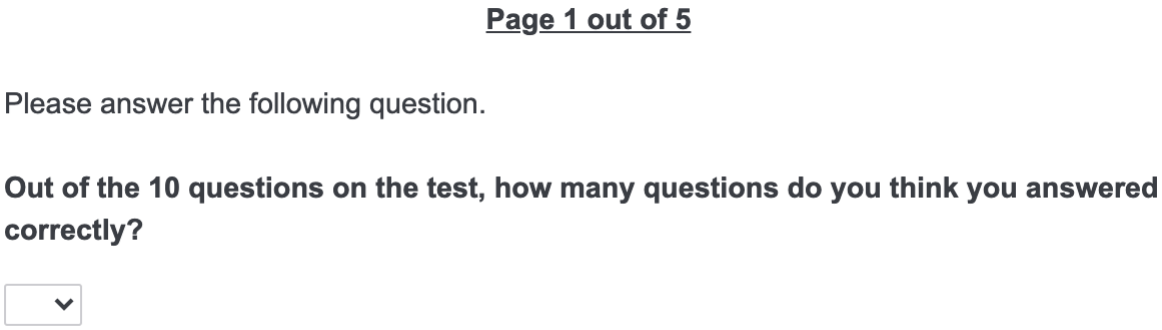


Figure D.24: Self-Evaluation Questions in Student Study

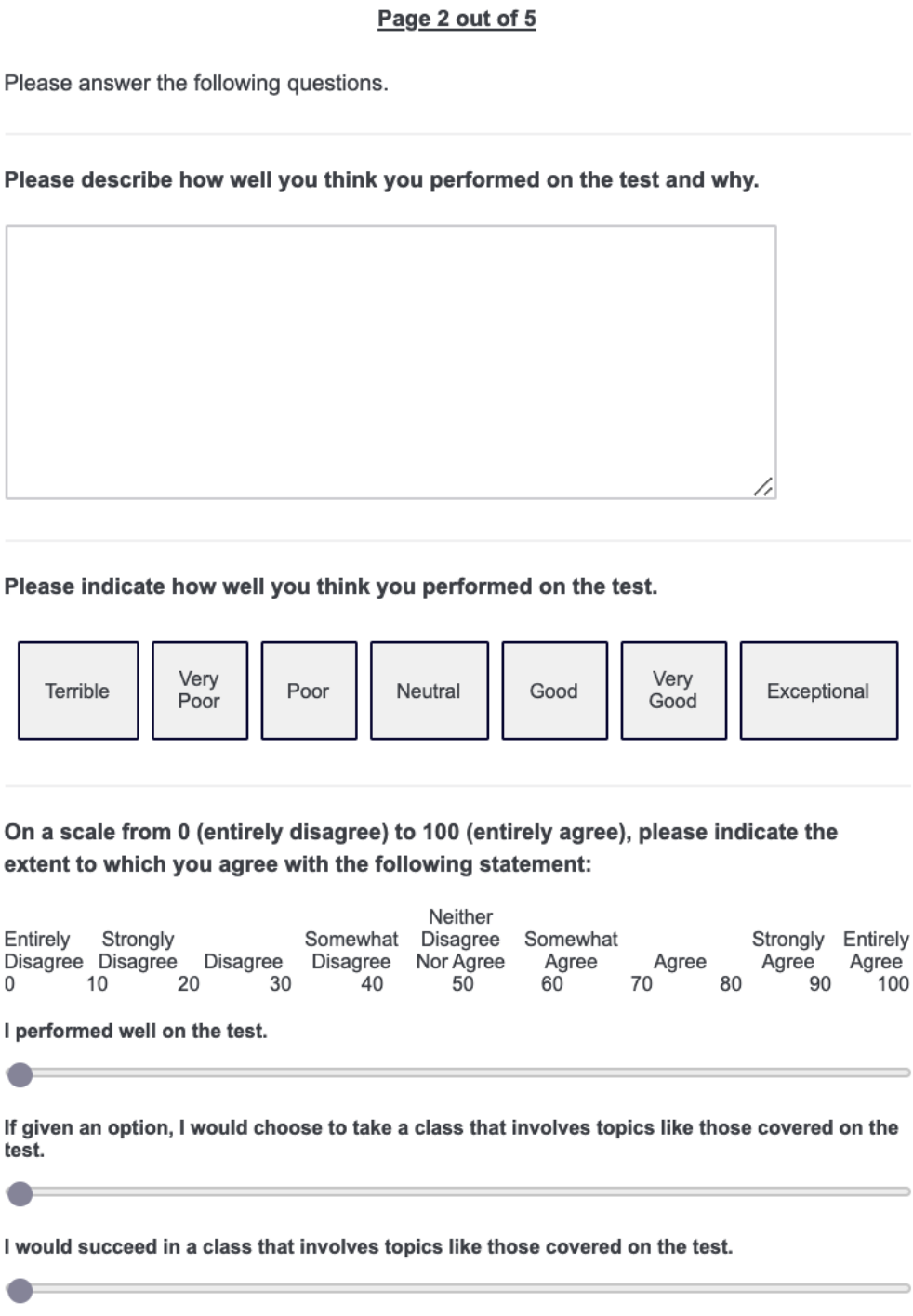


Figure D.25: Absolute Performance Information in Student Study

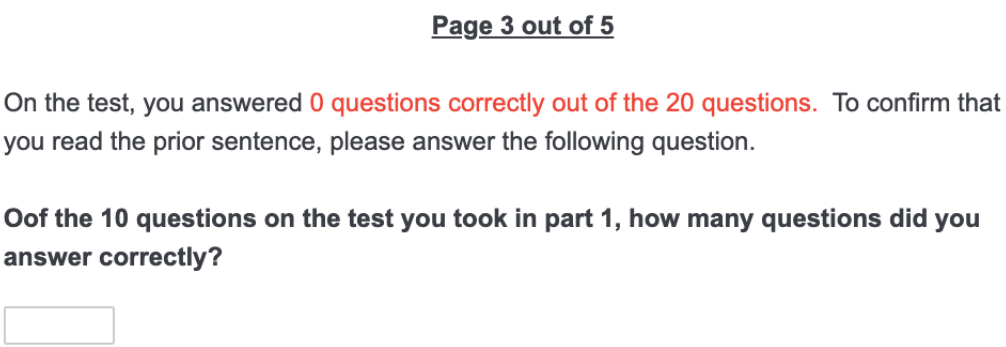


Figure D.26: Informed Self-Evaluation Questions in Student Study

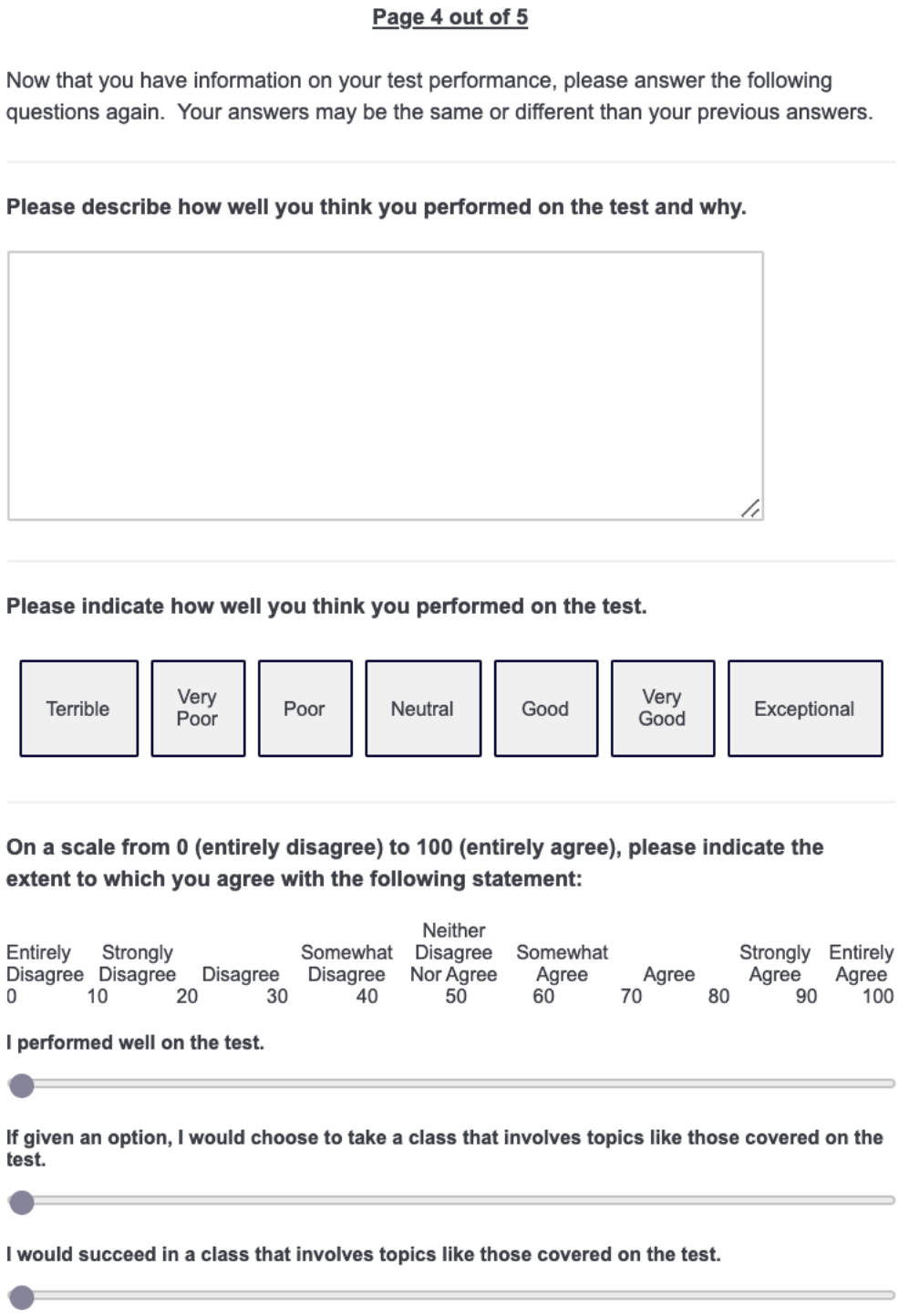


Figure D.27: Screenshot of Gender Question in the Student Study