Performance Feedback and Gender Differences in Persistence

Maria Kogelnik*

December 27, 2022

Abstract

I use an experiment and a field study to explore gender differences in persistence and the mechanisms driving this novel phenomenon. In the experiment subjects can either continue or drop out of an environment that involves ego-relevant feedback and solely rewards high performance. I find that men are on average 10 percentage points (15%) more likely to continue than women who performed equally well and received the same feedback. This gender gap in persistence can *not* be explained by preferences for competition or risk. Instead, I detect two novel mechanisms: First, men are more confident about their future performance even when compared to women who are similarly confident about their past performance – both in the experiment and a classroom field study. Second, experimental findings suggest that men seek, while women avoid exposure to additional feedback. Together, these mechanisms can explain roughly two-thirds of the gender gap in persistence.

Keywords: Gender, persistence, feedback, beliefs, information avoidance, economics experiment, field study.
JEL Codes: C91, D91, D83, J16, J24.

*Contact: University of Amsterdam (CREED) and Tinbergen Institute. Email: kogelnik.maria@gmail.com.
Acknowledgements: I am extremely grateful to Ryan Oprea and Sevgi Yuksel for their invaluable guidance and support. I thank Kelly Bedard, Javier Birchenall, Katherine Coffman, Michael Cooper, Florian Ederer, Ignacio Esponda, Erik Eyster, Peter Kuhn, Dominic Leggett, Shelly Lundberg, Aniko Öry, Rohini Pande, Heather Royer, Philipp Strack, Emanual Vespa, and participants at the 2021 All-California Labor Economics Conference, the 2021 ESA North American Meetings, the 2021 SOCAE Conference, and seminar participants at JILAEE, LMU Munich, Nottingham, NYU Abu Dhabi, UC Santa Barbara, Vienna, and Yale for many helpful comments and suggestions. **Funding:** Funding from the UCSB Economics Department and the UCSB Broom Center of Demography is gratefully acknowledged. **Ethics Approval:** This study obtained IRB approval at UCSB.

1 Introduction

The representation of women in stratified careers often resembles a "leaky pipeline:" the higher the hierarchical level, the lower the share of women in corporate management, academia and politics tends to be.[1] Suggested explanations include gender differences in job-related investments, maternity-related career interruptions and preferences over work conditions. In this paper, I present evidence in support of an additional channel: gender differences in persistence in response to performance feedback. Making one's way in stratified career paths typically involves exposure to feedback, i.e., information about one's past performance. If men and women differ in how they interpret or value this feedback, men could be more likely than equally performing women to persist, that is, to continue on these career trajectories rather than dropping out.

This paper studies gender differences in persistence and the mechanisms driving this phenomenon using a controlled laboratory experiment and a field study. The experiment is designed to investigate (i) whether men are more likely than women to persist in an environment that rewards solely high performance and involves exposure to feedback, and – if so – (ii) what mechanisms are driving this gender gap in behavior. The experimental design randomizes feedback conditional on performance, shuts down competition, and allows us to obtain novel insights on how people form beliefs about their future performance, as well as their preferences for additional feedback exposure. A classroom field study complements this experiment by testing the external validity of the belief formation patterns documented in the lab.

Using a controlled experiment to study gender differences in persistence has multiple advantages. First, any differences in the outside options or returns to persisting that men and women may face in the field can be shut down in the lab. Second, the feedback that people receive is perfectly observed, and it can be ensured that there is no gender bias in how the feedback is given, as well as no gender differences in selecting or expecting a certain kind of feedback. Furthermore, by exogenously varying the feedback, the effect of positive versus negative feedback can be explored across the performance distribution. Finally, understanding what mechanisms are driving the gender gap in persistence requires the measurement of variables that are unobserved in naturally occurring

[1] For example, see the Women in the Workplace 2021 report by McKinsey and LeanIn.org, as well as Bertrand and Hallock (2001) for a corporate context; the She Numbers 2018 report of the European Commission for research and innovation; Lundberg and Stearns (2019) for economics; and the Women in Politics 2019 report by the Inter-Parliamentary Union for politics.

data, such as beliefs about the future, or preferences to avoid or receive additional feedback.

The idea that men are more likely to persist in response to feedback than equally performing women is consistent with a recent empirical literature. Women have been found to be less likely than men to continue in STEM and economics majors in response to poor grades (Katz et al., 2006; Rask and Tiefenthaler, 2008; Kugler et al., 2021; Astorne-Figari and Speer, 2019), less likely to participate again in prestigious math exams, math olympiads, Rubik's Cube competitions, or college entry exams after scoring low previously (Ellison and Swanson, 2018; Franco, 2018; Buser and Yuan, 2019; Fang et al., 2021; Kang et al., 2021), less likely to submit an article to the largest economics conference in Brazil following a previous rejection (Pereda et al., 2020), and less likely to re-run for office after barely losing an election (Wasserman, 2021).[2] In the field, gender differences in persistence may be easier detectable in response to negative feedback (when many people drop out of a career trajectory), however it is also conceivable that positive feedback has a more encouraging effect on men to persist than on women. To better understand the effect of feedback on persistence, studying both positive and negative feedback is relevant.

The experiment at hand was designed to accomplish two goals. The first goal is to create a setting that captures the essential features of the decision of interest: a choice between persisting or dropping out of an environment that involves feedback and rewards solely high performance. Importantly, this feedback should be ego-relevant in the sense that people may care about feedback beyond it being instrumental to their choices – a natural feature of many stratified professions. The second goal of the experimental design is to explore the mechanisms behind this gender gap in persistence – focusing on the role of beliefs, preferences for additional feedback, and risk preferences.

In the *Baseline* treatment, subjects are asked to perform a challenging and ego-relevant task (an IQ test), which they can either pass or fail. Then they receive feedback; an informative signal about their past performance that is either positive or negative. To explore the effect of positive versus negative feedback across the performance distribution, this feedback is randomized conditional on having passed or failed, and of known accuracy. Subjects then face two options: If they *continue*, they get additional feedback (i.e., they learn if they *really* passed or failed the first IQ test), have to take a second IQ test (henceforth labelled the "future IQ test"), and receive a high

[2] In contrast, Thomsen (2018) and Bernhard and de Benedictis-Kessner (2021) do not find gender differences in politician persistence following election losses.

bonus payment if they pass the future IQ test, but nothing otherwise. Alternatively, if they *quit*, they get no additional feedback, complete an easy test, and receive a fixed minimum payment that does not depend on their performance. Since neither the compensation nor the feedback provided depend on the performance of other participants, a potential gender gap in persistence in this setting therefore does not reflect the well-studied gender differences in the willingness to compete (e.g., see the seminal work of Niederle and Vesterlund, 2007).

My first main finding is that women are about 10 percentage points less likely than men to continue in this environment when controlling for subjects' performance, the feedback they received, as well as self-reported characteristics. For men, the average probability of continuing is roughly 60%, while for women it is only about 50%. Notably, conditional on past test scores, women who received *positive* feedback are similarly likely to continue as men who received *negative* feedback. To shed light on the mechanisms driving this gender gap in behavior, I study the role of beliefs, preferences for additional feedback, and risk preferences.

The first mechanism I explore is how people form beliefs about their future performance. Recall that continuing is only financially rewarding for those who achieve a high performance in the future (i.e., subjects who pass the future IQ test). While previous experiments have largely focused on gender differences in beliefs about people's *past* performance (see below), it is worth pointing out that men and women may differ in their perceptions of how predictive the past is of their future performance, and they could adjust these beliefs differently in response to ego-relevant feedback. A novel feature of my experimental design is that it allows us to differentiate if gender differences in confidence about one's future performance are present at the stage of initial beliefs before feedback, arise when people update their beliefs in response to feedback, or arise because men and women make different inferences about their future, given their beliefs about their past performance. This is achieved by eliciting subjects' beliefs about their past *and* future performance both before *and* after receiving feedback. Reporting true beliefs is incentivized.

I find that women are less confident about passing the future IQ test both before and after receiving feedback, relative to men who performed equally well on the first IQ test. Remarkably, men are more confident about passing the future IQ test *even when* compared to women who performed equally well on the first test *and* are similarly confident about their past performance. This novel insight suggests that men might discount how predictive their previous failures are, or

over-weigh how predictive their previous successes are of their future performance, relative to women who previously performed equally well. Consequently, men's expected returns from persisting are higher. I find no evidence of gender differences in updating beliefs in response to feedback, however. Roughly one-third of the gender gap in persistence is attributable to gender differences in beliefs about one's future performance.

To examine the outside validity of the gender differences in beliefs documented in the lab in a less artificial context, I conduct a classroom field study. In this study, undergraduate students are asked to report beliefs about their past and future performance on midterm exams after completing the first midterm, but before learning their grades. Findings in the field are remarkably similar to the lab not only qualitatively but also in terms of the effect size. Not only are men more confident about their future performance than women who performed equally well on the first midterm exam; Men also make more optimistic projections of their future performance even when compared to women who performed equally well *and* are similarly confident about their past performance.

The second mechanism I explore concerns gender differences in preferences for additional feedback. One natural feature of the *Baseline* treatment is that persisting involves exposure to additional feedback, while quitting does not. If women dislike being exposed to ego-relevant feedback, or if men enjoy getting additional feedback, this could help explain the gender gap in persistence. Note that this additional feedback is not valuable for decision making in the experiment, as subjects can only learn if they passed the first IQ test *after* deciding between continuing and quitting; But men and women may derive a different consumption value from this additional feedback (e.g., see Köszegi (2006) and Karlsson et al. (2009) for anticipatory utility models). To explore this idea, the design includes one treatment where subjects receive additional feedback (i.e., they learn if they really passed or failed the first IQ test) regardless of whether they continue or quit. Other than that, all features of this *AlwaysInfo* treatment and the *Baseline* are identical. A between-design is used, i.e., all subjects participate in either the *Baseline* or the *AlwaysInfo* treatment. Comparing behavior between these two treatments then allows us to assess to what extent the gender gap in persistence is attributable to feedback avoidance and feedback seeking.

I find suggestive evidence that gender differences in feedback avoidance account for almost 30% of the gender gap in persistence. Directionally, this is driven both by men who continue in order to receive additional feedback, and by women who quit in order to avoid additional feedback.

These estimates of the *AlwaysInfo* treatment effect control for gender differences in confidence, that is, these estimates do not reflect gender differences in the expectations about what additional feedback they might receive.

As continuing constitutes a risky payoff structure while quitting guarantees a fixed minimum payment, quitting might be relatively more attractive for women if they are more averse to taking risks, all else equal. The design allows us to explore the role of risk preferences on the gender gap in persistence by including a task that resembles the continuation decision but is stripped from all features other than the compensation scheme and risk. No gender differences in risk aversion are detected in this setting, however, and controlling for subjects' risk preferences essentially has no impact on the estimated gender gap in persistence.

Performance feedback mechanisms may contribute to a gender gap in ability within organizations if low-performing men are more likely to persist, or if high-performing women are less likely to continue. In the experiment, men are adversely selected into continuing when taking past performance as a measure of ability: Conditional on having failed the first IQ test, men are almost 40% more likely to continue than women. But since people's performance can improve over time, this does not imply that women's continuation decisions necessarily better predict their future performance; By dropping out, women forgo the opportunity of learning that they may improve and that persisting could pay off for them despite initial setbacks.

Contribution. This paper makes three main contributions to the literature. First, to my knowledge, this is the first paper to document gender differences in persistence in a controlled setting, and to explore through which channels receiving positive versus negative feedback affects persistence. The presented findings do not reflect gender differences in the willingness to compete, as competition is shut down by design in the experiment. Related experiments have studied how feedback about one's *relative* performance affects gender differences in choosing a hard over an easy mazes task (Niederle and Yestrumskas, 2008), in setting goals for one's future performance on an adding numbers task (Buser, 2016), and in choosing a competitive over a piece-rate payment scheme in adding numbers tasks (Berlin and Dargnies, 2016; Buser and Yuan, 2019), as well as in verbal and math quizzes (Coffman et al., 2021). In contrast, this paper studies persistence, i.e., the behavior of continuing rather than dropping out of in an environment that rewards high performance and

involves ego-relevant feedback.

Second, this paper presents the novel insight that men – even when compared to women who performed similarly *and* are similarly confident about their past performance – tend to be more confident about their future performance; both before and after receiving feedback. Previous studies have largely focused on gender differences in beliefs regarding subjects' past performance: Controlling for actual performance, women have been found to be less confident about their past performance (e.g., Deaux and Farris, 1977; Lundeberg et al., 1994; Falk et al., 2006; Niederle and Yestrumskas, 2008; Mobius et al., 2014; Coffman et al., 2019; Thaler, 2021; Coffman and Klinowski, 2022), and to update more conservatively (Mobius et al., 2014; Coutts, 2018) and more pessimistically (Berlin and Dargnies, 2016) in response to feedback. Other studies, however, find no gender gap in confidence (Ertac, 2011; Berlin and Dargnies, 2016; Coutts, 2018). Furthermore, gender differences in both initial beliefs and information processing have been found to vary with the gender-congruence of quiz domains (Coffman, 2014; Bordalo et al., 2019; Coffman et al., 2019, 2021). The only study I am aware of that elicits beliefs about one's future (but not past) performance before and after feedback is Alan and Ertac (2019), who examine the gender gap in competitiveness among children, and thus also elicit beliefs about their opponents. In contrast, eliciting beliefs about both the past *and* the future allows me to detect that men and women differ in how they extrapolate from the past when forming beliefs about their future performance, as well as the role of these beliefs on persistence.

Finally, by presenting an experimental design that allows us to isolate the role of gender differences in feedback avoidance on persistence, this paper contributes to a relatively under-studied literature on how preferences for information affect economic behavior. Golman et al. (2017) provide an excellent review of the literature on information avoidance, but do not mention gender. Buser and Yuan (2019) find that information avoidance can explain the gender gap in competition in the first, but not in later rounds of an adding numbers task. Eil and Rao (2011) and Mobius et al. (2011) find no gender differences in the average willingness to pay for performance feedback and ego-relevant information, but note that women are more likely than men to require a compensation to receive this information.[3] In a recent comprehensive study involving different tasks, difficulty levels and feedback conditions, Coffman and Klinowski (2022) find no gender differences in the demand for

[3] In Eil and Rao (2011), these differences are not statistically significant.

feedback. Contrary to their setup where subjects do not get any feedback before deciding whether to demand feedback, I examine preferences for *additional* feedback exposure. My experiment aims at exploring the role of these preferences for the gender gap in persistence at the aggregate rather than studying information avoidance at the individual level.

The remainder of this paper is organized as follows. Section 2 describes the experimental design and implementation. Section 3 presents the experimental results on gender differences in persistence, as well as the mechanisms driving this gender gap. Section 4 introduces the classroom field study and demonstrates that the belief formation pattern documented in the lab replicate well in the field. Section 5 discusses whether gender differences in persistence contribute to a gender gap in ability within organizations. Finally, Section 6 concludes.

2 Experimental Design

Design goals and overview. The experiment was designed to accomplish two goals. The first goal is to create a setting that allows us to study gender differences in persistence in response to ego-relevant feedback. This requires mimicking some essential features of a stratified career: A challenging task where solely high performance is rewarded; the provision of ego-relevant feedback; and a choice between either *continuing* or *dropping out*. The second goal is to explore through what mechanisms gender differences in persistence may operate – particularly those that cannot be identified in naturally occurring data: beliefs (and how these respond to feedback), preferences for additional feedback, and risk preferences.

The experiment consists of four main parts that are described below. To eliminate income effects and incentives to hedge, one of the four main parts was randomly drawn for payment at the end. In addition to a show-up fee of \$5, subjects earned a bonus payment that could range between \$0 and \$22 in the part drawn for payment. To credibly implement both treatments, subjects were never told which part was drawn for payment.

A timeline of the main parts of the experiment is provided in Figure 1.[4] Instructions clarified how to earn money before each part, but subjects were not told what would happen in later parts of

[4]See the Supplementary Appendix for a detailed overview of the order in which instructions etc. where presented to subjects. Additional design elements that are not essential for understanding the main results (such as a survey at the end) are described there as well.

the experiment. Subjects had to correctly answer comprehension quizzes at different points of the experiment before moving on. A between-design was used, i.e., all subjects participated in either the *Baseline* or the *AlwaysInfo* treatment. The only component that differs across treatments is what happens if subjects quit in Part 3 of the experiment, see below. Instructions and screenshots of the experimental interface (including comprehension quizzes) are provided in the Supplementary Appendix.

Part 1: IQ test. Subjects were asked to take an IQ test, consisting of seven Raven (1973)'s Progressive Matrices, including a range from relatively easy to relatively difficult questions. Raven's matrices have been frequently used in economics experiments to generate an environment where ego utility is at stake (e.g., Zimmermann, 2020; Oprea and Yuksel, 2022). Subjects were told that this test is frequently used to measure intelligence. (Note that while some studies find that men on average perform slightly better on this IQ test, if anything more subjects in my experiment perceived women to perform better, than men to perform better.[5])

Before taking the IQ test, subjects were informed that they would either *pass* or *fail* this test. To pass, at least five of the seven questions had to be solved correctly. If Part 1 was drawn for payment, subjects earned a bonus of \$20 if they passed, and \$0 if they failed the IQ test. It was pointed out that whether they passed or failed did not depend on the performance of other participants. Subjects had 90 seconds to answer each question, and a timer indicated how much time was left. Wrong answers were not penalized, and unanswered questions were counted as wrong.

Part 2: Performance feedback and beliefs.

Feedback. Feedback was conveyed in the form of a binary signal: Subjects got to see one card that either said that they passed, or that they failed the IQ test, as depicted in Figure 2. This feedback was randomized and matched the true state of having passed or failed with a known accuracy of two-thirds. In other words, subjects who passed the IQ test were twice as likely to see a card saying that they passed, than seeking a fake card telling them that they failed, and

[5] Lynn and Irwing (2004) conduct a meta analysis and conclude that while there is no gender gap in performance in Raven's matrices from age 6-14, males on average perform slightly better starting at age 15. In an unicentivized question at the end survey (after completing the experiment) I asked subjects whether they think that men or women perform better on the IQ test used in the experiment, see Screenshot 107 in the Supplementary Appendix. Among all female (male) participants, roughly 28% (21%) thought that women perform better, 14% (8%) thought that men perform better, and 58% (72%) thought that there is no gender gap in performance.

vice versa. Randomizing feedback has the advantage that the effect of receiving positive versus negative feedback can be explored across the performance distribution. Providing feedback through this known process ensures that there is no gender bias in what feedback is given, and that men and women cannot endogenously affect what kind of feedback they are seeking.

Beliefs. To investigate the role of beliefs for gender differences in persistence, the following two questions were asked both before and after the provision of feedback, yielding a set of four elicited beliefs per subject.[6] Before the second question, subjects were informed/reminded that they might be asked to take a "future IQ test" of a similar level of difficulty later in the experiment.

1. How likely (out of 100) do you think it is that you passed the IQ test?

   – *Announcement of future IQ test.* –

2. How likely (out of 100) do you think it is that you could pass the future IQ test?

A novel advantage of eliciting these two beliefs both before and after feedback is that this allows us to explore gender differences (i) in how people form beliefs about their future, given beliefs about their past performance; and (ii) how these beliefs respond to feedback. If Part 2 was drawn for payment, subjects earned a bonus of either \$20 or \$0, determined by the crossover method (Mobius et al., 2014). This mechanism implies that subjects maximize their chance of winning \$20 by always reporting their true beliefs, which was emphasized in the instructions.[7]

Part 3: Continue or quit. The main outcome of interest in the experiment is how subjects choose between the two options of *continuing* and *quitting*. Subjects' continuation probabilities serve as a measure of persistence. The two options vary in terms of (i) the additional feedback subjects get, (ii) the difficulty of the next task, and (iii) the compensation scheme of the next task – see panel (b) of Figure 1. Subjects were informed in detail about what each option entailed and had to correctly answer comprehension questions before making their decision. Choices could *not* be reversed later, e.g., after receiving additional feedback. It was further emphasized that quitting does not imply leaving the experiment early.

[6] When eliciting beliefs the second time, i.e., after the provision of feedback, the card conveying the feedback was displayed next to the questions. See Screenshots 64 and 66 in the Supplementary Appendix.

[7] The crossover mechanism requires the assumption of monotonic preferences, but not expected utility preferences or risk neutrality to be truth-inducing.

Continue. This option aims to mimic the consequences of persisting on a stratified career trajectory. If subjects continued, they first received additional feedback by learning if they really passed or failed the first IQ test.[8] Then they were asked to take a second IQ test that resembled the first IQ test in terms of style and difficulty. (The information of having passed or failed was further displayed next to each question of the second IQ test in order to create frequent feedback exposure.) If Part 3 was drawn for payment, subjects who continued earned a bonus of $20 if they passed, and $0 if they failed the second IQ test. Consequently, continuing was only financially rewarding for subjects who could pass the second test.

Quit. Quitting serves as a natural outside option for those who "drop out" of the career path they had encountered before. Subjects who quit were asked to complete an "easy test," consisting of seven very easy Raven's Matrices.[9] If Part 3 was drawn for payment, subjects who quit received a fixed minimum payment, described below in more detail.

The only feature distinguishing the *Baseline* from the *AlwaysInfo* treatment is whether or not subjects who *quit* learn if they really passed or failed the first IQ test. In the *Baseline*, only subjects who *continue* receive this additional feedback. That is, *quitting* allows subjects in the *Baseline* to avoid finding out if they really passed or failed.[10] In contrast, subjects in the *AlwaysInfo* treatment learn their first test result regardless of whether they continue or quit, i.e., Step 1 in panel (b) of Figure 1 is the same for all subjects in the *AlwaysInfo* treatment. This treatment thus shuts down preferences for additional feedback as a motive for continuing or quitting.[11] Comparing behavior across the two treatments therefore allows us to isolate the role of feedback avoidance and feedback seeking for the gender gap in persistence.

Part 4: Risk task. Part 4 was designed to enable the estimation of risk preferences in the context most relevant to the decision of interest, as recommended by Niederle (2014). Subjects faced two

[8] In addition, subjects learned if they had guessed most boxes right or wrong in a trivial "Guessing Game" that had been administered before the first main part of the experiment. This information was by design orthogonal to subjects' IQ test performance, had no consequences on their earnings, and was held constant across treatments. The sole purpose of providing this information was to give the researcher the option of running an additional treatment arm at a later point in time. See Appendix B for details.

[9] Having an easier task as an outside option feels natural and keeps opportunity costs of time similar across the two options.

[10] As subjects were not told which part was drawn for payment in the end, they could not infer this information from their final earnings in the experiment either.

[11] One could argue that the experience of taking another IQ test might convey additional feedback even if one does not learn the test result. With this in mind, the *AlwaysInfo* treatment effect can be thought of as a lower bound of the effect of preferences for additional feedback on persistence.

options that were analogous to the two options in Part 3 (continuing versus quitting), but stripped from all features other than the compensation scheme and risk involved. Analogously to continuing, one option constituted a lottery that paid \$20 with some probability p, and \$0 with some probability $100 - p$, where p was tailored to each subject's previously reported belief about passing the second IQ test after having received feedback in Part 2.[12] Analogously to quitting, subjects could receive a fixed minimum payment, see below.

Measurement of persistence and BDM mechanism. In Part 3 and Part 4, rather than asking subjects to directly choose one of the two options, an incentive-compatible BDM procedure (Becker et al., 1964) was used to elicit subjects' preferred *switch point* – defined as the lowest secure payment for quitting so that they would prefer quitting over continuing.[13] The higher this requested minimum payment for quitting, the higher was the chance that they would continue and vice versa. The BDM was implemented in a purposely understandable and intuitive way, see Appendix B.

Using a BDM is appealing in this context for multiple reasons: First and foremost, subjects' switch points allow us to compute their ex-ante desired probability of continuing, which can be used as a measurement of persistence (see Section 3). Moreover, having a continuous measure of persistence allows us to explore if people's continuation decisions are a good predictor of their future performance, and if there are gender differences therein (see Section 5). Finally, conditional on a reported switch point, it is random who actually continues and who quits in the experiment. This allows us to compute the counterfactual earnings of subjects who continued, had they quit (see Appendix D).

2.1 Implementation

The experiment was implemented using Qualtrics code programmed by the author, and subjects made decisions on a computer. Roughly one third of all sessions was conducted in the EBEL laboratory at the University of California, Santa Barbara, in February and March of 2020. Due to

[12] For example, if a subject assessed the probability of passing the second test to be 70% after seeing their card, they later faced a lottery that paid \$20 with a chance of 70%, and \$0 with a chance of 30%. Recall that at the time when beliefs were elicited, subjects were not informed about what would happen in later parts of the experiment, and thus did not have incentives to report a higher belief of passing the future test in order to encounter a lottery with more favorable odds. Note that it was not deceptive to tell subjects that they would maximize their chance of winning \$20 by always reporting their true beliefs if Part 2 was drawn for payment.

[13] The interpretation in Part 4 is analogous to this, i.e., the switch point in Part 4 corresponds to the lowest minimum payment that subjects would prefer over the lottery.

the Covid-19 pandemic, the data collection had to be paused and was eventually moved online. The remaining sessions were conducted over Zoom in the summer of 2020. All features of the experiment were kept as similar as possible between in-person and Zoom sessions. Instructions were displayed on slides on the screen and read out loud by the experimenter in both in-person and Zoom sessions. Subjects were asked to keep their video turned on throughout the experiment in Zoom sessions. To preserve anonymity, the name of subjects in Zoom sessions was changed to numbers before admitting participants from the waiting room. Subjects then received a link to the experiment in the Zoom chat, and stayed in the Zoom meeting throughout the experiment.

All subjects were recruited from the EBEL subject pool using the Online Recruitment System for Economic Experiments (ORSEE) recruiting software (Greiner, 2015). Subjects signed up to participate in an experiment "on the economics of decision making," and gender was neither mentioned during the recruitment process nor in the instructions. The same number of men and women were invited to each session, so the gender composition of each session was roughly balanced. Subjects self-reported their gender identity in a survey at the end of the experiment, see Appendix B. Payments were made in cash at the end of in-person sessions, and via Venmo within 24 hours following Zoom sessions. Experimental sessions lasted around 80 minutes, and average payments were approximately $18 (with a minimum payment of $5 and a maximum payment of $27).

3 Experimental Results

3.1 Data overview

Sample. A total of 205 subjects participated in the experiment, out of which 102 identified as *Male*, and 103 identified as *Female*. This sample excludes participants that reported *Other* as their gender identity or had comprehension issues in the experiment.[14] Of this sample, 94 subjects (43 men and 51 women) were assigned to the *Baseline* treatment, and 111 (59 men and 52 women) were assigned to the *AlwaysInfo* treatment.

[14]Six subjects reported *Other* as their gender identity. Subjects had to answer all comprehension questions correctly to move on. A shortcoming of the experimental software written by the author is that one cannot identify subjects that needed multiple attempts to answer all comprehension questions correctly. Instead, a survey question at the end asked subjects to self-report if they "understood all instructions in this experiment," and if not, to explain what was not clear. 15 female and 16 male subjects indicated that "not everything was clear," and most of them reported comprehension issues associated with the BDM. These 31 subjects were excluded from the analysis.

As Table 1 shows, men and women in the *Baseline* sample differ along a few dimensions. Men were significantly more likely to pass the first IQ test ($p = 0.003$), and on average could solve almost one more question of the seven questions on the test correctly ($p = 0.007$). In terms of self-reported characteristics, women on average reported a slightly higher GPA than men ($p = 0.004$).[15] Furthermore, while the share of subjects who reported a STEM field or Economics/Accounting as their major or intended major is directionally higher for men than for women, these differences are not statistically significant. To account for these gender differences in self-reported characteristics, unless otherwise noted, regressions in this paper control for all self-reported characteristics listed in Table 1, as well as a dummy variable for whether sessions were conducted in person or over Zoom.

Gender differences in persistence in the raw data. As a measurement of persistence, a subject's ex-ante desired probability of continuing is used, which can be derived directly from their reported switch point in Part 3 of the experiment.[16] To get a first intuition for gender differences in persistence in the raw data, Figure 3 shows an empirical CDF of subjects' continuation probabilities in the *Baseline* treatment, separately for men and women. In the raw data, i.e., before controlling for subjects' performance and the feedback they received, men's empirical CDF first-order stochastically dominates the empirical CDF of women. The vertical lines in Figure 3 depict that men's average continuation probability in the *Baseline* treatment is 61%, while for women it is only 49%, thus constituting a gender gap in persistence of about 12 percentage points in the raw data. This does not imply that there are gender differences in persistence, however, as the distribution of performance on the first IQ test is substantially different for men and women, see Table 1. To resolve this confound, in what follows regressions are presented to study if there are gender differences in persistence when controlling for subjects' performance, the feedback they received, as well as self-reported characteristics.

3.2 Formal analysis of gender differences in persistence

Aggregate results. To explore more formally if there is a gender gap in persistence, Table 2 presents OLS estimates of the probability to continue in the *Baseline* treatment. As a reference,

[15] One female subject reported a GPA of 362. This was considered a typo and was re-coded as 3.62.

[16] The BDM involves 23 questions, see Appendix B. A subject's ex-ante probability of continuing increases linearly with their reported switch point. More specifically, $SwitchPoint_i/23$ is the probability that subject i continues.

column (1) shows that absent of controls, women are about 12 percentage points less likely to continue than men, corresponding to the average gender gap in the raw data shown in Figure 3. When controlling for past performance (measured as scores on the first IQ test), the feedback that subjects received, as well as self-reported characteristics, the estimated gender gap in persistence amounts to roughly 10 percentage points ($p = 0.016$), see column (2). Given that the average probability of continuing for men who received positive feedback is 68% in the *Baseline*, women are one average about 15% less likely to continue than men. It is worth noting that relative to men who received positive feedback, the estimated effect sizes of "being female" and of negative feedback on persistence are similar. Put differently, women who received positive feedback are on average not more likely to continue than men who received negative feedback.

This estimated gap is robust when controlling for whether subjects passed the first IQ test (column 3) or when allowing for an interaction of the *Female* dummy with the test score (column 4). To put the estimated gender gap of this experiment into perspective, note that it is similar in magnitude to some studies that are using naturally occurring data.[17] That being said, gender differences in persistence naturally vary greatly by context.

Result 1. *In the Baseline treatment, women are on average about 10 percentage points (15%) less likely to continue than men when controlling for their past performance, the feedback they received, as well as self-reported characteristics.*

Heterogeneity by feedback and first IQ test performance. Does the effect of receiving negative versus positive feedback vary by gender in this controlled environment? As column (5) of Table A1 shows, this idea is not supported in the data, as the interaction effect of the *Female* dummy with the negative feedback dummy is statistically insignificant. In other words, negative feedback does not appear to have a more discouraging effect on women's decision to persist than it has for men. Similarly, positive feedback does not appear to have a more encouraging effect on men than on women. Directionally, men are more likely to continue regardless of what feedback they received. The estimated gender gap in persistence among those who received positive feedback is

[17] For example, Buser and Yuan (2019) find a 10-20 percentage point gender gap in participating again in a math olympiad after missing the cutoff to the second round previously. Pereda et al. (2020) document a 5.9 gender gap in the likelihood of re-submitting an article to an economics conference after a previous rejection. Wasserman (2021) find that women are about 10 percentage points (or 50%) less likely than men to re-run for office after having lost an election previously.

15 percentage points ($p = 0.012$), and about twice as big as the gender gap in response to negative feedback, which is only about 7 percentage points and not statistically significant ($p = 0.236$). Section 3.3 will discuss that this may in part be driven by gender differences in feedback avoidance in response to positive feedback. The gender gap in persistence is further driven by subjects who failed the first IQ test. Men who performed poorly on the first IQ test are thus over-represented in the sample that continues, relative to women who performed poorly. Details and implications of this adverse selection of men will be discussed in Section 5.

3.3 Mechanisms behind the gender gap in persistence

What can explain this documented gender gap in persistence? The experimental design allows us to explore how beliefs, preferences for additional feedback, and risk aversion shape persistence. In what follows, I will argue that a large share of the gender gap in persistence is attributable to beliefs and preferences for feedback, while the role of risk preferences is negligible.

Mechanism 1: Beliefs about passing the future IQ test. If women are less confident about their future performance, and thus expect lower returns from persisting than men, it is rational for them to quit more often, all else equal. Implications are different, however, if men are initially more confident, if they extrapolate differently from the past when forming beliefs about the future, or if gender differences in beliefs arise or are reinforced in response to feedback. My design can differentiate these channels. When analyzing beliefs, data are pooled across treatments to increase power.[18]

Initial beliefs before feedback. Compared to men who performed equally well on the first IQ test, women are initially less confident not only about having passed the first test (consistent with much of the literature), but also about passing the future IQ test, see columns (1) and (2) of Panel A, Table 5. If anything, the gender gap in confidence about one's future performance is directionally even more pronounced than about one's past performance (7 versus 10 percentage points). This is worth noting as the literature has largely focused on the gender gap in confidence about people's previous performance, despite the fact that a range of economic decisions, including

[18] Recall that no design elements differ across treatments until after the belief elicitation, see Section 2.

the decision to persist, are arguably a function of beliefs about the future rather than the past.[19] To be as confident as men about passing the future IQ test, women on average need to score more than one standard deviation higher on the first IQ test.

Interestingly, men and women make different inferences about their future performance, given their beliefs about the past. As column (3) of Panel (A) in Table 5 shows, women are less confident about passing the future IQ test *even when* controlling for beliefs about having passed the first test. Put differently, even when comparing men and women that performed equally well *and* are similarly confident about having passed the first IQ test, men are on average substantially more confident about passing the future IQ test. Figure 4 illustrates this gender gap in subjects' projections of their future performance, given their beliefs about their past. One explanation for this gap could be that men interpret previous failures as less predictive, or previous successes as more predictive of their future success than women. Notably, Table A2 shows that while past test scores are a strong predictor of passing the future IQ test (column 1), there is no significant gender difference in how predictive past scores are of people's future performance (column 2).[20] But when men discount how predictive previous failures are, or over-weigh how predictive previous successes are, relative to women, this allows them to be more confident about their future, and consequently men have higher perceived returns from persisting than women who performed equally well.

Updating in response to feedback. If men respond stronger to positive feedback, or if women respond stronger to negative feedback when updating about their future performance, this could further amplify the gender gap in initial confidence documented above. To explore this possibility, note that Bayesian updating in this setting can be written in log-form as

$$ln\left(\frac{p}{1-p}\right) = ln\left(\frac{p_0}{1-p_0}\right) + \mathbf{1}\{pos.\} * ln\left(\frac{\phi}{1-\phi}\right) + \mathbf{1}\{neg.\} * ln\left(\frac{1-\phi}{\phi}\right), \quad (1)$$

[19] See Section 1 for a review of the literature on gender differences in beliefs. If the finding that the gender gap in confidence with regard to future events is directionally bigger than with respect to past events also applies to other settings, then beliefs might explain a larger fraction of gender differences in behavior (e.g., the willingness to compete) than previously thought. For example, to control for beliefs when estimating the gender gap in choosing a tournament payment scheme, Niederle and Vesterlund (2007) use subjects' guesses of their *past* tournament performance, not their predicted guesses of their *future* tournament performance.

[20] As the subset of subjects for which their future performance is observed is selected in the experiment, Heckman regressions are used Table A2, discussed in more detail in Section 5. In the field study where future performance is observed for all participants, there is no gender difference in how predictive the past performance is of the future either, see Section 4 and Table A9.

where p denotes the posterior belief, p_0 denotes the prior belief, $\mathbf{1}\{pos.\}$ and $\mathbf{1}\{neg.\}$ denote indicator functions of receiving positive or negative feedback, respectively; ϕ denotes the probability that the cards conveying the feedback reveal the true state, e.g., of having passed or failed the first IQ test.[21] With this in mind, linear regressions of the following form can be estimated (Mobius et al., 2014):

$$ln\left(\frac{p_i}{1-p_i}\right) = \alpha * ln\left(\frac{p_{0i}}{1-p_{0i}}\right) + \beta_p * \mathbf{1}\{pos.\} * ln\left(\frac{\phi}{1-\phi}\right) + \beta_n * \mathbf{1}\{neg.\} * ln\left(\frac{1-\phi}{\phi}\right) + \epsilon_i. \tag{2}$$

Note that for a perfect Bayesian agent, $\alpha = \beta_p = \beta_n = 1$. Further, and $\beta_p = \beta_n$ indicates putting the same weight on positive and negative feedback when updating.[22] Gender differences in updating can be estimated by looking at the interaction of the β coefficients and a female dummy.

As Table A3 shows, however, men and women on average do not update significantly differently in response to feedback, see columns (2) and (4). Specifically, while there is some over-reaction to negative feedback when updating on their past performance, men and women place similar weights on negative as well as positive feedback when updating about their future performance. This suggests that how people's beliefs respond to feedback plays no important role for gender differences in persistence.

Beliefs after feedback and their effect on persistence. After having received performance feedback, the gender gap in beliefs about people's future performance remains, but the gender gap in beliefs about having passed the first test closes, as columns (1) and (2) of panel B in Table 3 show. Controlling for past test scores and beliefs about having passed the first IQ test, men are on average roughly 7 percentage points more confident about passing the future IQ test than women ($p = 0.005$), see column (3). That is, the insight that men and women make different inferences about their future given their past performance holds even after people receive feedback. Figure A1 illustrates that this gender gap in how people make make inferences about

[21] By design, $\phi = \frac{2}{3}$ when updating about the state of having passed the first IQ test. There is no universally true value of ϕ when updating about the future, however: Depending on the (unobserved) beliefs that people may hold about how informative their past performance – and thus the past feedback – is for their future performance, it might be rational for different people to put different weights on the positive and negative feedback. That being said, one can still assess whether there is a gender gap in how much weight subjects put on positive and negative feedback when updating beliefs about their future performance.

[22] Similarly, β_p or β_n bigger (smaller) than 1 would indicate over-reaction (under-reaction) to the positive or negative feedback, respectively; $\alpha < 1$ would indicate base-rate neglect and $\alpha < 1$ would imply that subjects are updating too conservatively.

their future performance, given their beliefs about the past, emerges following both positive and negative feedback.

How much of the gender gap in persistence can be attributed to gender differences in beliefs about one's future performance? Recall that in the *Baseline* treatment, the gender gap in persistence amounts to about 10 percentage points. When controlling for subjects' posterior beliefs of passing the future IQ test – the beliefs that subjects report after having received feedback and directly before their continuation decision – this gap drops to 6.7 percentage points ($p = 0.072$), see column (2) of Table A4. While this estimate is not statistically distinguishable from the "original" gender gap presented in column (1), this suggests that roughly one-third of the gender gap in persistence can be explained by gender differences in beliefs about performing well in the future.

Result 2. *Women are less confident about passing the future IQ test than men who performed equally well on the first test and who are similarly confident about their past performance – both before and after getting feedback. This gender gap in beliefs accounts for roughly one-third of the gender gap in persistence.*

Mechanism 2: Avoiding and seeking additional feedback. Persisting in stratified careers such as corporate management or academia naturally involves exposure to frequent performance feedback. If women dislike this exposure more so than men, or if men enjoy receiving additional feedback more so than women, this could help explain the under-representation of women in these careers. To explore this possibility, I compare subjects' behavior across the two treatments of my experiment. Recall that if subjects are more (less) likely to continue in the *AlwaysInfo* treatment than in the *Baseline*, this can be interpreted as evidence supporting the idea that feedback avoidance (seeking) affects persistence. That is, if the estimated treatment effect is positive (negative), this indicates feedback avoidance (seeking). Figure 5 compares average continuation probabilities for men and women between the two treatments. In the raw data, the gender gap in persistence shrinks substantially in the *AlwaysInfo* treatment relative to the *Baseline*. This is driven by two forces: On average, women avoid, and men seek exposure to the additional feedback of learning if they really passed or failed the first IQ test.

One caveat of analyzing the *AlwaysInfo* treatment effect more formally is that although subjects were randomized into treatments, not all observables are perfectly balanced across the two

treatments, see Table A6. In particular, subjects who got assigned to the *AlwaysInfo* treatment on average reported a slightly higher GPA, and women (but not men) who got assigned to the *AlwaysInfo* were more likely to report a non-white race identity, and to report US citizenship, than women who got assigned to the *Baseline* treatment. When estimating the treatment effect, controls for these self-reported characteristics are included. In addition, controls for the beliefs that subjects reported after getting feedback are included.[23]

Aggregate estimates of the *AlwaysInfo* treatment effect are directionally consistent with the idea that women avoid additional feedback while men seek it, see column (1) of Table 4: Men are on average 6.5 percentage points less likely to continue in the *AlwaysInfo* treatment than the *Baseline* ($p = 0.080$), which suggests that learning if they really passed or failed the first IQ test is a motive for them to continue in the *Baseline*, i.e., the prospect of getting additional feedback potentially makes persisting more attractive for men. For women, the estimated *AlwaysInfo* effect is directionally consistent with feedback avoidance, but not significantly different from zero in the aggregate sample ($p = 0.433$).

It is possible that the estimates presented in column (1) of Table 4 mask some heterogeneity of preferences for additional feedback exposure. For example, subjects who got negative feedback might want to avoid learning their test outcome hoping that the negative feedback was wrong, or they might prefer finding out their test result to prove the negative feedback wrong, and there could be gender differences therein. Perhaps surprisingly, columns (2)-(3) show that women on average engage in feedback avoidance if they received positive feedback ($p = 0.031$), but not if they received negative feedback. One possible explanation for this could be that women might shy away from learning their true test result if they actually failed but received positive (wrong) feedback. This would suggest that some women prefer not to go after opportunities in order to avoid finding out that they are not as talented as they had hoped.[24] In contrast, men tend to exhibit a preference for learning their true test result especially when they received negative feedback ($p = 0.091$), but estimates are not significant following positive feedback. Interestingly, women who under-reacted to

[23] Beliefs about subjects' past performance are controlled for to account for potential gender differences in expectations about what feedback they would receive upon continuing. Beliefs about their future performance are controlled for in order to avoid "double-counting" of the confidence mechanism described above.

[24] Indeed, women who failed but received positive feedback are on average 15 percentage points more likely to continue in the *AlwaysInfo* than in the *Baseline* treatment. But as the sample size per cell would be very small when looking at gender differences by treatment, positive versus negative feedback and separately by having passed or failed the first IQ test, exploring this more formally is not possible with the data at hand.

feedback when updating about the past test are especially prone to engage in feedback avoidance, as column (3) of Table A7 shows. Contrary to this, no significant treatment effects are found when looking at the sub-groups of subjects who over-reacted (column 2) or updated too optimistically or too pessimistically in response to the feedback (columns 4 and 5).

What fraction of the gender gap in persistence is attributable to gender differences in feedback avoidance and feedback seeking? When weighting all estimates of column (1) in Table 4 by the fraction of men and women in the *Baseline* treatment, 45.8% of the gender gap would be explained by preferences for additional feedback.[25] But since the estimated treatment effect on women at the aggregate is not statistically different from zero, a more conservative approach would be to only count the effect on men, while considering the effect on women to be zero. This more conservative back-of-the-envelope calculation yields that 28.8% of the gender gap can be explained by preferences for additional feedback. That being said, at the aggregate the estimated *AlwaysInfo* treatment effect is not significant at the 5% level, thus caution is warranted when interpreting this estimate.

Result 3. *There is suggestive evidence that on average men seek, while women avoid exposure to additional feedback. These preferences account for roughly* 30% *of the gender gap in persistence when considering estimates that are significant at least at the* 10% *confidence level.*

Ultimately, it is worth noting that even though subjects do not learn if they passed or failed the future IQ test if they continue, the experience of taking this additional test may convey some "internal feedback." For example, women might have a stronger distaste than men to have the experience of not knowing how to solve the questions on the future IQ test. With this in mind, the estimated *AlwaysInfo* treatment effect could be regarded as a lower bound of how much of the gender gap in persistence is attributable to gender differences in preferences for feedback avoidance.[26]

Mechanism 3: Risk preferences. Pursuing a stratified career is a risky choice if doing so is only financially rewarding when accomplishing a high performance, while dropping out involves a

[25] In the *Baseline* treatment, 46% of subjects are men and 54% are women. Thus, the gender gap in the *AlwaysInfo* treatment is $6.5 * 0.46 + 3.2 * 0.54 = 4.72$ percentage points smaller than in the *Baseline*, where the estimated gender gap in persistence is 10.3 percentage points. Thus, $4.72/10.3 = 45.8\%$ of the gap in persistence can be explained by gender differences in feedback preferences.

[26] This holds under the assumption that gender differences in avoiding or seeking "internal feedback" (e.g., the feedback that is conveyed by experiencing how challenging the IQ test is) go into the same direction as avoiding or seeking the additional feedback of learning if their result on the first IQ test.

secure compensation. This feature was mimicked in the experiment: Continuing only pays off if subjects pass the future IQ test, while quitting guarantees a minimum payment. To investigate whether gender differences in risk preferences affect the gender gap in persistence, Part 4 of the experiment allows us to estimate risk parameters, see Appendix A for details.

As Table A5 shows, women are on average not more risk averse than men in this experiment, which is consistent with some previous studies.[27] This result is obtained when estimating CRRA or CARA risk parameters, with or without controls for beliefs and performance. It is therefore not surprising that the estimated gender gap in persistence is essentially unaffected when controlling for risk parameters, see columns (3)-(6) of Table A4. This suggests that risk preferences do not constitute an important channel for explaining gender differences in persistence in this setting.

Other possible mechanisms. The preceding analysis suggests that gender differences in beliefs and preferences for additional feedback together account for roughly two-thirds of the gender gap in persistence, while the role of risk preferences is negligible. This raises the question of what can explain the remaining third. One could speculate that gender differences in seeking challenges are at play, since continuing involves a more challenging task than quitting. This is possible, but it would be somewhat inconsistent with the literature.[28]

Differences in how men and women are being socialized might affect their decision to persist beyond channels that this experiment can identify. If anything, to hint at whether the gender gap in persistence might be more pronounced for subjects of a traditional family background or those with conservative attitudes, a small set of questions was included in the end survey (see Supplementary Appendix, Screenshots 108-112). Table A8 presents the corresponding results. Column (2) indicates that the gender gap in persistence appears to be especially pronounced (at about 17 percentage

[27] The evidence on gender differences in risk preferences is mixed. Niederle (2014) points out that while some studies do find that women are more averse to take risks, these differences are often small in magnitude, and largely vary by elicitation methods. She also notes that the literature on gender differences in risk aversion might suffer from a publication bias. Eckel and Grossman (2008) review 13 lab and field economics experiments, out of which 8 find women to be more risk averse than men at the 10% confidence level or higher, while 5 either find no gender difference in risk taking or are less conclusive. They stress that many of these studies fail to account for important controls such as wealth. Croson and Gneezy (2009) review 10 economics experiments and conclude that while 8 of them document women to be more risk averse than men, in 2 of them the evidence is mixed. Byrnes et al. (1999) conduct a meta-analysis of 150 psychology studies and conclude that in most studies, men are found to be significantly more likely to take risks than women.

[28] Niederle and Yestrumskas (2008) experimentally study gender differences in seeking challenges. They find that a gender gap in choosing a challenging versus an easy mazes task closes when subjects receive information about whether the challenging task is likely payoff-optimal for them. The authors interpret this as evidence against the idea that there are gender differences in preferences for the characteristics of the hard versus the easy task.

points, $p = 0.006$) among subjects that reported more conservative attitudes on gender roles.[29] Furthermore, the gap is marginally smaller (and insignificant) for the subgroup of subjects that consider their parents' occupations to be typical for men/women of their generation, and slightly bigger for subjects that reported that their father worked more hours for pay than their mother during their childhood, see columns (2) and (3). It is beyond the scope of this paper to explore more thoroughly how attitudes and socialization may affect persistence.

4 Field Study and Outside Validity of Belief Mechanism

One novel finding of the laboratory experiment is that men tend to make more optimistic projections of their future performance even when compared to women who are similarly confident about their past performance. As this finding has several implications, it is important to understand if these gender differences in how people extrapolate from the past when forming beliefs about their future also arise in more natural environments and in domains other than the Raven's IQ test. Examining the external validity of the belief formation patterns detected in the laboratory requires a setting where people perform a challenging task at least twice, similar to the two IQ tests in the experiment. That way, beliefs about both one's past and future performance can be elicited. Furthermore, there needs to be some uncertainty about people's past performance – just like in the experiment where subjects do not know for sure if they passed or failed the first IQ test.

With this in mind, the following field study was conducted. Students at an introductory economics course at UCSB ("Econ 1") were asked to participate in a short research study on beliefs about future success. Econ 1 is general education course and over two-thirds of all students who take this course do *not* end up majoring in economics. The course involves multiple midterm exams, and students were asked two questions after completing the first midterm exam, but before learning how many points they earned on the first exam. Details of the course, subject pool and implementation are provided in Appendix C. The following two questions were asked:

1. How likely (out of 100) do you think it is that you answered at least 12 of 15 questions correctly on the first Econ 1 midterm quiz?

[29] The line for this category was drawn arbitrarily to capture roughly one half of the sample. It includes subjects who either (strongly) disagreed that "women should pay their own way on dates," or who did not strongly disagreed that "wives with a family have no time for outside employment."

2. How likely (out of 100) do you think it is that you will answer at least 12 of 15 questions correctly on the second Econ 1 midterm quiz?

Note that each midterm exam consisted of 15 questions (and in preceding years roughly half of all students could solve at least 12 questions correctly). These questions were chosen to be as similar as possible to the elicited beliefs in the laboratory experiment. Students' answers to these questions were then matched with their actual exam scores. This allows us to directly test if Result 2 from the experiment replicates in the field: Do men hold more confident beliefs about their future performance *even when* compared to women who are similarly confident about their past performance *and* performed equally well?

Table 5 compares gender differences in beliefs about one's past and future performance in the laboratory experiment (panel A) and the field study (panel B). It is striking that the estimates from the laboratory replicate in the field not only qualitatively but also in terms of the effect size. In particular, column 3 shows that men are more confident than women to achieve a high score on the future midterm exam / IQ test even when holding constant their past performance *and* beliefs about their past performance. Figure A2 visualizes this gender gap in extrapolating from the past when forming beliefs about the future, and the pattern is qualitatively very similar in the experiment (a) and the field study (b). Furthermore, just as in the experiment, while men appear to perceive how predictive their past performance is of their future success differently than women, empirically there are no gender differences in how predictive past test results are of scoring above the cutoff on the future midterm exam, see Table A9.

In sum, the field study presents reassuring evidence that gender differences in how people interpret the past when forming beliefs about their future are not unique to the artificial environment of the lab, but might be more systematic. Considering that the decision to persist arguably depends on people's beliefs about their future performance rather than their past, this finding highlights the role of beliefs for understanding the under-representation of women in stratified careers.

Result 4. *The novel insight from the laboratory that men hold more confident beliefs about their future performance even when compared to women who are similarly confident about their past performance replicates remarkably well in a classroom field setting.*

5 Efficiency of the Different Self-selection of Men and Women

Do gender differences in persistence contribute to a gender gap in performance within organizations? This would be the case, for example, if performance feedback mechanisms deter high-performing women from continuing more so than men, or if they deter low-performing men less from continuing than women. A natural feature of the experiment is that people's past performance is not necessarily a perfect predictor of their future performance. That being said, gender differences in the efficiency of subjects' self-selection in the experiment can be assessed along two dimensions: First, does past performance predict continuation decisions? Second, do continuation decisions predict future performance?

In the experiment, men who persist are adversely selected relative to women when taking subjects' past performance as a measure of ability: In fact, the gender gap in persistence is entirely driven by subjects who failed the first IQ test, as Figure 6 visualizes. Men who failed are on average 15 percentage points more likely to continue than women who failed ($p = 0.035$); In contrast, among subjects who passed, the gender gap in persistence is negligible in magnitude and statistically indistinguishable from zero, see columns (2)-(3) of Table 6. Furthermore, when looking at the total sample, the marginal effect of having scored one standard deviation higher on the first IQ test on the probability of continuing is directionally about twice as big for women than it is for men, see column (4). As column (5) shows, these differences do not vary at the treatment level, however, which suggests that preferences for additional feedback do not affect how predictive men's and women's past performance is of their continuation decisions.

The fact that men who failed the first IQ test are more likely to self-select into continuing does not necessarily imply that women's continuation decisions better predict their future performance, however. This is because past test scores naturally are no perfect predictor of subjects' future test scores: The correlation coefficient between IQ test scores is 0.46 in the experiment. (To put this number in perspective, the correlation of the first two midterm exam scores is 0.41 in the classroom field study, where the performance on the future midterm exam is observed for all participants.) This further highlight the role of individuals' perceptions on how predictive their previous performance is of the future, as discussed in Sections 3 and 4.

When estimating if men's or women's continuation decisions better predict their future per-

formance, note that the sample of subjects who continue – and thus the sample for which the second IQ outcome can be observed – is selected. To account for this sample selection, Table A11 presents Heckman regressions to explore if the switch point in Part 3 of the experiment (which directly translates into the probability of continuing, see Section 2), is more predictive of the performance on the second IQ test for men or for women. Step 1 estimates the self-selection into continuing using subjects' switch points in Part 3 and Part 4 of the experiment.[30] Step 2 estimates what factors can predict the performance on the second IQ test. Columns (2) and (4) indicate that there is no significant gender difference in how predictive continuation probabilities are of subjects' likelihood of passing the second IQ test, or of their future test scores. Furthermore, note that once controlling for past performance, higher continuation probabilities are not associated with a significantly higher future performance, suggesting that the extent to which individuals' choices predict their future performance in this setting appears to be limited.

Summing up, while men who continued in the experiment are indeed adversely selected when taking the first test performance as a measure of ability, this does *not* imply, however, that the differential self-selection of men and women results in an adverse selection of men in terms of their future performance. To a large extent, this may be the case because the empirical relationship between past and future performance is naturally noisy. In addition, given that the sub-sample of subjects who continue in the experiment is positively selected, this study may be under-powered to detect gender differences in how predictive subjects' continuation decisions are of their future performance.

Result 5. *Men are adversely selected into persisting when taking past performance as a measure of ability. This does not imply, however, that women's continuation decisions are a better predictor of their future performance.*

A related question is whether subjects' continuation decisions maximized their earnings in the experiment. Appendix D documents that while most subjects who continued did achieve small positive returns relative to their counterfactual earnings for quitting, those who fail the second IQ test often face substantial opportunity costs; There is no gender difference therein.

[30] Recall that the latter represents subjects' preference for continuing, stripped from all features except payoffs and risk. The correlation of the two switch points in the aggregate sample is $\rho = 0.579$.

6 Discussion

Using a controlled laboratory experiment, I detect a novel channel that may contribute to the under-representation of women in stratified professions: gender differences in persistence in response to performance feedback. My results indicate that this gender gap in behavior is largely attributable to two novel mechanisms: differences in how men and women form beliefs about their future given past experiences, and differences in preferences for exposure to additional feedback. In a classroom field study designed to test the outside validity of the belief formation patterns documented in the laboratory, my findings replicate remarkably well.

It is worth noting that a gender gap in persistence was detected in the experiment when looking solely at a one-time decision in response to a one-time provision of feedback, which suggests that the compound effect of feedback mechanisms in stratified careers where people are frequently exposed to performance feedback, and have to decide between persisting and dropping out along many steps of the career ladder, may be even larger. Furthermore, the study population in this paper may already be positively selected in terms of their persistence (UCSB students), and it is conceivable that gender differences in persistence are perhaps even more pronounced in less selected populations.

From a policy perspective, an important implication of my results is that women may be under-represented in stratified professions in part because men and women tend to have different perceptions of how predictive their previous performance is for being successful on a given career path later on. How can institutions address this gender disparity in perceptions, and ultimately persistence? Performance feedback is usually given on people's past performance, leaving it up for interpretation what can be inferred from this feedback about one's future performance. If people instead received information on how predictable (or rather unpredictable) people's past performance tends to be of their future success in these careers, it is conceivable that the gender gap in perceptions could be mitigated. Such an information intervention might also improve the efficiency of people's self-sorting into persisting versus dropping out of stratified careers. Furthermore, one could speculate that the mechanisms I identify point to the possibility that environments that involve a long-term exposure to feedback before decisions to persist are made (e.g., school tracks where sorting into specialized tracks takes place at a later time) might reduce the gender gap in persistence

twofold: First, by giving people an opportunity to learn that one's performance can improve over time; And second, by mitigating how much people care about an exposure to additional feedback.

I see several directions in which this research agenda can be advanced. It could be interesting to explore if the gender gap in persistence is amplified by factors that I intentionally shut down in this experiment, but that might be prevalent in natural settings such as the workplace. For example, the role of social signaling and an urge to comply with social gender norms was probably limited in the experiment, as subjects' decisions could not be directly observed by others (recall that decisions were made on a computer and data were collected in an anonymized way). It is conceivable that the feedback avoidance and belief mechanisms are exacerbated if outcomes can be observed by other people. Related to this, while Ludwig et al. (2017) find that women report less confident beliefs about their past performance when their actual performance is observed by others, Buser et al. (2021) find that the gender gap in the willingness to enter tournaments is unaffected by whether tournament entry decisions are publicly observable. It further stands to reason that competition or feedback that entails social comparison exacerbates the gender gap in persistence, considering that we have ample evidence of gender differences in the willingness to compete. In addition, it could be interesting to explore if the mechanisms driving persistence vary by task domains, especially in gender-congruent versus gender-incongruent domains (which would be consistent with Coffman, 2014; Bordalo et al., 2019; Coffman et al., 2019, 2021).

Finally, aside from the gender dimension, my findings allude to the broad research topic of how people form beliefs about future events, and to what extent these beliefs are adapted and potentially distorted in a self-serving way in response to (ego-relevant) information and feedback on past events. Broadly speaking, we seem to lack a systematic understanding of how people extrapolate from information on past events (e.g., grades in school) when forming beliefs about future events (e.g., the chance of succeeding in a particular job). Yet, how people form such "mental models" on how these events are connected is of great relevance for understanding economic behavior. And while my findings suggest that men and women might apply different models when interpreting their own past performance to form beliefs about their own future performance, it is also possible that they apply different models to other people depending on their attributes. For example, committees might interpret past accomplishments differently when evaluating male or female candidates. Such behavior could point to additional channels fueling the under-representation

of women in stratified professions.

Maria Kogelnik, Univeristy of Amsterdam & Tinbergen Institute

Table 1: Summary Statistics, Baseline Treatment.

	Men	Women	p-value
IQ Test Performance			
Avg. Score 1. Test	4.40	3.63	0.007
Passed 1. Test	0.60	0.29	0.003
Self-reported Characteristics			
Average GPA	3.09	3.67	0.004
STEM Major	0.42	0.31	0.294
Econ / Accounting Major	0.21	0.10	0.133
Non-White	0.70	0.84	0.093
English First Language	0.79	0.71	0.350
US Citizen	0.81	0.78	0.723
Observations			
Baseline Treatment	43	51	-
AlwaysInfo Treatment	59	52	-
Total	102	103	-

Notes: The panels on IQ test performance and self-reported characteristics show data of the *Baseline* treatment. P-values refer to a Wilcoxon-Mann-Whitney Test testing the hypothesis that the distribution of a characteristic is the same for men and women.

Table 2: OLS Estimates, Probability of Continuing, Baseline Treatment.

	Probability of Continuing			
	(1)	(2)	(3)	(4)
Female	-0.120***	-0.103**	-0.0883**	-0.100**
	(0.0424)	(0.0422)	(0.0405)	(0.0413)
Z-Score 1. IQ Test		0.0601***	0.00330	0.0378*
		(0.0151)	(0.0267)	(0.0193)
Negative Feedback		-0.106***	-0.0901***	-0.103***
		(0.0281)	(0.0281)	(0.0281)
Passed 1. IQ Test			0.150***	
			(0.0540)	
Female * Z-Score 1. IQ Test				0.0487*
				(0.0275)
Additional Controls	-	✓	✓	✓
Mean Reference Group	0.61	0.68	0.55	0.68
Observations Baseline	94	94	94	94
Observations Total	205	205	205	205

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This table is an abbreviation of Table A1, displaying only estimates that are relevant to the *Baseline* treatment. Robust standard errors in parentheses. Additional controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity). Constants not displayed. The mean of the reference group shows the average probability of continuing for all men (column 1), men who received positive feedback (columns 2 and 4), and men who received positive feedback but failed the first IQ test (column 3) in the *Baseline*.

Table 3: OLS Estimates of Initial Beliefs and Beliefs After Feedback.

	Belief: Passed 1. IQ Test	Belief: Will Pass 2. IQ Test	
	(1)	(2)	(3)
Panel A: Initial Beliefs (Priors)			
Female	-6.909**	-9.584***	-4.993**
	(3.362)	(3.070)	(2.140)
Z-Score 1. IQ Test	10.92***	7.903***	0.645
	(1.621)	(1.555)	(1.174)
Prior 1. IQ Test			0.665***
			(0.0510)
Additional Controls	✓	✓	✓
Mean Reference Group	55.57	66.61	66.61
Observations	205	205	205
Panel B: Beliefs After Feedback (Posteriors)			
Female	-1.196	-7.561**	-6.802***
	(3.256)	(3.126)	(2.405)
Z-Score 1. IQ Test	10.80***	8.760***	1.905
	(1.578)	(1.615)	(1.458)
Neg. Feedback	-32.98***	-18.55***	2.389
	(3.276)	(3.079)	(2.563)
Posterior 1. IQ Test			0.635***
			(0.0587)
Additional Controls	✓	✓	✓
Mean Reference Group	69.94	73.69	73.69
Observations	205	205	205

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Constants not displayed. Additional controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity). Data from both treatments combined. The mean of the reference group in panel (A) refers to men's average initial beliefs, and in panel (B) refers to men's average beliefs after feedback, conditional on having received positive feedback.

Table 4: AlwaysInfo Treatment Effect.

	Probability of Continuing		
	(1) **All**	(2) Positive Feedback	(3) Negative Feedback
Estimated Treatment Effect			
Men	-0.065*	-0.049	-0.110*
	(0.037)	(0.055)	(0.065)
Women	0.032	0.128**	-0.060
	(0.041)	(0.058)	(0.061)
Controlling for Beliefs	✓	✓	✓
Additional Controls	✓	✓	✓
$H_0 : \text{TME}_{\text{Men}} = \text{TME}_{\text{Women}}$	0.051	0.010	0.478
Observations	205	97	108

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This table presents estimates of the impact of the *AlwaysInfo* treatment on the probability of continuing relative to the *Baseline* treatment, separately for men and women. Positive (negative) point estimates correspond to feedback avoidance (feedback seeking). Controlling for scores on the first IQ test and beliefs about past and future IQ test performance reported after feedback. (Columns (2)-(3) further control for having passed or failed.) Additional controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity). The second last line reports p-values testing the hypothesis that the treatment effect is the same for men and women.

Table 5: OLS Estimates of Initial Beliefs - Laboratory vs. Field Study.

	Belief: Passed 1. IQ Test	Belief: Will Pass Future IQ Test	
	(1)	(2)	(3)
Panel A: Laboratory Experiment			
Female	-6.909**	-9.584***	-4.993**
	(3.362)	(3.070)	(2.140)
Z-Score 1. IQ Test	10.92***	7.903***	0.645
	(1.621)	(1.555)	(1.174)
Prior 1. IQ Test			0.665***
			(0.0510)
Additional Controls	✓	✓	✓
Mean Reference Group	55.57	66.61	66.61
Observations	205	205	205
	Belief: 1. Midterm	Belief: Future Midterm	
Panel B: Classroom Field Study			
Female	-6.498***	-7.744***	-4.302***
	(2.199))	(1.738)	(1.320)
Z-Score 1. Exam	7.926***	1.820**	-2.380**
	(1.343)	(0.920)	(1.013)
Prior 1. Exam			0.530***
			(0.0541)
Mean Reference Group	78.09	81.18	81.18
Observations	368	368	368

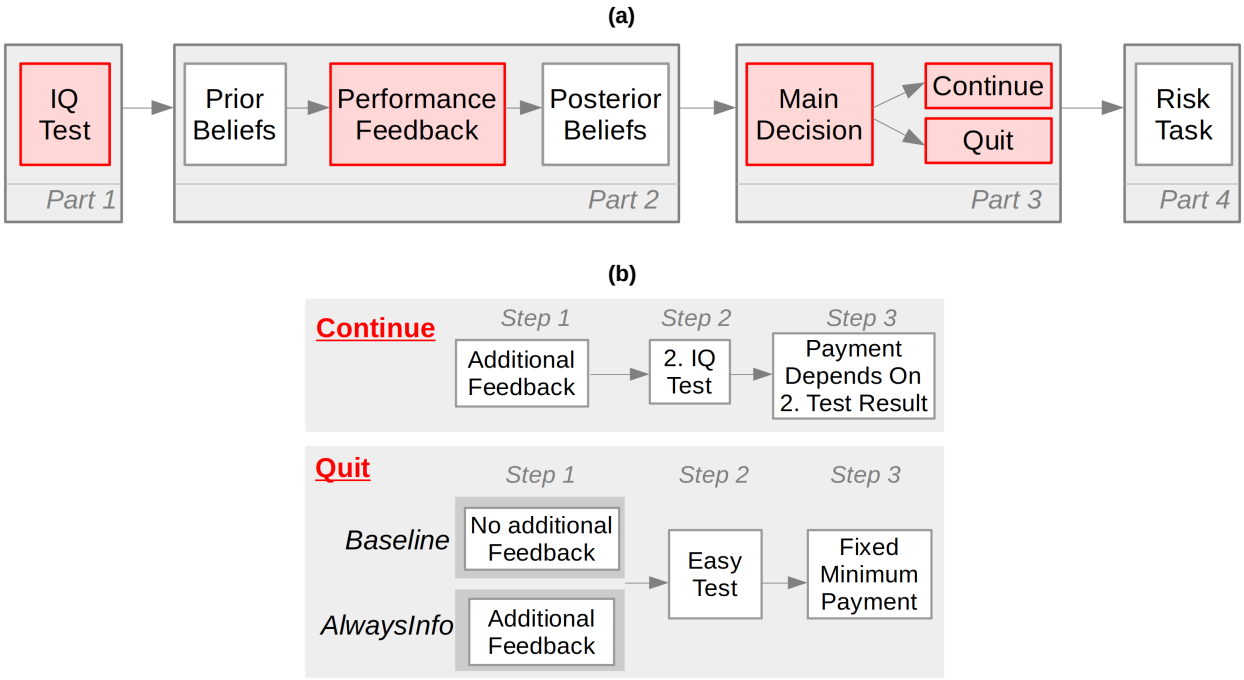
Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Constants not displayed. Column (1) presents estimates of people's initial beliefs (before receiving feedback) about their past performance. Columns (2) and (3) present estimates of people's initial beliefs about their future performance. The mean of the reference group refers to men's average initial beliefs. Panel A reports initial beliefs in the laboratory experiment and is identical to Panel A in Table 3. Additional controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity). Data from both treatments combined. Panel B reports beliefs of the classroom field study, controlling for self-reported race identity.

Table 6: Probability of Continuing by 1. IQ Test Performance, Baseline Treatment.

	All	Passed	Failed	**All**	
	(1)	(2)	(3)	(4)	(5)
Female	-0.103**	-0.00738	-0.153**	-0.100**	-0.0987**
	(0.0422)	(0.0514)	(0.0713)	(0.0413)	(0.0412)
Z-Score 1. IQ Test	0.0601***	0.0512	-0.00938	0.0378*	0.0380*
	(0.0151)	(0.0527)	(0.0295)	(0.0193)	(0.0194)
Neg. Feedback	-0.106***	-0.137***	-0.0718	-0.103***	-0.103***
	(0.0281)	(0.0417)	(0.0434)	(0.0281)	(0.0280)
Female * Z-Score 1. IQ Test				0.0487*	0.0550*
				(0.0275)	(0.0325)
Female * Z-Score 1. IQ Test * AlwaysInfo					-0.0127
					(0.0393)
Additional Controls	✓	✓	✓	✓	✓
Mean Reference Group	0.68	0.76	0.55	0.68	0.68
Observations Baseline	94	41	53	94	94
Observations Total	205	84	121	205	205

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This table displays estimates relevant to the *Baseline* treatment. Robust standard errors in parentheses. Constants not displayed. Additional controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity).

Figure 1: Timeline of the experiment: 4 main parts.

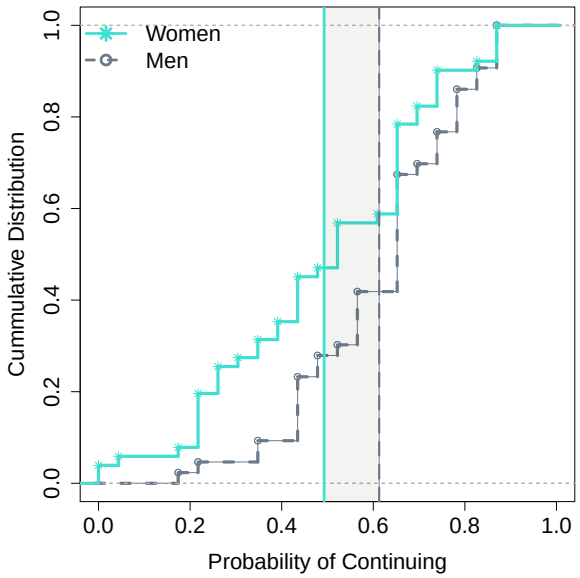


Panel (a) depicts the four main parts of the experiment, one of which was randomly drawn for payment at the end. Panel (b) provides a more detailed overview of what happens if subjects continue or quit, corresponding to Part 3 in panel (a). The only feature distinguishing the *Baseline* from the *AlwaysInfo* treatment is whether or not subjects who quit receive additional feedback, i.e., whether they learn if they passed of failed the IQ test in Part 1. The "continue" option does not vary across treatment arms.

Figure 2: Cards shown to subjects to convey feedback.

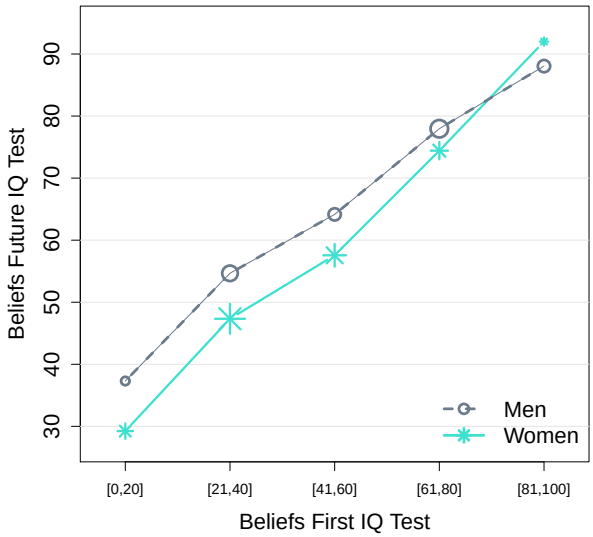
This figure displays the cards shown to subjects to convey feedback in Part 2 of the experiment. Subjects either received positive feedback (a card saying that they passed), or negative feedback (a card saying that they failed the IQ test), randomized conditional on their actual performance (having passed or failed).

Figure 3: Probability of Continuing by Gender, Raw Data, Baseline Treatment.



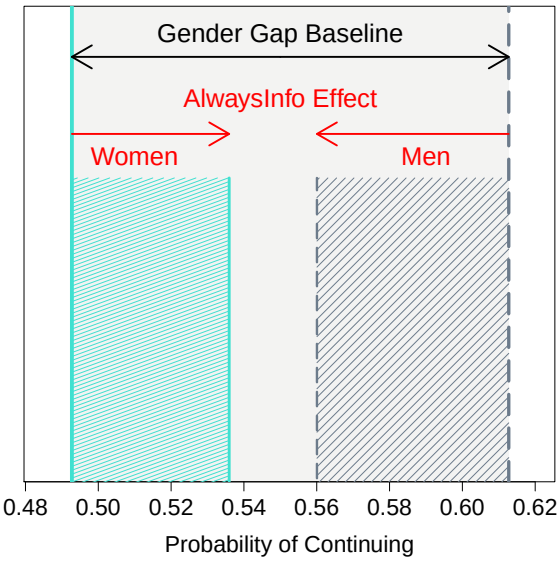
This figure shows empirical cumulative distribution functions of subjects' continuation probabilities, separately for men and women. The vertical lines represent the means of each group, and the gray shaded area highlights the gender difference in average probabilities of continuing, i.e., the gender gap in persistence. Raw data from the *Baseline* treatment are plotted, i.e., without controls for performance or feedback.

Figure 4: Beliefs About One's Future Performance, Given Beliefs About the Past (Before Feedback).



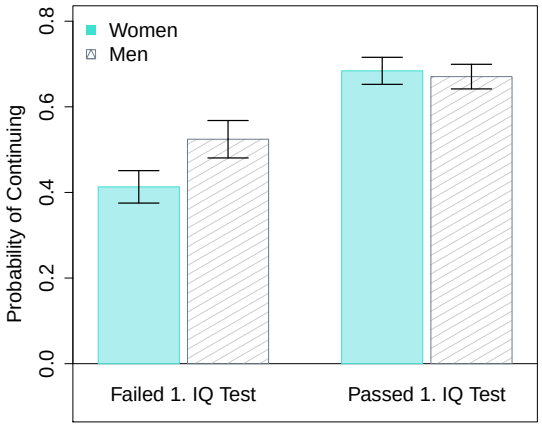
This figure plots gender differences in average beliefs about passing the future IQ test (y-axis), given beliefs about having passed the first IQ test (x-axis). Initial beliefs are plotted, i.e., beliefs before getting feedback. The size of the points represents the relative share of observations in a given bin category of the x-axis. Data from both treatments combined.

Figure 5: AlwaysInfo Treatment Effect Relative to Baseline Treatment.



This figure shows compares the average probability of continuing between the *AlwaysInfo* treatment and the *Baseline*, separately for men and women.

Figure 6: Probability of Continuing by Test Result and Gender, Baseline Treatment.



Bars represent the average probabilities of continuing depending on having failed or passed the first IQ test, separately for women and men, alongside the standard errors of each group, in the *Baseline* treatment.

Appendices

A Estimation of Risk Parameters

The following discusses how risk parameters are estimated for each subject. Recall that in Part 4 of the experiment, subjects were asked to choose between some fixed payment and a lottery $\mathcal{L}$ that pays \$20 with probability p and \$0 with probability $1-p$. Subjects reported a switch point s such that they (weakly) prefer getting paid \$$s$ with certainty over getting the lottery. Assuming monotonic preferences, this also implies that they (weakly) prefer the lottery to getting paid \$$(s-1)$ with certainty.

Under the assumption of narrow framing, i.e., that subjects do not consider their wealth outside of the experiment when making their decision in Part 4, as well as expected utility preferences, subject i's reported switch point in Part 4 therefore implies that

$$U(s_i) \quad \geq \quad U(\mathcal{L}_i) = p_i * U(20) \quad \geq \quad U(s_i - 1). \tag{3}$$

Equation 3 yields an upper and a lower bound for subject i's risk parameter r_i, which can be estimated by imposing a functional form $U(.)$ such as CRRA or CARA.[31] In the relevant regression estimates presented in this paper, risk parameters are computed as the mean of that interval, separately under the assumption of CRRA and CARA utility functions.

[31] Under the assumption of CRRA (Constant Relative Risk Aversion) preferences, $U(x,r) = \frac{x^{1-r}}{1-r}$ if $r \neq 1$, and $U(x,r) = ln(x)$ if $r = 1$. Under the assumption of CARA (Constant Absolute Risk Aversion) preferences, $U(x,r) = \frac{e^{-rx}}{r}$.

B Details BDM Mechanism

The following describes details of the incentive-compatible BDM procedure (Becker et al., 1964) used to implement Part 3 and Part 4 of the experiment. See the Supplementary Appendix for the instructions and screenshots of all steps of the experimental interface. The instructions used to explain the BDM to subjects are largely based on Healy (2020).

Figure A3 shows a screenshot of how the BDM was presented to subjects in Part 3 of the *Baseline* treatment. (Part 4 and the *AlwaysInfo* treatment were conducted in an analogous manner.) There was a list of 23 questions, and in each question subjects were presented with a choice between *Option A (to quit)* or *Option B (to continue)*. The only feature varying across questions was the amount of *earn_quit*; *earn_quit* is the fixed minimum payment associated with *Option A (quitting)*, and the amount of *earn_quit* increases from \$0 to \$22 in one-dollar-increments across the 23 questions.

Subjects were told that it was assumed they would prefer *Option A (quitting)* in the first few questions (i.e., when *earn_quit* is high), but at some point would switch to *Option B (continuing)*. Instead of answering all 23 questions, subjects were asked to report their "switch point" – the dollar value of *earn_quit* at which they would like to switch from *Option A* to *Option B*. Put differently, the switch point is the lowest bonus payment associated with quitting for which a subject (weakly) prefers quitting over continuing. Crucially, one of the questions was randomly drawn after subjects reported their switch point, and their choice in this given question was implemented. This mechanism is therefore incentive-compatible, as a higher requested payment for quitting goes along with a higher chance of continuing. It is important to note that a subject's reported switch point in Part 3, divided by 23, can be interpreted as their chosen ex-ante probability of continuing, and thus serves as a measurement of persistence.

Emphasis was put on making the BDM intuitive to use through a number of visual and interactive features. The colors of the two options (orange for *Option A* and purple for *Option B*) in the list of questions and the instructions correspond to the colors of the slider with which subjects did report their switch point, see Figure A3. If a subject reported a relatively low switch point, they had a relatively high chance of ending up with *Option A*, and the slider bar had a relatively larger orange than purple fraction, and vice versa. An interactive interface further ensured that after

bringing the slider bar into a position, subjects could see what their current switch point implies before submitting their choice. Moreover, subjects had the opportunity of familiarizing themselves with how the BDM works through a practice round that involved a generic *Option A* and *Option B*. That is, subjects could "play around" with the interactive slider of the BDM before learning about the two options of continuing and quitting.

C Field Study: Sample and Implementation Details

The classroom field study was conducted within the context of an introductory undergraduate course (Econ 1) at UC Santa Barbara. This is a general education course and usually the first economics course that students take at UCSB. More than half of all students enrolled in Econ 1 are freshmen students, and approximately $25-30\%$ of students that complete this course end up majoring in economics. Roughly 45% of Econ 1 students at UCSB are women.

All students enrolled in Econ 1 in the 2021 fall quarter were invited to participate in what was labelled a "short research survey." Students were informed that the purpose of this study was to investigate people's beliefs about future success. For completing this study (which took students slightly less than 4 minutes on average), they earned 0.5 bonus points that counted towards their final grade in Econ 1, which accounted for roughly 12.5% of the point difference between two letter grades.[32] In addition to the bonus points, students who completed the survey could earn a $50 prize by reporting accurate beliefs.[33] It was pointed out in both the announcement emails and the instructions that their Econ 1 instructor and TA were not involved as researchers in this study. Screenshots of the materials used to conduct the field study can be found in the Supplementary Appendix.

The study was administered online and conducted on October 15, 2021 during a pre-announced time window of a few hours following the first Econ 1 midterm exam, before test scores were released. The instructor sent students a link to the survey via email. The compliance rate to participate in the research study was 63%. The final sample used for the analysis consists of 368 observations – 184 men and 184 women; This sample excludes observations that could not be matched as well as students who did not identify as either male or female.[34]

[32]The maximum point score students could achieve in the course was 100. The average point gap between most letter grades was 4 points, and thus the 0.5 bonus points accounted for roughly 12.5% of the gap between grades. Students were also given the option to complete a "research-alternative task" to earn the same 0.5 bonus points, which took roughly the same time to complete, and consisted of ten slider tasks.

[33]They were informed that they could maximize their chance of winning a $50 prize by making accurate assessments. To award these prizes, students were randomly selected and the same crossover mechanism as in the main experiment was used (see Section 2). To keep the survey as short as possible, the details of this mechanism were not explained, however participants were informed that they could email the researcher if they had questions about the compensation mechanism. No such inquiries were made.

[34]Nine students who participated did not identify as either male or female. Observations could not be matched if at least one of the following criteria applied: (i) the survey was not completed (which counted as withdrawal from voluntary participation), (ii) the survey was filled out multiple times (with different answers; two students did this), (iii) the Econ 1 course was not completed (in which case the exam grades were not obtained), (iv) no student identifier or an invalid ID was reported in the survey. Aside from the survey participants, 26 students opted to complete a

D Individual Returns to Continuing versus Quitting

Did subjects' continuation decisions maximize their earnings in the experiment, and are there gender differences therein? Assessing whether continuing paid off at the individual level requires the computation of counterfactual outcomes: How much would subjects who continued have earned, had they quit? Suppose that Part 3 of the experiment is drawn for payment. Subjects who continued earned \$20 if they passed and \$0 if they failed the second IQ test, and their expected counterfactual earnings for quitting are $\frac{s+22}{2}$ for a reported switch point of s, by construction of the BDM. For the subset of subjects who continued, define as a subject's "premium of continuing" their actual earnings in Part 3 of the experiment (\$20 or \$0) minus their expected counterfactual earnings for quitting. (Note that one cannot compute counterfactual earnings for subjects who quit, as their performance on the second IQ test is unobserved.)

Figure A4 shows a histogram of the "premium of continuing" for the subset of subjects that continued in the *Baseline* treatment. (Results are qualitatively similar when data are pooled across treatments.) This distribution is bimodal for both men and women; On the one hand, the sub-sample of subjects who continues is positively selected – in the *Baseline*, roughly three-quarters of all subjects that continue pass the second IQ test. The median premium of continuing is very small but positive for both men and women (\$1 and \$0.5, respectively). That is, most subjects that continued were financially marginally better off from doing so.

On the other hand, subjects who continue but fail the second IQ test forgo the secure minimum payment that quitting involves. For about one-quarter of all subjects who continue, these opportunity costs of continuing can be substantial, as the left tail of the distribution in Figure A4 shows. Consequently, the average premium of continuing is negative for both men and women, however this is driven by a rather small fraction of subjects that failed the second IQ test but requested a relatively high minimum payment for quitting.

"research-alternative task" instead of the research survey to earn the same 0.5 bonus points. Four students completed both the research study and the research-alternative task, and their responses were included in the analysis.

E Additional Tables and Figures

Table A1: OLS Estimates of the Probability to Continue

	Probability of Continuing				
	(1)	(2)	(3)	(4)	(5)
Female	-0.120***	-0.103**	-0.0883**	-0.100**	-0.140**
	(0.0424)	(0.0422)	(0.0405)	(0.0413)	(0.0553)
Z-Score 1. IQ Test		0.0601***	0.00330	0.0378*	0.0591***
		(0.0151)	(0.0267)	(0.0193)	(0.0152)
Neg. Feedback		-0.106***	-0.0901***	-0.103***	-0.111***
		(0.0281)	(0.0281)	(0.0281)	(0.0348)
Passed 1. IQ Test			0.150***		
			(0.0540)		
Female * Z-Score 1. IQ Test				0.0487*	
				(0.0275)	
Female * Negative Feedback					0.0661
					(0.0717)
AlwaysInfo	-0.0527	-0.0591	-0.0540	-0.0723*	-0.0561
	(0.0363)	(0.0427)	(0.0406)	(0.0432)	(0.0431)
AlwaysInfo * Female	0.0959	0.109*	0.102*	0.111**	0.174**
	(0.0585)	(0.0560)	(0.0547)	(0.0556)	(0.0681)
AlwaysInfo * Fem. * Neg. Feedback					-0.113
					(0.0824)
Additional Controls	-	✓	✓	✓	✓
Mean Reference Group	0.61	0.68	0.55	0.68	0.68
Observations Baseline	94	94	94	94	94
Observations Total	205	205	205	205	205

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This table is an extension of Table 2, displaying all estimates, i.e., not only those that are relevant to the *Baseline* treatment. Robust standard errors in parentheses. Constants not displayed. Additional controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity). The mean of the reference group shows the average probability of continuing for all men (column 1), men who received positive feedback (columns 2, 4, and 5), and men who received positive feedback but failed the first IQ test (column 3) in the *Baseline*.

Table A2: Heckman Probit: Predicting Future Performance With Past Performance.

	(1)	(2)
Step 1: Selection into Continuing		
Probability of Continuing	3.549***	3.552***
	(0.526)	(0.528)
Step 2: Passed 2. IQ Test		
Z-Score 1. IQ Test	0.428***	0.473*
	(0.142)	(0.246)
Female		-0.183
		(0.317)
Z-Score 1. IQ Test * Female		-0.262
		(0.296)
Additional Controls	✓	✓
Observations Continued	105	105
Observations Total	205	205

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Constants not displayed. Additional controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity). The switch point in part 3 translates into the ex-ante probability of continuing. The performance on the 2. IQ test is only observable conditional on continuing. Data from both treatments combined.

Table A3: OLS Estimates of Log-Likelihood Bayesian Updating.

	First IQ Test		Future IQ Test	
	(1)	(2)	(3)	(4)
α	0.834***	0.842***	0.922***	0.888***
	(0.0648)	(0.122)	(0.0624)	(0.114)
βp	1.227***	1.104***	0.839***	0.807***
	(0.156)	(0.259)	(0.140)	(0.207)
βn	1.672***	1.711***	1.104***	0.887***
	(0.159)	(0.257)	(0.145)	(0.260)
α * Female		-0.00537		0.0594
		(0.135)		(0.127)
βp * Female		0.260		0.127
		(0.317)		(0.276)
βn * Female		-0.0708		0.376
		(0.332)		(0.310)
$H_0 : \beta_p = \beta_n$	0.045	0.112	0.190	0.815
$H_0 : \beta_p * Female = \beta_n * Female$	-	0.482	-	0.562
Observations	205	205	205	205

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Variants of equation 2 are estimated. Prior beliefs of 100 and 0 were coded to 99 and 1, respectively, so that the log-likelihood was well defined for all subjects. Columns (1)-(2) estimate belief updating on the first IQ test, where $\phi = \frac{2}{3}$ by design. Columns (3)-(4) estimate updating on the future test for $\phi = 0.62$, for which the estimates of β_p and β_n were reasonably close to 1. (Note that different ϕ values would scale the estimates, but would not lead to a different conclusion when testing the hypotheses that $\beta_p = \beta_n$ or that $\beta_p * Female = \beta_n * Female$.) The second to third last rows show p-values associated with the corresponding hypothesis tests. Data from both treatments combined.

Table A4: OLS Estimates of the Probability to Continue.

	(1)	(2)	(3)	(4)
Female	-0.103**	-0.0673*	-0.104**	-0.114**
	(0.0422)	(0.0371)	(0.0428)	(0.0447)
Z-Score 1. IQ Test	0.0601***	0.0305**	0.0520***	0.0571***
	(0.0151)	(0.0153)	(0.0161)	(0.0164)
Neg. Feedback	-0.106***	-0.0418	-0.110***	-0.124***
	(0.0281)	(0.0292)	(0.0286)	(0.0291)
Posterior 2. IQ Test		0.00346***		
		(0.000698)		
CRRA Risk Parameter			-0.0305***	
			(0.00995)	
CARA Risk Parameter				-0.481**
				(0.197)
Additional Controls	✓	✓	✓	✓
Mean Reference Group	0.68	0.68	0.65	0.65
Observations Baseline	94	94	78	79
Observations Total	205	205	178	182

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. This table only displays estimates that are relevant to the *Baseline* treatment, but uses data from all treatments. Robust standard errors in parentheses. Constants not displayed. Additional controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity). Column (1) in this table corresponds to Column (1) of Table 2. CRRA and CARA risk parameters refer to the means of the risk parameter intervals computed under the assumption of narrow framing with a base wealth of 0. The number of observations in Columns (3) and (4) are lower as the risk parameters are not well-defined for all subjects. The mean of the reference group shows the average probability of continuing for men who received positive feedback in the *Baseline*. For columns (3) and (4), this average refers to the subset of subjects for which the risk parameters are well defined.

Table A5: OLS Estimates of Risk Parameters

	CRRA Risk Paramenter		CARA Risk Paramenter	
	(1)	(2)	(3)	(4)
Female	0.0103	0.105	-0.00805	0.00425
	(0.316)	(0.333)	(0.0161)	(0.0164)
Posterior 2. IQ Test		0.0115*		0.00128***
		(0.00587)		(0.000281)
Z-Score 1. IQ Test		-0.0825		-0.00369
		(0.239)		(0.0120)
Additional Controls	✓	✓	✓	✓
Mean Reference Group	-0.108	-0.108	-0.077	-0.077
Observations	178	178	182	182

Notes: Robust standard errors in parentheses. Constants not displayed. Additional controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity). The mean of the reference group refers to men's average estimated risk parameters. The number of observations refers to the number of subjects for which a respective risk parameter was well-defined. Data from both treatments combined.

Table A6: Summary Statistics: AlwaysInfo Treatment Relative to Baseline Treatment

	Baseline Averages			**AlwaysInfo Relative to Baseline**					
	Men	Women	All	Men		Women		All	
				Difference	p-value	Difference	p-value	Difference	p-value
1. IQ Test Performance									
Score 1. Test	4.40	3.63	3.86	-0.40	0.112	0.06	0.814	-0.12	0.503
Passed 1. Test	0.60	0.29	0.44	-0.16	0.104	0.03	0.702	-0.05	0.480
Self-reported Characteristics									
GPA	3.09	3.67	3.24	0.37	0.000	0.16	0.060	0.25	0.000
STEM Major	0.42	0.31	0.36	0.06	0.577	0.03	0.728	0.05	0.442
Econ / Accounting Major	0.21	0.10	0.15	0.06	0.475	0.11	0.114	0.09	0.093
Non-White	0.70	0.84	0.78	-0.05	0.573	-0.21	0.017	-0.14	0.033
English First Language	0.79	0.71	0.78	-0.01	0.894	0.12	0.148	0.06	0.330
US Citizen	0.81	0.78	0.80	0.03	0.656	0.16	0.020	0.09	0.062
Beliefs									
Prior 1. IQ Test	61.14	46.71	53.31	-9.63	0.045	-1.17	0.945	-4.60	0.208
Prior 2. IQ Test	68.95	55.82	61.83	-4.06	0.240	-0.19	0.963	-1.27	0.600
Posterior 1. IQ Test	52.58	45.45	48.71	-0.33	0.873	-0.84	0.835	-0.04	0.907
Posterior 2. IQ Test	64.26	51.84	57.52	-1.65	0.701	1.35	0.840	0.68	0.988
Risk Preferences									
CRRA Risk Parameter	-0.12	-0.15	-0.14	0.01	0.876	0.08	0.302	0.05	0.420
CARA Risk Parameter	-0.08	-0.09	-0.08	0.003	0.532	0.01	0.268	0.01	0.244

Notes: This table displays variables that by design should be unaffected by the treatment. Differences indicate the average of a variable in the *AlwaysInfo* treatment relative to the *Baseline*. P-values refer to a Wilcoxon-Mann-Whitney Test testing the hypothesis that the distribution of a characteristic is the same for both treatments.

Table A7: AlwaysInfo Treatment Effect, by Deviations from Bayesian Benchmark on 1. IQ Test.

	Probability of Continuing				
	(1) **All**	(2) Over-reacting	(3) Under-reacting	(4) Too optimistic	(5) Too pessimistic
Estimated Treatment Effect					
Men	-0.065*	0.018	-0.094	-0.039	-0.013
	(0.037)	(0.053)	(0.066)	(0.103)	(0.044)
Women	0.032	0.033	0.178**	0.133	-0.002
	(0.041)	(0.057)	(0.081)	(0.103)	(0.097)
Controlling for Beliefs	✓	✓	✓	✓	✓
Additional Controls	✓	✓	✓	✓	✓
$H_0 : \text{TME}_{\text{Men}} = \text{TME}_{\text{Women}}$	0.051	0.818	0.007	0.058	0.878
Observations	205	117	77	77	117

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This table presents estimates of the average treatment effect of the *AlwaysInfo* relative to the *Baseline* treatment, separately for men and women. Positive (negative) point estimates correspond to feedback avoidance (feedback seeking). Controlling for beliefs about past and future IQ test performance reported after feedback. Additional controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity). The second last line reports p-values testing the hypothesis that the treatment effect is the same for men and women. Columns (2)-(5) show estimates for sub-samples of subjects, depending on how they updated on their performance on the 1. IQ test relative to the Bayesian benchmark. Columns (2) and (3) display subjects that over-reacted and under-reacted to the feedback in Part 2 (i.e., who updated as if the feedback was more informative than it was by design). Column (4) displays subjects that either over-reacted to positive, or under-reacted in response to negative feedback (i.e., who updated too optimistically); and column (5) vice versa.

Table A8: OLS Estimates of Probability of Continuing, Subgroups by Family Background and Attitudes, Baseline Treatment.

	All	Parents Typ. Occupations	Dad Worked More	Conservative Attitudes
	(1)	(2)	(3)	(4)
Female	-0.103**	-0.0907	-0.122**	-0.166***
	(0.0422)	(0.0556)	(0.0567)	(0.0590)
Negative Feedback	-0.106***	-0.0879**	-0.0597	-0.101**
	(0.0281)	(0.0425)	(0.0462)	(0.0390)
Z-Score 1. IQ Test	0.0601***	0.0719***	0.0811***	
	(0.0151)	(0.0201)	(0.0246)	(0.0193)
Additional Controls	✓	✓	✓	✓
Mean Reference Group	0.68	0.64	0.67	0.68
Observations Baseline	94	56	49	53
Observations Total	205	119	104	112

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Constants not displayed. Only estimates relevant to the *Baseline* treatment are shown. Column (1) shows the total sample. Column (2) shows the subgroup of subjects that did *not* disagree / strongly disagree that both their mother's and father's occupation was "typical for a woman/man of her/his generation." Column (3) shows the subgroup of subjects that reported a strictly higher "hours worked for pay" for their father than mother in a "typical week" when they were a child. Column (4) shows the subgroup of subjects that either disagreed or strongly disagreed that "women should pay their own way on dates," *or* that did not strongly disagree that "a wife with a family has no time for outside employment." The mean of the reference group refers to the average continuation probability for men who received positive feedback in the *Baseline*. Additional controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity).

Table A9: Predicting Future Performance With Past Performance: Field Study.

	Above Cutoff 2. Midterm Exam	
	(1)	(2)
Z-Score 1. Exam	-0.464***	-0.350**
	(0.113)	(0.146)
Female		0.177
		(0.141)
Z-Score 1. Exam * Female		-0.271
		(0.192)
Additional controls	✓	✓
Observations	368	368

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Constants not displayed. Additional controls: Race identity.

Table A10: Probability of Continuing by 1. IQ Test Performance

	All	Passed	Failed	**All**	
	(1)	(2)	(3)	(4)	(5)
Female	-0.103**	-0.00738	-0.153**	-0.100**	-0.0987**
	(0.0422)	(0.0514)	(0.0713)	(0.0413)	(0.0412)
Neg. Feedback	-0.106***	-0.137***	-0.0718	-0.103***	-0.103***
	(0.0281)	(0.0417)	(0.0434)	(0.0281)	(0.0280)
Z-Score 1. IQ Test	0.0601***	0.0512	-0.00938	0.0378*	0.0380*
	(0.0151)	(0.0527)	(0.0295)	(0.0193)	(0.0194)
AlwaysInfo	-0.0591	-0.0890	-0.0770	-0.0723*	-0.0732*
	(0.0427)	(0.0633)	(0.0846)	(0.0432)	(0.0435)
AlwaysInfo * Female	0.109*	0.0338	0.182*	0.111**	0.108**
	(0.0560)	(0.0938)	(0.0925)	(0.0556)	(0.0548)
Female * Z-Score 1. IQ Test				0.0487*	0.0550*
				(0.0275)	(0.0325)
AlwaysInfo * Female * Z-Score 1. IQ Test					-0.0127
					(0.0393)
Mean Reference Group	0.68	0.76	0.55	0.68	0.68
Observations Baseline	94	41	53	94	94
Observations Total	205	84	121	205	205

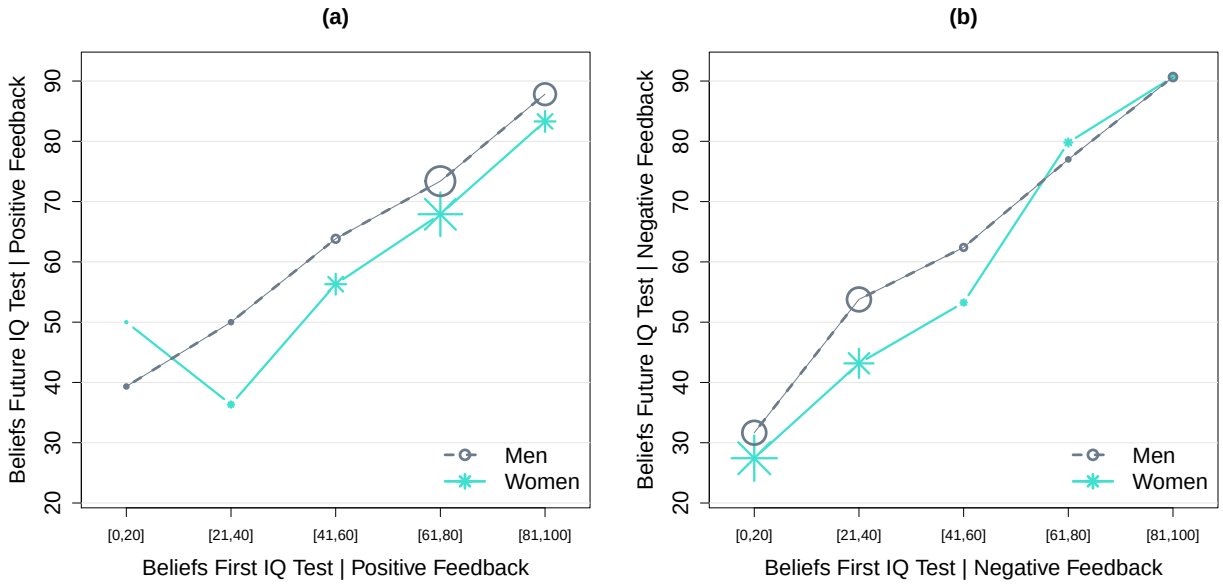
Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Constant not displayed. Additional controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity). This table is the extended version of Table 6, displaying all estimates, not just those that are relevant to the *Baseline* treatment.

Table A11: Performance 2. IQ Test by Ex-ante Probability of Continuing

	Heckman Probit Passed 2. IQ Test		Heckman Z-Score 2. IQ Test	
	(1)	(2)	(3)	(4)
Step 1: Selection into Continuing				
Switch Point Part 3	0.175***	0.170***	0.170***	0.168***
	(0.0251)	(0.0266)	(0.0275)	(0.0266)
Switch Point Part 4	-0.0255	-0.0196	-0.0214	-0.0170
	(0.0181)	(0.0220)	(0.0259)	(0.0225)
Step 2: Performance 2. IQ Test				
Switch Point Part 3	0.135***	0.106	0.0777**	0.0406
	(0.0262)	(0.0658)	(0.0344)	(0.0405)
Female		0.897		0.557
		(1.265)		(0.818)
Female * Switch Point Part 3		-0.0709		-0.0343
		(0.0802)		(0.0499)
Z-Score 1. IQ Test		0.325**		0.379***
		(0.144)		(0.108)
Additional Controls	-	✓	-	✓
Observations Continued	105	105	105	105
Observations Total	205	205	205	205

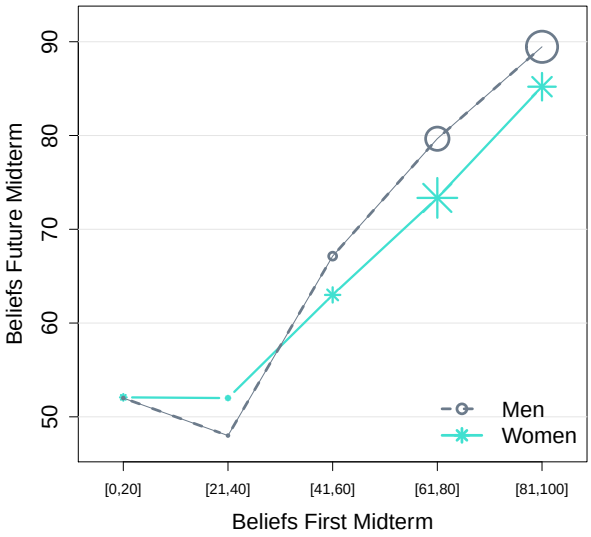
Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Constants not displayed. Additional controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity). The switch point in part 3 translates into the ex-ante probability of continuing. The switch point in part 4 translates into the ex-ante probability of getting the lottery in the risk task. The performance on the 2. IQ test is only observable conditional on continuing. Data from both treatments combined.

Figure A1: Beliefs About One's Future Performance, Given Beliefs About the Past (After Feedback).



This figure plots gender differences in average beliefs about passing the future IQ test (y-axis), given beliefs about having passed the first IQ test (x-axis). Posterior beliefs are plotted, i.e., beliefs after getting feedback. The size of the points represents the relative share of observations in a given bin category of the x-axis. Panel (a) shows this relationship conditional on having received positive, while panel (b) shows this relationship conditional on having received negative feedback. Data from both treatments combined.

Figure A2: Beliefs About One's Future Performance, Given Beliefs About the Past (Field Study).



This figure plots gender differences in average beliefs about answering at least 12 questions correctly on the the future midterm exam (y-axis), given beliefs about having answered at least 12 questions correctly on the first midterm exam (x-axis). The size of the points represents the relative share of observations in a given bin category of the x-axis.

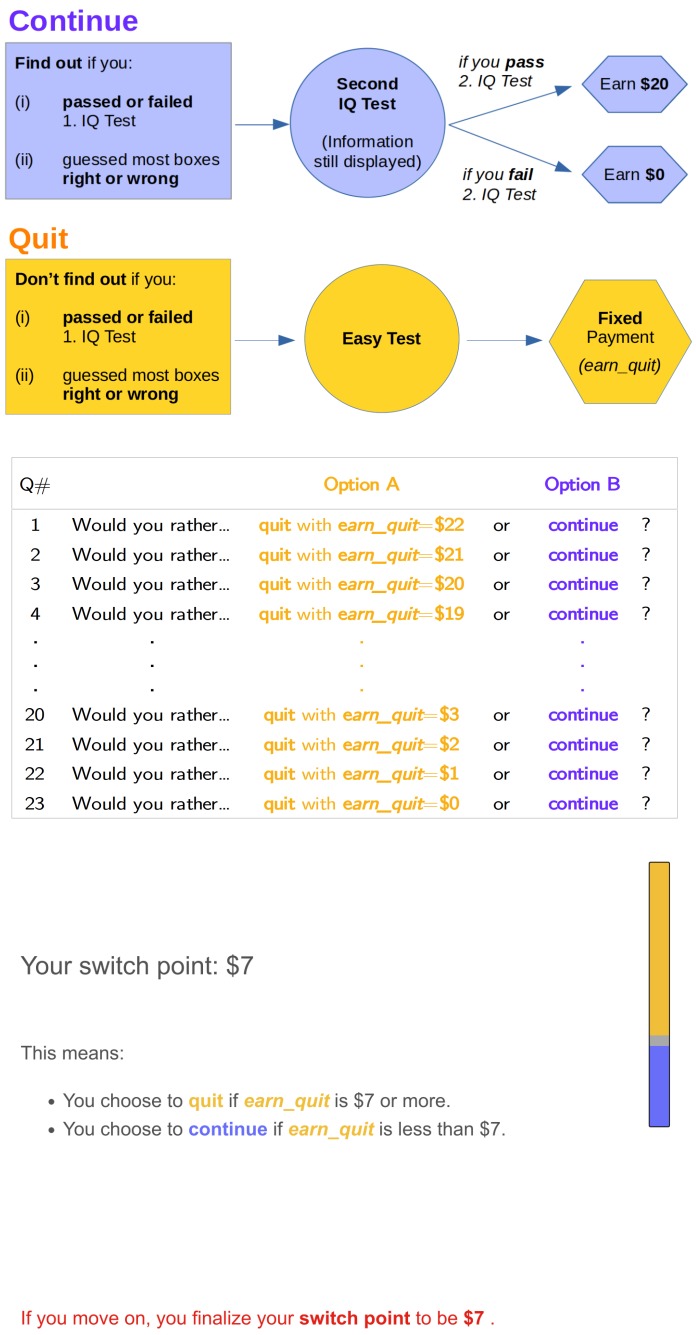
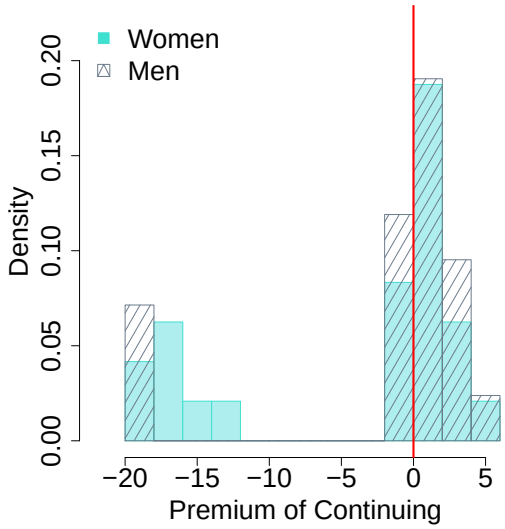


Figure A3: Screenshot of the BDM decision interface in the *Baseline* treatment. Subjects see an overview of what happens if they continue or quit, a list of questions referring to their preferences for either option under different quitting payments, and a slider to report the switch point after which they would like to switch from Option A to Option B.

Figure A4: Premium of Continuing Relative to Expected Payments for Quitting.



This figure shows, separately by gender, a histogram of the premium payment of continuing in US Dollars, defined as a subject's actual bonus payment from continuing in Part 3 of the experiment ($20 if they passed and $0 if they failed the second IQ test) minus their expected counterfactual earnings for quitting (implied by their switch point in the BDM in Part 3), for the subset of subjects that continued in the *Baseline* treatment. Positive values (i.e., the mass right of the red vertical line) indicate that a subject's actual earnings from continuing were higher than their expected counterfactual earnings from quitting, given their switch point.

References

Alan, S. and S. Ertac (2019). Mitigating the gender gap in the willingness to compete: Evidence from a randomized field experiment. *Journal of the European Economic Association 17*(4), 1147–1185.

Astorne-Figari, C. and J. D. Speer (2019). Are changes of major major changes? the roles of grades, gender, and preferences in college major switching. *Economics of Education Review 70*, 75–93.

Becker, G. M., M. H. DeGroot, and J. Marschak (1964). Measuring Utility by a Single-response Sequential Method. *Behavioral Science 9*(3), 226–232.

Berlin, N. and M.-P. Dargnies (2016). Gender Differences in Reactions to Feedback and Willingness to Compete. *Journal of Economic Behavior & Organization 130*, 320–336.

Bernhard, R. and J. de Benedictis-Kessner (2021). Men and Women Candidates are Similarly Persistent after Losing Elections. *Proceedings of the National Academy of Sciences 118*(26).

Bertrand, M. and K. F. Hallock (2001). The Gender Gap in Top Corporate Jobs. *ILR Review 55*(1), 3–21.

Bordalo, P., K. B. Coffman, N. Gennaioli, and A. Shleifer (2019). Beliefs about Gender. *American Economic Review 109*(3), 739–773.

Buser, T. (2016). The Impact of Losing in a Competition on the Willingness to Seek Further Challenges. *Management Science 62*(12), 3439–3449.

Buser, T., E. Ranehill, and R. van Veldhuizen (2021). Gender differences in willingness to compete: The role of public observability. *Journal of Economic Psychology 83*, 102366.

Buser, T. and H. Yuan (2019). Do Women give up Competing more easily? Evidence from the Lab and the Dutch Math Olympiad. *American Economic Journal: Applied Economics 11*(3), 225–52.

Byrnes, J. P., D. C. Miller, and W. D. Schafer (1999). Gender Differences in Risk Taking: A Meta-analysis. *Psychological Bulletin 125*(3), 367.

Coffman, K., M. Collis, and L. Kulkarni (2019). Stereotypes and Belief Updating. *Working Paper*.

Coffman, K. B. (2014). Evidence on Self-stereotyping and the Contribution of Ideas. *The Quarterly Journal of Economics 129*(4), 1625–1660.

Coffman, K. B., P. U. Araya, and B. Zafar (2021). A (dynamic) investigation of stereotypes, belief-updating, and behavior. *Working Paper*.

Coffman, K. B. and D. Klinowski (2022). Gender and preferences for performance feedback. *Working Paper*.

Coutts, A. (2018). Good News and Bad News are Still News: Experimental Evidence on Belief Updating. *Experimental Economics 22*, 369–395.

Croson, R. and U. Gneezy (2009). Gender Differences in Preferences. *Journal of Economic Literature 47*(2), 448–74.

Deaux, K. and E. Farris (1977). Attributing Causes for One's Own Performance: The Effects of Sex, Norms, and Outcome. *Journal of Research in Personality 11*(1), 59–72.

Eckel, C. C. and P. J. Grossman (2008). Men, Women and Risk Aversion: Experimental Evidence. *Handbook of Experimental Economics Results 1*, 1061–1073.

Eil, D. and J. M. Rao (2011). The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself. *American Economic Journal: Microeconomics 3*(2), 114–38.

Ellison, G. and A. Swanson (2018). Dynamics of the Gender Gap in High Math Achievement. *Working Paper*.

Ertac, S. (2011). Does Self-relevance affect Information Processing? Experimental Evidence on the Response to Performance and Non-performance Feedback. *Journal of Economic Behavior & Organization 80*(3), 532–545.

Falk, A., D. Huffman, and U. Sunde (2006). Self-confidence and Search. *Working Paper*.

Fang, C., E. Zhang, and J. Zhang (2021). Do Women give up Competing more easily? Evidence from Speedcubers. *Economics Letters*, 109943.

Franco, C. (2018). How does Relative Performance Feedback affect Beliefs and Academic Decisions? Evidence from a Field Experiment. *Working Paper*.

Golman, R., D. Hagmann, and G. Loewenstein (2017). Information Avoidance. *Journal of Economic Literature 55*(1), 96–135.

Greiner, B. (2015). Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE. *Journal of the Economic Science Association 1*(1), 114–125.

Healy, P. J. (2020). Explaining the BDM - or any random Binary Choice Elicitation Mechanism - to Subjects. *Working Paper*.

Kang, L., Z. Lei, Y. Song, and P. Zhang (2021). Gender Differences in Reactions to Failure in High-Stakes Competition: Evidence from the National College Entrance Exam Retakes. *Working Paper*.

Karlsson, N., G. Loewenstein, and D. Seppi (2009). The ostrich effect: Selective attention to information. *Journal of Risk and Uncertainty 38*(2), 95–115.

Katz, S., D. Allbritton, J. Aronis, C. Wilson, and M. L. Soffa (2006). Gender, Achievement, and Persistence in an Undergraduate Computer Science Program. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems 37*(4), 42–57.

Köszegi, B. (2006). Ego Utility, Overconfidence, and Task Choice. *Journal of the European Economic Association 4*(4), 673–707.

Kugler, A. D., C. H. Tinsley, and O. Ukhaneva (2021). Choice of Majors: Are Women Really Different from Men? *Economics of Education Review 81*, 102079.

Ludwig, S., G. Fellner-Röhling, and C. Thoma (2017). Do women have more shame than men? an experiment on self-assessment and the shame of overestimating oneself. *European Economic Review 92*, 31–46.

Lundberg, S. J. and J. Stearns (2019). Women in Economics: Stalled Progress. *Journal of Economic Perspectives 33*(1), 3–22.

Lundeberg, M. A., P. W. Fox, and J. Punćcohaŕ (1994). Highly Confident but Wrong: Gender Differences and Similarities in Confidence Judgments. *Journal of Educational Psychology 86*(1), 114.

Lynn, R. and P. Irwing (2004). Sex Differences on the Progressive Matrices: A Meta-analysis. *Intelligence 32*(5), 481–498.

Mobius, M. M., M. Niederle, P. Niehaus, and T. S. Rosenblat (2011). Managing Self-Confidence: Theory and Experimental Evidence. *Working Paper*.

Mobius, M. M., M. Niederle, P. Niehaus, and T. S. Rosenblat (2014). Managing Self-Confidence. *Working Paper*.

Niederle, M. (2014). Gender. *Handbook of Experimental Economics 2*, 481–462.

Niederle, M. and L. Vesterlund (2007). Do Women Shy Away from Competition? Do Men Compete Too Much? *The Quarterly Journal of Economics 122*(3), 1067–1101.

Niederle, M. and A. H. Yestrumskas (2008). Gender Differences in Seeking Challenges: The Role of Institutions. *Working Paper*.

Oprea, R. and S. Yuksel (2022). Social Exchange of Motivated Beliefs. *Journal of the European Economic Association 20*(2), 667–699.

Pereda, P. C., L. Matsunaga, M. D. M. Diaz, B. P. Borges, J. Mena-Chalco, F. Rocha, R. D. T. Narita, and C. Brenck (2020). Are Women Less Persistent? Evidence from Submissions to a Nationwide Meeting of Economics. *Working Paper*.

Rask, K. and J. Tiefenthaler (2008). The role of grade sensitivity in explaining the gender imbalance in undergraduate economics. *Economics of Education Review 27*(6), 676–687.

Raven, J. C. J. C. (1973). *Advanced Progressive Matrices*. London: H.K. Lewis.

Thaler, M. (2021). Gender Differences in Motivated Reasoning. *Journal of Economic Behavior & Organization 191*, 501–518.

Thomsen, D. M. (2018). Gender differences in candidate reemergence. *Working Paper*.

Wasserman, M. (2021). Gender Differences in Politician Persistance. *Review of Economics and Statistics*.

Zimmermann, F. (2020). The Dynamics of Motivated Beliefs. *American Economic Review 110*(2), 337–361.