Performance Feedback and Gender Differences in Persistence

Maria Kogelnik*

This Draft: November 15, 2021
Click here for an updated version.

Abstract

The decision to persist in stratified career trajectories is often dynamic in nature: people receive performance feedback and decide whether to persist or to drop out. I show experimentally that men are on average 10 percentage points more likely to persist in an environment that rewards high performance than equally performing women who received the same feedback. Roughly 30% of this gender gap in persistence can be explained by men seeking, and women avoiding exposure to additional feedback. Another 30% can be explained by women being less confident about their future performance, as men tend to consider previous failures to be less predictive of their future than women.

Keywords: Gender differences, persistence, performance feedback, beliefs, future confidence, information avoidance, economics experiments.
JEL Codes: C91, D91, D83, J16, J24.

*Contact: Economics Department, University of California at Santa Barbara. Email: kogelnik.maria@gmail.com.
Acknowledgements: I am extremely grateful to Ryan Oprea and Sevgi Yuksel for their invaluable guidance and support. This paper has benefited greatly from feedback from Peter Kuhn and Heather Royer. I thank Kelly Bedard, Javier Birchenall, Katherine Coffman, Michael Cooper, Florian Ederer, Ignacio Esponda, Erik Eyster, Shelly Lundberg, Aniko Oery, 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 UC Santa Barbara 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.

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, STEM, and politics tends to be.[1] Making one's way in these career trajectories usually involves frequent exposure to performance feedback. How people respond to this feedback - what beliefs they form about their future performance and whether they enjoy being exposed to additional feedback - will ultimately affect who chooses to persist in these environments. If men and women respond differently to feedback, for example if men become overly optimistic about their future performance following positive feedback, or if women drop out to avoid being exposed to negative feedback, this could help explain the gender differences in persistence we observe.

This paper presents a laboratory experiment designed to study (i) whether there is a gender gap in the decision to persist in an environment that rewards high performance and involves exposure to feedback, and, if so, (ii) what channels are driving this gender difference in behavior.

Using a controlled experiment to study these applied questions has two advantages. First, an experiment allows us to shut down any differences in the opportunity costs and returns to persisting that men and women may face in the field. Furthermore, we can control the feedback and ensure that there is no gender bias in how the feedback is given to men and women, as well as no gender differences in what kind of feedback people seek or expect to get. Second, exploring what channels are driving the gender gap in persistence requires the measurement of variables that are unobserved in naturally occurring data, such as people's beliefs about their future performance or their preference to avoid or receive additional feedback. In the lab, we can shed light on the underlying mechanisms shaping behavior, which can inform field studies and point to interesting avenues for policy interventions.

Why would we expect that men and women might respond differently to feedback on their performance? To begin with, a recent empirical literature is consistent with the idea that *negative* feedback has a more discouraging effect on women than on men: 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 initially missing the cutoff, scoring low on practice exams, or having lost 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

[1] See Bertrand and Hallock (2001) for a corporate context, Lundberg and Stearns (2019) for women in economics, and the *She Numbers* report of the European Commission for research and innovation: https://op.europa.eu/en/publication-detail/-/publication/9540ffa1-4478-11e9-a8ed-01aa75ed71a1).

election (Wasserman, 2021).[2] Furthermore, it is conceivable that *positive* performance feedback has a more encouraging effect on men than on women. Men - relative to equally performing women - might become more optimistic about their future performance in response to positive feedback, or might have a stronger preference to receive additional feedback in the future.

The first goal of the experimental design 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 rewards high performance, and involves exposure to feedback. Feedback is often times either positive or negative, and ego-relevant in the sense that people may care about performance feedback for reasons beyond its pure instrumental value.

In the *Baseline* treatment of the experiment, subjects are asked to perform a challenging, ego-relevant task - an IQ test - which they can either pass or fail. After taking the test, they receive either positive or negative performance feedback that is randomized conditional on their true performance, and of known accuracy. Randomizing the feedback allows us to explore the effect of receiving positive versus negative feedback across the performance distribution. Subjects then face two options: If they *continue*, they are exposed to additional feedback, take another IQ test, and receive a high bonus if they pass the second test, but nothing otherwise. Alternatively, if they *quit*, they receive no more performance feedback, complete an easy task, and receive a fixed payment regardless of their performance on the easy task.

The first main finding is that women are about 10 percentage points less likely 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%. This gender gap in persistence is driven by subjects who received positive feedback on their performance, as well as subjects who failed the first IQ test. Notably, women who received *positive* feedback are one average less likely to continue that men who received *negative* feedback, albeit this difference is not statistically distinguishable.

The second goal of the design is to explore what channels may be driving this gender gap in persistence. More specifically, the design is equipped to shed light on the role of beliefs and how beliefs respond to feedback, the role of feedback avoidance, as well as risk preferences in explaining gender differences in persistence.

Beliefs about passing the second IQ test likely play an important role for the continuation decision, as continuing is only financially rewarding for subjects who can pass the second IQ test. To investigate the role of confidence on persistence, subjects' beliefs about passing the future IQ test, as well as the first IQ test, are elicited both before and after receiving feedback, and reporting true beliefs is incentivized. This allows us to explore (i) how these beliefs respond to performance feedback, and (ii) how subjects project beliefs about their past performance onto their future per-

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

formance. It is worth pointing out that beliefs about past events need not directly translate into beliefs about a similar event in the future. To form beliefs about the future, people need to have some concept of how predictive past outcomes are of future outcomes. Especially in motivated contexts where ego utility is at stake, it is conceivable that people distort these beliefs in response to feedback, and there might be gender differences therein.

Before receiving feedback, women are found to be less confident both about having passed the first IQ test and about passing the second IQ test, relative to equally performing men. Interestingly, men and women differ in how they project beliefs about their past performance onto beliefs about their future performance. Controlling for past performance and beliefs about their past performance, men are still significantly more confident about passing the future IQ test than women. This is driven by subjects who failed the first IQ test, i.e. men tend to consider past failures to be less predictive of their future performance than women. In response to feedback, there is no gender difference in updating, however: both men and women update their beliefs about their past and future performance upwards in response to positive, and downwards in response to negative feedback. The resulting gender differences in posterior beliefs - after having received feedback - are therefore driven by gender differences in prior beliefs. Women are less confident about passing the future IQ test both after getting positive and negative feedback, and one reason for this is that men - relative to equally performing women - tend to discount how predictive previous failures are of their future. Roughly 30% of the gender gap in persistence can be explained by gender differences in beliefs about passing the future test.

The second channel we explore are gender differences in feedback avoidance. Recall that in the *Baseline* treatment, subjects who continue are exposed to additional feedback, while subjects who quit do not receive any additional feedback. Thus, if women have a stronger preference to avoid exposure to additional feedback, or if men have a stronger preference to receive additional feedback, this could help explain to the documented gender gap in persistence. To study the role of feedback avoidance, the design includes one treatment arm where subjects receive the same performance feedback regardless of whether they continue or quit. This *AlwaysInfo* treatment thus shuts down preferences to receive or avoid additional feedback as a motive for continuing or quitting. A between-design is used, i.e. all subjects participate in either the *Baseline* or the *AlwaysInfo* treatment.

Comparing behavior across the two treatments suggests gender differences in information avoidance account for roughly 30% of the gender gap in persistence. This is driven both by women who quit in order to avoid additional feedback, and men who continue in order to receive additional feedback. Notably, these estimates of the *AlwaysInfo* treatment effect control for beliefs to ensure that gender differences in information avoidance do not solely reflect the documented gender differences in confidence.

The design further allows us to control for subjects' risk preferences. 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. No gender differences in risk aversion are found in this setting, however, and controlling for subjects' estimated risk preferences essentially has no impact on the gender gap in persistence.[3]

To what extent is the documented gender difference in persistence concerning from an efficiency perspective? Policy makers might want to intervene if the self-sorting of men and women into different career trajectories entails that "able" individuals are not continuing, or that "unable" individuals are not quitting. In the controlled environment of the experiment, women who are ex-ante more likely to continue are also ex-post more likely to pass the second IQ test. This is not the case for men, however, i.e. the self-selection of women appears to be more efficient than the self-selection of men, as their continuation decisions better predict their future performance.

Contribution to the literature. To the best of my knowledge, this is the first paper documenting gender differences in persistence in a controlled setting, and to explore through which channels receiving positive versus negative feedback affects persistence and gender differences therein. The role of relative performance feedback on gender disparities in labor supply decisions has been studied by Berlin and Dargnies (2016) and by Buser and Yuan (2019), who explore the role of such feedback on choosing a competitive over a piece-rate payment scheme. Niederle and Yestrumskas (2008) study how feedback affects the gender gap in choosing a hard versus an easy task. Related to this, Buser (2016) document how feedback affects the gender differences in setting goals for one's future performance.

In the paper at hand, competition and social comparison are shut down by design. Subjects are evaluated, rewarded, and receive feedback on their absolute, and not their relative performance. This allows us to shed light on whether gender differences in persistence arise independent of the well-studied gender differences in the willingness to compete (see the seminal work of Niederle and Vesterlund, 2007), which may call for different institutional responses than policies targeting the gender gap in entering tournaments.[4]

[3] While it seems to be a common perception among economists that women are more risk averse than men, a careful look at the literature paints a more nuanced picture. In her handbook chapter, 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 further 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.

[4] Note that some environments that are commonly described as "competitive" often do not involve direct tournaments: Journal editors accept papers which they consider "good enough" in absolute terms, rather than the best $X\%$ of papers that are submitted. Similarly, companies interested in hiring qualified candidates might decide to make multiple offers provided they can attract suitable candidates, but none if no candidate is "above the bar." It is therefore of interest if gender differences in persistence arise even when there is no direct competition.

To the best of my knowledge, this is also the first paper studying how gender differences in feedback avoidance affect persistence. Broadly speaking, gender differences in preferences for feedback appear to be an under-explored research area. Golman et al. (2017) provide an excellent review of the theoretical and empirical literature on information avoidance, but do not mention gender. Buser and Yuan (2019) investigate how avoiding the information of having won or lost affects gender differences in choosing a competitive payment scheme in a repeated adding number task, and find that information avoidance can explain why women choose the piece-rate over the competitive payment scheme in the first, but not in later rounds. Eil and Rao (2011) and Mobius et al. (2011) elicit subjects' willingness to pay (WTP) for ego-relevant information, but do not explore the consequences of these preferences for economic decisions. Both studies find no gender differences in the average WTP, but note that women are more likely than men to require a subsidy for this information.[5]

This paper also contributes to a rich literature on gender differences in beliefs. Women have often been found to have less confident prior beliefs about their performance than men when controlling for actual 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)), however other studies do not find any gender gaps in prior confidence (e.g. Ertac (2011), Berlin and Dargnies (2016), Coutts (2018)). Coffman (2014) and Bordalo et al. (2019) document that gender differences in prior confidence are especially pronounced in gender-congruent domains. Similarly, Coffman et al. (2019) document over-reaction to information that refers to a gender-congruent domain. Mobius et al. (2014) and Coutts (2018) find that women update more conservatively. In contrast, Berlin and Dargnies (2016) document over-reaction to feedback for both men and women, and women are found to update more pessimistically than men.

An insight that the experiment at hand adds to this literature is that men and women differ in how they project their beliefs about their past performance onto their future performance. Controlling for past performance and beliefs about one's past performance, poorly performing men report substantially more optimistic beliefs about passing the future IQ test. In other words, men - relative to women - discount how predictive previous failures are of their future performance. This is the case both before and after receiving feedback. Furthermore, the point estimate of the gender difference in prior beliefs about passing the *future* IQ test is substantially bigger than about having passed an IQ test in the past. While this difference is not statistically distinguishable, this suggests that beliefs might explain a larger fraction of gender differences in behavior (e.g. the willingness to compete) than previously thought, as many experimental studies control for beliefs about a past state, but not the relevant future state, when studying economic decisions.

To the best of my knowledge, this is the first experiment eliciting beliefs not only about a past, pre-determined state (e.g. having passed the first IQ test), but also about a *future* state

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

(e.g. passing the future IQ test). This gap in the literature is surprising, considering that many economically relevant decisions are modelled as a function of beliefs about the future, rather than the past. By documenting how beliefs about the future respond to ego-relevant information about a past state, this paper also contributes to a growing literature on motivated belief formation.[6]

2 Experimental Design

Design goals. The experimental design aims to accomplish two goals. The first goal is to create a setting that allows us to study whether women are more likely to drop out of an environment that rewards high performance and involves exposure to ego-relevant feedback, relative to equally performing men who received the same feedback. This requires a challenging, ego-relevant task with incentives to perform well; feedback that can be positive or negative, but is exogenous conditional on performance; and two options to choose from in response to this feedback: (i) the option to *continue*, which involves another challenging task, additional feedback, and a high payment conditional on performing well, and (ii) the option to *quit*, which involves an easy task as an outside option, no more exposure to feedback, and a fixed payment that does not depend on one's performance.

The second goal is to explore what channels may be driving gender differences in persistence. The experimental design should allow us to shed light on channels that cannot be isolated using naturally occurring data. These include (i) beliefs about performing well in the future, and how these beliefs respond to feedback, (ii) preferences to receive or avoid additional feedback, and (iii) risk preferences.

Overview. The experiment consists of four main parts that are described below in more detail. 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 of the experiment, subjects were not told which part was drawn for payment.

A timeline of the experiment is provided in Figure 1. Instructions clarified how to earn money before each part, however subjects were not told what was going to happen in later parts of the experiment. That is, each set of instructions in Figure 1 only specified the steps until the next set of instructions. 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, explained below in more detail.

[6]The literature on motivated beliefs typically studies situations where people form beliefs about past, predetermined states (e.g. having scored at the top half on an IQ test) in response to signals about these same states that are either of known accuracy (Ertac (2011), Eil and Rao (2011), Mobius et al. (2014), Zimmermann (2020)), of unknown accuracy (Oprea and Yuksel, 2021), or uninformative (Thaler, 2021).

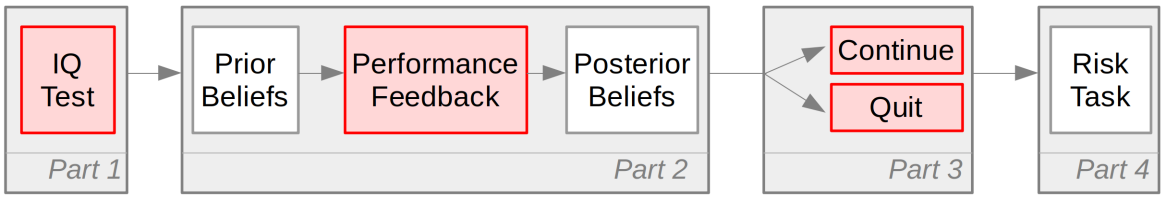


Figure 1: Timeline of the experiment. Arrows indicate the order in which subjects moved through the experiment. One of the four parts highlighted in gray was randomly drawn for payment at the end. The only feature distinguishing the *AlwaysInfo* treatment from the *Baseline* are the consequences of quitting (Part 3).

Part 1: Challenging, ego-relevant task (IQ test). Subjects were asked to take an IQ test, consisting of seven Raven's Progressive Matrices, including a range from relatively easy to relatively difficult matrices. Raven's matrices are frequently used in economics experiments to generate an environment where ego utility is at stake, e.g. see Zimmermann (2020) or Oprea and Yuksel (2021). To emphasize the ego-relevant component of this task, subjects were told that this test is frequently used to measure intelligence. Figure A1 shows an example of one easy and one difficult IQ test question used in the experiment.

Before taking the IQ test, subjects were told that they will either *pass* or *fail* this test. To pass, subjects knew they had to solve at least five of the seven questions 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.

To ensure that any potential gender differences in persistence in this experiment do not reflect gender differences in the willingness to compete, it was highlighted to subjects that whether they passed or failed did depend not on the performance of other participants. Subjects had 90 seconds to answer each question, and a timer on the screen indicated how much time was left. Wrong answers were not penalized, and unanswered questions were counted as wrong.

Part 2: Beliefs and performance feedback. Subjects were asked to report how likely (out of 100) they think it is (i) that they passed the first IQ test, and (ii) that they could pass a similar IQ test in the future. To document how beliefs about one's past and future performance respond to feedback, these two questions were asked before and after the provision of feedback. Eliciting beliefs about the first IQ test in addition to beliefs about the future IQ test allows us to study if men and women make different projections of their future, given their beliefs about the past.

If Part 2 was drawn for payment, subjects earned either \$20 or \$0. The crossover method (Mobius et al., 2014) ensured that subjects maximized their chance of winning \$20 by always reporting their true beliefs, and this was emphasized in the instructions.[7] This method has the advantage

[7] In this mechanism, a reported belief (e.g. of having passed the first test), X, is compared with a uniform random draw between 0 and 100, Y. If $Y \geq X$, subjects were paid \$20 with a chance of Y%, and \$0 with a chance of

that it only requires monotonic preferences, but does not require expected utility preferences or risk neutrality to be truth-inducing.

Prior beliefs (before feedback). Before getting any direct feedback on their test performance, subjects' prior beliefs of having passed the first test were elicited, see the upper panel of Figure A2. Subjects then learned that they might be asked to take another IQ test of the same format later in the experiment, consisting of similar questions of a similar level of difficulty. After providing this information, subjects' prior beliefs of passing this future test were elicited.[8]

Performance feedback. Subjects then received a binary signal - a card saying either that they passed, or that they failed the first IQ test - which was randomized matched the true state of having passed/failed with a known accuracy of two-thirds.[9] Randomizing the feedback has the advantage that the effect of receiving positive versus negative performance feedback can be explored across the performance distribution. Furthermore, since the feedback is generated by a computer following a known process, the design ensures that there is no gender bias in providing feedback, nor can men and women endogenously affect what type of feedback they are getting.

Posterior beliefs (after feedback). Following the feedback, subjects' beliefs about passing the first IQ test and about passing a similar IQ test in the future were elicited a second time. The card conveying the performance feedback was still displayed while eliciting posteriors, see the lower panel of Figure A2.

Part 3: Main decision of interest (continue or quit). The main outcome of interest in the experiment is how subjects choose between the two options of *continuing* and *quitting*. These two options vary in terms of (i) the additional feedback provided, (ii) the task, and (iii) the payment scheme. This decision aims to capture some essential features of the decision to persist in stratified career trajectories in the field. The consequences of each option were explained in detail, and subjects had to correctly answer comprehension questions about what each option entailed before making their decision. It was emphasized that quitting does *not* imply leaving the experiment early.

$(100-Y)\%$. If $Y < X$, subjects were paid \$20 if the situation in the question occurs (e.g. if they passed the first test), and \$0 otherwise. If Part 2 was drawn for payment, one of the four belief elicitation questions - two prior beliefs and two posterior beliefs - was randomly drawn for payment. If a subject did not *continue* in Part 3 of the experiment, and their future performance was thus unobserved, only the beliefs referring to their past performance were eligible for payment.

[8] Note that while strategically reporting pessimistic beliefs about the future might be a concern ex-ante, the data show that the vast majority of subjects are initially more confident about passing the future test, than about having passed the first test, discussed in more detail in Section 3.

[9] The feedback was implemented by telling subjects that the computer would be generating three cards - two of which say the truth of whether they passed or failed the first IQ test, and one of which is fake and says the contrary. Out of these three cards, the computer would randomly pick one card and show it to them. It was pointed out to subjects that this means that it will be twice as likely that the card they will see tells the truth, than to be fake. Figure A3 depicts the cards used to convey the feedback.

Continue. Subjects who continued first learned if they passed of failed the first IQ test.[10] They were then asked to take another IQ test of the same format as the first IQ test, but with different questions. While taking the second test, the information of having passed or failed the first test was still displayed in the center of each question, with the aim to imitate an environment where people are frequently exposed to performance feedback. If Part 3 was drawn for payment and a subject continued, they earned \$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. Whether or not subjects receive additional feedback if they quit is the only feature that distinguishes the *Baseline* from the *AlwaysInfo* treatment.

***Baseline* treatment.** If subjects quit in the *Baseline*, they did not get any additional feedback on their performance, and could thus avoid the information of having passed or failed the first IQ test.[11] Quitters completed an "easy test" consisting of seven very easy Raven's Matrices, and it was pointed out to them that it was very likely that they could solve all questions of the easy test. Having an easy outside option seems natural and has the advantage of keeping opportunity costs of time comparable across the two options.[12] If Part 3 was drawn for payment and subjects quit, they received a fixed payment, described below in more detail.

***AlwaysInfo* treatment.** Subjects who got assigned to the *AlwaysInfo* treatment always learned if they passed or failed the first IQ test, regardless of whether they continued or quit.[13] This treatment thus shuts down preferences to receive or avoid additional feedback as a motive to continue or quit. Comparing the two treatments therefore allows us to isolate the role of information avoidance and information seeking in driving the gender gap in persistence. The other features of quitting - the easy task and the fixed minimum payment scheme - were the same in the *Baseline* and *AlwaysInfo* treatment.

Part 4: Risk task. The risk task was framed as a choice between two options: a fixed payment and a lottery that paid \$20 with some probability p, and \$0 with some probability $100 - p$. This was implemented in a way such that the two options in the risk task (lottery vs. fixed payment) were identical to the options presented in the original decision problem (continue vs. quit), stripped

[10] In addition, they learned if they guessed most boxes right or wrong in a trivial *"Guessing Game."* See Appendix B for details.

[11] Since subjects were not told which part was drawn for payment in the end, they could not infer their test outcome from their earnings in the experiment.

[12] If quitters did not have to take an easy test, the outside option would have been to sit and wait for subjects who chose to continue. For some subjects, the associated boredom might be more costly than taking the challenging IQ test, and there could be gender differences therein. In addition, ensuring that subjects had to spend about the same time on the easy and the challenging IQ test shut down potential social preferences not to let other subjects wait.

[13] More specifically, subjects learned if they passed or failed the first test *after* making their decision, but before taking the second IQ test or the easy test, respectively. While taking the second IQ test or the easy test, that information was displayed next to each question. An overview of what it means to continue or quit in the *AlwaysInfo* treatment is depicted in Figure A4.

from all features other than payoffs and risk.[14] This allows us to estimate risk preferences in the context most relevant to the decision of interest, as recommended by Niederle (2014).

BDM mechanism used in Part 3 and Part 4. Rather than asking subjects to directly choose one of the two options in Part 3 and Part 4, an incentive-compatible BDM procedure (Becker et al., 1964) was used to elicit subjects' *switch point*, i.e. the lowest payment for the quitting option they were willing to accept to prefer quitting over continuing.[15] The higher this requested minimum payment for quitting, the higher was the chance that they would continue, and vice versa. Special emphasis was put on implementing the BDM in an understandable and intuitive way, see Appendix B.

Using a BDM has two advantages in this context: 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. Second, 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 a subject who continued, had they quit, which will be discussed in Appendix E.

2.1 Implementation Details

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 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 comparable as possible between in-person and Zoom sessions. As mentioned in Section 3, a Zoom dummy variable is included in all relevant regressions, however this variable is not statistically significant in any of the presented results.

Instructions were displayed 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 exper-

[14] The probability p was tailored to each subject's individual posterior belief of passing the second IQ test. 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 of what would happen in later parts of the experiment, and thus did not have incentives to report a high posterior belief of passing the future test in order to encounter a lottery with more favorable odds. Note it was not deceptive to tell subjects that they would maximize their chance of winning \$20 by always reporting their true belief if Part 2 was drawn for payment.

[15] The interpretation in Part 4 is analogous to this, i.e. the switch point corresponds to the lowest safe payment such that subjects prefer the safe option over the lottery, etc.

iment.

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 in the recruitment process nor 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 Results

This section is structured as follows: Section 3.1 provides an overview of gender differences in the raw data. Section 3.2 documents that there is a gender difference in persistence even when controlling for performance, feedback, and self-reported characteristics, and analyzes what groups are driving this difference. Section 3.3 shows that gender differences in beliefs and feedback avoidance account for a large share of the gender gap in persistence, while risk preferences play a negligible role. Section 3.4 discusses whether the gender gap in persistence is concerning from an efficiency perspective.

3.1 Data Description

Sample overview. 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.[16] As Table 1 shows, men and women in the sample differ along a few dimensions. Men were significantly more likely to pass the first IQ test ($p = 0.003$), and on average correctly solved almost one question more out of the seven questions on the test ($p = 0.007$). Across treatments, roughly 40% of all subjects continued. Conditional on continuing, there is no statistically significant gender difference in performance on the second IQ test. In terms of self-reported characteristics, women in the sample on average report a slightly higher GPA than men ($p = 0.004$).[17] Furthermore, the share of subjects who report STEM or Economics/Accounting as their major or intended major is higher for men than

[16] 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.

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

for women, however the differences are not statistically significant.

To account for these gender differences in self-reported characteristics, unless otherwise noted, regressions in this paper include a set of controls with the characteristics listed in Table 1, as well as a dummy variable for whether sessions were conducted in person or over Zoom.

Table 1: Summary Statistics

	Men	Women	p-value
Treatment			
Baseline	0.42	0.50	0.292
AlwaysInfo	0.58	0.50	0.292
IQ Test Performance			
Avg. Score 1. Test	4.40	3.63	0.007
Avg. Score 2. Test if Cont.	5.52	4.67	0.026
Passed 1. Test	0.60	0.29	0.003
Passed 2. Test if Cont.	0.86	0.71	0.236
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

Notes: This table displays the shares of subjects for which a given characteristic applies, separately for men and women. 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.

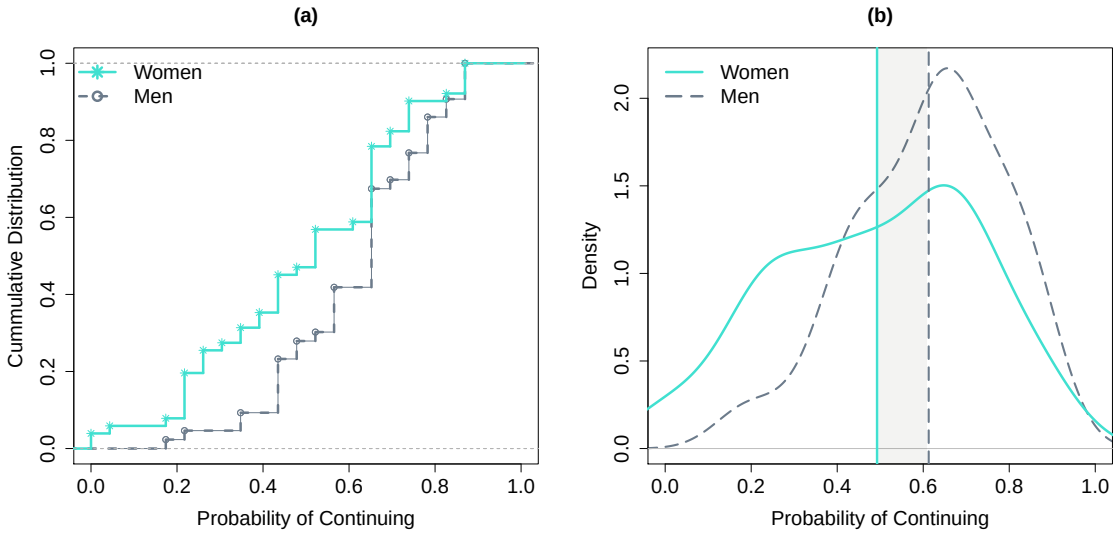
Gender differences in persistence in the raw data. The central question of interest is whether women are significantly less likely to continue than men, conditional on their performance on the first IQ test and the feedback they received. Recall that the switch point in the first BDM directly translates into a subject's ex-ante desired probability of continuing, which serves as a measurement of persistence.[18]

[18]The BDM involves 23 questions, see Appendix B. A subject who reports a switch point of 0 will quit with certainty. By construction of the BDM, the switch point cannot be higher than 22, and thus the probability of continuing cannot exceed 96% (= 22/23). We can thus interpret $SwitchPoint_i/23$ as the probability that subject i continues.

To get a first intuition for gender differences in persistence in the raw data, Figure 2 shows an empirical CDF alongside a density plot of the probability to continue 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, the distribution of continuation probabilities for men first order stochastically dominates the distribution of women, as can be seen in panel (a). Men on average have a 61% chance to continue in the *Baseline* treatment, while women on average only have a 49% chance to continue ($p = 0.006$), see panel (b).

This figure does not imply that there are gender differences in persistence, however, as the distribution of ability - measured in performance on the first IQ test - is substantially different for men and women, as Table 1 illustrates. To resolve this confound, in what follows regressions are presented to analyze gender differences in persistence while controlling for subjects' performance, the feedback they received, as well as self-reported characteristics.

Figure 2: Probability of Continuing by Gender, Raw Data, Baseline Treatment.



Panel (a) displays the empirical cumulative distribution function, separately for men and women. Panel (b) displays the empirical density function, separately for men and women. The vertical lines in panel (b) represent the means of each group, and the gray shaded area highlights the gender difference in average probabilities of continuing. This figure presents raw data from the *Baseline* treatment, i.e. without controls for performance or feedback.

3.2 Formal Analysis of Gender Differences in Persistence

Aggregate results. To formally explore if there is a gender gap in persistence when controlling for ability (measured in test scores on the first IQ test), performance feedback, as well as self-reported characteristics, Table 2 displays OLS regressions estimating the probability to continue using a rich set of controls. Column (1) shows that on average, women are indeed about 10 p.pt. less likely to

continue when controlling for these factors. In other words, the gender gap documented in Figure 2 does not primarily reflect gender differences in confounding factors such as IQ test performance, but rather an actual gender difference in behavior.

Result 1. *In the Baseline treatment, women are on average about 10 p.pt. less likely to continue than men when controlling for their past performance, the feedback they received, and self-reported characteristics.*

Table 2: OLS Estimates of the Probability to Continue.

	Probability of Continuing					
	(1)	(2)	(3)	(4)	(5)	(6)
	All	**All**	Passed	Passed	Failed	Failed
Z-Score 1. IQ Test	0.0612***	0.0604***	0.0522	0.0736	-0.00956	-0.0116
	(0.0154)	(0.0156)	(0.0537)	(0.0569)	(0.0301)	(0.0311)
Female	-0.103**	-0.145**	-0.00738	-0.0628	-0.153**	-0.226**
	(0.0422)	(0.0571)	(0.0514)	(0.0513)	(0.0713)	(0.104)
Neg. Feedback	-0.106***	-0.120**	-0.137***	-0.187***	-0.0718	-0.0668
	(0.0281)	(0.0486)	(0.0417)	(0.0568)	(0.0434)	(0.0967)
Neg. Feedback * Female		0.0753		0.123		0.100
		(0.0797)		(0.0937)		(0.128)
Mean Reference Group	0.68	0.68	0.76	0.76	0.55	0.55
Observations	205	205	84	84	121	121

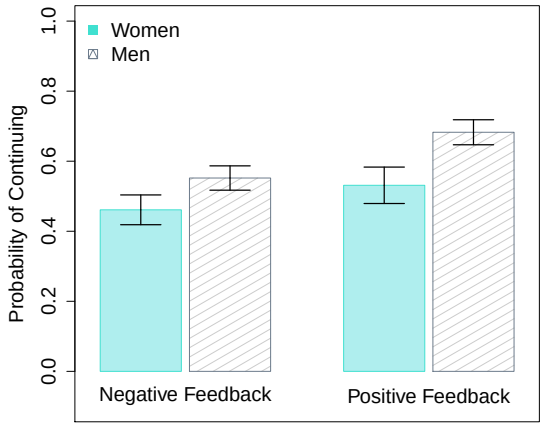
Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This table is an abbreviation of Table A1, displaying only estimates that are relevant to the *Baseline* treatment. Robust standard errors in parentheses. Controls include a dummies for Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity). Constants not displayed. "Neg. Feedback" is an indicator variable taking on 1 if subjects received a card saying that they failed the first IQ test, and 0 otherwise. *All* refers to all subjects, *Passed* and *Failed* refers to the sub-samples that passed or failed the first IQ test. The mean of the reference group shows the average probability of continuing for men who received positive feedback in the *Baseline*.

Negative versus positive feedback. It is possible that negative or positive feedback affects the continuation decision of men and women differently, for instance if women get more discouraged following negative feedback on their performance, or if men have a stronger preference to continue following positive feedback. Figure 3 plots subjects' average continuation probabilities by the sign of the feedback and gender. A perhaps surprising pattern is revealed: Men are on average more likely to continue regardless of whether they received negative or positive feedback, and the gender gap in persistence is bigger for subjects that received positive feedback.

To explore gender differences in responding to the sign of the feedback more formally, column (2) of Table 2 allows for an interaction effect of the female indicator and a negative feedback indicator. This interaction effect is statistically insignificant, i.e. the effect of the sign of the feedback does not depend on gender in this controlled environment.

Relative to equally performing men who also received positive feedback, women are about 15 p.pt. less likely to continue ($p = 0.012$). This effect is sizable: Looking at the point estimates, women who received positive feedback are on average more likely to quit than men who received negative feedback, all else equal, albeit this difference is not statistically distinguishable. Among those who received negative feedback, on the other hand, the estimated gender gap in continuing is only about half as large (7 p.pt.) and not statistically significant ($p = 0.236$).[19] We will later discuss what can explain the differences in the effect of positive versus negative feedback on the gender gap.

Figure 3: Probability of Continuing by Sign of Feedback and Gender, Baseline Treatment.



Bars represent the average probabilities of continuing, broken down by gender and negative vs. positive feedback, alongside the standard errors of each group, in the *Baseline* treatment.

Differences by first IQ test outcome. As can be seen in columns (3)-(6) of Table 2, subjects who failed the first IQ test are driving the gender gap in persistence. When controlling for the feedback but no interaction with the female indicator, Column (5) indicates that women who failed are 15 p.pt. less likely to continue than men ($p = 0.035$). Just as in the aggregate sample, this gender gap is especially pronounced for subjects who got positive feedback, conditional on which women are about 23 p.pt. less likely to continue ($p = 0.033$). The magnitude of this effect is

[19] When weighting the estimates in column (2) of Table 2 by the fraction of subjects who received positive or negative feedback, the overall gender gap in persistence in estimated to be 10.5 p.pt., i.e. essentially the same as in the simple model displayed in column (1).

especially stark when put in perspective: Among those who failed the first test, men who received *negative* feedback are more than three times as likely to continue than women who received *positive* feedback.

Result 2. *The gender gap in persistence is driven by subjects who received positive feedback and subjects who failed the first IQ test.*

3.3 Channels driving the documented gender gap in persistence

What can explain this gender gap in persistence? And why are women - relative to equally performing men - less likely to continue in this environment in response to *positive* performance feedback? The design is equipped to explore a number of different channels that might be driving this effect. In what follows, the roles of beliefs, preferences for additional feedback, and risk aversion are analyzed.

Channel 1: Gender differences in future confidence. Since continuing is only financially rewarding for subjects who pass the future IQ test, beliefs about one's probability of passing the future test will likely impact subjects' continuation decisions. One advantage of the experimental design is that we get to explore at what point in time gender differences in beliefs arise. Men might start off with more confident prior beliefs than equally performing women. This has been documented in the literature with respect to people's past performance (see Section 1), however it is unclear if men and women tend to make different projections of their future performance, given beliefs about their past performance. Furthermore, we would like to know how beliefs about the future respond to past performance feedback. In what follows, gender differences in prior beliefs, in updating, and in posterior beliefs are discussed.

Prior beliefs (before receiving feedback). Before receiving feedback on their performance, women on average are not only less confident than men about having passed the first IQ test; they are even less confident about passing the future IQ test. Panels (a) and (b) of Figure 4 show gender differences in initial confidence in the aggregate. When controlling for past performance and self-reported characteristics, women are about 7 p.pt. less confident to have passed the first IQ test, see column (1) in panel (A) of Table 3. If anything, this gender gap is even more pronounced when looking at subjects' prior confidence regarding their future performance: Women are on average almost 10 p.pt. less confident about passing the future IQ test than equally performing men, see column (2). Notably, to be as confident as men about their future performance, women on average have to score more than one standard deviation higher on the first IQ test.

One possible explanation for why men are systematically more confident about passing the future IQ test could be that they apply a more optimistic mapping when extrapolating from their beliefs about their past to their future performance. Indeed, men are more confident about passing

the future IQ test even when holding beliefs about having passed the first test constant, see column (3) of Table 3. This effect is driven by subjects who failed the first test, as Table A2 shows. This suggests that men consider previous failures to be less indicative of their future performance than women. One reason for why men who performed poorly are more likely to persist than women who performed poorly could thus be that they believe having failed previously is less predictive of their future performance.

Updating in response to feedback. Aside from the gender gap in prior confidence, do men and women update their beliefs differently after receiving performance feedback? To explore this possibility, note that Bayesian updating in this setting can be written in log-form as

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

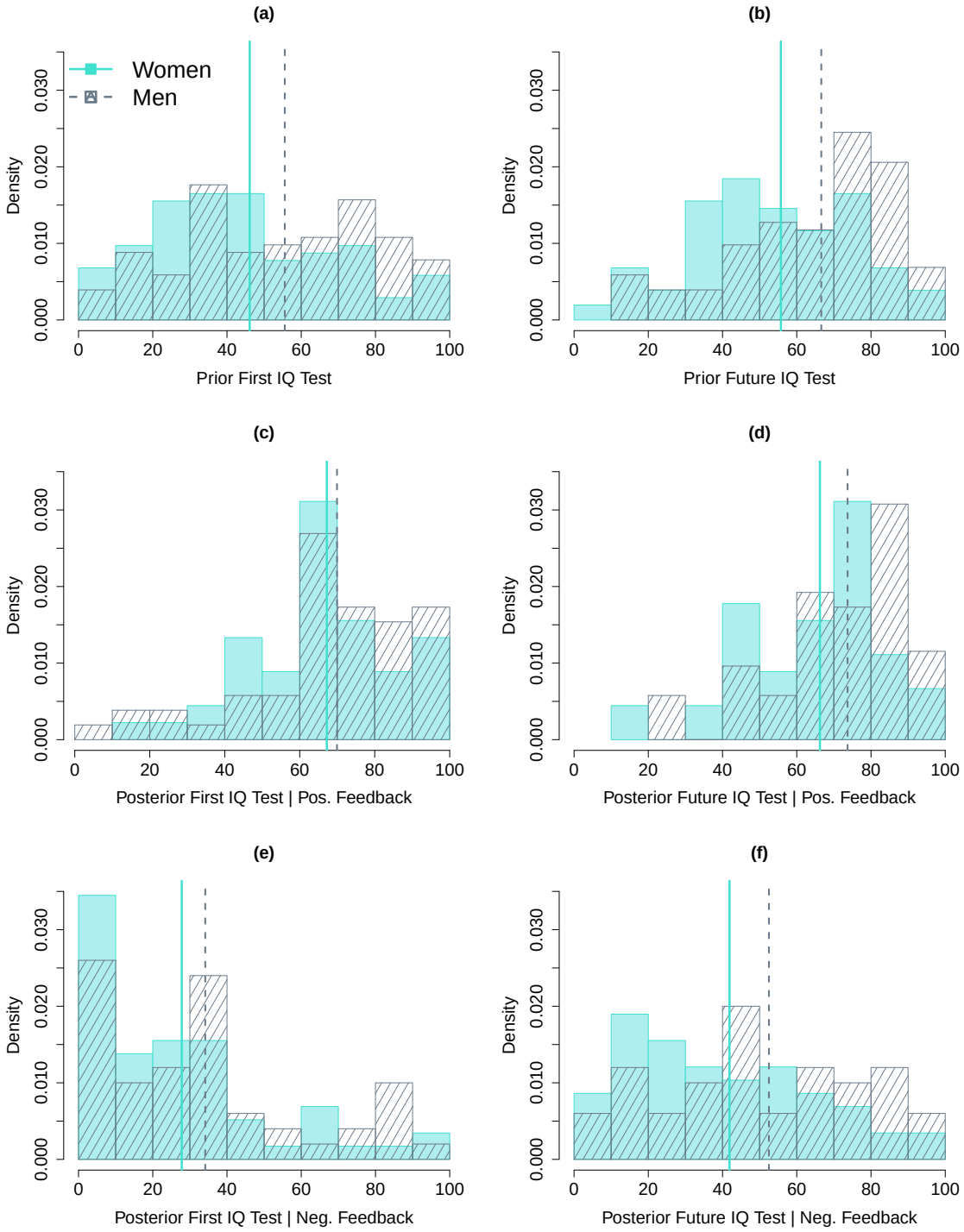
where p denotes the posterior belief, p_0 denotes the prior belief, $\mathbf{1}\{pos.\}$ and $\mathbf{1}\{neg.\}$ denote indicator functions of receiving positive or negative feedback, respectively, and ϕ denotes the probability with which the cards conveying the feedback reveal the true state of having passed or failed the first IQ test. When updating about the first IQ test, $\phi = \frac{2}{3}$ by design, however there is no objectively true value of ϕ when updating about the future.[20] Writing Bayesian updating in log form allows me to estimate linear regressions of the following form:

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

For a perfect Bayesian agent, $\alpha = \beta_p = \beta_n = 1$. If subjects put the same weight on positive and negative feedback when updating, $\beta_p = \beta_n$. Similarly, β_p or β_n bigger (smaller) than 1 would indicate over-reaction (under-reaction) to the positive or negative feedback, respectively. Finally, $\alpha < 1$ would indicate base-rate-neglect, i.e. that subjects do not place enough weight on their prior, while $\alpha > 1$ would indicate that subjects are updating conservatively, putting too much weight on their priors.

[20] In this setting, there is no Bayesian benchmark for how rational subjects should update their beliefs about their future performance in response to the past feedback. 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 forming beliefs about their future performance. In Table A3 and Table A4, estimates for $\phi = 0.62$ are shown for the future test, which fit the data in the sense that estimates for β_p and β_n were reasonably close to 1. In results not reported, values of ϕ slightly smaller or bigger yield very similar results.

Figure 4: Gender Differences in Past and Future Confidence.



Panels (a), (c), and (e) depict gender differences in past confidence, i.e. reported beliefs about having passed the first IQ test. Panels (b), (d), and (f) depict gender differences in future confidence, i.e. reported beliefs about passing the future IQ test. Panels (a) and (b) show prior beliefs, i.e. beliefs that subjects reported after taking the test, but before getting feedback on their performance. Panels (c) and (d) show posterior beliefs after having received positive, and panels (e) and (f) show posterior beliefs after having received negative feedback. Vertical lines depict the mean beliefs of each group.

Table 3: OLS Estimates of Prior and Posterior Beliefs

	First Test	Future Test	
	(1)	(2)	(3)
Panel A: Prior Beliefs			
Z-Score 1. IQ Test	11.13***	8.052***	0.658
	(1.652)	(1.585)	(1.196)
Female	-6.909**	-9.584***	-4.993**
	(3.362)	(3.070)	(2.140)
Prior 1. IQ Test			0.665***
			(0.0510)
Mean Reference Group	55.57	66.61	66.61
Observations	205	205	205
Panel B: Posterior Beliefs			
Z-Score 1. IQ Test	10.98***	8.892***	1.921
	(1.605)	(1.657)	(1.496)
Female	-0.754	-6.803*	-6.324**
	(4.201)	(3.759)	(3.096)
Neg. Feedback	-32.57***	-17.85***	2.825
	(4.675)	(4.494)	(3.558)
Neg. Feedback * Female	-0.850	-1.458	-0.918
	(6.324)	(6.018)	(4.683)
Posterior 1. IQ Test			0.635***
			(0.0589)
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. Controls include a dummies for 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. Data from *Baseline* and *AlwaysInfo* combined. Prior beliefs were reported after taking the test, but before receiving feedback. Posterior beliefs were reported after receiving feedback on their past performance. The mean of the reference group in panel (A) refers to men's average prior beliefs, and in panel (B) refers to men's average posterior beliefs conditional on having received positive feedback.

In the aggregate sample, there is no evidence that men and women place different weights on positive or negative feedback when updating about their past or future performance, see Table A3. Both men and women overreact to negative feedback when updating about the first test, but not when updating about the future IQ test. As there are no gender differences in how much weight is put on positive or negative performance feedback when updating about passing the future test, how people adjust their beliefs in response to feedback probably has a negligible role on gender differences in persistence.

Posterior beliefs (after receiving feedback). After receiving performance feedback, the gender gap in future confidence remains, but the gender gap in confidence regarding the first test closes, as panel B of Table 3 shows. More specifically, when controlling for performance on the first test and demographic characteristics, women who received positive feedback are on average about 7 p.pt. less confident about their future performance than men ($p = 0.072$), while women who received negative feedback are about 8 p.pt. less confident ($p = 0.088$), see column (2).

The gender gap in posterior future confidence is similar in magnitude, and statistically significant at the 5% level when controlling for posterior beliefs of having passed the first test in addition, see column (3): After receiving feedback, women are roughly 7 p.pt. less confident about passing the future IQ test, relative to equally performing men, and this applies to both positive and negative feedback. Similarly to before, this gender gap is primarily driven by subjects who failed the first IQ test, as panel B of Table A2 shows. Gender differences in posterior confidence at the aggregate level are visualized in panels (c)-(f) of Figure 4. It can be seen that women are on average less confident about both their first and their future test performance, both before and after receiving either positive or negative performance feedback.

How much of the gender gap in persistence can be attributed to gender differences in future confidence? Recall that in the *Baseline* treatment, women were on average about 10 p.pt. less likely to continue. Once controlling for subjects' posterior beliefs of passing the future IQ test, this gap is reduces to about 7 p.pt., see column (1) of Table A5. That is, the gender gap in future confidence can explain about 30% of the gender gap in persistence.[21]

Result 3. *Women on average are less confident about passing the future IQ test than equally performing men who receiving the same feedback. Men appear to consider previous failures to be less predictive of their future than women, however there are no gender differences in updating after receiving feedback. Beliefs account for roughly 30% of the gender gap in persistence.*

[21] In a model that allows for gender differences in responding to positive vs. negative feedback, the gender gap in future confidence can explain about 25% of the gender gap in persistence conditional on having received positive, but over 50% of the (statistically insignificant) gender gap among subjects who received negative feedback, see column (2) of Table A5. When weighting these estimates by the fraction of subjects who received positive and negative feedback, about 38% of the gender gap in persistence can be explained by gender differences in future confidence.

Channel 2: Gender differences in avoiding and seeking additional feedback. In the *Baseline* treatment, subjects only learn if they passed or failed the first IQ test if they continue, but not if they quit. In contrast, this information is provided regardless of whether subjects continue or not in the *AlwaysInfo* treatment. Therefore, if subjects are more (less) likely to continue in the *AlwaysInfo* treatment than in the *Baseline*, this can be interpreted as evidence of information avoidance (seeking). Recall that we are interested in this as persisting in challenging, ego-relevant careers often involves being exposed to additional performance feedback. If there are gender differences in preferences to get additional ego-relevant feedback, this could play a role for explaining the gender gap in persistence.

Gender differences in information preferences might partially explain the gender gap in persistence in the *Baseline* treatment if women quit in order to avoid additional feedback, if men continue to receive additional feedback, or both. As panel (a) of Figure 5 illustrates, both forces appear to be at play: in the raw data, women are on average about 4 p.pt. *more* likely to continue, and men are about 5 p.pt. *less* likely to continue if they learn their test result even as they quit (as in the *AlwaysInfo* treatment).

Panel (b) of Figure 5 breaks this behavioral response down by the sign of the feedback in the raw data. Men on average have a preference to learn their test outcome regardless of whether they received positive or negative feedback on their performance. For women, however, the average response is more asymmetric: women who received positive feedback have a strong preference not to learn their true test result. They are about 11 p.pt. more likely to continue in the *AlwaysInfo* treatment than in the *Baseline*, which is consistent with the idea that they want to hold on to the "glimmer of hope" that the positive feedback conveys. Women who receive negative feedback, on the other hand, on average do not act differently in the *Baseline* and the *AlwaysInfo* treatment, i.e. on average they neither have a preference to avoid, nor to receive additional feedback.

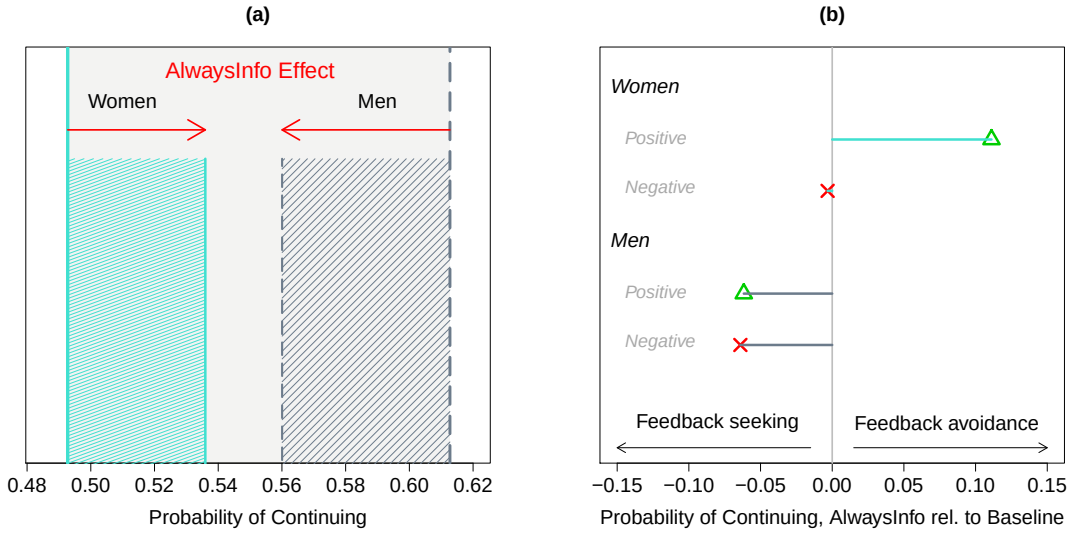
When controlling for subjects' test performance, feedback, and self-reported characteristics, estimates are directionally consistent with the idea that women prefer to avoid additional feedback, while men embrace it, see column (1) of Table A7. With the exception of women who received positive feedback (for which $p = 0.056$), however, this effect is not statistically significant. It is possible that the estimates presented for the total sample mask some heterogeneity of information preferences. For example, women who received negative information and failed the first test might want to avoid learning their test outcome in order to hold on to the "glimmer of hope" that the negative feedback might have been wrong after all. Similarly, women who received negative information but passed the first test might prefer learning their test outcome to prove the negative feedback wrong. These effects would cancel each other out, thus masking information preferences of women who received negative feedback at the aggregate.

To analyze if the *AlwaysInfo* treatment effect varies by subjects' actual test results, columns (2)-(3) of Table A7 provide estimates separately for subjects who passed and failed the first IQ test.

Men who received positive feedback and passed the first IQ test are on average 13.5 p.pt. less likely to continue under the *AlwaysInfo* than the *Baseline* treatment ($p = 0.046$), which suggests that seeking additional confirmation of their high performance can be a motive for men to continue.

In contrast, getting positive feedback can be a motive to quit for women who failed the first test; Women who failed the first test but received positive performance feedback are 25 p.pt. more likely to quit in the *Baseline* treatment, which is consistent with the idea that women have a strong preference not to "ruin the good state" by learning that the positive feedback was actually wrong and that they failed the first test. For subjects who received negative feedback, on the other hand, the sub-group exhibiting a sizable effect, albeit only statistically significant at the 10% confidence level, are women who passed the first test. They are about 18 p.pt. less likely to continue if they passed the first test ($p = 0.056$), which weakly supports the idea that women who passed but got negative feedback are more likely to drop out if quitting allows them to avoid the exposure to additional (potentially negative) performance feedback. Put differently, women might decide not to got after opportunities in the field in order to indirectly avoid finding out that they might not be as talented or skilled as they had hoped.

Figure 5: AlwaysInfo Treatment Effect Relative to Baseline Treatment



Panel (a) shows how the probability of continuing changes for women and men in the *AlwaysInfo* treatment, relative to the *Baseline*. The gray shaded area in the background represents the gender gap in the *Baseline* and corresponds to the difference in means in 3panel (b) of Figure 2. Panel (b) shows the differences in means of the *AlwaysInfo* treatment and the *Baseline*, separately for subjects who received positive and negative feedback by gender.

It is worth noting that preferences to receive or avoid additional feedback may in part be driven by beliefs. That is, the estimated treatment effect displayed in columns (1)-(3) of Table A7 might to some extent reflect gender differences in future confidence, rather than "pure" informational

preferences. To account for this possibility, columns (4)-(6) of Table A7 display estimates of the *AlwaysInfo* treatment effect that further control for subjects' posterior beliefs of passing the future IQ test. Qualitatively, not much changes when controlling for these beliefs, i.e. the treatment effect is not primarily driven by gender differences in future optimism.

To what extent can gender differences in information avoidance and information seeking - independent of beliefs - explain the gender gap in persistence? To answer this question, consider the following back-of-the-envelope calculation: Weighting the estimates of column (4) in Table A7 by the fraction of the respective groups in the *Baseline* treatment suggests that slightly more than 40% of the documented gender gap in persistence can be explained by gender differences in feedback avoidance.[22] Note, however, that not estimates of the *AlwaysInfo* effect are statistically significant from zero. A more conservative approach of estimating the impact of the *AlwaysInfo* treatment effect would thus be to only consider subgroups for which the effect is statistically significant at the 10% level or higher, and assume that the treatment has no effect on the other sub-groups in the sample. Using this conservative approach, still about 30% of the gender gap can be explained by gender differences in information preferences.

[22]For example, consider panel (A) of Table 4. In the *Baseline* treatment, 46% of subjects are men and 54% are women. Thus, the gender gap in the *AlwaysInfo* treatment is $6.7 * 0.46 + 2.5 * 0.54 = 4.42$ p.pt. smaller than in the *Baseline*, where there is a 10.3 p.pt. gender gap in persistence. Thus, $4.42/10.3 = 42.9\%$ of the gap in persistence can be explained by gender differences in information preferences. Essentially the same result is obtained if weighting the estimates of panels (B) and (C) by their corresponding fractions in the *Baseline*.

Table 4: Effect of AlwaysInfo Treatment

	Probability of Continuing		
	(1)	(2)	(3)
	All	Passed	Failed
Panel A: Aggregate			
Men	-0.067*	-0.100	-0.109
	(0.037)	(0.062)	(0.069)
Women	0.025	-0.019	0.045
	(0.042)	(0.061)	(0.064)
Panel B: Positive Feedback			
Men	-0.058	-0.121*	-0.103
	(0.045)	(0.064)	(0.087)
Women	0.106*	0.028	0.189**
	(0.059)	(0.070)	(0.095)
Panel C: Negative Feedback			
Men	-0.071	-0.032	-0.124
	(0.051)	(0.104)	(0.085)
Women	0.029	-0.121	-0.007
	(0.056)	(0.092)	(0.077)
Controlling for Beliefs	Yes	Yes	Yes

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*, separately by gender, first IQ test outcome, and sign of the feedback. All estimates control for subjects' posterior beliefs of passing the future IQ test, a dummy for Zoom vs. in-person sessions, and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity).

Result 4. *Gender differences in preferences to receive additional performance feedback account for roughly* 40% *of the documented aggregate gender gap in persistence. On average, women avoid, and men seek the information of having passed the first IQ test.*

Channel 3: Gender differences in risk preferences. Quitting implies a positive minimum payment while continuing only pays off if subjects pass the second IQ test. If women are more averse to take risks than men, this could constitute another channel driving the documented gender differences in persistence.

Appendix D provides details on how risk parameters are estimated. As Table A6 shows, women are on average not more risk averse than men in this experiment. It is therefore not surprising that the estimated gender gap in persistence in the total sample is essentially unaffected when including risk parameters as a control, see columns (3)-(6) in Table A5.

3.4 Efficiency considerations

So far, this paper has documented that women are more likely than men to drop out of a challenging environment when controlling for past performance and feedback, and that gender differences in beliefs and preferences for additional performance feedback can explain a large fraction of this gap. The following analysis discusses to what extent the gender gap in persistence is inefficient, and concludes that the self-selection of women better predicts their performance. In addition, Appendix E discusses payoff implications of the differential sorting of men and women at the individual level.

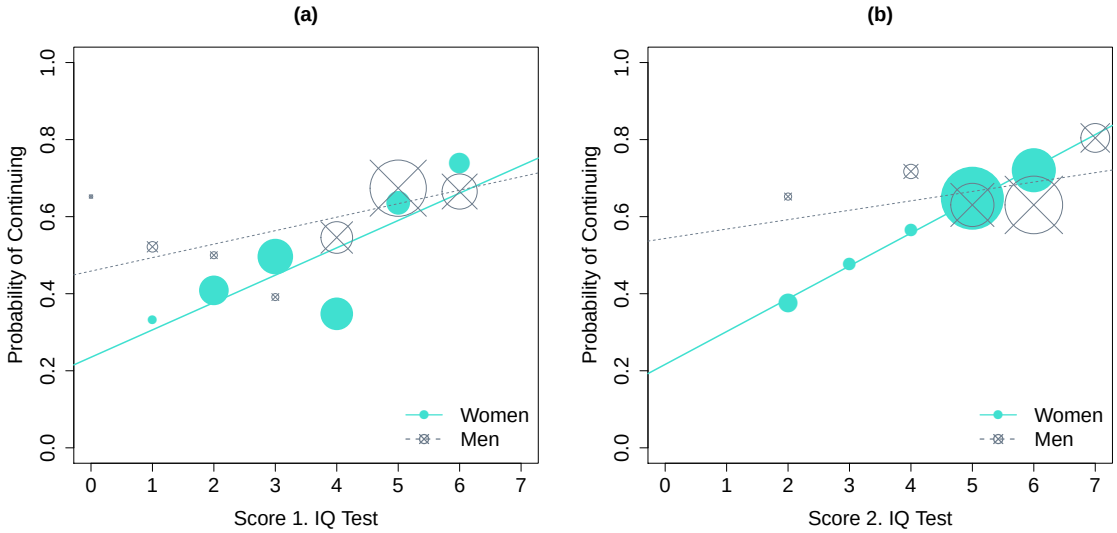
If it is beneficial for society that the highest performing individuals persist in challenging career paths, a natural question to ask is whether "more able" individuals are indeed more likely to do so. In the context of this experiment, ability is measured twice; through the score on the first IQ test, and - conditional on continuing - the score on the second IQ test. These measures are different, as a subject's first IQ test outcome is no perfect predictor of their future test outcome.[23] Conditional on continuing and having passed (failed) the first IQ test, 85% (67%) pass the second IQ test in the *Baseline* treatment, and the correlation coefficient between test scores is 0.49.

Are subjects who perform better on the first IQ test more likely to continue? When taking past performance as a measure of ability, the self-selection of women that persist in the challenging environment appears to better match their ability. As shown in panel (a) of Figure 6, women who score one point higher on the first IQ test are on average about 7 p.pt. more likely to continue ($p = 0.001$), while men who score one point higher are only about 4 p.pt. more likely to continue ($p = 0.017$). Furthermore, conditional on having failed the first test, women on average continue with a probability of 41%, while men continue with a probability of 52%. That is, if subjects' first IQ test outcome is taken as a measurement of their ability, "less able" men are on average more likely to persist than "less able" women ($p = 0.059$), i.e. men are more negatively selected. Conditional on having passed the first test, however, the average probability of continuing is essentially the same for men and women (67% and 68%, respectively; $p = 0.753$).

[23] It is a natural feature of many real world environments of interest that people's performance is not perfectly correlated over time. For example, students do not receive the same test scores on each exam, or a researcher's publications do not always end up at similarly ranked journals.

When considering future performance as a measure of ability, the pattern that women tend to better self-select info continuing is even more pronounced: As panel (b) of Figure 6 shows, on average women who score one point higher on the future IQ test were ex ante 9 p.pt. more likely to continue ($p = 0.001$), but men who scored one point higher on the future IQ test were ex ante not significantly more likely to continue ($p = 0.349$).

Figure 6: Probability of Continuing by Test Scores and Gender



Data from the *Baseline* treatment are visualized. Points represent the mean probability of continuing for each score on the first IQ test, separately for men and women. The straight lines represent the fitted OLS regression lines of each group. The size of the points reflects the relative share of the number of subjects with a certain score of each group. Panel (a) displays the relationship between performance on the first IQ test and decision to continue. Panel (b) displays the relationship between performance on the second IQ test and decision to continue, for the sub-sample of subjects that continued.

Table 5: Prediction of Passing 2. IQ Test

	(1) Probit	(2) Heckman Probit
Selection: Continue		
Probability of Continuing		3.224**
		(1.293)
Female		0.175
		(0.925)
Prob. of Cont. * Female		0.174
		(1.510)
Neg. Feedback		0.489*
		(0.292)
Passed 2. IQ Test		
Probability of Continuing	-1.036	-2.289*
	(1.875)	(1.215)
Female	-4.060**	-2.885***
	(1.664)	(1.056)
Prob. of Cont. * Female	5.934**	4.029**
	(2.551)	(1.714)
Mean Reference Group	66.04	66.04
Observations	45	94

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Constant not displayed. Data from the *Baseline* treatment are presented. Column (1) presents estimates of a Probit regression model. Column (2) presents estimates of a Heckman Probit regression model that accounts for sample selection. The probability of continuing captures a subject's switch point in the BDM of the main decision task, relative to the support of possible switch points. The mean of the reference group captures the average probability of passing the second IQ test for men who continued.

One shortcoming of taking subjects' performance on the second IQ test as a measure of ability is that the sample of subjects who continue - and thus the sample for which the second IQ test outcome can be observed - is selected. To address this issue, Table 5 presents two different estimates

of the probability of passing the second IQ test: Column (1) displays estimates of a conventional Probit regression, while column (2) displays estimates of a Heckman Probit regression that accounts for sample selection. More specifically, the selection of subjects who continue is estimated as a function of their reported switch point in Part 3 of the experiment (which directly translates into the theoretical probability of continuing), an indicator for being female, an interaction of the two, and an indicator for having received negative feedback. A comparison of the estimates presented in the two columns indicates that while being female is positively correlated with the selection of subjects that continued, the estimates are qualitatively similar. That is, (i) women who perform better on the second IQ test have a higher ex-ante probability of continuing, and (ii) that this does not apply to men for reasons beyond selection.

Result 5. *Women with a higher ex-ante probability of continuing are on average more likely to pass the second IQ test. This is not the case for men.*

4 Discussion

Revisiting the effects of positive versus negative performance feedback. One of the findings of this paper is that while there is a significant gender gap in persistence following positive feedback, this gap is insignificant following negative feedback. Another perspective to look at this pattern is the following: While men are significantly more likely to *continue* if they receive positive than if they receive negative feedback - all else equal - this is not the case for women. In other words, women who received positive feedback are *not* significantly more likely to continue than women who received negative feedback.

This homogeneous response of women in response to positive versus negative feedback is especially surprising considering that women - just like men - update their beliefs upwards following positive, and downwards following negative feedback, as discussed in Section 3.3. Positive feedback appears to provoke two opposite dynamics for women; on the one hand, this feedback makes them more optimistic about their performance, but on the other hand they shy away from learning that the positive feedback might have been wrong. In other words, the belief channel and the information avoidance channel operate into different directions. While men also become more optimistic about their performance after receiving positive feedback, they do not shy away from learning that they might have failed the first test. In sum, this generates a gender gap in persistence following positive performance feedback.

Other factors potentially contributing to gender differences in persistence. Together, gender differences in beliefs and preferences for additional feedback account for roughly 70% of the gender gap in persistence in this controlled environment. That being said, an interesting question is what could be driving the part of the gender gap in persistence that cannot be explained by

these factors. As discussed in Section 3.3, estimated risk parameters do not affect the gender gap in persistence.

One possibility is that there are gender differences in seeking challenges, however this would be inconsistent with the literature: Niederle and Yestrumskas (2008) 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 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 task.

Another possibility is that the decision to persist is to some extent governed by self-image concerns. The option to *continue* might be more attractive than the option to *quit* for someone who gets utility from holding a self-image of "not giving up easily," all else equal. It is possible that men on average attribute a higher value to holding such a self-image than women, but identifying this channel is beyond the scope of this experiment.

5 Conclusion

Using a controlled laboratory experiment, this paper has documented that men - relative to equally performing women - are more likely to persist in an environment that rewards high performance and involves exposure to ego-relevant performance feedback. To a large extent, this effect can be explained by two channels. First, women are systematically less confident about their future performance, and conditional on these beliefs their expected returns from persisting are lower. Second, men on average enjoy receiving additional feedback that is revealed upon continuing, while women prefer to avoid it.

What are the implications of this experiment for understanding gender differences in persistence in the field? Consistent with previous applied work, this experiment finds gender differences in persistence following feedback in a controlled setting, where it can be ruled out that men and women face different opportunity costs, different returns to persisting, or do receive or self-select into different feedback.

Recall that this experiment focuses on a one-time decision in response to a one-time provision of feedback, and finds substantial gender differences in persistence in this controlled environment. Considering that the decision to persist or drop out of stratified career trajectories such as academia, politics, or corporate management is made many times along the way of pursuing such a career path, the compound effect of gender differences in persistence following feedback potentially plays a notable role in explaining the under-representation of women in these fields. That being said, the people persisting are more and more selected with each step up the career ladder, e.g. how a female full professor responds to performance feedback is unlikely representative of women in general.

References

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. 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.

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*.

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.

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.

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.

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 (2021). Social Exchange of Motivated Beliefs. *Journal of the European Economic Association*.

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.

Thaler, M. (2021). The Fake News Effect: Experimentally Identifying Motivated Reasoning Using Trust in News. *Working Paper*.

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.

Appendices

A Additional Figures and Tables

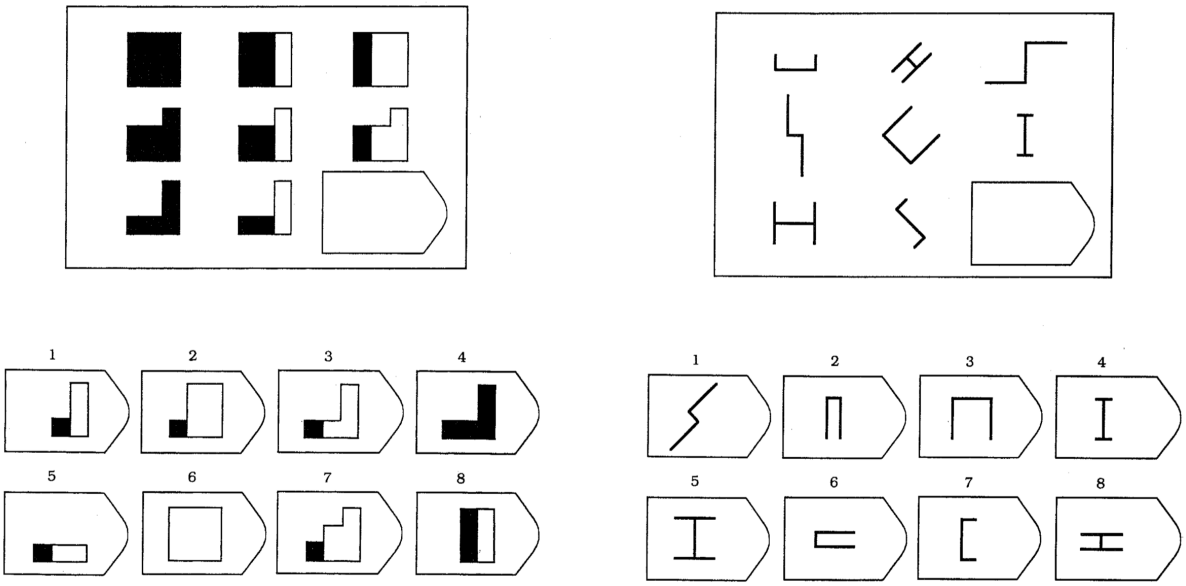


Figure A1: Example of one relatively easy and one difficult Raven's Progressive Matrix used in the experiment. Subjects were asked to find the piece which completes the pattern.

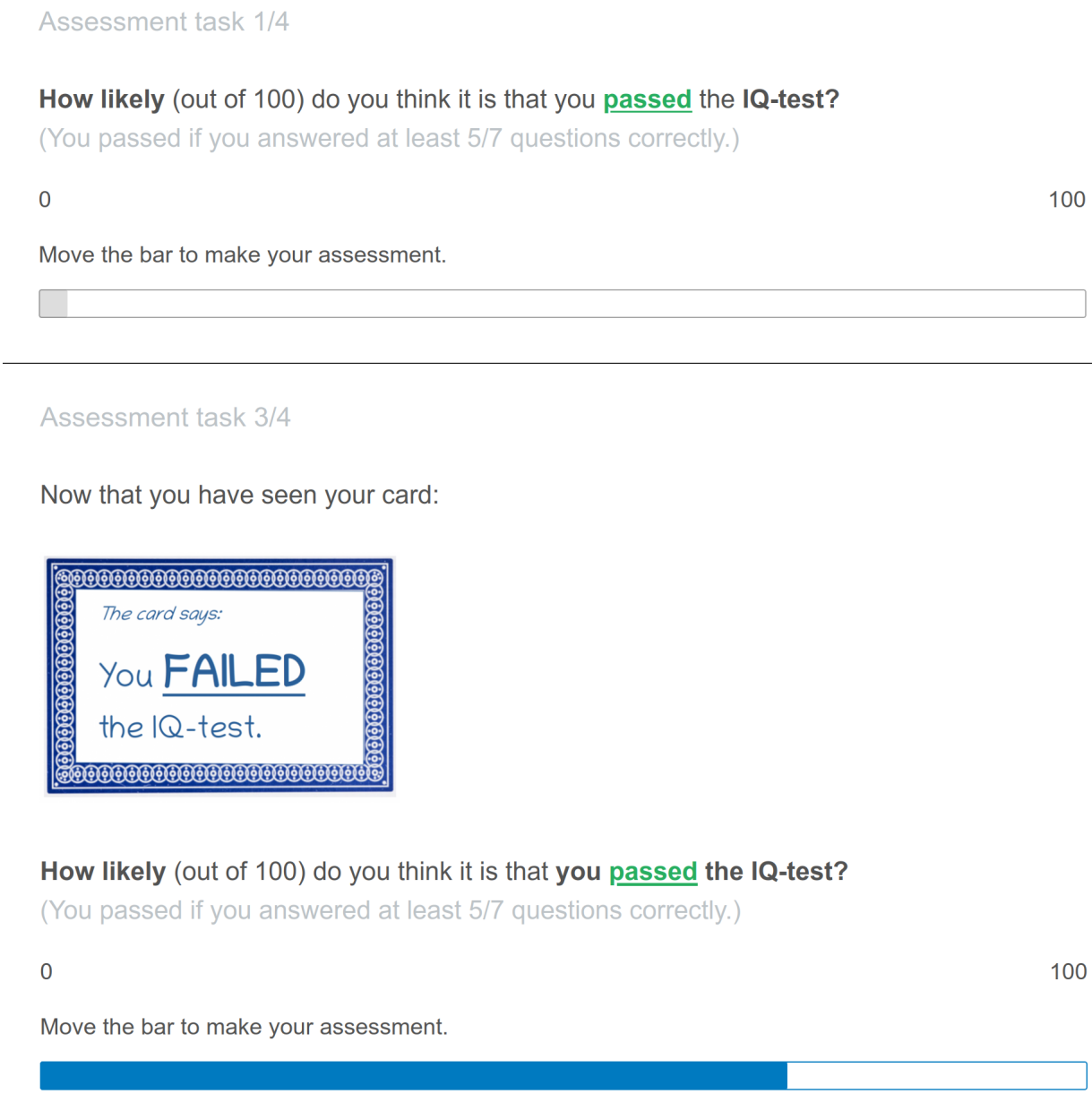


Figure A2: The upper panel shows a screenshot of how the prior belief of having passed the first IQ test was elicited, before clicking on the bar to report a number from 0 to 100. When clicking on the bar, the number showed up next to the position of the screen. The lower panel shows a screenshot of how the posterior belief of having passed the first IQ test was elicited, after having received negative performance feedback. The initial position of the bar represents the initial position of the reported prior belief on the same question. Subjects could bring the bar in a new position in response to the feedback.

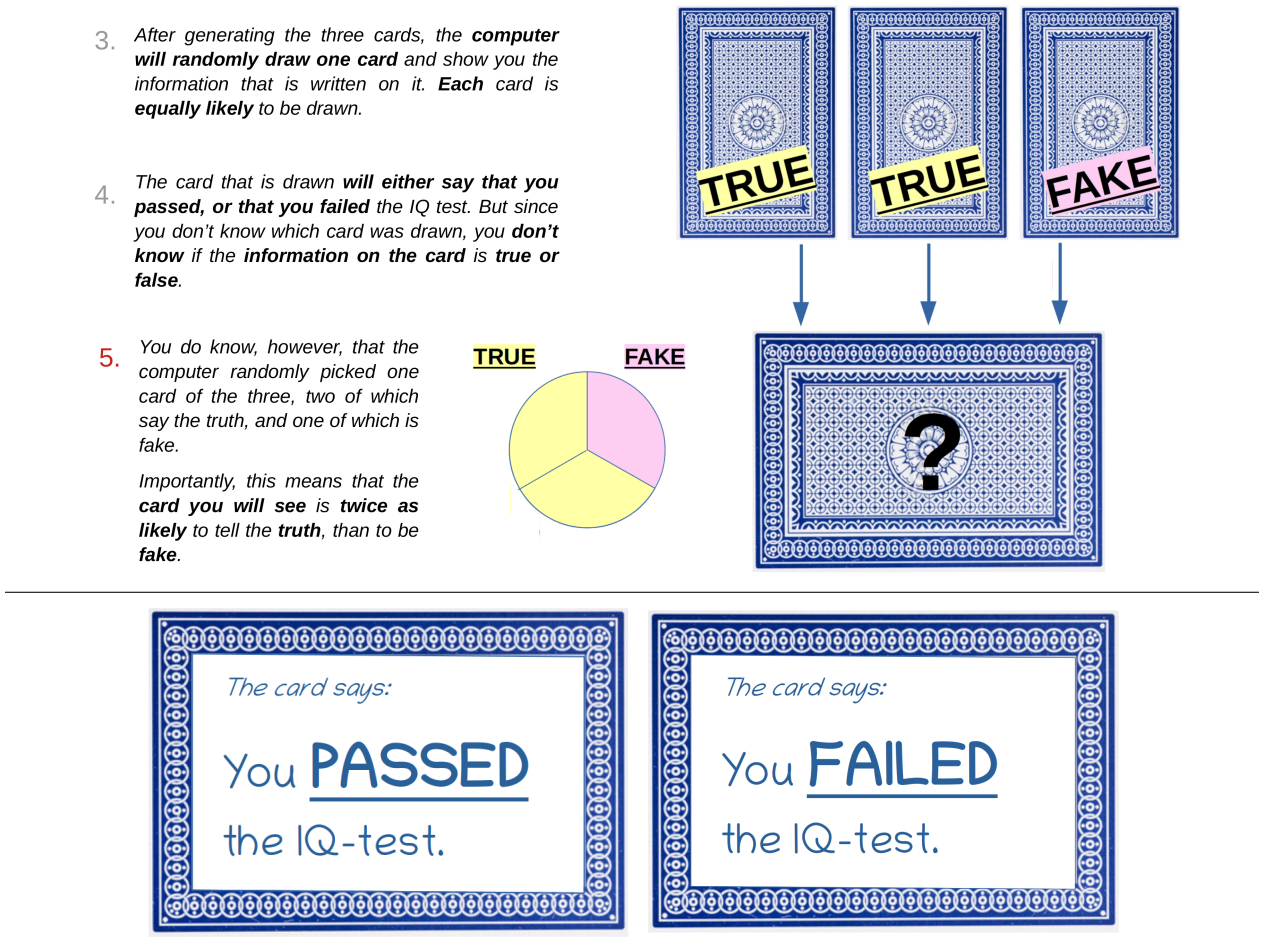


Figure A3: The upper panel shows a screenshot of part of the instructions used to explain the feedback in Part 2. The lower panel displays the cards shown to subjects; Subjects either saw a card saying that they passed, or that they failed the IQ test.

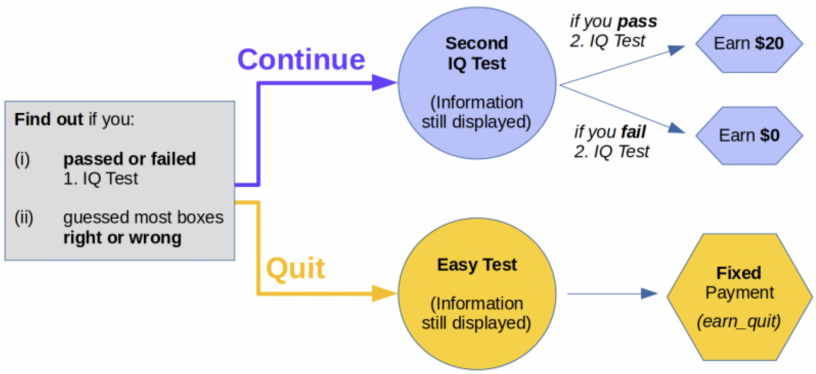


Figure A4: Screenshot of the overview of what happens if subjects continue or quit in the *AlwaysInfo* treatment.

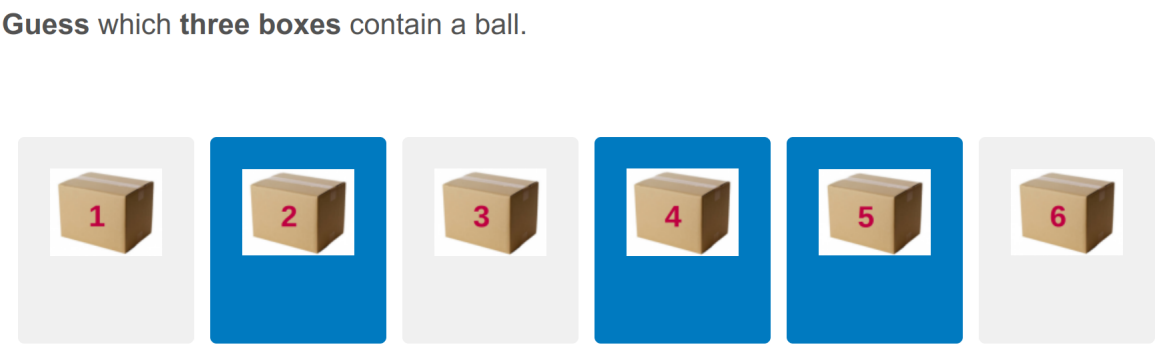


Figure A5: Screenshot displaying the *Guessing Game* subjects were asked to take at the beginning of the experiment. Subjects had to guess which 3 of the 6 boxes contain a ball. After clicking on a box, the background color of the box changed from gray to blue.

Table A1: OLS Estimates of the Probability to Continue

	Probability of Continuing					
	(1)	(2)	(3)	(4)	(5)	(6)
	All	**All**	Passed	Passed	Failed	Failed
Z-Score 1. IQ Test	0.0612***	0.0604***	0.0522	0.0736	-0.00956	-0.0116
	(0.0154)	(0.0156)	(0.0537)	(0.0569)	(0.0301)	(0.0311)
Female	-0.103**	-0.145**	-0.00738	-0.0628	-0.153**	-0.226**
	(0.0422)	(0.0571)	(0.0514)	(0.0513)	(0.0713)	(0.104)
Neg. Feedback	-0.106***	-0.120**	-0.137***	-0.187***	-0.0718	-0.0668
	(0.0281)	(0.0486)	(0.0417)	(0.0568)	(0.0434)	(0.0967)
Neg. Feedback * Female		0.0753		0.123		0.100
		(0.0797)		(0.0937)		(0.128)
AlwaysInfo	-0.0591	-0.0639	-0.0890	-0.135**	-0.0770	-0.0631
	(0.0427)	(0.0512)	(0.0633)	(0.0663)	(0.0846)	(0.0980)
AlwaysInfo * Female	0.109*	0.182**	0.0338	0.138	0.182*	0.313**
	(0.0560)	(0.0712)	(0.0938)	(0.0882)	(0.0925)	(0.136)
AlwaysInfo * Neg. Feedback		0.0157		0.144		-0.0327
		(0.0702)		(0.132)		(0.111)
AlwaysInfo * Female * Neg. Feedback		-0.129		-0.324*		-0.167
		(0.109)		(0.164)		(0.164)
Mean Reference Group	0.68	0.68	0.76	0.76	0.55	0.55
Observations	205	205	84	84	121	121

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Controls include a dummies for 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. "Neg. Feedback" is an indicator variable taking on 1 if subjects received a card saying that they failed the first IQ test, and 0 otherwise. *All* refers to all subjects, *Passed* and *Failed* refers to the sub-samples that passed or failed the first IQ test. The first four lines of this table correspond to Table 2.

Table A2: OLS Estimates of Prior and Posterior Beliefs

	First Test		Future Test			
	(1)	(2)	(3)	(4)	(5)	(6)
	Passed	Failed	Passed	Failed	Passed	Failed
Panel A: Prior Beliefs						
Z-Score 1. IQ Test	2.605	6.000**	1.311	5.129	-0.205	0.633
	(7.949)	(2.865)	(6.299)	(3.296)	(4.571)	(2.271)
Female	2.188	-7.327	0.669	-12.90***	-0.605	-7.413**
	(5.690)	(4.436)	(4.329)	(4.306)	(3.311)	(2.848)
Prior 1. IQ Test					0.582***	0.749***
					(0.0625)	(0.0714)
Mean Reference Group	66.25	44.46	73.29	59.66	73.29	59.66
Observations	84	121	84	121	84	121
Panel B: Posterior Beliefs						
Z-Score 1. IQ Test	-0.132	5.966**	0.835	5.147	0.914	1.273
	(8.379)	(2.733)	(7.362)	(3.326)	(5.713)	(2.744)
Female	3.543	2.776	1.231	-11.87*	-0.890	-13.68**
	(5.835)	(8.143)	(5.224)	(7.084)	(3.820)	(5.931)
Neg. Feedback	-24.38***	-34.84***	-8.079	-25.01***	6.514	-2.387
	(7.602)	(7.186)	(6.840)	(7.083)	(4.396)	(5.784)
Neg. Feedback * Female	3.396	-2.505	1.719	4.954	-0.313	6.581
	(13.40)	(9.728)	(10.96)	(9.191)	(7.088)	(7.414)
Posterior 1. IQ Test					0.599***	0.649***
					(0.0785)	(0.0756)
Mean Reference Group	77.23	58.25	77.22	68.05	77.22	68.05
Observations	84	121	84	121	84	121

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Controls include a dummies for 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. Data from *Baseline* and *AlwaysInfo* combined. Panel (A) shows prior beliefs, i.e. beliefs that subjects reported after taking the test, but before receiving feedback. Panel (B) shows posterior beliefs, i.e. beliefs that subjects reported after receiving feedback on their past performance. *Passed* and *Failed* refers to the sub-samples that passed or failed the first IQ test. The mean of the reference group in panel (A) refers to men's average prior beliefs, and in panel (B) refers to men's average posterior beliefs conditional on having received positive feedback.

Table A3: OLS Estimates of Log-Likelihood Bayesian Updating

	First Test		Future Test	
	(1)	(2)	(3)	(4)
	All	All	All	All
	$\phi = \frac{2}{3}$	$\phi = \frac{2}{3}$	$\phi = 0.62$	$\phi = 0.62$
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)
$H_0 : \alpha = 1$	0.011	0.196	0.211	0.328
$H_0 : \beta_p = 1$	0.148	0.690	0.251	0.351
$H_0 : \beta_n = 1$	0.000	0.006	0.475	0.664
$H_0 : \beta_p = \beta_n$	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. Columns (1)-(2) estimate belief updating on the first IQ test, where $\phi = \frac{2}{3}$ by design. Columns (3)-(3) estimate equation for the future test for $\phi = 0.62$, as the true ϕ - how informative the first signal is on the future test - is neither known to the subjects nor the experimenter. When a belief of 100 (0) was reported, this was coded as 99 (1) so that the log likelihood was well defined for all subjects. The second to sixth last rows show p-values associated with the corresponding hypothesis tests.

Table A4: OLS Estimates of Log-Likelihood Bayesian Updating, by First Test Result

	First Test				Future Test			
	(1) Passed $\phi = \frac{2}{3}$	(2) Failed $\phi = \frac{2}{3}$	(3) Passed $\phi = \frac{2}{3}$	(4) Failed $\phi = \frac{2}{3}$	(5) Passed $\phi = 0.62$	(6) Failed $\phi = 0.62$	(7) Passed $\phi = 0.62$	(8) Failed $\phi = 0.62$
α	0.650***	0.975***	0.597***	1.208***	0.858***	0.935***	0.769***	0.982***
	(0.0933)	(0.104)	(0.137)	(0.198)	(0.112)	(0.0822)	(0.207)	(0.130)
βp	1.625***	1.083***	1.719***	0.679***	1.018***	0.791***	1.309***	0.300
	(0.246)	(0.165)	(0.356)	(0.182)	(0.209)	(0.207)	(0.366)	(0.296)
βn	0.995***	1.683***	0.986***	1.700***	0.605	1.216***	0.422	1.007***
	(0.243)	(0.228)	(0.323)	(0.401)	(0.402)	(0.166)	(0.681)	(0.303)
α * Female			0.134	-0.328			0.179	-0.0267
			(0.155)	(0.226)			(0.211)	(0.163)
βp * Female			-0.230	0.732**			-0.607	0.900**
			(0.472)	(0.304)			(0.421)	(0.386)
βn * Female			-0.0161	-0.0272			0.269	0.349
			(0.455)	(0.483)			(0.741)	(0.357)
$H_0 : \alpha = 1$	0.000	0.809	0.004	0.297	0.209	0.434	0.268	0.890
$H_0 : \beta_p = 1$	0.013	0.614	0.047	0.081	0.932	0.317	0.400	0.020
$H_0 : \beta_n = 1$	0.984	0.003	0.966	0.083	0.329	0.196	0.398	0.981
$H_0 : \beta_p = \beta_n$	0.116	0.043	0.190	0.023	0.442	0.107	0.354	0.036
$H_0 : \beta_p * Female = \beta_n * Female$	-	-	0.768	0.208	-	-	0.397	0.234
Observations	84	121	84	121	84	121	84	121

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Variants of equation 2 are estimated. When updating about the first IQ test, $\phi = \frac{2}{3}$ by design. When a belief of 100 (0) was reported, this was coded as 99 (1) so that the log likelihood was well defined for all subjects. The second to sixth last rows show p-values associated with the corresponding hypothesis tests.

Table A5: OLS Estimates of the Probability to Continue.

	(1) All	(2) All	(3) All	(4) All	(5) All	(6) All
Z-Score 1. IQ Test	0.0311**	0.0291*	0.0530***	0.0513***	0.0581***	0.0561***
	(0.0156)	(0.0156)	(0.0164)	(0.0166)	(0.0167)	(0.0169)
Female	-0.0673*	-0.109**	-0.104**	-0.150**	-0.114**	-0.153**
	(0.0371)	(0.0502)	(0.0428)	(0.0591)	(0.0447)	(0.0600)
Neg. Feedback	-0.0418	-0.0410	-0.110***	-0.115**	-0.124***	-0.118**
	(0.0292)	(0.0448)	(0.0286)	(0.0554)	(0.0291)	(0.0568)
Neg. Feedback * Female		0.0760		0.0810		0.0684
		(0.0723)		(0.0818)		(0.0840)
Posterior 2. IQ Test	0.00346***	0.00354***				
	(0.000698)	(0.000690)				
CRRA Risk Parameter			-0.0305***	-0.0296***		
			(0.00995)	(0.00941)		
CARA Risk Parameter					-0.481**	-0.460**
					(0.197)	(0.186)
Mean Reference Group	0.68	0.68	0.68	0.68	0.68	0.68
Observations	205	205	178	178	182	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. Controls include a dummies for 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. Column (1) in this table corresponds to Column (1) of Table 2 as a reference. 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*.

Table A6: OLS Estimates of Risk Parameters

	CRRA Risk Paramenter			CARA Risk Paramenter		
	(1)	(2)	(3)	(4)	(5)	(6)
	All	Passed	Failed	**All**	Passed	Failed
Female	0.0103	-0.137	-0.0898	-0.00805	-0.00452	-0.0129
	(0.316)	(0.279)	(0.441)	(0.0161)	(0.0144)	(0.0222)
Mean Reference Group	-0.109	-0.126	-0.093	-0.073	-0.067	-0.079
Observations	178	66	112	182	67	115

Notes: Robust standard errors in parentheses. Controls include a dummies for 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 refers to men's average estimated risk parameters.

Table A7: Effect of AlwaysInfo Treatment

	Probability of Continuing					
	(1)	(2)	(3)	(4)	(5)	(6)
	All	Passed	Failed	**All**	Passed	Failed
Panel A: Aggregate						
Men	-0.059	-0.089	-0.077	-0.067*	-0.100	-0.109
	(0.043)	(0.063)	(0.085)	(0.037)	(0.062)	(0.069)
Women	0.050	-0.055	0.105	0.025	-0.019	0.045
	(0.050)	(0.062)	(0.076)	(0.042)	(0.061)	(0.064)
Panel B: Positive Feedback						
Men	-0.064	-0.135**	-0.063	-0.058	-0.121*	-0.103
	(0.051)	(0.066)	(0.098)	(0.045)	(0.064)	(0.087)
Women	0.118*	0.002	0.250**	0.106*	0.028	0.189**
	(0.064)	(0.070)	(0.105)	(0.059)	(0.070)	(0.095)
Panel C: Negative Feedback						
Men	-0.048	0.009	-0.096	-0.071	-0.032	-0.124
	(0.060)	(0.124)	(0.103)	(0.051)	(0.104)	(0.085)
Women	0.005	-0.178*	0.050	-0.029	-0.121	-0.007
	(0.064)	(0.091)	(0.088)	(0.056)	(0.092)	(0.077)
Controlling for Beliefs	No	No	No	Yes	Yes	Yes

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*. All estimates control for a dummies for Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity). Columns (4)-(6) further control for subjects' posterior beliefs of passing the future IQ test. Panel A shows the estimated treatment effect for the model where no interaction effect of the female indicator and the negative feedback is included, e.g. columns (1), (3), and (5) in Table 2 and Table A1. Panels B and C correspond to the model allowing for such interaction effects, i.e. columns (2), (4), and (6) of the same tables.

B Additional Design Elements

Mechanism Used to Implement Main Decision Task and Risk Task. Subjects were given two options in the main decision task (continue vs. quit), as well as the risk task (lottery vs. fixed payment). Rather than asking subjects to directly choose one of the two options, the minimum fixed payment for which they preferred quitting over continuing (in Part 3), and the fixed payment over the lottery (in Part 4) were elicited, using an incentive-compatible BDM procedure (Becker et al., 1964). The instructions to implement the BDM in this experiment are largely based on Healy (2020).

Figure A6 shows a screenshot of how the BDM was presented to subjects in Part 3 of the *Baseline* treatment. There was a list of 23 questions, and in each question subjects could choose between *Option A (to quit)* or *Option B (to continue)*. The only feature varying across questions was the amount of *Earn_A* - the fixed payment associated with *Option A* - which increased from \$0 to \$22 in one-dollar-increments. Subjects were told that it was assumed they would prefer *Option A* in the first few questions (i.e. when *Earn_A* was high), but at some point would prefer *Option B*. Subjects were then asked to report their "switch point" - the dollar value of *Earn_A* at which they would like to switch from *Option A* to *Option B*. As one of the questions was randomly drawn after subjects reported their switch point, this mechanism is incentive-compatible. Note that a subject's reported switch point in the main decision task, divided by 23, can be interpreted as their preferred ex-ante probability of continuing.

Using a BDM has two advantages in this context: First and foremost, subjects' valuation of quitting relative to continuing can be observed, yielding richer data than a binary choice of whether to continue or quit. Second, 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 a subject who continued, had they quit, which is important for individual welfare considerations, see Appendix E.

Emphasis was put on implementing the BDM in a way that is understandable and intuitive for subjects. To familiarize subjects with how the BDM works and how their decision affects their outcome, a practice BDM was introduced before explaining the actual decision task.[24] A number of visual and interactive features made the BDM especially intuitive to use.[25]

[24] The practice BDM consisted of two generic options - *Option A* and *Option B*. While Option A implied to take *Path A* and earn some fixed amount *Earn_A*, Option B implied to take *Path B* with no fixed payment. Subjects were told that they would later learn what all of these mean.

[25] The colors of the two options (orange for *Option A* and purple for *Option B*) in the list of questions and the instructions corresponded to the colors of the slider. 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 ensured that after bringing the slider bar into a position, subjects could see what their current switch point implies before submitting their choice.

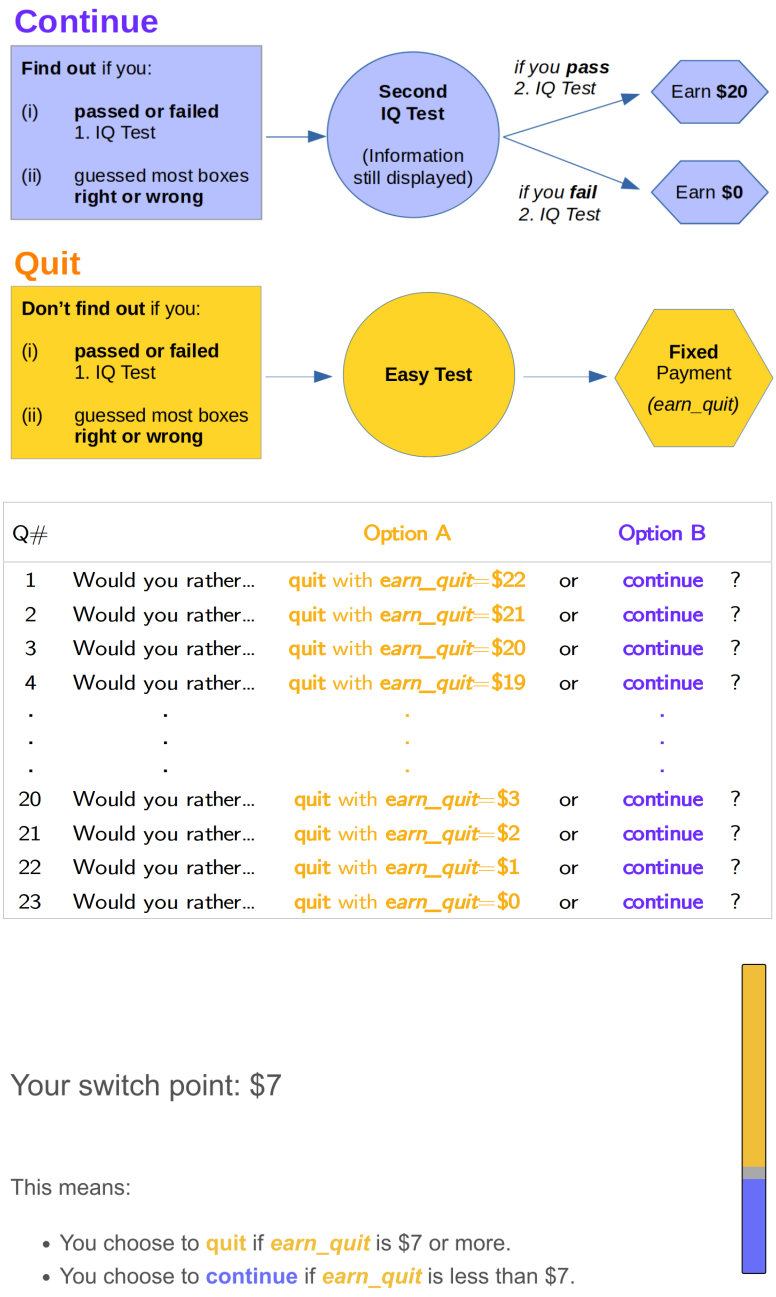


Figure A6: Screenshot of the BDM decision interface in the *Baseline* treatment. Subjects see an overview of what happens if they continue or quit, a list of questions referring to their preferences for either option under different quitting payments, and a slider to report the switch point after which they would like to switch from Option A to Option B.

Guessing Game at the Beginning. Before the main part of the experiment began, a trivial *"Guessing Game"* was conducted. This game is not meaningful in the *Baseline* or the *AlwaysInfo* treatment. The reason for including it was to keep things consistent with a third treatment for which the data may be collected in the future.[26]

Survey at the End. After completing the risk task, subjects filled out a short survey. This survey included demographic questions such as gender and race, academic information such as chosen major and GPA, as well as some open-form qualitative questions.

C Experimental Instructions

Instructions at the very beginning of the experiment: here.

Instructions before the IQ test: here.

Instructions before eliciting prior beliefs: here.

Instructions before providing feedback: here.

Instructions before eliciting posterior beliefs: here.

Instructions before practice BDM: here.

Instructions before main decision (continue/quit) - *Basline* treatment: here.

Instructions before main decision (continue/quit) - *AlwaysInfo* treatment: here.

Instructions before risk task: here.

D 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, and 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 the experiment when making their decision in Part 4, subject i's reported switch point in Part 4 therefore implies that

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

[26] In the *"Guessing Game"*, subjects had to guess which 3 out of 6 closed boxes contain a ball, see Figure A5. Correct guessed were not rewarded financially, and subjects were not told the correct answer. After subjects submitted their guesses, it was announced that the main experiment would begin.

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 such as CRRA or CARA.[27] In what follows, risk parameters are computed as the mean of that interval, separately under the assumption of CRRA and CARA utility functions.

E Individual returns to continuing versus quitting.

Computing the returns to continuing at the individual level requires estimating counterfactual outcomes: How much would have subjects who continued earned, had they quit, and vice versa?

Recall that conditional on reporting the same switch point in Part 3 of the experiment, it is random who continues and quits. This does not imply, however, that subjects who reported the same switch point provide a valid counterfactual for one another, as switch points also reflect individuals' preferences and beliefs. In particular, it is unclear what subjects who quit would have earned, had they continued. We do know, however, how much subjects who continued and reported a switch point s would have earned, had they quit: By construction of the BDM, their expected earnings of quitting are $\frac{s+22}{2}$, while their actual earnings of continuing are \$20 if they passed, and \$0 if they failed the second IQ test. With this in mind, for each switch point one can compare the average earnings of subjects who continued with their counterfactual earnings, had they quit. This addresses the question of whether, among the subjects who continued, those who had a higher ex-ante probability of continuing earned more from continuing relative to their expected earnings for quitting, and whether this differs by gender.

As Figure A7 shows, subjects that continued on average would have earned more money in Part 3 of the experiment, had they quit. In this figure, subjects are grouped by quintiles of their probability of continuing, separately by gender.[28] The average premium of continuing is computed as the difference between a quintile's average earnings for continuing and a quintile's average (theoretical) earnings for quitting. Figure A7 illustrates that for women, the average premium of continuing tends to increase with their ex-ante probability of continuing, however on average the theoretical earnings of quitting exceed the achieved earnings of continuing across the distribution. More specifically, women who continue on average lose between $\$1 - \$7$ in experimental earnings relative to their expected earnings for quitting. For men, a slightly different picture emerges: Among those who continue, the 20% with the lowest probability of continuing (i.e. Quintile 1) on average earned about \$3 more from continuing than if they had quit. Most other men who continued, however, could have increased their expected earnings by quitting more often.

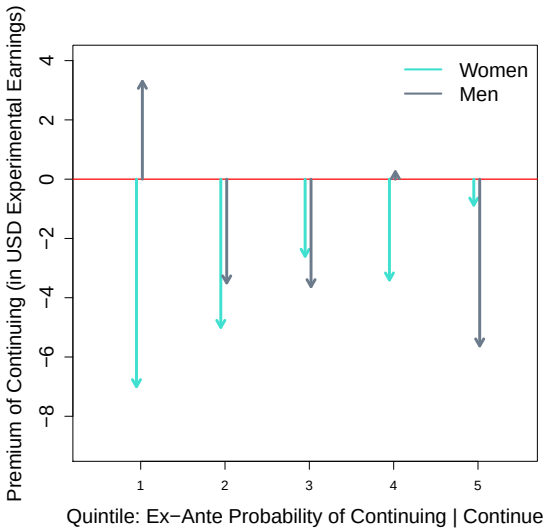
In sum, this back-of-the-envelope calculation suggests that on average, subjects who continued in the experiment would have earned more by quitting. This insight may be surprising considering that among those who continued, the majority (78%) passed the second IQ test. When taking subjects' outside option into consideration, however, those who continued but failed forwent substantial earnings associated with quitting, so that the average premium of continuing is negative for most subjects, including subjects who had a high ex-ante probability of continuing, e.g. subjects in

[27] 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, $U(x,r) = \frac{e^{-rx}}{r}$.

[28] That is, after ranking all subjects that continued by their probability of continuing (separately by gender), Quintile 1 captures the 20% of subjects with the lowest probability of continuing, etc.

Quintile 5.

Figure A7: Average Premium of Continuing by Quintiles: Probability of Continuing



Data from the *Baseline* treatment are visualized for the subset of subjects that continued. The premium of continuing is computed as the difference between a group's average earnings for continuing and a group's average (theoretical) earnings for quitting.