

UC Santa Barbara

UC Santa Barbara Electronic Theses and Dissertations

Title

Essays in Experimental and Applied Microeconomics

Permalink

<https://escholarship.org/uc/item/26k5z3q5>

Author

Kogelnik, Maria

Publication Date

2022

Peer reviewed|Thesis/dissertation

University of California
Santa Barbara

Essays in Experimental and Applied Microeconomics

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Economics

by

Maria Kogelnik

Committee in charge:

Professor Ryan Oprea, Chair
Professor Peter Kuhn
Professor Heather Royer
Professor Sevgi Yuksel

September 2022

The Dissertation of Maria Kogelnik is approved.

Professor Peter Kuhn

Professor Heather Royer

Professor Sevgi Yuksel

Professor Ryan Oprea, Committee Chair

June 2022

Essays in Experimental and Applied Microeconomics

Copyright © 2022

by

Maria Kogelnik

to Vroni

Acknowledgements

I am grateful to everyone who has supported me during my PhD studies. I am deeply indebted to my advisors for their guidance and encouragement. I thank Ryan Oprea and Sevgi Yuksel for being incredibly generous with their time, and for inspiring and encouraging what researcher I want to become. I thank Heather Royer for being a truly wonderful mentor and coauthor, and Peter Kuhn for advocating for me from early on.

Two chapters of this dissertation are based on joint work with my coauthors. I thank Mireille Jacobson and Heather Royer; Hazem Alshaikhmubarak, Dave Hales, Molly Schwarz and Kent Strauss; and - in ongoing work - Florian Hoffmann, Thomas Lemieux and Mirko Titze; for letting me work with you and learn from you.

I have been very fortunate to do my PhD at UC Santa Barbara. In lieu of everyone who has contributed to a welcoming and supportive environment at the economics department and the graduate program, I thank Kelly Bedard, Dick Startz, and Mark Patterson for all their work.

The best part of my graduate studies have been the friends I met along the way. Thank you for all the shared memories, discussions and laughter.

Finally, danke Philipp, for always believing in me, and for always being there for me and with me.

Curriculum Vitæ

Maria Kogelnik

Education

- 2022 Ph.D. in Economics (Expected), University of California, Santa Barbara.
- 2017 M.A. in Economics, University of California, Santa Barbara.
- 2016 M.A. in Economics, University of British Columbia.
- 2015 M.Sc. in Applied Economics, Leopold-Franzens University Innsbruck.
- 2013 B.Sc. in Management and Economics, Leopold-Franzens University Innsbruck.

Publications

“Holiday, Just One Day Out Of Life: Birth Timing and Post-natal Outcomes” with Mireille Jacobson and Heather Royer. *Journal of Labor Economics* (2021), 39(S2): S651 – S702.

Permissions and Attributions

1. Chapter 2 (and Appendix B) is the result of a collaboration with Mireille Jacobson and Heather Royer, and has previously appeared in the *Journal of Labor Economics*, 39(S2), S651-S702. The published version is available at <https://www.journals.uchicago.edu/doi/10.1086/712493>. The journal permits republication in dissertations with proper citation.
2. Chapter 3 (and Appendix C) is the result of a collaboration with Hazem Al-shaikhbumarak, David Hales, Molly Schwarz and Kent Strauss.

Abstract

Essays in Experimental and Applied Microeconomics

by

Maria Kogelnik

The first chapter studies gender differences in persistence in response to performance feedback. The decision to persist in stratified career trajectories is often dynamic in nature: people receive 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. About one-third of this gap can be explained by gender differences in beliefs about the future. In the laboratory as well as a field study, men are more optimistic about their future performance even when compared to women who are similarly confident about their past performance. Another 30% of the gender gap in persistence is attributable to men seeking, and women avoiding exposure to additional feedback.

The second chapter is based on joint work with Mireille Jacobson and Heather Royer. We observe that fewer births occur on major US holidays than would otherwise be expected. We use California data to study the nature and health implications of this birth date manipulation, and document 18 percent fewer births on the day of and just after a holiday. C-sections account for roughly half of the decline. “Missing” holiday births are moved to the roughly two-week window both before and after the holiday. High-risk births are more likely to be re-scheduled than low-risk births. Despite the documented change in timing, we find little evidence of any adverse health consequences for babies born around a holiday.

The third chapter is based on joint work with Hazem Alshaikhmubarak, David Hales, Molly Schwarz and Kent Strauss. We experimentally study how mutual payoff information affects play in strategic settings. Subjects play the Prisoner's Dilemma or Stag Hunt game against randomly re-matched opponents under two information treatments. In our partial-information treatment subjects are shown only their own payoffs, while in our full-information treatment they are shown both their own and their opponent's payoffs. In both treatments, they receive feedback on their opponent's action after each round. We find that mutual payoff information initially facilitates reaching the Pareto-efficient outcome in both games. While play in the Prisoner's Dilemma converges toward the unique Nash equilibrium of the game under both information treatments, mutual payoff information has a substantial impact on the equilibrium selection in the Stag Hunt throughout all rounds of the game. Using a belief-learning model and simulations of play, we provide evidence that these effects are driven not only by initial play but also by the way subjects learn. We propose that strategic uncertainty is a probable channel through which payoff information affects play.

Contents

Curriculum Vitae	vi
Abstract	viii
1 Performance Feedback and Gender Differences in Persistence	1
1.1 Introduction	1
1.2 Experimental Design	8
1.3 Results	15
1.4 Efficiency of the Different Self-selection of Men and Women	25
1.5 Discussion	28
2 Holiday, Just One Day Out Of Life: Birth Timing and Post-natal Outcomes	42
2.1 Introduction	42
2.2 Background on Medical Delivery Interventions	49
2.3 Conceptual Framework: How Holidays affect Birth Timing	53
2.4 Empirical Approach	56
2.5 Results	65
2.6 Robustness and Sensitivity Analysis	79
2.7 Conclusions	84
3 Knowing Me, Knowing You: An Experiment on Mutual Payoff Information and Strategic Uncertainty	106
3.1 Introduction	106
3.2 Experimental Design	113
3.3 Descriptive Results	117
3.4 Estimating a Learning Model and Simulations	121
3.5 Discussion	130
3.6 Conclusion	133

A	Appendix to Chapter 1	145
A.1	Additional Figures and Tables	146
A.2	Additional Design Elements	156
A.3	Instructions and Experimental Interface	159
A.4	Classroom Field Study	233
A.5	Estimation of Risk Parameters	235
A.6	Individual Returns to Continuing versus Quitting	236
B	Appendix to Chapter 2	239
B.1	Additional Figures and Tables	239
C	Appendix to Chapter 3	255
C.1	Material Presented to Subjects during the Experiment	255
C.2	Additional Tables	265

Chapter 1

Performance Feedback and Gender Differences in Persistence

1.1 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. Such feedback – information that people receive about their past performance – is often times either positive or negative, and ego-relevant in the sense that people may care about this feedback for reasons beyond its instrumental value. If men and women respond differently to this information on their past performance, for instance if men become overly optimistic about their future following positive feedback, or if women drop out to avoid being exposed to negative

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

feedback, this could help explain the gender differences in persistence we observe.

This paper presents a laboratory experiment designed to study (i) whether men are more likely than women to persist in an environment that rewards high performance and involves exposure to feedback, and – if so – (ii) what channels are driving this gender gap in behavior. A focus is put on the roles of beliefs about the future as well as preferences for additional feedback exposure. In addition, this paper presents a classroom field study designed to test the external validity of the belief formation patterns documented in the lab.

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

The idea that men and women may respond differently to feedback on their performance is consistent with a recent empirical literature. Women have been found to be less likely than men to continue in STEM and economics majors in response to poor grades (Katz et al., 2006; Rask and Tiefenthaler, 2008; Kugler et al., 2021; Astorner-Figari and Speer, 2019), less likely to participate again in prestigious math exams, math olympiads, Rubik’s Cube competitions, or college entry exams after scoring low previously (Ellison and Swanson, 2018; Franco, 2018; Buser and Yuan, 2019; Fang et al., 2021; Kang et al., 2021), less likely to submit an article to the largest economics conference in Brazil fol-

lowing a previous rejection (Pereda et al., 2020), and less likely to re-run for office after barely losing an election (Wasserman, 2021).² Gender differences in persistence may be easier detectable in response to negative feedback, when many people leave a career trajectory at the same time. It is also conceivable, however, that positive feedback has a more encouraging effect on men than on women, which is equally relevant to understand gender differences in persistence, and may have different policy implications.

The first goal of the experiment 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 ego-relevant feedback. In the *Baseline* treatment, subjects are asked to perform a challenging and ego-relevant task (an IQ test), which they either pass or fail. They then receive feedback – an informative message that is either positive or negative. To explore the effect of positive versus negative feedback across the performance distribution, this feedback is randomized conditional on having passed or failed, and on known accuracy. Subjects then face two options: If they *continue*, they are exposed to additional feedback, take a second IQ test, and receive a high bonus if they pass this future test, but nothing otherwise. Alternatively, if they *quit*, they receive no more performance feedback, complete an easy test, and receive a fixed payment that does not depend on their performance on the easy test.

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%. Men who received negative feedback are just about as likely to continue as men who received positive

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

feedback.

The second goal of the experiment is to explore what channels are driving this gender gap in persistence. As continuing is only financially rewarding for subjects who pass the second IQ test, the first channel of interest is how people form beliefs about their future performance. Gender differences therein may be present at the stage of prior beliefs before feedback, may arise when people update their beliefs in response to feedback, or both. Furthermore, men and women may differ in how they extrapolate from past experiences when forming beliefs about their future; They could hold different beliefs about whether their past performance is predictive of their future success, and they could adjust these beliefs differently in response to ego-relevant feedback. A novel feature of the design is that it allows us to disentangle these mechanisms by eliciting subjects' beliefs about their past and future performance both before and after receiving feedback. Reporting true beliefs is incentivized.

Women are found to be less confident about passing the future IQ test both before and after receiving feedback, relative to equally performing men. Interestingly, men make more optimistic projections of their future performance even compared to women who are similarly confident about their past performance. This suggests that when men form beliefs about their future, they may discount how predictive previous failures are, or over-weigh how predictive previous successes are of their future – relative to equally performing women. Consequently, men's expected returns from persisting are higher. In response to feedback, however, there is no evidence of gender difference in updating. Roughly one-third of the gender gap in persistence is attributable to gender differences in beliefs about passing the future test.

To examine the outside validity of the gender differences in beliefs documented in the lab, a classroom field study is conducted. In this study, undergraduate students are asked to report beliefs about their past and future performance on midterm exams after taking

the first exam, but before learning their grade. Findings in the field are remarkably similar to the lab not only qualitatively but also in terms of the effect size of the gender gap in confidence. Controlling for past exam scores, women are less confident both about their past and future performance. Importantly, men also make more optimistic projections of their future, given their beliefs about their past performance.

The second channel of interest concerns gender differences in preferences for feedback. Persisting on a career path often involves exposure to additional feedback. If women avoid this feedback, or if men seek it, this could help explain gender differences in persistence. Recall that in the *Baseline* treatment, subjects receive additional feedback if they continue, but not if they quit, which makes quitting relatively more attractive for subjects who want to avoid exposure to additional feedback. To study gender differences in feedback avoidance and feedback seeking, the design includes one treatment arm where subjects receive additional feedback regardless of whether they continue or quit. This *AlwaysInfo* treatment thus shuts down preferences for 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 that gender differences in information avoidance may account for almost 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. These estimates of the *AlwaysInfo* treatment effect control for beliefs to ensure that gender differences in preferences for feedback do not reflect the documented gender differences in confidence.

The design further allows us to explore the role of risk preferences on the gender gap in persistence. 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 has essentially no impact on the estimated gender gap in persistence.³

Performance feedback mechanisms may contribute to a gender gap in ability within organizations if low-performing men are more likely to persist, or if high-performing women are less likely to continue. In the experiment, men are adversely selected when taking past performance as a measure of ability. As people's past performance is naturally no perfect predictor of their future performance, however, this does not imply that women's continuation decisions better predict their performance. By dropping out, women forgo the opportunity of learning that their performance may improve over time, and that persisting may pay off later on.

This paper makes three main contributions to the literature. First, to 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 absolute performance feedback affects persistence and gender differences therein. As there is no competition or feedback on one's relative performance in this experiment, this finding does not reflect gender differences in the willingness to compete (e.g., see the seminal work of Niederle and Vesterlund, 2007).⁴ Related experiments have studied how feedback on one's relative performance affects gender differences in choosing a hard over an easy task (Niederle and Yestrumskas, 2008), in setting goals for one's future performance

³Niederle (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.

⁴While stratified careers are commonly described as competitive, they often do not involve direct tournaments. It is therefore of interest if gender differences in persistence arise even when there is no direct competition.

(Buser, 2016), in choosing a competitive over a piece-rate payment scheme (Berlin and Dargnies, 2016; Buser and Yuan, 2019), and in choosing compensation schemes across different quiz domains (Coffman et al., 2021).

Second, this paper presents the first evidence that men – even when compared to women who are similarly confident about their *past* performance – tend to be substantially more confident about their *future* performance, both before and after receiving feedback. This insight contributes to the literature on gender differences in confidence and belief formation in response to feedback.⁵

Finally, by presenting an experimental design that allows us to isolate the role of gender differences in feedback avoidance on persistence, this paper contributes to a relatively under-studied literature on how preferences for information can affect economic behavior. 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) find that gender differences in avoiding the information of having won or lost previously can explain the gender gap in competition in the first, but not in later rounds of an adding numbers task. Eil and Rao (2011) and Mobius et al. (2011) 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 aver-

⁵With the exception of Alan and Ertac (2019) – who study how children's beliefs about their future, but not their past performance, respond to feedback – this literature largely focuses on beliefs about subjects' past performance: Women have often been found to have less confident prior beliefs about their past 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. The evidence on gender differences in updating is mixed; While Mobius et al. (2014) and Coutts (2018) find that women update more conservatively, Berlin and Dargnies (2016) document over-reaction to feedback for both men and women, and find that women are updating more pessimistically than men. Furthermore, Coffman et al. (2019) find over-reaction to information that refers to a gender-congruent domain. Looking at a dynamic setting, Coffman et al. (2021) find that one week after receiving negative feedback, women are more pessimistic about their past performance than men, holding constant initial beliefs.

age WTP, but note that women are more likely than men to require a subsidy for this information.⁶

The remainder of this essay is organized as follows. Section 1.2 describes the experimental design. Section 1.3 presents evidence on gender differences in persistence, and analyzes what channels are driving this gender gap. Section 1.4 discusses whether gender differences in persistence contribute to a gender gap in ability within organizations. Finally, Section 1.5 concludes.

1.2 Experimental Design

Design goals and overview. The first goal of the design is to create a controlled setting to study gender differences in persistence. Such a setting should mimic the essential features of a stratified career trajectory that rewards high performance and involves exposure to ego-relevant feedback, as well as of an outside option. This requires a challenging task with incentives to perform well; ego-relevant feedback that can be positive or negative, but is exogenous conditional on past 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 reward 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 of the design is to explore what channels may be driving gender differences in persistence, with a focus on channels that cannot be isolated using naturally occurring data. These channels are (i) beliefs about one's future performance, and how these beliefs respond to past feedback, (ii) preferences to receive or avoid additional feedback, and (iii) risk preferences.

⁶In Eil and Rao (2011), these differences are not statistically significant.

The experiment consists of four main parts that are described below. To eliminate income effects and incentives to hedge, one of the four main parts was randomly drawn for payment at the end. In addition to a show-up fee of \$5, subjects earned a bonus payment that could range between \$0 and \$22 in the part drawn for payment. To credibly implement both treatments of the experiment, subjects were not told which part was drawn for payment. Aside from the four main parts, the design included some additional elements such as a survey at the end, see Appendix A.2.

A timeline of the experiment is provided in Figure 1.1. Instructions clarified how to earn money before each part, however subjects did not know what would happen in later parts of the experiment. Subjects had to correctly answer comprehension quizzes at different points of the experiment before moving on. A between-design was used, i.e., all subjects participated in either the *Baseline* or the *AlwaysInfo* treatment. The only component that differs across treatments is what happens if subjects quit in Part 3 of the experiment, see below. Instructions and screenshots of the experimental interface (including comprehension quizzes) can be found in Appendix A.3.

Part 1: IQ test. Subjects were asked to take an IQ test, consisting of seven Raven (1973)'s Progressive Matrices, including a range from relatively easy to relatively difficult matrices. Raven's matrices are frequently used in economics experiments to generate an environment where ego utility is at stake (e.g., Zimmermann, 2020; 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.

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 *not* depend 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: Performance feedback and beliefs. Feedback was conveyed in form of a binary signal. All subjects got to see one card that either said that they passed, or that they failed the IQ test, as depicted in Figure 1.2. This feedback was randomized and matched the true state of having passed/failed with a known accuracy of two-thirds. In other words, subjects who passed the IQ test were twice as likely to see a card saying that they passed, than seeing a fake card telling them that they failed, and vice versa. Randomizing feedback has the advantage that the effect of receiving positive versus negative feedback can be explored for all subjects, regardless of how they performed. Furthermore, providing feedback through this known process ensures that there is no gender bias, and that men and women cannot endogenously affect what kind of feedback they are getting.

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

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

(Announcement of future IQ test.)

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

What is novel about eliciting these two beliefs both before and after feedback is that doing so allows us to explore gender differences in (i) how people form about their future, given beliefs about their past performance; and (ii) how these beliefs respond to feedback.

If Part 2 was drawn for payment, subjects either earned a bonus of \$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 that it only requires monotonic preferences, but does not expected utility preferences or risk neutrality to be truth-inducing.

Part 3: Continue or quit. The main outcome of interest in the experiment is how subjects choose between the two options of *continuing* and *quitting*. Subjects' continuation probabilities serve as a measure of persistence in the experiment, see Section ?? for details. The two options vary in terms of (i) the additional feedback subjects get, (ii) the difficulty of the task they face, and (iii) the payment scheme. 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.

Continue. This option aims to mimic the consequences of persisting on a career path that rewards high performance and involves frequent exposure to ego-relevant feedback. Subjects first learned if they really passed or failed the first IQ test.⁸ They were

⁷In 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 $(100 - Y)\%$. 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.

⁸In addition, subjects learned if they guessed most boxes right or wrong in a trivial "Guessing Game." This information was included to give the researcher the option of running an additional treatment arm at a later point in time, and is held constant across the *Baseline* and *AlwaysInfo* treatment. See Appendix A.2 for details.

then asked to take a second IQ test that resembled the first IQ test in terms of style and difficulty. The information of having passed or failed was further displayed next to each question of the second IQ test in order to create frequent feedback exposure. If Part 3 was drawn for payment, subjects who continued earned a bonus of \$20 if they passed, and \$0 if they failed the second IQ test. Consequently, continuing was only financially rewarding for subjects who could pass the second test.

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

The only feature that distinguishes the *Baseline* from the *AlwaysInfo* treatment is whether or not subjects who quit learn if they really passed or failed the first IQ test. In the *Baseline*, subjects who quit did not learn if they passed or failed the first IQ test, and thus could avoid this additional feedback by quitting.¹⁰ In contrast, subjects in the *AlwaysInfo* treatment were exposed to this additional feedback regardless of whether they continued or quit, as the name suggests.¹¹ This treatment thus shuts down preferences for additional feedback as a motive for continuing or quitting. Comparing behavior across the two treatments therefore allows us to isolate the role of information avoidance and information seeking on the gender gap in persistence.

Part 4: Risk task. If Part 4 was drawn for payment, subjects either received a fixed minimum payment (see below) or a lottery that paid \$20 with some probability p , and \$0

⁹Having an easier outside option feels natural and helps to keep opportunity costs of time similar across the two options.

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

¹¹They 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.

with some probability $100 - p$. These two options (a lottery versus a fixed payment) were analogous to the two options in Part 3 (continuing versus quitting), but stripped from all features other than payoffs and risk.¹² 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' preferred *switch point* – defined as the lowest payment for quitting so that they would prefer quitting over continuing.¹³ The higher this requested minimum payment for quitting, the higher was the chance that they would continue, and vice versa. Special emphasis was put on implementing the BDM in an understandable and intuitive way, see Appendix A.2.

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. 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, see Appendix A.6.

¹²The 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 that it was not deceptive to tell subjects that they would maximize their chance of winning \$20 by always reporting their true beliefs if Part 2 was drawn for payment.

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

1.2.1 Implementation

The experiment was implemented using Qualtrics code programmed by the author, and subjects made decisions on a computer. Roughly one third of all sessions was conducted in the EBEL laboratory at the University of California, Santa Barbara, in February and March of 2020. Due to the Covid-19 pandemic, the data collection had to be paused and was eventually moved online. The remaining sessions were conducted over Zoom in the summer of 2020. All features of the experiment were kept as similar as possible between in-person and Zoom sessions. Instructions were displayed on slides on the screen and read out loud by the experimenter in both in-person and Zoom sessions. Subjects were asked to keep their video turned on throughout the experiment in Zoom sessions. To preserve anonymity, the name of subjects in Zoom sessions was changed to numbers before admitting participants from the waiting room. Subjects then received a link to the experiment in the Zoom chat, and stayed in the Zoom meeting throughout the experiment.

All subjects were recruited from the EBEL subject pool using the Online Recruitment System for Economic Experiments (ORSEE) recruiting software (Greiner, 2015a). Subjects signed up to participate in an experiment “on the economics of decision making,” and gender was neither mentioned during 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 A.2. 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).

1.3 Results

1.3.1 Data overview

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

As Table 1.1 shows, men and women in the *Baseline* sample differ along a few dimensions. Men were significantly more likely to pass the first IQ test ($p = 0.003$), and on average could solve almost one more question of the seven questions on the test correctly ($p = 0.007$). In terms of self-reported characteristics, women on average reported a slightly higher GPA than men ($p = 0.004$).¹⁵ Furthermore, while the share of subjects who reported a STEM field or Economics/Accounting as their major or intended major is directionally higher for men than for women, these differences are not statistically significant. To account for these gender differences in self-reported characteristics, unless otherwise noted, regressions in this paper control for all self-reported characteristics listed in Table 1.1, as well as a dummy variable for whether sessions were conducted in person or over Zoom.

Gender differences in persistence in the raw data. As a measurement of persistence, a subject’s ex-ante desired probability of continuing is used, which can be derived

¹⁴Six 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.

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

directly from their reported switch point in Part 3 of the experiment.¹⁶ To get a first intuition for gender differences in persistence in the raw data, Figure 1.3 shows an empirical CDF of subjects' probability of continuing in the *Baseline* treatment, separately for men and women. In the raw data, i.e., before controlling for subjects' performance and the feedback they received, men's empirical CDF first-order stochastically dominates the empirical CDF of women. The vertical lines in Figure 1.3 depict that men's average continuation probability in the *Baseline* treatment is 61%, while for women it is only 49%, thus constituting a gender gap in persistence of about 12 percentage points in the raw data.

This does not imply that there are gender differences in persistence, however, as the distribution of performance on the first IQ test is substantially different for men and women, see Table 1.1. To resolve this confound, in what follows regressions are presented to study if there are gender differences in persistence when controlling for subjects' performance, the feedback they received, as well as self-reported characteristics.

1.3.2 Formal analysis of gender differences in persistence

Aggregate results. To explore if there is a gender gap in persistence more formally, Table 1.2 presents OLS estimates of the probability to continue in the *Baseline* treatment. As a reference, column (1) shows that absent of controls, women are about 12 percentage points less likely to continue than men, corresponding to the average gender gap in the raw data as shown in Figure 1.3. When controlling for past performance (measured as the score on the first IQ test), the feedback that subjects received, as well as self-reported characteristics, the estimated gender gap in persistence amounts to roughly 10 percentage points, see column (2), and this gap is statistically significant ($p = 0.016$). Given that

¹⁶The BDM involves 23 questions, see Appendix A.2. A subject's ex-ante probability of continuing increases linearly with their reported switch point. More specifically, $SwitchPoint_i/23$ is the probability that subject i continues.

the average probability of continuing for men who received positive feedback is 68% in the *Baseline*, women are on average about 15% less likely to continue than men. This estimated gap is robust when controlling for whether subjects passed the first IQ test (column 3) or an interaction of the *Female* dummy with the test score, see column (4). It is worth noting that the average continuation probability of women who received *positive* feedback is 53% – which is not larger than the average continuation probability of men who received *negative* feedback (55%).

To put the estimated gender gap of the experiment into perspective, note that it is similar in magnitude to some of the gender differences in persistence that have been documented in the field.¹⁷

Result 1.1. *In the Baseline treatment, women are on average about 10 percentage points (or 15%) less likely to continue than men when controlling for their past performance, the feedback they received, as well as self-reported characteristics.*

Heterogeneity by feedback and first IQ test performance. Does the effect of receiving negative versus positive feedback vary by gender in this controlled environment? As column (5) of Table A.1 shows, this hypothesis is not supported in the data, as the interaction effect of the *Female* dummy with the negative feedback dummy is statistically insignificant. In other words, negative feedback does not appear to have a more discouraging effect on women’s decision to persist than it has for men. Similarly, positive feedback does not appear to have a more encouraging effect on men than on women. Directionally, men are more likely to continue regardless of what feedback they received. The estimated gender gap in persistence among those who received positive feedback is

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

15 percentage points ($p = 0.012$), and about twice as big as the gender gap in response to negative feedback, which is only about 7 percentage points and not statistically significant ($p = 0.236$). Section 1.3.3 will discuss that this may in part be driven by gender differences in feedback avoidance in response to positive feedback.

The gender gap in persistence is further driven by subjects who failed the first IQ test. Men who performed poorly on the first IQ test are thus over-represented in the sample that continues, relative to women who performed poorly. Details and implications of the adverse selection of men will be discussed in Section 1.4.

1.3.3 Channels driving the gender gap in persistence

What can explain this gender gap in persistence? The experimental design allows us to explore the roles of beliefs, preferences for additional feedback, and risk aversion, and these channels are analyzed in what follows.

Channel 1: Gender differences in beliefs about passing the future IQ test. If women are less confident about their future performance, and thus expect lower returns from persisting than men, it is rational for them to quit more often, all else equal. The following analysis explores at what instance gender differences in beliefs about one's future performance arise, differentiating beliefs before feedback from how beliefs respond to feedback. To increase power, data are pooled across treatments. (Recall that no design elements differ across treatments until after the belief elicitation, see Section 1.2.)

Gender differences in beliefs before feedback - evidence from the lab and the field. After taking the first test but before receiving feedback, women – relative to equally performing men – are on average less confident both about their past and their future performance, see columns (1) and (2) of of Panel (A) in Table 1.3. If anything,

the gender gap in confidence regarding the future IQ test is even more pronounced (at 10 percentage points) than the gap in confidence regarding the past IQ test (at 7 percentage points). Both differences are highly significant. 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.

Interestingly, men and women appear to differ in how they extrapolate from their beliefs about their past performance when forming beliefs about their future. As column (3) of Panel (A) in Table 1.3 shows, women are less confident about their future performance even when controlling for beliefs about their past. Put differently, even when comparing men and women that are similarly confident about having passed the first IQ test, men are on average substantially more confident about passing the future IQ test, as Figure 1.4 illustrates. One explanation for this could be that men perceive previous failures as less predictive, or previous successes as more predictive of their future than women, and consequently they are more confident moving forward.

To examine the outside validity of this gender difference in extrapolating from the past when forming beliefs about the future, a field classroom study was conducted, details of which are provided in Appendix A.4. In this field study, undergraduate UCSB students who just finished their first Econ 1 midterm exam were asked to report two beliefs, very similar to the ones elicited in the experiment: (i) how likely they think it is that they scored above a certain cutoff on the first midterm exam, and (ii) that they will score above this cutoff on the next midterm exam. Panel (B) of Table 1.3 shows that the gender differences in belief formation from the lab replicate remarkably well in the field – both qualitatively and in terms of the effect size.

Result 1.2. *Before receiving feedback, men are on average more confident about their future performance than women, even when controlling for their past performance and be-*

liefs about their past performance. This insight from the laboratory experiment replicates well in a field classroom study.

No gender differences in updating in response to feedback. If men respond stronger to positive feedback, or if women respond stronger to negative feedback when updating about their future performance, this could further exacerbate the documented gender gap in confidence about the future. To explore this possibility, note that Bayesian updating in this setting can be written in log-form as

$$\ln \left(\frac{p}{1-p} \right) = \ln \left(\frac{p_0}{1-p_0} \right) + \mathbf{1}\{pos.\} * \ln \left(\frac{\phi}{1-\phi} \right) + \mathbf{1}\{neg.\} * \ln \left(\frac{1-\phi}{\phi} \right), \quad (1.1)$$

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; ϕ denotes the probability with which the cards conveying the feedback reveal the true state of having passed or failed the first IQ test, which by design equals two-thirds when updating about the first IQ test.¹⁸ Linear regressions of the following form can thus be estimated:

$$\begin{aligned} \ln \left(\frac{p_i}{1-p_i} \right) = & \alpha * \ln \left(\frac{p_{0i}}{1-p_{0i}} \right) \\ & + \beta_p * \mathbf{1}\{pos.\} * \ln \left(\frac{\phi}{1-\phi} \right) + \beta_n * \mathbf{1}\{neg.\} * \ln \left(\frac{1-\phi}{\phi} \right) + \epsilon_i. \end{aligned} \quad (1.2)$$

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

¹⁸There is no objectively true value of ϕ when updating about the future, however. Put differently, there is no objective and observable Bayesian benchmark for how rational subjects should update their beliefs about their future performance in response to the past feedback. In Table ??, estimates for $\phi = 0.62$ are shown for the future test, for which the estimates of β_p and β_n were reasonably close to 1. Different values would scale the estimates, but would not lead to a different conclusion when testing the hypothesis that men and women update differently.

feedback, respectively.

Table A.3 shows that men and women on average do not place different weights on positive or negative feedback when updating about their past or future performance, as the interaction of the β terms and a female dummy is not statistically distinguishable from zero. Thus, how people adjust their beliefs in response to feedback arguably plays no important role for explaining the gender gap in persistence.

Gender differences in beliefs after feedback and their effect on persistence.

After having received performance feedback, the gender gap in beliefs about people's future performance remains, but the gender gap in beliefs about having passed the first test closes, as columns (1) and (2) of panel B in Table A.2 show. Notably, following feedback, it is still the case that men tend to make more optimistic projections of their future performance even when compared to women who are similarly confident about their past performance; Controlling for past test scores and beliefs about having passed the first IQ test, men on average are about 7 percentage points more confident about passing the future IQ test ($p = 0.005$) than women, see column (3). Figure A.1 illustrates that these gender differences in how people interpret the past when forming beliefs about the future are notable both in response to positive and negative feedback.

How much of the gender gap in persistence can be attributed to gender differences in confidence about one's future performance? Recall that in the *Baseline* treatment, the gender gap in persistence amounts to about 10 percentage points. When controlling for subjects' posterior beliefs of passing the future IQ test – the beliefs that subjects report about their future performance directly before their continuation decision – this gap drops to 6.7 percentage points ($p = 0.072$), see column (2) of Table A.4. While this estimate is not statistically distinguishable from the “original” gender gap presented in column (1), this suggests that roughly one-third of the gender gap in persistence is

attributable to gender differences in confidence.

Result 1.3. *After receiving feedback, women remain less confident about passing the future IQ test, relative to equally performing men who received the same feedback. This gender gap in future confidence accounts for roughly one-third of the gender gap in persistence.*

Channel 2: Gender differences in avoiding and seeking additional feedback.

Persisting in careers such as corporate management or academia typically involves exposure to frequent performance feedback. If women dislike this exposure more so than men, or if men enjoy receiving additional feedback more so than women, this could help explain gender differences in persistence. To explore this possibility, subjects' continuation probabilities between the *AlwaysInfo* and the *Baseline* treatment are compared. Recall that if subjects are more (less) likely to continue in the *AlwaysInfo* treatment than in the *Baseline*, this can be interpreted as evidence of feedback avoidance (seeking).

Figure 1.5 compares average continuation probabilities for men and women between the two treatments. In the raw data, the gender gap in persistence shrinks substantially in the *AlwaysInfo* treatment, relative to the *Baseline*. This is driven by two forces: On average, women avoid, and men seek exposure to the additional feedback of learning if they passed or failed the first IQ test.

One caveat of analyzing the *AlwaysInfo* treatment effect more formally is that although subjects were randomized into treatments, not all observables that should be orthogonal to the treatment assignment are perfectly balanced across the two treatments, see Table A.6. In particular, subjects on average reported a slightly higher GPA in the *AlwaysInfo* treatment, and women (but not men) who got assigned to the *AlwaysInfo* treatment were more likely to report a non-white race identity, and to report

US citizenship, than women who got assigned to the *Baseline*. Just as before, controls for these self-reported variables as well as subjects' beliefs about passing the first and future IQ test after receiving feedback are included in the presented regressions below. The latter ensures that the estimated treatment effect does not reflect gender differences in expectations about what feedback they would receive upon continuing.

Aggregate estimates of the *AlwaysInfo* treatment effect are directionally consistent with the idea that women avoid additional feedback, while men seek it. As column (1) of Panel A in Table 1.4 shows, men are on average 6.5 percentage points less likely to continue in the *AlwaysInfo* treatment than the *Baseline* ($p = 0.080$), which suggests that men on average prefer to learn if they passed or failed the first IQ test. This makes persisting in an environment that involved additional feedback exposure relatively more attractive for them. For women, the estimated *AlwaysInfo* effect is directionally consistent with feedback avoidance, but not significantly different from zero in the aggregate sample ($p = 0.433$).

It is possible that the estimates presented in column (1) of Panel A, Table 1.4, mask some heterogeneity of preferences for additional feedback exposure. For example, following negative feedback, subjects might want to avoid learning their test outcome so that they can hold on to the “glimmer of hope” that the negative feedback was wrong, or they might prefer finding out their test result to prove the negative feedback wrong. To explore this idea, columns (2)-(5) of Panel A in Table 1.4 presents separate estimates for sub-groups of subjects by the feedback they received, as well as their first IQ test result.

Men who passed the first IQ test and received positive feedback are on average 11.6 percentage points less likely to continue in the *AlwaysInfo* treatment than the *Baseline* ($p = 0.079$), which suggests that seeking additional positive 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 IQ test; they are on average

16.9 percentage points more likely to continue in the *AlwaysInfo* treatment ($p = 0.083$), which is consistent with the idea that women might decide not to go after opportunities in order to avoid finding out that they are not as talented or skilled as they had hoped after receiving some positive feedback initially. For subjects who received negative feedback, in contrast, average continuation probabilities are not statistically different between the two treatments, as columns (4) and (5) indicate.

To what extent can gender differences in feedback avoidance and feedback seeking – independent of beliefs – explain the gender gap in persistence? When weighting all estimates of Panel A in Table 1.4 by the fraction of the respective groups in the *Baseline* treatment, around 50% of the gender gap in persistence in the *Baseline* is attributable to gender differences in preferences for additional feedback, see line (i) of Panel B.¹⁹ But since not all estimates of the *AlwaysInfo* effect are statistically significant from zero, a more conservative approach of estimating the impact of the *AlwaysInfo* treatment effect would be only consider subgroups for which the effect is statistically significant at the 10% level or higher. Using this approach, estimates suggest that slightly less than one-third of the gender gap in persistence can be explained by gender differences in preferences for additional feedback, see line (ii) of Panel B. That being said, none of the estimates presented in panel A are significant at the 5% level, so while this exercise may provide a first intuition for the role of feedback avoidance in explaining persistence, caution is warranted in interpreting these estimates.

Result 1.4. *Gender differences in preferences for additional feedback exposure may account for roughly one-third of the documented aggregate gender gap in persistence. On*

¹⁹For example, consider column (1) of Panel A in Table 1.4. In the *Baseline* treatment, 46% of subjects are men and 54% are women. Thus, the gender gap in the *AlwaysInfo* treatment is $6.5 * 0.46 + 3.2 * 0.54 = 4.72$ percentage points smaller than in the *Baseline*, where there is a 10.3 percentage points gender gap in persistence. Thus, $4.72 / 10.3 = 45.8\%$ of the gap in persistence can be explained by gender differences in feedback preferences. A qualitatively similar result is obtained if weighting the estimates of columns (2)-(5) of Panel A by their corresponding fractions in the *Baseline*.

average, women avoid, and men seek additional feedback on whether they passed or failed the first IQ test.

Channel 3: Gender differences in risk preferences. Persisting in stratified career trajectories may be a risky choice if doing so is only financially rewarding for people who can achieve a high performance. In the experiment, quitting guarantees a minimum payment, while continuing only pays off if subjects pass the second IQ test. If women are more averse to taking risks than men, this could constitute another channel driving the documented gender gap in persistence.

Appendix A.5 provides details on how risk parameters are estimated in this study. As Table A.5 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 is essentially unaffected when including risk parameters as a control, see columns (3)-(6) in Table A.4. This suggests that risk preferences do not constitute an important channel for explaining gender differences in persistence in this setting.

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

Do gender differences in persistence contribute to a gender gap in performance within organizations? This would be the case, for example, if performance feedback mechanisms deter high-performing women from continuing more so than men, or if they deter low-performing men less from continuing than women. The following analysis explores if there are gender differences in the efficiency of people's self-selection in the sense that (i) high-performing individuals are more likely to continue, and that (ii) continuation decisions predict people's future performance.

In the experiment, the gender gap in persistence is driven by subjects who failed the first IQ test, as panel (a) of Figure 1.6 indicates. Conditional on having failed, women are about 15 percentage points less likely to continue ($p = 0.035$), while for subjects who passed the first test, the gap is negligible in magnitude and statistically indistinguishable from zero, as columns (1)-(3) of Table 1.5 show. Furthermore, when looking at the total sample, the marginal effect of scoring one standard deviation higher on the first IQ test on the probability of continuing is roughly twice as high for women than it is for men, see column (4). In sum, men are adversely selected relative to women when taking subjects' past performance as a measure of ability. As column (5) shows, this effect does not vary at the treatment level, as the interaction with the *AlwaysInfo* dummy is small in magnitude and statistically insignificant. This suggests that preferences for additional feedback exposure do not affect how efficient the self-selection of women relative to men is when taking past performance as a measure of ability.

A natural feature of this experiment is that people's past performance is not a perfect predictor of their future performance. Conditional on continuing and having passed (failed) the first IQ test, 85% (67%) passed the second IQ test in the *Baseline* treatment, and the correlation coefficient between test scores is 0.49. (To put this number in perspective, the correlation coefficient of the first two midterm exam scores is 0.41 in the Econ 1 classroom field study, see Appendix A.4.) With this in mind, a relevant question to ask is how predictive subjects' continuation decisions are of their future performance.

Do the continuation decisions of men or women better predict their future performance? As the sample of subjects who continue - and thus the sample for which the second IQ outcome can be observed - is selected, Heckman regressions are presented in Table 1.6 to account for sample selection. The main purpose of this table is to investigate if the switch point in Part 3 of the experiment (which directly translates into a subject's probability of continuing, see Section 1.2), is more predictive of the performance on the

second IQ test for men or for women. In step 1, the self-selection into continuing is estimated using subjects' switch points in Part 3 and Part 4 of the experiment (which are correlated, $\rho = 0.55$ in the *Baseline*).²⁰ In step 2, it is estimated what factors can predict the performance on the second IQ test. Columns (1) and (3) show that absent of controls, the ex-ante probability of continuing (i.e., the switch point in part 3) is a significant predictor both for passing the second IQ test, and for the score on the second IQ test. Once controls are included, however, this regressor is no longer statistically significant; That is, on average subjects' continuation decisions do not predict their future performance very well.

Estimates further suggest that there are no gender differences in how predictive the ex-ante probability of continuing is for subjects' future performance, as the interaction of the *Female* dummy with the reported switch point is small in magnitude and statistically insignificant, see columns (2) and (4) of Table 1.6. Note that this conclusion does not vary at the treatment level. Taken together, while the raw data suggest that among subjects who continued and failed the 2. IQ test, men had a higher ex-ante probability of continuing and are thus adversely selected (see panel (b) of Figure 1.6), this idea is not supported in the more formal exploration presented in Table 1.6. Considering that over three-quarters of the subjects who continued in the *Baseline* treatment passed the second IQ test, the sample is probably too small to pick up significant gender differences in the efficiency of subjects' self-selection with respect to their future performance.

Summing up, while men who continued in the experiment are adversely selected when taking the first test performance as a measure of ability, this does *not* imply, however, that the differential self-selection of men and women results in an adverse selection of men in terms of their future performance. To a large extent, this may be the case

²⁰The estimates presented in Table 1.6 are qualitatively robust to different specifications of the selection equation, such as using the sign of the signal or gender as a predictor of continuing.

because the empirical relationship between past and future performance is naturally noisy. In addition, given that the sub-sample of subjects who continue in the experiment is positively selected, this study may be under-powered to detect gender differences in how predictive subjects' continuation decisions are of their future performance.

Result 1.5. *When taking past performance as a measure of ability, there is an adverse selection of men into persisting, relative to women. The relationship of past and future performance is noisy, however, and there is no detectable gender gap as to how predictive continuation decisions are of subjects' future performance.*

A related question is whether subjects' continuation decisions maximized their earnings in the experiment. Appendix A.6 explores this question, and concludes that on average both men and women who continued would have had higher expected earnings, had they quit.

1.5 Discussion

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. Gender differences in beliefs and preferences for additional feedback together account for roughly two-thirds of the gender gap in persistence, and while the role of risk preferences is negligible. This raises the question of what can explain the remaining third of the gender gap in persistence in this controlled setting.

One possibility is that there are gender differences in seeking challenges. As continuing involves another IQ test, while quitting involves an easy test, continuing might be relatively more attractive for men if they enjoy performing challenging tasks more

than women, all else equal. One study by Niederle and Yestrumskas (2008) investigates this hypothesis, and finds that a gender gap in choosing a challenging versus an easy mazes task closes when subjects receive information about whether the challenging task is likely payoff-optimal for them. The authors interpret this as evidence against the idea that there are gender differences in preferences for the characteristics of the hard versus the easy task. Furthermore, note that subjects know that they will face the same task type regardless of whether they continue or quit, i.e., quitting does not allow them to avoid Raven’s matrices all-together.

Another possibility is that women have a stronger distaste than men to perform a task they are not good at. While subjects know they will not receive any direct feedback on the second IQ test if they continue, the experience of taking another IQ test may already convey an unpleasant feeling if subjects do not know how to solve the questions. Put differently, anticipating this “internal negative feedback” may deter women from continuing more so than men. With this in mind, the estimated *AlwaysInfo* treatment effect may be regarded as a lower bound of how much of the gender gap in persistence is attributable to gender differences in feedback avoidance, broadly speaking. Women could have a stronger preference to avoid negative “internal feedback” that is potentially conveyed while taking the second IQ test, in addition to avoiding the “external feedback” that is provided when learning their first test result.

Finally, it is worth noting that subjects have been socialized as men or women for about two decades before participating in the experiment. They may have adapted gender-congruent heuristics that could affect their decisions in this controlled setting. The experiment was not designed to identify such channels, however some survey questions at the end relate to whether subjects’ parents were fulfilling gender-traditional roles, as well as their own attitudes toward gender roles, see Appendix A.3. If anything, this limited set of questions can hint at whether the gender gap in persistence appears to

be especially pronounced among subjects of more traditional backgrounds or of more conservative attitudes.

Table A.8 presents estimates of the gender gap in persistence in the *Baseline* treatment that account for subjects' self-reported family background and personal attitudes. When looking at the total sample, the estimated gender gap in persistence is robust to controlling for subjects' reported parental characteristics and personal attitudes on gender roles, see columns (1)-(5). It is further similar in magnitude for the sub-sample of subjects who did not disagree that their parents' occupations were typical for men/women of their generation, however for this group the gap is not statistically significant ($p = 0.106$), perhaps because of power issues, see column (6).

For subjects that reported that their father used to work more hours for pay during their childhood than their mother, the estimated gender gap in persistence is marginally bigger at 12 percentage points ($p = 0.035$), see column (7). Furthermore, column (8) shows that for subjects with more conservative attitudes – those who either (strongly) disagreed that “women should pay their own way on dates,” or who did not strongly disagree that “wives with a family have no time for outside employment” – the estimated gender gap is almost 17 percentage points and highly significant ($p = 0.006$). These estimates are not statistically distinguishable from the gender gap of the total sample, however. Taken together, more work will be needed to explore the role of family backgrounds and attitudes on gender differences in behavior more carefully.

To what extent can the gender differences in persistence documented in this study help explain the under-representation of women in stratified careers? It is worth pointing out that gender differences in persistence were detected in the experiment despite the absence of competition or feedback that entails social comparison. Moreover, as subjects' decisions could not be observed by others, the role of social signaling and an urge to comply with social gender norms was probably limited. It is left to future research to

study whether these factors interact with and potentially exacerbate the gender gap in persistence.

Furthermore, note that a gender gap in persistence was detected even when looking solely at a one-time decision in response to a one-time provision of feedback. When pursuing a stratified career, however, people are frequently exposed to performance feedback, and have to decide between persisting and dropping out along many steps of the career ladder. The compound effect and its implications for education and labor market outcomes may therefore be larger. And while the sample of people who persist on a career trajectory is getting more and more selected with position seniority, note that gender differences in persistence in this experiment have been documented among UCSB students – people who may already be positively selected in terms of their persistence.

An insight of this study that has several implications is how men and women differ when forming beliefs about their future performance. First, recall that the gender gap in confidence about passing the future IQ test is directionally much bigger than the gender gap in confidence with respect to the past test. 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 past events, but not the relevant future event, when studying economic decisions.

Furthermore, even though there is no gender difference in how predictive people's past performance is of their future success in this study, men and women appear to perceive the underlying statistical relationship of their past and future performance differently. If women who initially perform poorly are overly deterred from persisting because they perceive their past performance to be more predictive of their future success than men, they forgo the opportunity of learning that they might improve over time, and that persisting could be rewarding for them in the long run despite initial setbacks. A fruitful area for future research could be to study if providing information on how (un-)predictive

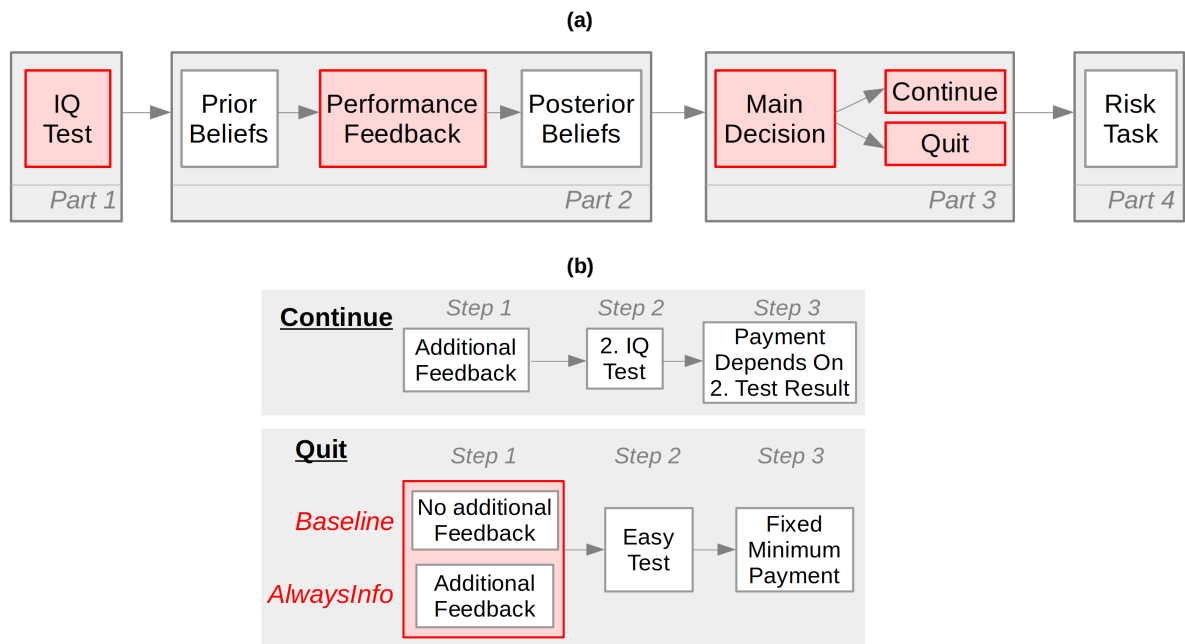
past outcomes are of future successes can help reduce the gender gap in confidence, and ultimately in persistence.

Table 1.1: Summary Statistics, Baseline Treatment.

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

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.

Figure 1.1: Timeline of the experiment.



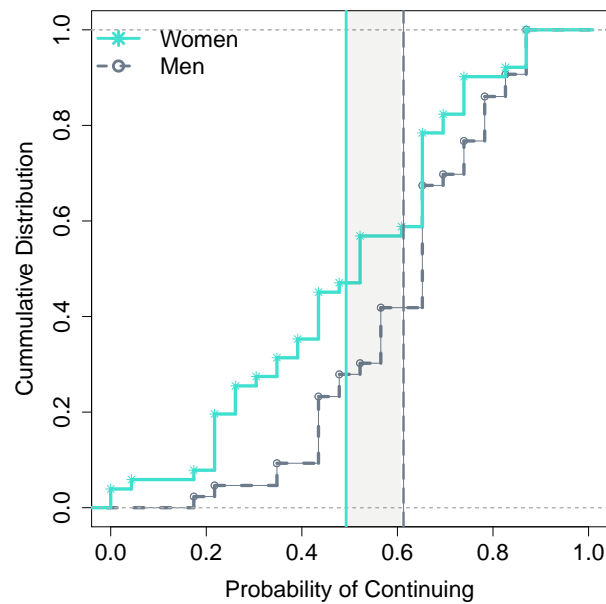
Panel (a) depicts the four main parts, one of which was randomly drawn for payment at the end. Panel (b) provides an overview of what happens if subjects continue or quit, corresponding to Part 3 in panel (a). The only feature distinguishing the *Baseline* from the *AlwaysInfo* treatment is whether subjects receive additional feedback on the first IQ test if they quit.

Figure 1.2: Cards shown to subjects to convey feedback.



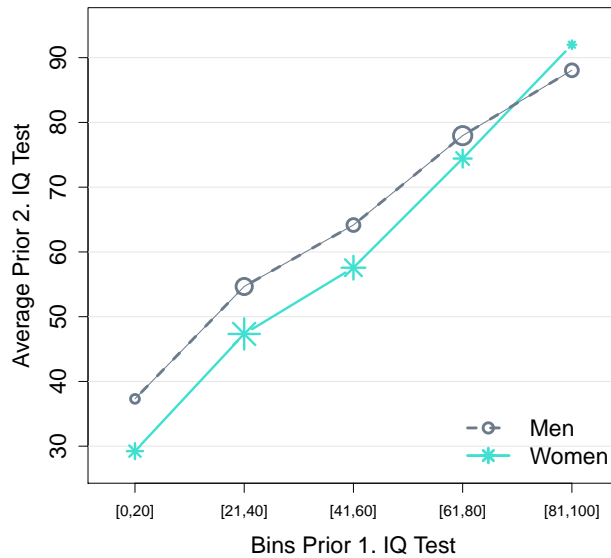
This figure displays the cards shown to subjects to convey feedback in Part 2 of the experiment. Subjects either received positive feedback (a card saying that they passed), or negative feedback (a card saying that they failed the IQ test), randomized conditional on their performance (having passed or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

Figure 1.3: Probability of Continuing by Gender, Raw Data, Baseline Treatment.



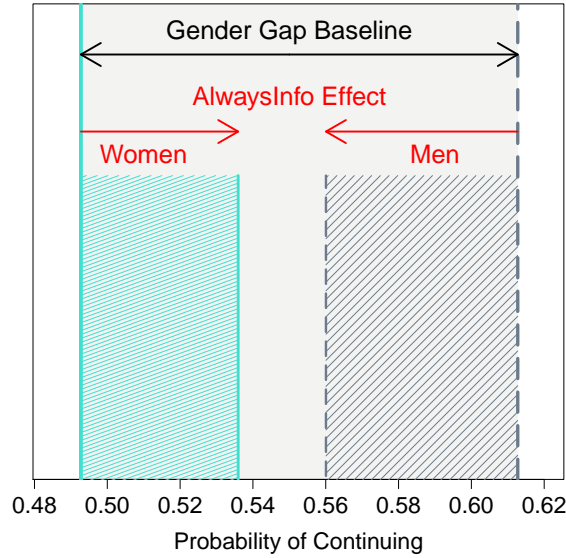
This figure shows the empirical cumulative distribution function of subjects' continuation probabilities, separately for men and women. The vertical lines represent the means of each group, and the gray shaded area highlights the gender difference in average probabilities of continuing. Raw data from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

Figure 1.4: Gender Differences in Beliefs About the Future, Given Beliefs About the Past.



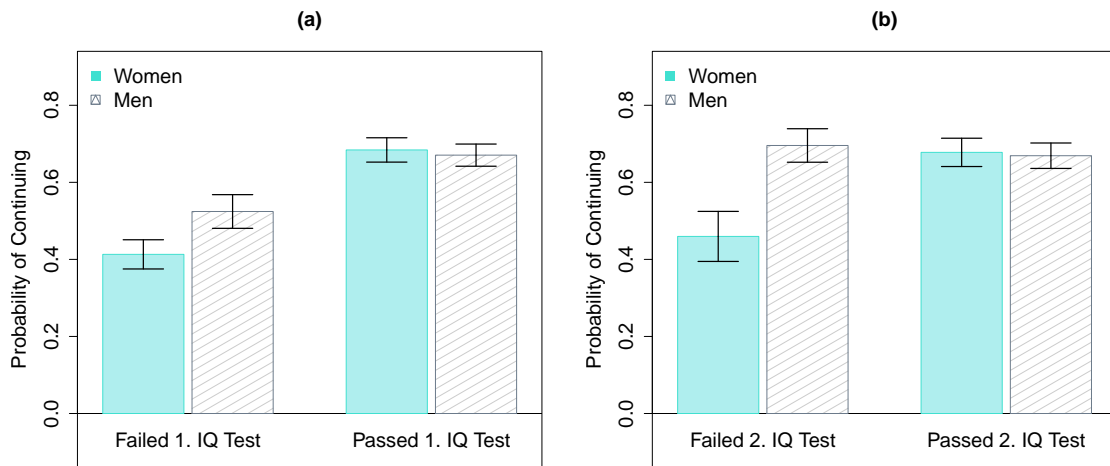
This figure plots gender differences in prior beliefs about passing the 2. IQ test, given prior beliefs about the 1. IQ test. The size of the points represents the relative share of observations in a given bin category of prior beliefs about the 1. IQ test. On average, men are more optimistic than women about passing the future IQ test, given their beliefs about having passed the first IQ test.

Figure 1.5: AlwaysInfo Treatment Effect Relative to Baseline Treatment.



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

Figure 1.6: Probability of Continuing by Test Results and Gender, Baseline Treatment.



Bars represent the average probabilities of continuing, separately for women and men, alongside the standard errors of each group, in the *Baseline* treatment. Panel (a) breaks this comparison down by the 1. IQ test results, and panel (b) breaks it down by the 2. IQ test result.

Table 1.2: OLS Estimates, Probability of Continuing, Baseline Treatment.

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

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

Table 1.3: OLS Estimates of Prior Beliefs (Before Feedback) - Laboratory vs. Field Study.

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

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Constants not displayed. The mean of the reference group refers to men's average prior beliefs. Panel A reports prior beliefs (before receiving feedback) in the laboratory experiment. Additional controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity). Panel B reports beliefs of the Econ 1 classroom field study, controlling for self-reported race identity.

Table 1.4: AlwaysInfo Treatment Effect.

	Probability of Continuing				
	(1) All	(2) Pos. Feedback Passed	(3) Failed	(4) Neg. Feedback Passed	(5) Failed
<i>Panel A: Estimated Treatment Effect</i>					
Men	-0.065* (0.037)	-0.116* (0.065)	-0.101 (0.090)	-0.047 (0.107)	-0.125 (0.086)
Women	0.032 (0.041)	0.037 (0.070)	0.169* (0.079)	-0.093 (0.096)	-0.006 (0.076)
<i>Panel B: Fraction of Gender Gap in Persistence Explained</i>					
(i) All	0.46		0.55		
(ii) At least 10% confidence level	0.29		0.26		

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Panel A presents estimates of the impact of the *AlwaysInfo* treatment on the probability of continuing relative to the *Baseline* treatment, separately for men and women. Controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity). Positive (negative) point estimates correspond to feedback avoidance (feedback seeking). Panel B shows what fraction of the gender gap in persistence of the *Baseline* treatment can be explained by gender differences in feedback avoidance and feedback seeking, weighting (i) all estimates, (ii) all estimates that are statistically significant at the 10% level.

Table 1.5: Probability of Continuing by 1. IQ Test Performance, Baseline Treatment.

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

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

Table 1.6: Performance 2. IQ Test by Ex-ante Probability of Continuing

	Heckman Probit		Heckman	
	Passed 2. IQ Test		Z-Score 2. IQ Test	
	(1)	(2)	(3)	(4)
Step 1: Selection into Continuing				
Switch Point Part 3	0.175*** (0.0251)	0.170*** (0.0265)	0.170*** (0.0275)	0.168*** (0.0266)
Switch Point Part 4	-0.0255 (0.0181)	-0.0185 (0.0218)	-0.0214 (0.0259)	-0.0169 (0.0224)
Step 2: Performance 2. IQ Test				
Switch Point Part 3	0.135*** (0.0262)	0.101 (0.0649)	0.0777** (0.0344)	0.0427 (0.0402)
Female		0.914 (1.336)		0.485 (0.815)
Female * Switch Pt. Part 3		-0.0391 (0.0878)		-0.0187 (0.0509)
Z-Score 1. IQ Test		0.394*** (0.153)		0.391*** (0.104)
AlwaysInfo		-0.533 (0.549)		-0.453 (0.356)
Female * Switch Pt. Part 3 * AlwaysInfo		-0.0466 (0.0382)		-0.0157 (0.0235)
Additional Controls	-	✓	-	✓
Observations Continued	105	105	105	105
Observations Total	205	205	205	205

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

Chapter 2

Holiday, Just One Day Out Of Life: Birth Timing and Post-natal Outcomes

2.1 Introduction

The role of medical intervention in childbirth has risen over time. The use of Cesarean section (i.e., the delivery of a baby via a surgical procedure) has increased steadily from a rate of 21 percent in 1996 to 31.9 percent in 2016.¹ Similarly, rates of induction and stimulation of labor (two methods of precipitating a birth) have grown considerably. The use of labor inductions among singleton births, for example, was under 10 percent in 1990 and over 23 percent in 2012 (Osterman and Martin, 2014). With the growth of these techniques, medical professionals are able to time deliveries quite precisely. In the case of Cesarean section, the birth can be timed to a particular day and hour.

¹Source: Centers for Disease Control and Prevention: <https://www.cdc.gov/nchs/fastats/delivery.htm>

The appropriate use of Cesarean sections (C-sections) is a matter of considerable interest, with many payers and policymakers embarking on quality improvement initiatives to reduce elective use of these procedures. These efforts stem from a growing body of evidence demonstrating the harms from C-section use in low-risk pregnancies (e.g., see Card et al., 2018). Less fully understood are the implications of birth timing through labor induction or stimulation. Moreover, conditional on use, the optimal timing of these interventions remains mostly unclear. In 2013, in response to research documenting a critical development phase between 37 and 39 weeks, the American College of Obstetrics and Gynecologists (ACOG) revised its method of categorizing births based on gestational age.² The goal was, at least in part, to discourage “unnecessary deliveries” before 39 weeks of gestation (39-40 weeks is considered a full-term birth). The decision of whether or not and when to intervene is all the more important given the association of neonatal health with longer term outcomes such as infant mortality, education, and earnings (Almond et al., 2005; Black et al., 2007; Royer, 2009; Figlio et al., 2014).

Because of selection – the set of women who undergo C-sections, labor induction or labor stimulation is nonrandom – it is challenging to understand how these interventions affect outcomes. We argue that holidays provide a useful quasi-experiment in this regard. Figure 2.1 documents the distribution of the number of births across three types of days: major holidays, weekends, and neither major holidays or weekends for two samples – the United States 1968-1988 and California 2000-2016.³ The number of births per day is approximately 2 standard deviations lower on major holidays than on non-holiday, non-weekend days. Similar to the holiday drop in births, births decline over the weekend, when

²Source: <https://www.acog.org/About-ACOG/ACOG-Departments/Deliveries-Before-39-Weeks?> The new classification was as follows: 1) early term: 37 weeks through 38 weeks and 6 days, 2) full term: 39 weeks through 40 weeks and 6 days, 3) late term: 41 weeks through 41 weeks and 6 days, and post-term: 42 weeks and later.

³The inclusion of the entire United States in this figure is for illustrative purposes as we focus our attention on California for 2000 to 2016. Publicly available data for the United States only include exact date of birth for years prior to 1989, starting in 1969.

fewer medical practitioners schedule deliveries. The drop in births as a consequence of holidays is well-documented elsewhere (Borst and Osley, 1975; Macfarlane, 1978; Rindfuss et al., 1979; Mangold, 1981; Cohen et al., 1983; Hawe et al., 2001; Hong et al., 2006; Goodman et al., 2005; Bauer et al., 2013; Gelman et al., 2013; Martin et al., 2018). For the most part, however, this literature does not characterize how holiday births are displaced over time or study the consequences of this displacement for delivery or birth outcomes.

To uncover the effects of holidays, we utilize data on the universe of California births between 2000 and 2016, which amounts to almost 9 million observations. These data allow us to detail the effects of holidays not only on birth timing, but also on delivery method (e.g., vaginal versus C-section) and birth outcomes (e.g., birth weight, Apgar scores, delivery complications). For 2016, the use of C-section in California matched the national average.⁴

We begin by systematically documenting the depression of births around major US holidays. This requires careful consideration of the appropriate counterfactual, both because births exhibit regular seasonal and within-week patterns and because holidays impact the distribution of births in a broad region of the event. Comparing Monday holidays to births on Wednesday, when births are commonly scheduled, or on Sunday, when they are rarely scheduled, would bias our estimates. In short, taking into account day-of-the-week effects is imperative to appropriately quantifying the impact of holidays on births. Furthermore, because holidays cause births to be displaced, comparisons of birth counts on the holiday versus, for example, the days just before the holiday are likely to be biased if births are systematically elevated before the holiday.

Using insights from the tax bunching and test score manipulation literature (Saez, 2010; Chetty et al., 2011; Kleven and Waseem, 2013; Diamond and Persson, 2016; Dee

⁴Source: https://www.cdc.gov/nchs/pressroom/sosmap/cesarean_births/cesareans.htm

et al., 2019), we determine what we call a “manipulation window.” The manipulation window is the time period around a holiday during which some births are shifted due to the holiday. In other words, the manipulation window captures the start and end dates of the period over which births are retimed due to the holiday. Unlike analyses of test score or income tax manipulation, which push observations in one direction (e.g., above a test score threshold as in Dee et al. (2019) or before a tax deadline as in Dickert-Conlin and Chandra (1999)), holidays can shift births both before and after they would otherwise occur. As the “missing mass” of births during the holiday period (i.e., fewer births than otherwise would be expected) must be counteracted with an “excess mass” of births (i.e., more births than otherwise would be expected), our manipulation window is the period of time around a holiday for which the sum of the missing and the excess mass is closest to zero.

Using this zero net mass heuristic, our estimated manipulation window spans the period from 11 days prior to 16 days after the holiday.⁵ This implies that the manipulation of births around holidays occurs over 4 weeks. However, three-quarters of the manipulation is contained within $+/- 1$ week of the holiday. To create a counterfactual set of births, we match births within this manipulation window to the closest day just outside the window that falls on the same day of the week. For days prior to the holiday, this means we compare births on day d with births on day $d - 14$. For days following the holiday, each day d is contrasted with day $d + 21$. These nearest day, day of week matched controls enable us to isolate the holiday effect from strong seasonal and within-week cyclical patterns in births.

Relative to this counterfactual, we estimate that about 500 births or 18 percent of births on the day of the holiday and the day just after are shifted to other days within

⁵The total number of births moved to the period before versus after the holiday is roughly equal, although the pre and post periods themselves are unequal such that displacement of a similar number of births occurs over more post than pre-holiday days.

the manipulation window. About 50 percent of the decline across these two days is due to a reduction in C-sections. The remainder is roughly split between spontaneous vaginal births and vaginal births after an induction or stimulation of labor. The relative reduction in births on the holiday is larger for high-risk births, defined as births to a mom with a prior C-section, a breeched birth, a multiple birth pregnancy, or with an infection such as HIV. About 65 percent of high-risk births are shifted away from the holiday.

Although the births retimed due to holidays are selected, we can use outcomes for all births in the holiday manipulation window to understand the reduced-form impact of holiday birth timing manipulation. Using logic invoked in Diamond and Persson (2016), who study the effect of test score manipulation on earnings, we compare birth outcomes for days in the manipulation window to counterfactual days just outside the manipulation window. Following this methodology, we find little evidence of adverse effects on outcomes in the holiday manipulation window. Our reduced-form holiday effects are a reduction of 2 grams in birth weight (off of a mean of about 3300 grams). We find no other meaningful changes in a host of other outcomes, including newborn conditions, labor complications, low Apgar scores, admission to the neonatal intensive care unit (NICU) or the use of assisted ventilation. These estimates are small from a health perspective even if we inflate them by reasonable approximations of the fraction of retimed births (e.g., scaling by the roughly 500 births out of 40,000 that are retimed over the 28-day holiday manipulation period implies a roughly 150 gram or 4.5% reduction in birth weight) to arrive at an implied IV estimate (i.e., the effect of a birth being retimed). Even among high-risk pregnancies, we find little impact of holiday-related birth timing manipulation on infant health. For example, for high-risk births, our implied IV estimates suggest reductions in birth weight on the order of 50 grams. These findings contrast with the adverse effects found in papers on C-sections (e.g., Card et al., 2018), although those

papers contrast the starker effect of using the procedure versus not.

A key distinction of our work is that it largely compares the effects of changes in the scheduling of procedures. While we cannot rule out compositional changes in the use of interventions, with, for example, some births that would have been spontaneous retimed via C-section and other births that would have been C-section delayed to the point of becoming spontaneous, we find no net changes in delivery type within the manipulation window. Because we find no clear evidence of extensive margin changes in intervention use, our results speak largely to the effect of the optimal timing of these interventions. In contrast, the literature on the adverse health impacts of C-sections compares the effects of using these procedures versus not using them.

This paper adds to the extensive existing literature on the effect of holidays on the number of births. Most of this earlier work focuses on international settings or uses data from earlier years, when both C-sections and inductions were far less frequent. Using US births, Gelman et al. (2013) and follow-on analysis at fivethirtyeight.com, establish similar patterns of birth counts on holidays as we do in this paper.⁶ Martin et al. (2018), in an analysis of UK births, probe deeper into understanding the holiday effect by examining changes in delivery type on holidays.⁷ Like Gelman et al. (2013), Martin et al. (2018) show neither how holiday births are displaced over time, nor what the consequences of this displacement are for delivery or birth outcomes.⁸ Our work fills these gaps by exploring the dynamics of retimed births, i.e., how births are temporally

⁶Source: <https://fivethirtyeight.com/features/some-people-are-too-superstitious-to-have-a-baby-on-friday-the-13th/>

⁷Source: <https://fivethirtyeight.com/features/some-people-are-too-superstitious-to-have-a-baby-on-friday-the-13th/>

⁸Previous studies document higher rates of perinatal and/or neonatal mortality on holidays (Macfarlane, 1978; Stephansson et al., 2003; Hong et al., 2006). These effects, which do not consider the outcomes of the displaced births, may be due to the selection of births on holidays and/or the experience and level of hospital staffing on holidays. Similar increases in mortality rates have been found on weekends relative to weekdays (Gould et al., 2003; Hendry, 1981; Mathers, 1983; Hamilton and Restrepo, 2006; Pasupathy et al., 2010; Restrepo et al., 2018).

displaced around holidays; the manner of the re-timing, i.e., how delivery types change on holidays and across the period of displacement; and the health consequences of these changes.

More broadly, this paper provides insights on how the timing of delivery through medical interventions impacts infant health. Based on scheduling manipulations due to holidays, the re-timing of interventions within plus or minus two weeks has only very small (and likely medically and economically insignificant) effects on infant health outcomes. This finding is noteworthy given that ACOG's consensus recommendations generally fall within a few weeks of a due date, depending on the precise condition of the baby or mom (see Spong et al., 2011). These results are also instructive for hospitals setting policies about delivery timing and staffing around the holidays. Furthermore, our results are informative for researchers intending to leverage regression discontinuities generated by birth dates. Regression discontinuity thresholds on or around holidays may create complications as shifts in births can occur as far as 2.5 weeks away from a holiday. Thus, for robustness, our estimates would suggest that donut regression discontinuity estimation (Barreca et al., 2016) excluding up to 3 weeks before and after a regression discontinuity cutoff would be worthy of estimation. However, with such a wide period removed from the analysis, the regression discontinuity estimates may lose some of their appeal – focusing on contrasts potentially less ex-ante similar.

This paper adds to the economics literature on the relationship between birth timing manipulations and post-natal health outcomes. Schulkind and Shapiro (2014), studying the effect of tax deductions, find that such deductions lead to lower birth weight and lower Apgar scores. Borra et al. (2016) study the effect of the revocation of a baby bonus in Spain that led to reductions in birth weight and increases in hospitalization rates. Our paper is distinct from these two in that our main estimates, which exclude January 1st as a holiday of interest, represent responses to holidays independent of financial incentives.

Moreover, in contrast to tax incentives or baby bonuses, which require specific knowledge to take advantage of them, major holidays likely affect the timing decisions of a larger population.

The remainder of this chapter is organized as follows. Section 2.2 provides some background on the medical interventions that enable birth timing. Section 2.3 describes a conceptual model for understanding how holidays may affect birth timing. The empirical approach is presented in Section 2.4, results in Section 2.5 and robustness checks in Section 2.6. Some concluding thoughts are offered in Section 2.7.

2.2 Background on Medical Delivery Interventions

The timing of birth dates is possible because of several medical interventions. For reference, “full term” gestational length is 39-40 weeks after a woman’s last menstrual period. Below we discuss three of the more commonly-used methods of intervention: C-section, labor induction, and labor stimulation.

At one extreme, a birth can be timed quite precisely, down to the exact day, via a C-section. A C-section is the surgical delivery of a child through the mother’s abdomen. C-sections are further distinguished as planned or emergency. Most planned C-sections occur without a trial of labor first. However, some experts suggest first laboring and then performing a C-section is beneficial even in the case of a planned C-section (Black et al., 2015).

Indications for a planned C-section include mechanical obstructions like placenta previa (i.e., the placenta covers the cervix), breech position (i.e., the fetus is not in the head down position), multiple births (e.g., triplets), or maternal infections, such as HIV, which has increased transmission risk during active labor (Berghella, 2018). Since the late 1980s, one of the most important reasons to plan a C-section is a prior delivery

via C-section (Oster, 2018). It is standard practice across many hospitals to prohibit vaginal births after a previous C-section (VBAC births) to reduce the risks of a uterine rupture. In 2017, 87 percent of pregnant women in the US who had a previous C-section subsequently delivered via C-section.⁹

The scheduling of planned C-sections varies by the underlying risk factors. The American College of Obstetricians and Gynecologists (ACOG) recommends that most C-sections take place at 39 weeks or later, as earlier C-sections are associated with adverse respiratory and neonatal outcomes (Tita et al., 2009). A classic obstetrics textbook (Edmonds, 2011) recommends planning the delivery of twins at 37 to 38 weeks. Even with multiple acute risks present, consensus recommendations typically suggest C-section scheduling after 36 weeks of gestation, although recommendations vary by condition (see Spong et al., 2011). Over the time period we study, ACOG guidelines have become more cautious about early planned C-sections.

On the other hand, the use of emergency C-sections is more reflective of imminent risks. These risks include fetal distress, the lack of blood and oxygen flow through the umbilical cord, placenta abruption (the disconnecting of the placenta from the uterine wall), stalled labor, and a baby's size being too large for the birth canal.

The optimal timing and usage of C-sections is a source of considerable debate. Planned C-sections are largely timed weighing the underlying health risks precipitating a C-section against worries about insufficient respiratory development. Non-labor deliveries lead to changes in stress operation, immune response, and altered epigenetic functioning based on experiments in non-human species (Black et al., 2015). Individuals born via C-section have higher rates of infant hospitalization, obesity, and type 1 diabetes (Black et al., 2015). Even low-risk pregnancies are not immune to some of these risks (Card et al., 2018).

⁹Source is authors' calculation of the 2017 Detailed Natality Files.

At the other extreme in terms of both invasiveness and timing precision are induction and stimulation of labor. Induction of labor refers to techniques to stimulate uterine contractions prior to the onset of spontaneous labor (Grobman et al., 2018). Similarly, stimulation of labor, also known as augmentation of labor, refers to techniques that help labor progress after the onset of spontaneous labor. Clinical guidelines often use the terms induction and stimulation interchangeably.

Inductions enable the timing of a birth within a 1 to 3-day window.¹⁰ Typical indications for labor induction and stimulation include a post-term pregnancy and premature labor rupture of membranes (“water breaking”) along with several of the precursors to C-sections, including hypertension, preeclampsia, and maternal diabetes (Grobman et al., 2018). Clinicians generally consider induction/stimulation medically indicated when the risks of continued pregnancy outweigh the maternal and fetal risks of delivery earlier than may have occurred spontaneously.

Unless the fetus or mother is at risk, ACOG does not recommend inducing a birth before 39 weeks of gestation.¹¹ For late-term pregnancies (pregnancies in the 41st week), induction is to be considered and in the post-term pregnancies (pregnancies in the 42nd week or later), induction is recommended.¹² The process of inducing or stimulating labor is usually achieved through administering some drugs. Other methods of precipitating

¹⁰Analysis of nearly 11,000 women undergoing labor induction finds that over 65 percent end the latent phase, i.e., reach active labor or 5-cm dilation, within 6 hours of induction and over 96 percent within 15 hours (Grobman et al., 2018). Median time to delivery after the active phase is 4 hours and the 95th percentile is 13 hours for nulliparous women with epidural analgesia; time to delivery decreases slightly with parity and without epidural analgesia (Zhang et al., 2010). Consequently, inductions allow the timing of births to roughly a 1 to 3-day window.

¹¹Source: <https://www.acog.org/Patients/FAQs/Induction-of-Labor-at-39-Weeks?>

¹²Several observational studies suggest that C-section rates are higher among induced births (Bailit et al., 2015; Luthy et al., 2004). The experimental literature finds otherwise. Hannah et al. (1992) randomized women with uncomplicated pregnancies at 41 weeks or more duration into induction of labor or watchful waiting (or “expectant management”) and found no differences in outcomes, with the exception of lower C-section delivery rates in the induced group. Nielsen et al. (2005), which randomized women between 39 and 42 weeks of gestation and favorable pre-labor cervix (or “Bishop scores”) into induction versus expectant management, found no differences in C-section rates. This finding is supported by meta-analyses (e.g., Saccone and Berghella, 2015; Walker et al., 2015).

labor include membrane stripping in an attempt to detach the fetal membranes from the cervix, nipple stimulation, and acupuncture (Boulvain et al., 2003; Kavanagh et al., 2005; Smith et al., 2013; Modlock et al., 2010).

The risks associated with induction are minimal, particularly in comparison to those associated with C-sections. Grobman et al. (2018) in a recent randomized-controlled evaluation of the consequences of labor induction among low-risk first-time mothers document reduced rates of C-sections without adverse perinatal outcomes. That said, ACOG lists infection, uterine rupture, increased risk of C-section, and fetal death as possible risks of induction.¹³

Use of these interventions has increased significantly over time. In 2016, nearly one third of all births in the United States were delivered by C-section, while in the mid 1990s, only about 20 percent of all births were delivered by C-section.¹⁴ C-section rates were highest in 2009 and have subsequently declined. Inductions have followed a similar pattern. In 1990, the overall rate of labor induction was just under 10 percent and by 2012, it had increased to 23.3 percent. More recently, the fraction of early term births (before 39 weeks) induced has fallen from a peak of 21.0 percent in 2005 to 18.4 percent in 2012.¹⁵ This is likely a consequence of a change in ACOG's recommendations to prevent early-term inductions (Oster, 2018).

¹³Source: <https://www.acog.org/Patients/FAQs/Labor-Induction>

¹⁴Source: https://www.cdc.gov/nchs/data/nvsr/nvsr66/nvsr66_01.pdf

¹⁵Source: https://www.cdc.gov/nchs/data/databriefs/db155_table.pdf#1

2.3 Conceptual Framework: How Holidays affect Birth Timing

Why might holidays affect the use of medical interventions to time births? The reasons are multifold and include preferences for leisure at the time of the holiday among both patients and health care providers, financial incentives, as well as cultural beliefs.

For many holidays (e.g. 4th of July, Thanksgiving), both medical providers and parents may have a strong disutility for a birth during that period. As holidays are often associated with social and family gatherings, working on that day or giving birth on that day may incur large utility costs. Hospitals often compensate their staff via extra pay or extra time off if they work on holidays, and therefore have incentives to reduce their staff on holidays.

In the case of New Year's, parents expecting a new child around January 1st face incentives to expedite the birth of their child to qualify for a tax exemption in the current year. In the US, the timing of births around January 1st is sensitive to the annual child tax benefit, as these benefits are not prorated (LaLumia et al., 2015; Schulkind and Shapiro, 2014; Dickert-Conlin and Chandra, 1999).¹⁶ To disentangle the effect of holidays from financial incentives, we focus specifically on holidays that are likely to coincide with time off and social gatherings, but not with financial incentives. This is why we exclude New Year's from our main analysis.

Furthermore, being born on a certain day may be associated with good or bad luck. In countries and among ethnic groups that follow the lunar calendar, for example, births are more common on auspicious days and less common on inauspicious days (Lo, 2003; Lin et al., 2006; Almond et al., 2015). Likewise, in the US, births are more common on

¹⁶Two other well-documented policies provide incentives for birth timing manipulation: baby bonuses (Gans and Leigh, 2009; Brunner and Kuhn, 2014; Borra et al., 2016) and family leave (Neugart and Ohlsson, 2013; Tamm, 2013; Jürges, 2017).

Valentine’s Day and less common on Halloween (Levy et al., 2011). In addition, parents might prefer that their child’s birthday does not coincide with a fixed day holiday, such as Christmas or July 4th. Similar to inauspicious dates, some dates, such as September 11 after 2001 and Friday the 13th have “negative” connotations. Because such days are unlikely to enter medical providers’ utility function, we use these dates later to isolate demand-side influencers (separate from supply-side shifters) in birth timing to shed light on the potential role of demand versus supply side channels in holiday birth timing.

Irrespective of the motivation, providers can intervene to time a birth through the use of a C-section, induction, or stimulation of labor. These actions are not costless, however. As detailed above, C-sections in particular, which offer the most precise way to time a birth, can increase the health risks for a newborn baby and his/her mother. Conditional on using these techniques, the optimal timing of these procedures is largely unknown and based primarily on consensus. Governing bodies like ACOG recognize and warn against the potential risks of untimely delivery. In the presence of a due date near the holiday, however, medical providers and patients may decide that the utility of intervening outweighs the utility of not intervening.

An intervention to alter a birth’s timing in response to a holiday can have one of two effects on the delivery process: (i) the delivery method and timing can be altered, or (ii) the delivery timing alone can be changed. In the first scenario, a birth may have been intended to be spontaneous in absence of the holiday, but then is delivered via C-section, induction or stimulation as a result of the holiday. In the second scenario, a birth may have been planned to happen via C-section, induction, or stimulation even in the absence of the holiday, but the timing of that procedure may be altered as a result of it. We refer to holiday-related changes in delivery type as extensive margin changes and changes in the timing of a delivery type as intensive margin changes.

While our data do not allow us to fully separate these two scenarios, as described later,

we can estimate the overall effect of holidays on the prevalence of the different delivery types. This information allows us to answer the question of whether a holiday increases the overall incidence of C-sections. A priori, we would expect the second scenario would be more common as it does not conflict with ACOG guidelines discouraging the use of C-sections for uncomplicated pregnancies and implies only a rescheduling of procedures among people already intending to use them.

Given the decision to time a birth to avoid a holiday, when are such births likely to be (re)scheduled? Consider a birth with a due date near a holiday. To avoid the holiday, providers can decide to schedule the birth before or after the holiday. Having a birth early increases the health risks to newborns and moms. But scheduling the birth after the holiday, particularly among pregnancies that are nearly full term, raises the risk that a woman goes into spontaneous labor, a process that for some complications (e.g., breech birth, eclampsia), providers want to avoid. Moreover, for births past the due date, risks of delaying birth include a high birth weight baby (which is a precursor for diabetes) and perinatal mortality (Galal et al., 2012). For births very near term, the probability of imminent birth is high. Among spontaneous vaginal births in the United States in 2017, only 8 percent of births occurred before 38 weeks of gestation and 68 percent of births were between 38 and 40 weeks of gestation inclusive; 24 percent of births occur after 40 weeks.¹⁷ For births far from term, the risk of labor naturally starting is lower, although high-risk births may have higher probabilities of early delivery. In sum, it is ex-ante plausible that births may be shifted to occur before or after a holiday.

So far, this discussion has raised the possibility that holidays impact maternal and infant health through the timing and delivery method. Another channel through which holidays may affect birth outcomes is through the supply of medical professionals. Staffing may be reduced on holidays. Also, holidays may affect the quality of staffing. If work-

¹⁷Based on authors' calculation using the 2017 Detailed Natality Files.

ing on a holiday is undesirable, less experienced workers with lower seniority may be requested to work on a holiday. Junior staff may also get a higher marginal utility from enhanced holiday pay (e.g., time and a half), and may therefore self-select to work holiday shifts. To test the importance of the supply (i.e., medical profession driven) versus demand (i.e., patient driven) channel, we conduct a sensitivity analysis in which we separately examine births around September 11th (after 2001) and Friday the 13th, as these days might arguably affect the demand to (re)schedule, while leaving the supply side of medical professionals unaffected. In an attempt to shut down the demand channel and focus on the impact of restricted supply, we also perform analyses of birth timing around the annual ACOG meetings (available upon request). We found little impact of these meetings on birth timing, a factor that may be attributable to the fact that only about 3,600 out of about 36,000 active OB/GYNs attend these meetings.¹⁸

2.4 Empirical Approach

Data. Our primary data source is the restricted-access 2000-2016 California Birth Statistical Master Files. These data cover the universe of California births during this period and come from birth certificate information that the parents and medical provider fill out at the time of birth. These data include demographic information (e.g., age, education) for the parents, health conditions/outcomes of the mother and infant (e.g., gestational diabetes, birth weight, gestational length), and the use of medical interventions (e.g., C-section, induction, and stimulation). Critical to our approach, these data include the exact date of birth of the infant. With such information, we can document precisely the

¹⁸The 3,600 figure is based on the authors' phone conversations with ACOG staff calculation for 2016, using the ACOG annual meetings demographics: <https://annualmeeting.acog.org/wp-content/uploads/2018/04/2017-ACSM-Demographics.pdf>.

displacement of births across the holiday period.¹⁹

Table 2.1 provides some basic descriptive statistics about the interaction of holidays with births and delivery methods. There are on average 1442 births per day, but births are systematically lower on holidays and weekends (with a mean of about 1100 births per day) than on other days. The data on delivery mode make clear that this is a result of scheduling. The number of C-section deliveries is nearly 50 percent lower on holidays and weekends than on other days. Induced/stimulated births are about 28 percent lower. Spontaneous vaginal births are also lower on holidays (by about 15 percent), although they account for a much higher share of births on holidays (52 percent) than on other non-weekend days (44 percent).

Selection of Holidays. Not all holidays are likely to impact the timing of births. Moreover, it is not a priori obvious which holidays should matter most for patients or providers. As a uniform selection rule aimed at isolating holidays that patients or medical providers may want to avoid because of the joint utility of leisure (i.e., I enjoy the holiday because my friends and families also have that day off), we focus on federal holidays for which salaried workers typically get paid time off. This set contains New Year's Day, Presidents' Day, Memorial Day, Independence Day, Labor Day, Thanksgiving, and Christmas. We exclude Martin Luther King Day and Veterans Day from the analysis because many private sector workers do not get paid time off on these days. In our main analysis sample, we exclude Christmas and New Year's Day because their timing overlaps with changes in tax incentives and the pattern of birth timing manipulation looks quite different from that of other holidays. Specifically, the decline spans a much wider interval. This unusual pattern is likely due to the fact that many individuals take time off during the week between Christmas and New Year's. We provide supplementary estimates that

¹⁹Our goal is to estimate the effect of holidays on the timing of delivery, not on the timing of conception.

include Christmas and New Year’s as holidays to show that our qualitative findings are not materially affected by their exclusion.

Determining the Holiday Manipulation Window. To determine the effect of a holiday, we first need to establish which days around a holiday are impacted by it. To do this, we borrow the insight from the public finance bunching and test score manipulation literature (e.g., Kleven, 2016; Diamond and Persson, 2016; Dee et al., 2019) that within the manipulation region, the missing mass – in our case, the drop in births around the holiday – must equal the excess mass – in our case, the rise in births away from the holiday.²⁰

The estimation of the holiday effect on births necessitates specifying an appropriate counterfactual distribution of births. That is, we must estimate how births would be distributed across days surrounding a holiday in absence of this holiday. In the bunching literature, this counterfactual distribution is estimated via the inclusion of high order polynomials in the “running” variable (i.e., the variable determining treatment, which is date of birth in our setting). For example, the literature on income manipulation due to taxes specifies a polynomial function in pre-tax income (the running variable in that setting) to predict what the distribution of pre-tax income would look like in the manipulation region in the absence of a change in tax incentives.

We depart from this literature somewhat to take a more flexible approach in specifying the counterfactual. We do this because in our context, births in the weeks surrounding the holiday window are likely to be a valid counterfactual (i.e., within a reasonable time frame around a holiday, births are uniformly distributed across weeks). In contrast, the income distribution exhibits considerable curvature, so higher order polynomials are

²⁰Unlike in the case of tax bunching, this equality must hold. Tax incentives (e.g., discrete changes in marginal tax rates) may cause some individuals to not report income – resulting in a non-equivalence of the missing and excess masses.

needed to properly characterize the counterfactual distribution. For completeness, in results not reported, we have followed the previous bunching literature's approach without substantive changes in our conclusions.

Broadly speaking, our estimation strategy involves matching each date in the manipulation window with a day outside of the manipulation window that falls on the same day of the week. We match on day of the week because it is an important predictor of the number of births, as seen in Figure 2.1. Days preceding a holiday (inclusive of the holiday) are matched with days that precede the manipulation window and days following a holiday are matched with days that follow the manipulation window.

Our estimation strategy is best illustrated via a graphical representation. Panel A of Figure 2.2 provides a hypothetical example for the sake of illustration. In this figure, the manipulation region is -3 to 3, i.e., holidays lead to a displacement of births 3 days before and up through 3 days after the holiday. To create our counterfactual, each day in the manipulation window is matched to the closest day outside that window that occurs on the same day of the week. In this hypothetical example, the day before the holiday (-1) is a Wednesday and is thus matched with the next Wednesday in the counterfactual region (-8). The date of the holiday is matched with a day 7 days before the holiday. The day 2 days after the holiday is matched with a day 9 days after the holiday, and so on. Note that when the manipulation region spans a partial week, our estimation procedure excludes some dates near the holiday. For example, in the -3 to 3 manipulation region example, we do not include dates [4,7] and [-4,-6] in our estimation of the holiday effect. We do this at a cost of having counterfactual dates further from the holiday (e.g., 0 is matched to -7 rather than -4), but at the benefit of making comparisons across the same day of the week. As we discuss in more detail below, when the manipulation region becomes larger, the matched dates are further from the holiday. Importantly, we present sensitivity analyses to show that our conclusions are largely unchanged when we slightly

widen or narrow the manipulation region used.

The example in Panel A has a balanced manipulation window, but we allow the window to be unbalanced (i.e., wider or narrower before the holiday). This implies that we do not constrain the control period to be the same distance from the holiday on either side of the holiday. For example, with a manipulation window of $[-3,10]$, we would compare days before the holiday to days one week earlier, whereas days after the holiday would be compared to days two weeks later.

To determine our manipulation region, we calculate (excess mass – missing mass) for each possible manipulation window – starting with a 3-day minimum period on each side of the holiday working towards a 21-day maximum period on each of the holiday. In practice, excess mass is defined as occurring when $(\text{births} - \text{expected births}) > 0$. Similarly, missing mass is defined as $(\text{births} - \text{expected births}) < 0$. Ex ante, there are $19 * 19 = 361$ possible manipulation periods. Our chosen manipulation window is such that the excess mass – missing mass is closest to 0.

Our empirical procedure amounts to estimating regressions of the form:

$$Y_{it} = \alpha_0 + \sum_{j=1}^l \beta_j \mathbf{1}(\text{holiday}_i) * D_{ij} + \sum_{j=1}^l \gamma_j D_{ij} + \nu_{jt} + \sum_{k=1}^6 \delta_k DOW_i + \epsilon_{it}, \quad (2.1)$$

where Y_{it} is the count of the number of births on calendar day i and year t (e.g., September 3rd, 2000). l denotes the total number of days in the manipulation window. D_{ij} are indicator variables for each of the matched pairs (i.e., the matching of a day in the manipulation window with day outside of it). For the example laid out in Panel A of Figure 2.2, l would equal 7 as there are 7 days in the manipulation region. $\mathbf{1}(\text{holiday}_i) * D_{ij}$ is the interaction between the pair dummies and an indicator for whether a specific day in the pair is in the holiday interval (e.g., in our hypothetical example, $\mathbf{1}(\text{holiday}_i)$ equals 1 for days -3 to 3, and 0 otherwise). ν_{jt} are holiday period by year fixed effects (e.g., Labor

Day Period 2005) and DOW_i are day-of-the week dummies. The β_j s are the parameters of interest. They are interpreted as the excess births (if positive) or missing births (if negative) occurring on that day as a result of the holiday.

The possibility of overlap in the manipulation windows across holidays further complicates this regression. For example, in regressions that include Christmas and New Year’s as holidays, the post-Christmas period will coincide with the pre-New Year’s period. We address this by allowing each date to contribute to the estimate of multiple β_j s. Additionally, we control for other holidays that do not fall into our “paid time off”-heuristic but that may also affect birth timing, specifically Halloween and Valentine’s Day. We refer to these as “nuisance holidays” and control for these in all specifications, e.g., with and without Christmas and New Year’s. To control for these nuisance holidays, we include a full set of dummy variables for each day relative to the holiday. The length of the relative-date fixed effects spans the length of the optimal window (e.g., in the [-3,3] example, there would be 7 separate dummy variables for each holiday).

Our manipulation window selection algorithm chooses an optimal window spanning from -11 to +16 days around the holiday. Panel B of Figure 2.2 details the matching procedure for our chosen window. Note, as discussed above, as the window is relatively wide, the counterfactual dates are as far as 37 days away from the holiday. To address potential comparability concerns, we perform supplementary analyses that (i) widen or narrow the manipulation region (shown below) or (ii) use higher order polynomial functions in date of birth. Both of these analyses generate remarkably similar qualitative conclusions.

For our “optimal” window, excess mass – missing mass (i.e., the sum of the estimated β_j s throughout the manipulation region) is equal to 2.2 births. Given over 40,000 births in a typical 28-day period, 2.2 births is small and meaningfully close to 0.

Analysis of the Holiday Effect. With our chosen manipulation window and control days, we now fix l to 28, the total number of days manipulated, in equation 2.1 and estimate:

$$Y_{it} = \alpha_0 + \sum_{j=1}^{28} \beta_j \mathbf{1}(\text{holiday}_i) * D_{ij} + \sum_{j=1}^{28} \gamma_j D_{ij} + \nu_{jt} + \sum_{k=1}^6 \delta_k DOW_i + \epsilon_{it}. \quad (2.2)$$

Our main set of analyses documenting the displacement of births across the holiday period provides estimates of the β_j s (i.e., how births in the holiday period compare to those outside the holiday period).

In addition to births overall, we analyze the number of births by delivery mode – C-section, induced/stimulated vaginal birth, spontaneous vaginal birth – to understand how birth timing manipulation occurs. This is done via estimating regressions of the form of equation 2.2 but replacing the dependent variable with counts of the number of births delivered by one of those modes (e.g., C-section). We also perform the same type of analysis but with counts of births by term length category (e.g., pre-term vs. full term) and separately consider average gestational age by day to characterize the relationship between holiday birth timing manipulation and gestational length.

Using the same analytic sample and basic specification, we also consider the nature of the selection across the holiday period by analyzing the age, race and education of moms as well as delivery payment sources by delivery date. Finally, we consider outcomes such as birth weight, APGAR scores, and NICU use. Again this is done considering each of these outcomes as separate dependent variables.

Regression equations 2.1 and 2.2 describe the nature of the displacement of births due to the holiday by day. To understand the overall impact of the holiday on outcomes, we run a more aggregated regression that considers the average effect across the manipulation window rather than the day-by-day effects. Specifically, we contrast births in the

manipulation window with the counterfactual births outside the manipulation window to derive a reduced-form effect of holiday. This approach acknowledges the selection within the manipulation period (e.g., births on holidays are selected such that comparisons of the outcomes of births within the window would lead to biased estimates of the effect of holidays), while assuming no selection into the optimal holiday manipulation window. We can partially test the no selection assumption by comparing the observable background characteristics of mothers within the manipulation window to those from counterfactual days. Diamond and Persson (2016) in their study of the long-run effects of test score manipulation adopt a similar logic by comparing outcomes of students in the manipulated region with the outcomes of students just outside of the manipulated region.

To estimate the total reduced-form impact of holidays, we adopt an estimating equation analogous to equations 2.1 and 2.2 but aggregate the per-day holiday effect into one holiday effect. That is, we estimate the following equation:

$$Y_{it} = \alpha_0 + \beta_1 * \mathbf{1}(\text{holidayinterval}_{it}) + \sum_{j=1}^{28} \gamma_j D_{ij} + \nu_{jt} + \sum_{k=1}^6 \delta_k DOW_i + \epsilon_{it}, \quad (2.3)$$

where Y_{it} is a dependent variable of interest for a calendar date i in a particular year t (e.g., number of births, number of C-section births, mean gestation, mean birth weight). The key variable of interest is $\mathbf{1}(\text{holidayinterval}_{it})$, an indicator for the 28 day period around a holiday (i.e., a birth occurring between -11 and 16 days around a holiday). The coefficient on this indicator, β_1 , captures the reduced form effect of birth timing manipulation on mean daily outcomes over the holiday period.

Holidays will not impact all births in the manipulation window. Thus, to get a sense of the effect of a birth timing manipulation, we want to scale our reduced-form effects by the fraction of births in the manipulation window whose timing is manipulated. To

do this, we divide the estimates from equation 2.3 by an estimate of the fraction of births in the manipulation region that are retimed due to the holiday. We provide such implied IV estimates mainly as a benchmark to gauge the size of our effects because it is challenging to credibly identify the exact fraction of manipulated births. These implied IV estimates are intended to capture the size of the effect of a birth being manipulated as opposed to the reduced-form effects, which capture the effect of a birth being in the holiday manipulation window.

One reasonable estimate of the fraction manipulated is the ratio of the dip in births very proximate to the holiday (the day of the holiday + the day after) to the number of total births occurring in the manipulation window. This scaling factor implicitly assumes that the manipulation is local to the holiday and that it is the manipulated births alone that are impacted by the rescheduling. If, for example, a holiday shifts a birth from 1 day before the holiday to occurring 2 days after, such a manipulation would not be captured in our calculation of the manipulated fraction. As we will discuss later in more detail, this possibility would imply that the fraction of manipulated births by which we inflate our estimates can therefore be regarded as a lower bound of the effect of a birth manipulation. Thus, any scaled-up estimates of the effect of birth-timing manipulation on the outcomes studies here are upper bounds.

As discussed above, holiday-related shifting may also impose externalities on births that are not retimed because of the holidays. For example, a disproportionate shifting of births to the few days before or a few days after the holiday could cause congestion at some hospitals. To the extent that any such externality affects measured outcomes for births that are not directly manipulated, this would be captured in our aggregate outcome analysis.

Across all outcomes, we estimate equations 2.1, 2.2, and 2.3 using linear regression models. Because we analyze counts of births and deliveries by type, however, we also per-

form sensitivity checks using Poisson regression models for these outcomes. Qualitatively, our results are insensitive to model choice.

2.5 Results

Our results follow in several steps. First, we document the effect of the holidays on the timing of births using the precise date of birth information from the California birth files. Second, we examine how those effects vary by delivery type. Third, we study how the manipulation affects gestational length, which is directly impacted by birth timing manipulation. Fourth, we characterize the selection of births affected by holiday timing. Fifth, we finish with an analysis of the holiday effects on birth outcomes following, with modifications as discussed above, Diamond and Persson (2016). Finally, given that negative birth outcomes are relatively rare, we repeat the analysis for the sample of “high-risk” births (defined below). This sample has both a high probability of being scheduled and, independent of scheduling, of experiencing adverse outcomes.

Change in Births Around Holidays. Figure 2.3 plots the estimated β_j s from equation 2.2 or the change in daily births across the holiday period (ranging from 11 days before the holiday to 16 days after as dictated by our “optimal” window) relative to the counterfactual days either before or after the holiday period. Time zero is the holiday. As expected, the number of births is lower on the holiday itself than would otherwise be expected. The same is true of the day after the holiday, i.e., $t=+1$. We estimate a decline of roughly 511 births on the day of and just after a holiday. Given a mean of 1447 births per day in the analytic sample, this represents an almost 18 percent reduction in births over the 2-day period than would be otherwise expected.

Figure 2.3 makes evident that the holiday period is marked by a hollowing out of

the birth distribution across days. That is, missing holiday births are shifted to the roughly 2 weeks before and 2 weeks after the holiday interval. Holidays cause not just an expediting of some births but also a delay in some births, as postulated earlier.

The general pattern of displacement, i.e., the hollowing out of the birth distribution around the holiday, is quite similar when we include Christmas and New Year's in the analysis (see Figure B.1). One modest difference, however, is that births decline the day before as well as the day after the holiday relative to births on the counterfactual days. This is mostly driven by a reduction in births on Christmas Eve. The drop in births is slightly larger when Christmas and New Year's are included. Indeed, the effects of different holidays are quite heterogeneous. Figure B.2 shows, for each holiday separately, the average number of births on that holiday minus the average number of births on all days in the sample. The holiday drop in births is largest for Christmas Day, followed by New Year's and Memorial Day. In our main analysis, we exclude New Year's Day and Christmas from the holiday set because they also coincide with the tax incentives to time births.

Our interest is in holiday-related birth timing, not holiday + incentive-related timing. The holiday effect is also more pronounced in more recent decades. Figure 2.4 documents the holiday effect for California births for 4 different time periods.²¹ As the mean number of births varies considerably over time, we present Poisson estimates of equation 2.2 in this figure. In the early 1970s, the number of births on holidays was roughly 15 percent lower than expected. As medical delivery interventions became more common, this drop increased. Interestingly, however, the holiday effect is larger for the 2000-2002 period than for the 2014-2016 period. The fall in the holiday effect may be attributable to ACOG's 2013 guidance on what constitutes an early-term birth and in what instances such deliveries are warranted (of Obstetricians and Gynecologists), of Obstetricians and

²¹The California data prior to 2000 are taken from the National Vital Statistics data.

Gynecologists); Oster, 2018).²²

Table 2.2 (Panel A) supports the choice of the holiday window. The first column of regression estimates present estimates of equation 2.3 using total daily births as the outcome variable. Across the full 28-day holiday interval, we observe on average 4 fewer births per day.²³ As holidays affect the shifting of births (not their presence), the net effect on births should be very close to 0. Off a base of 1447 births per day, this 0.2 percent decline in births is neither statistically nor economically distinguishable from 0. The average net change in births when we consider Christmas and New Year’s (Panel B) is similarly small and statistically indistinguishable from zero – 1 more birth per day or a 0.08 percent increase in total births over the holiday period.

Change in Delivery Type Around Holidays. A crucial element in our documentation of the mechanisms driving the holiday effect is an analysis of delivery type. Figure 5 breaks out the change in the number of births by delivery type. We create three mutually exclusive categories: (1) births by C-section, (2) vaginal births after the induction/stimulation of labor and (3) vaginal births after spontaneous labor.²⁴ Note that the last two categories, by definition, exclude cases that end in a C-section. C-sections decline by about 251, primarily the day of and the day after the holiday. This represents a 2-day decline in C-sections of over 25 percent relative to a daily average of 442 C-sections. These C-sections are shifted to both before and after the holiday. The same basic patterns are found for spontaneous vaginal births and vaginal births after induction/stimulation

²²In 2013 ACOG redefined full term birth from deliveries between 37 and 42 weeks to deliveries between 39 and 40 weeks and reclassified deliveries from 37 to 38 weeks as “early term” births (Oster, 2018). ACOG also made specific recommendations on the risk factors that warrant early term delivery, with the implication that early (timed) deliveries should be avoided when specific risks are not involved (of Obstetricians and Gynecologists), of Obstetricians and Gynecologists); Oster, 2018).

²³Note this -4 births per day (or $28 * (-4) = -112$) is different from our 2.2 births from our manipulation window selection algorithm. This difference is attributable to different regression specification (equation 2.2 versus equation 2.3).

²⁴We classify births that are reported as both spontaneous and induced as induced births.

of labor. Specifically, on the day of and just after the holiday, spontaneous vaginal births decline by 138, representing a decline of about 10 percent over the 2-day period given an average day with 669 spontaneous vaginal births. Induced/stimulated vaginal births decline by 114 over the 2-day period, representing a drop of almost 20 percent. The reduction in C-sections on the holiday and the day after accounts for roughly 50 percent of the total decline in births. The remaining decline is split between spontaneous vaginal births (27 percent) and stimulated/induced vaginal births (27 percent).²⁵ These findings are consistent with Martin et al. (2018), who characterize the effect of bank holidays in the UK on delivery type. They report that the decline in holiday births is primarily due to a reduction in scheduled C-sections, followed by declines in induced vaginal births.²⁶

Measured across the whole 28-day holiday interval, we find no meaningful changes in delivery types (see Table 2.2, Panel A). Both C-section births and induced/stimulated vaginal births decline by about 2.5 per day. Although the estimated decline is statistically distinguishable from 0 for inductions, it amounts to a less than 1 percent decline. Thus, while births and delivery types are rescheduled across holiday intervals, the likelihood of a given delivery type (e.g., C-section delivery) changes only modestly. While we do not find evidence that holidays lead to changes in delivery methods, we cannot rule this possibility out either. For instance, it could be the case that births shift from spontaneous vaginal births to induced vaginal births before a holiday, and from induced vaginal births to spontaneous vaginal births after the holiday. These two effects would wash each other out and appear as though there was no impact on delivery type. However, as best we can gather, holidays appear to affect the timing of delivery and not the type of delivery

²⁵Note the sum of C-section, spontaneous vaginal, and induced/stimulated vaginal does not equal the total birth effect. That is because there is a small missing category (unclassified births). The holiday effect (i.e., the day of the holiday and the day after) for the unclassified birth category is a reduction of 8 births.

²⁶Unlike our work, Martin et al. (2018) do not look at the complete displacement of births around the holiday period, but instead focus on the impacts on the holiday day, the last week day before the holiday, and the first week day after the holiday.

(i.e., consistent with scenario 2 in our conceptual framework).

The general pattern of findings is quite similar when we include Christmas and New Year's in the analysis (Table 2.2, Panel B). The only notable difference is a statistically significant, albeit very small, increase in spontaneous vaginal births. The point estimate implies that vaginal births increase by about 0.6 percent as a result of holidays. In other words, the rescheduling of C-sections and inductions as a result of the holiday means that a few more births are delivered without medical intervention, a clear but small change along the extensive margin of delivery type.

Change in Gestational Length. A separate but related question to how holiday births are re-timed (i.e., using which interventions) is when a birth gets re-timed. To understand this issue, we analyze the number of deliveries by ACOG's definitions of "term pregnancy." Specifically, we analyze deliveries according to the following ACOG categories: (1) pre-term, which is prior to 37 weeks gestation; (2) early term, which is 37 0/7 weeks through 38 6/7 weeks of gestation; (3) full term, which is 39 0/7 weeks through 40 6/7 weeks of gestation; (4) late term, which is 41 0/7 weeks through 41 6/7 weeks of gestation; and (5) post-term, which is 42 weeks of gestation and beyond. We also consider the average length of gestation.

Figure 2.6 shows the change in mean gestational age by day (Panel A) and the change in the number of daily births according to the 5 ACOG term length categories: pre-, early, full, late and post-term (Panels B-F, respectively). Mean gestational age is about $\frac{1}{2}$ day lower on the day after a holiday, with much of the change on the holiday itself. Although statistically significant, this decline is neither medically nor economically meaningful relative to a mean gestational age of about 275 days.

While mean gestational age is largely unchanged across the holiday interval, the composition of births by term length does change across the window. If holidays only

affected the timing of “elective” intervention births, we would only expect impacts on births nearer to full-term. But births of all term lengths, with the exception of late term births, decline on the holiday itself relative to the counterfactual. Thus, the decline in total holiday births observed in Figure 2.1 is composed of births across the gestational age spectrum.

The rise in births before the holidays is mostly attributable to an increase in the number of full term births, whereas the increase after the holiday is largely due to a rise in the number of early term births. Because we cannot pinpoint exactly when these births would have happened in absence of the holidays, it is difficult to disentangle the exact mechanisms leading to these patterns. But these results are consistent with full-term births that would have been due on or near the holiday being moved earlier and with earlier term births that would have occurred proximate to the holiday being moved later. The asymmetry of this shifting may be because medical professionals worry about the adverse effects of altering early births and worry less about affecting the timing of full term births.

Measured across the whole 28-day holiday interval, we find no meaningful change in gestational length (see Table 2.3, Panel A). Mean gestational age is about 0.03 days longer across the holiday period, a result that is neither statistically, medically, nor economically significant. The number of full term births appears unchanged as a result of the holiday, while pre-term births increase by about 1.5 per day and early term births decrease by nearly 6 per day. Expressed relative to the daily means, this is an increase in pre-term births of about 1 percent and a decline in early term births of about 1.6 percent. On the other hand, late term births increase by about 3.5 per day or 2.6 percent relative to the mean and post-term births decline by 3.3 per day or 4 percent relative to the mean. Thus, on net, the timing of births is changed only quite modestly relative to what would occur absent a holiday. The term length results are quite different when Christmas and

New Year's are included – one reasonable hypothesis for why that is that Christmas and New Year's span a larger holiday period. Thus, there is possibly more demand for shifting births further away from these holidays. However, none of the estimates that include Christmas and New Year's is statistically different from zero.

Nature of Selection. Crucial to our interpretation of the impact of holiday-related birth retiming on outcomes is an understanding of who is affected by holidays. A priori, medical providers may be most willing to move births in cases where they expect little impact of this rescheduling on birth outcomes. To assess this possibility, we study how the characteristics of moms and babies differ across days within the holiday interval.

We first consider two measures of health risk: first, whether the pregnancy is “low risk;” and second, whether the mom is over age 35 – the cut-off traditionally used to define “advanced maternal age.” Low risk pregnancies are defined similar to Card et al. (2018) such that all of the following criteria apply to the birth: (i) singleton, (ii) not breech, (iii) gestation lasted at least 259 days, (iv) mother was at least 18 and no more than 35 years old, (v) mother did not have preeclampsia or eclampsia, (vi) mother had no more than 20 prenatal visits, (vii) baby had no intrauterine growth restriction, and (viii) mother had no previous C-section.²⁷

The fraction of births that are low risk spike on holidays as seen in panel A of Figure 2.7. This is not too surprising since high-risk pregnancies are more likely to be scheduled deliveries and thus explicitly shifted away from a holiday. However, the shift of high-risk (or non-low risk) births away from the holiday is only apparent for the holiday itself and not, for example, the day after the holiday. This pattern is corroborated by panel B of

²⁷Card et al. (2018) also exclude current C-sections. Since delivery type is one of our outcomes of interest, we do not make this restriction. Note that some of the conditions used in this measure, e.g., intrauterine growth restriction, are not in the data prior to 2007. So the low risk measure captures only the later part of our sample. In addition, as Card et al. focus on first births, we report our results on low risk first births as well.

Figure 2.7, which shows that moms of babies born on a holiday are less likely to be of “advanced maternal age,” a strong predictor of pregnancy complications (Fretts, 2018). On the other hand, as shown in panels C and D of Figure 2.7, women giving birth on a holiday are slightly less likely to be white and more likely to be teenagers. These patterns may indicate, aside from selection according to health risk, that socioeconomic status is a predictor of holiday birth timing, with the more advantaged births more likely to be moved.

In contrast to the timing of births within the holiday manipulation window (e.g., the comparison of births on the holiday to the day before), which is clearly endogenous, births across the manipulation period should not be subject to selection. That is, if we have chosen a valid counterfactual, our “holiday-treated” births (i.e., those inside the holiday manipulation window) should be ex-ante otherwise similar to our holiday “control” births (i.e., the counterfactual births in the control region, outside of the holiday manipulation window). This assumption, which is crucial to estimating the effect of holidays on birth outcomes, can be partially tested by estimating equation 2.3 using the background characteristics of mothers as dependent variables.

Across the holiday manipulation window, most characteristics are balanced (i.e., being within the manipulation window is not correlated with pre-determined characteristics), with the exception of maternal age and delivery type. Births are less likely to be from an older mother and more likely to be paid using public funds (see Table 2.4). The effect on the probability of the mother being older than 35 is -0.0013 off of a base of 0.14 (a 0.9 percent effect). Likewise, the share of deliveries that are public (private) pay increases (decreases) by about 1 percent. None of the other coefficients are statistically significant and, more importantly, all are quite small in magnitude. For example, the (positive) effect size implied by the estimated impact on the probability of the mother having a high school degree or less is only about 0.5 percent. This pattern holds in both

our main holiday analysis (Panel A of Table 2.4) and when we add Christmas and New Year's to the analysis (Panel B of Table 2.4). While the selection effects in Table 2.4 are, at face value, small in magnitude, holidays only affect a small fraction of births in the selection window. As such, one may want to scale those effects by the fraction of births manipulated. In calculations discussed below, a reasonable scaling is to multiply the Table 2.4 effects by 80 (i.e., 1.26 percent of births impacted). In that case, the selection effects are much larger.

Overall, however, the selection effects point to no clear direction of bias. For example, the estimates on older maternal age are negative (i.e., consistent with positively selection) whereas the impacts on the share of deliveries made via public payment are positive (i.e., consistent with negative selection). We recognize the possibility that a comparison of births within the manipulation window with those outside of the manipulation window may be clouded with some selection effects along these margins. With this in mind, we analyze birth outcomes both with and without controls for the background variables for which we find statistically-significant holiday effects.

Birth Outcomes. As we next look at birth outcomes, it is important to keep in mind the displacement patterns. In particular, holidays shift births both before the holiday and after the holiday. It is reasonable to assume that shifts in either of these directions could have heterogeneous effects. One plausible hypothesis, supported by the medical literature, is that moving a birth earlier may lead to an increase in adverse outcomes, whereas moving a birth later may not. The net effect on birth outcomes would be a weighted average of the effects on the births pushed earlier and those pushed later. In this case, the estimated effect on outcomes would not be very informative with regard to the effect of scheduling a birth early.

Another feature worth pointing out is the degree to which births are manipulated.

Most of the action happens within a short time period around the holiday. To a best approximation, manipulated births are shifted by a couple of days (not a couple of weeks). Ex-ante, one may presume that extending a pregnancy by a few days is likely to have minimal effects on outcomes. However, given the discussions amongst ACOG and the frequency with which birth timing is manipulated (e.g., weekends in addition to holidays), our estimates on birth outcomes are informative as to whether these common timing manipulations have deleterious effects. Moreover, ex ante, it is not evident whether the holiday drop occurs exclusively because of a change in the timing of births as opposed to a change in the timing of births and the mode of delivery. An alteration in the delivery method is likely to increase the likelihood of adverse outcomes more than a change in the timing of delivery by a few days.

Figure 2.8 shows how mean birth weight, the fraction of births with a newborn condition, and the fraction of births with a labor complication vary within the holiday period.²⁸ The fraction of births with newborn conditions or with labor complications appears to increase on a holiday by about 0.5 percentage points. These estimates imply increases of about 4 and 2.5 percent respectively, although neither estimate is statistically distinguishable from zero. Babies born on the day of or the day just before or after the holiday have slightly lower birth weight (between 10 and 20 grams lower off a mean of about 3300 grams) than would be predicted absent the holiday. This difference is very small and even more so when considered over the holiday period. The patterns we see here, particularly on holidays, are consistent with some of the selection effects we observed earlier (i.e., disadvantaged women are more likely to give birth on the holiday). But for the non-birth weight outcomes, which may be sensitive to staffing levels and quality,

²⁸Newborn conditions include conditions related to the central nervous system, respiratory system, digestive system, chromosomal anomalies, etc. There are a total of 75 possible conditions. Labor complications (which include delivery complications) include premature rupture of membrane (> 12 hours), cord prolapse, fetal distress, anesthetic complications, unsuccessful attempt at vaginal birth after C-section, maternal blood transfusion, etc. There are about 30 possible labor and delivery complications.

the estimates are also congruous with adverse impacts of reductions in the quantity or quality of staffing.

As shown in Panel A of Table 2.5, across the full 28-day holiday interval, we find a roughly 2-gram lower mean birth weight relative to the counterfactual days. Although statistically distinguishable from zero, this difference is neither medically nor economically significant. The fraction with any newborn conditions, however, decreases by about 0.08 percentage points or about 0.6 percent. The fraction with labor complications also decreases modestly – on the order of 0.4 percent, although the effect is statistically indistinguishable from zero in our main analytic sample.

Of course, these estimates are reduced-form impacts. They average across all births in the sample, but only a relatively small number of births are clearly manipulated by the holiday. To gauge the size of these effects better, we estimate the fraction of births within the manipulation window whose birth is manipulated. We do this by estimating the number of missing births on the holiday and the day after and dividing that by the number of total births in the manipulation window. This scaling corresponds to the ratio of 511 fewer births to the total of about 40,500 births across the 28-day manipulation window, or 0.013. Again this approach is subject to the caveat that we cannot measure the fraction manipulated directly and thus may be a poor approximation. For example, a C-section that would have been scheduled for 2 days after Thanksgiving but moved to the Tuesday before Thanksgiving would not be counted as a manipulated birth according to our back-of-the-envelope calculation. To the extent that we miss manipulations like this, we will be underestimating the fraction manipulated and thus, our scaling will deliver upward biased estimates. This scaling gives us a sort of IV estimate – the effect of manipulating the timing of a birth, provided that the usual IV assumptions are met, including that the occurrence of holiday only shifts births away from occurring on the holiday. But we are cautious in interpreting these scaled estimates as IV estimates since

our measure of the fraction manipulated is likely a very rough approximation.

The reduced-form birth weight effects are modest and statistically significantly different from zero. The “IV effect” for mean birth weight implies reductions of roughly 150 grams. When we control for potential selection in the manipulation window, specifically by including the share of moms over age 35, the share of private insurance delivery payments and the share of public insurance delivery payments in equation 2.3, the implied IV is about 104 grams (see Table B.1). For comparison, the effect of smoking on birth weight is approximately 250 grams (Almond et al., 2005) and in twin comparisons, a 150 gram difference in birth weight would imply a 0.0045 difference in high school completion and a 0.6 percent difference in earnings (Black et al., 2007). When New Year’s and Christmas are included in the analysis, the implied IV estimate is about one-third the size; holiday birth retiming reduces birth weight by about 50 grams. Since changes in birth weight at the mean may not as meaningful as those in the tails, we also analyze the share of births that are low birth weight, i.e., below 2500 grams. The reduced-form effects on low birth weight are small but when scaled by the fraction manipulated, are quite sizable (a 50 percent increase in the likelihood of low birth weight in Panel A and a 18 increase in Panel B).

In contrast, the effects on the other outcomes are all negative in sign (i.e., imply that holidays reduce the likelihood of those adverse outcomes), including outcomes such as the use of the neonatal intensive care unit, the use of assisted ventilation (an important outcome because of the late development of the respiratory system during pregnancy), the number of newborn conditions, and the number of labor complications (see Table B.2). For any newborn condition, the IV estimates suggest reductions in newborn conditions of about 7 percentage points or 76 percent relative to the mean of 9.2 percent of births with a newborn condition. The estimated impacts on any labor complications, are small and very imprecise. Labor complications decrease by about 0.2 percentage points. When

rescaled, this implies reductions of about 12 percentage points or 27 percent relative to a mean of 41 percent, although again these estimates are indistinguishable from zero. We estimate larger, more precise declines in newborn conditions and labor complications when we control for observables (see Table B.1). The IV estimates, once we control for pre-determined observables, imply declines of 8.4 percentage points in the share of newborn conditions and 18.6 percentage points in the share of labor complications. Both estimates are statistically distinguishable from zero. Holiday effect estimates for low birth weight and low Apgar score are small in magnitude and very imprecise, whether or not we control for observables.

While the reduction in births in proximity to the holiday is large, it is challenging to precisely estimate the effect of the holiday displacement on outcomes, since births get re-timed over a relatively wide window (28 days). Dividing a reduced-form estimate by a small number will likely lead to imprecise estimates. However, across a wide range of outcomes, aside from birth weight, there is limited evidence of adverse consequences of the re-timing of births as a consequence of holidays.

Sub-sample of High-Risk Pregnancies. A potentially important caveat to our conclusions above about the consequences of the holiday-related birth timing is that most pregnancies tend to be low risk. The low risk nature of most births may make it difficult to pick up any real effects of birth timing. To assess this issue, we redo the main analysis on the sample of births to women with high-risk pregnancies, where high-risk is defined as meeting at least one of the following criteria: mom had a prior C-section, baby is in the breeched position, multiple birth pregnancy, or mom has an infection such as HIV. In our sample, about 230 or 16 percent of births per day are from high-risk pregnancies. Almost all of these pregnancies result in a C-section birth (94.4 percent).

As in the overall sample, births from high-risk pregnancies drop on holidays (see

Figure 9 Panel A). The drop is considerable – 150 births off a base of 230 births from high-risk pregnancies. As expected, given that the high rate of (planned) C-section delivery among this group, the decline is almost entirely driven by C-sections (see Figure 2.9 Panel B).²⁹

Across the full holiday interval, we find a very small decline in the number of births for the high-risk group. Recall that ideally, this estimate would be as close to 0 as possible (i.e., within the holiday window, the net number of births is 0). But this effect is not sizable. The number of high risk births declines by about 2 per day or 0.7 percent over the holiday window (See Table 2.6). The decline, like the sample overall, is driven primarily by C-sections.

As in the full sample, early term high risk births that would have occurred on the holiday appear to be pushed later, and late term high risk births are pushed earlier (see Figure 2.10). However, we find no meaningful effect of birth retiming on the mean gestational age of high risk births across the holiday window (see Table 2.7 and Figure 2.10), an increase of 0.1 days off a base of 269 days. Across the holiday period, statistically-significant selection in terms of pre-determined characteristics is only present for the source of delivery payment (public or private). Specifically, the share of deliveries paid for through private insurance decreases by about 0.0036, which is offset by a similar size increase in the share paid for through public programs (see Table 2.8). These effects are large when considered relative to the roughly 2 percent of retimed high risk births – roughly 150 births out of over 6440 in the holiday interval. To address the possible concern about selection, we control for delivery payment variables in our birth outcome regressions as a robustness check.

In Table 2.9, we consider the impact of the holiday on birth outcomes for the high

²⁹To be consistent with the overall sample, we used the optimal manipulation window determined by the overall sample. If instead, the optimal manipulation window is calculated using only the sample of high-risk births, the window is [-18, 13].

risk sample (see Figure 2.11 for the day-by-day patterns in outcomes). With the lone exception of any labor complications, the estimates are consistent with a health-improving effect of holidays, at least in terms of sign. The effects on birth weight are opposite in sign but smaller in magnitude than for the full sample (see Table 2.9). The IV estimate implies an increase in birth weight on the order of 36 grams, although this estimate is quite imprecise. The IV effects for low birth weight translate to a 4 percentage point reduction (with or without controls – see Table B.3). The IV implied drop in the presence of any newborn condition is sizable (-9 percentage points off of a base of 16 percent) whereas the low Apgar score is not (-0.4 percentage points). Overall, in accordance with the results from the main sample, the mounting evidence does not point to sizable adverse effects of birth timing manipulation. However, because the fraction of births displaced as a result of the holiday is relatively small, precise conclusions about the size of the effects are not possible.

2.6 Robustness and Sensitivity Analysis

We perform numerous robustness and sensitivity checks. All support the finding that while holidays affect the timing of birth and, in some cases, mode of delivery, they have limited negative impacts on infant health.

Poisson Regression Models. To begin, we redo our analysis of daily births, overall and by delivery type, using Poisson Regression Models (PRM). The PRM is better suited to the count nature of these daily data. As shown in Figure B.3, estimates from the PRM also show a hollowing out of the birth distribution over the 28-day holiday period. Declines are observed on the holiday itself as well as the day after, with small daily increases offsetting the declines spread out across the rest of the holiday period.

The PRM estimates of the decline on the holiday and the day after are quite similar to what we find using a linear regression model of birth counts. Specifically, the PRM implies a decline in daily births of about 23 percent on the day of the holiday and 10 percent the day after, which is remarkably close to the 346 and 165 birth declines off a daily mean of 1447 births from the linear regression models (i.e., reductions of 24 percent and 11 percent, respectively).³⁰ Thus, the use of the linear regression model appears to have limited impact on our conclusions concerning the daily birth displacements.

Likewise, our estimates of the aggregate or net effect of holidays on births and delivery types across the full 28-day holiday period are remarkably similar across the two models. The estimates in Table B.4, using PRMs, imply a -0.21 percent decline in daily births, -0.4 percent decline in C-sections, 0.08 percent increase in spontaneous vaginal births and a -0.9 percent decline in induced/stimulated vaginal births. The analogous calculations from Table 2, imply a -0.28 percent decline in daily births, -0.5 percent decline in C-sections, 0.07 percent increase in spontaneous vaginal births and a -0.9 percent decline in induced/stimulated vaginal births relative to their means.

Separating Holiday Effects from Day of Week Effects. As another specification check, we restrict our analysis to those holidays that fall on fixed dates and thus varying days of the week across years. The restricted holiday set includes New Year's, July 4th and Christmas across 17 years. As day-of-the-week effects are strong, this restriction allows us to better separate out the effect of holidays from any day of the week effect. The pattern of displacement (see Figure B.4 and Table 2.5), however, follows the same hollowing-out pattern of the earlier figures. This suggests that our findings are unlikely to be driven by day of the week effects.

³⁰To interpret the PRM coefficients as percent change, we transform them as follows: $e^\beta - 1$. Thus a coefficient of -0.26 corresponds to a decline of 23 percent.

Sensitivity to the Holiday Manipulation Window. We next consider the sensitivity of our analysis to the manipulation window. The optimal manipulation window was chosen such that the reduction in births right around the holiday is fully offset with increases in births on either side of the holiday. As discussed above, however, the length of this manipulation window and the days captured by it have implications for the control period. Specifically, as the manipulation region becomes wider, the matched dates are further from the holiday. To address this issue, we test the robustness of our results to either narrowing or widening the holiday period by 3 days on either side. Thus, we consider as the treatment period either 8 days before and 13 days after (narrower window) or 13 days before and 19 days after (wider window) the holiday. Figure B.5 shows the evolution of the displacement of births using the narrower and wider treatment windows. The broad pattern of birth displacement is not materially affected by the window choice, which is expected given that the sharp decline in births occurs over only 2 days and the corresponding increase is spread out over many days on either side of the holiday.

When we consider the net change in births overall and by delivery type, however, the analysis makes clear why the optimal $[-11, 16]$ day window is preferred. Specifically, as shown in Table B.6, a smaller window (Panel B) implies a net drop in births of nearly 10 per day and a wider window a net increase in births of 12 per day. Both of these effects are statistically significant. It is a window in between the two that comes closer to capturing the region over which holidays displace births.

Despite this issue, the findings for selection are quite similar (see Table B.7). Across all windows, there is some evidence of selection in terms of the share of deliveries paid for via public programs or private insurance. The magnitude of the changes in deliveries paid for via public funds or private insurance is consistently smaller in the alternative windows. The effect on the likelihood of being an older mom is only statistically significant for the optimal window. Overall, the degree of selection appears to be dampened for the non-

optimal windows.

Table B.8 shows our estimates of birth outcome effects as we vary the window size. Holiday effect estimates for birth weight and labor conditions are sensitive to window size – both flip sign with the wider window and decrease in magnitude as we change window size. We continue to find declines in the fraction of births with newborn conditions, although the estimates are smaller in magnitude and very imprecise. The share of low birth weight births flips sign with the wider window, implying decreases in the fraction of low birth weight, but suggests an increasing fraction with the narrower window. Estimates for the low APGAR score continue to be statistically, economically and medically insignificant. On balance, these results support the conclusion that holiday birth retiming has limited negative impacts on newborn health.

Isolating Demand Side Effects. The conceptual framework outlined three mechanisms by which holidays could affect infant health: the mode of delivery, the timing of delivery, and the quality and/or quantity of staffing on holidays. It is conceivable, for instance, that some hospitals are understaffed on holidays, or that the staff in service are less experienced on average on those days. This might negatively affect neonatal health outcomes, and thus lead to an upward bias of the effect of birth timing on outcomes.

Unfortunately, we do not observe supply side characteristics such as which medical staff was present at birth in our data. To remove supply side factors from consideration, we focus on certain days that might have extensive or intensive margin effects on the scheduling of births, but little to no effects on staffing levels.

Specifically, we look at September 11th after 2001, and Friday the 13th, as parents may want to avoid births on these days.³¹ Mothers might avoid giving birth on September

³¹To induce supply side variation, we also investigated how the annual ACOG meetings affect birth timing. Past work (Gans et al., 2007) shows evidence that births fall in United States during the meetings by 1 percent but use national data and an earlier time period (1990-2003) along with a less-controlled regression specification. However, we did not find that the meetings affected birth timing even though

11th so that their child’s birthday is not associated with the terrorist attacks of 9/11. Furthermore, Friday the 13th is considered unlucky in Western culture. Provided that neither of these days has a direct impact on the staffing of hospitals, analyzing outcomes around these days can help us to better understand the effects of (re-)scheduled deliveries separately from supply side effects. However, to the extent that there is heterogeneity in the effects of holidays, it may be difficult to extrapolate from these two dates.

When estimating how September 11th and Friday 13th affect our outcomes of interest, we largely build on our main empirical strategy. That is, we consider the time window spanned by 11 days prior and 16 days after these days as manipulation region and control for all holidays. Put differently, September 11th and Friday 13th are what used to be our “holidays of interest,” and all holidays, as well as Halloween and Valentine’s Day, are considered to be possible confounders and “nuisance holidays.”

Results (not shown) suggest that approximately 50 fewer births occur on September 11th and Friday 13th on average, which is largely driven by a drop of C-sections by about 30. There are no significant changes in the number of births by other delivery types or other outcomes of interest such as birth weight or maternal characteristics but we must also note that the research design is less powerful for detecting these types of effects because the fraction of births manipulated is much smaller for these dates.

On the one hand, these results suggest that birth time manipulations around holidays might in part be driven by patients’ preferences. Considering that births get (re-)scheduled around events such as September 11th and Friday 13th – which presumably leave the quality and quantity of the supply side unaffected – demand side factors might account for some of the holiday birth timing manipulations. On the other hand, these findings reinforce the insight that there is little to no evidence of adverse health outcomes for babies whose birth date was manipulated by a few days.

many of the meetings occurred in California.

2.7 Conclusions

Consistent with previous studies, we find that births are less common on holidays. High-risk pregnancies are more likely to be shifted as a result of holidays. While the number of births displaced from holidays has generally increased over time, we observe a reduction in the holiday effect since at least 2014.

We find clear evidence that holidays shift the timing of births (i.e., that holidays affect birth timing on the intensive margin). In other words, among births scheduled via C-section or induction, holidays clearly affect the precise timing of these procedures. We find little support for the idea that holidays affect birth timing on the extensive margin, i.e., that more births are scheduled via C-section or induction as a result of holidays, given that there is no net change in delivery types. That said, extensive margin changes cannot be ruled out without knowing the delivery plan of mothers in our sample.

Across our sample, we find little evidence of unfavorable health consequences as a result of holiday-related birth timing. This is true for the overall sample and the sample of high risk births. That said, some caution is warranted in extrapolating our findings from the relatively small share of births (0.013) that get moved across the holiday period.

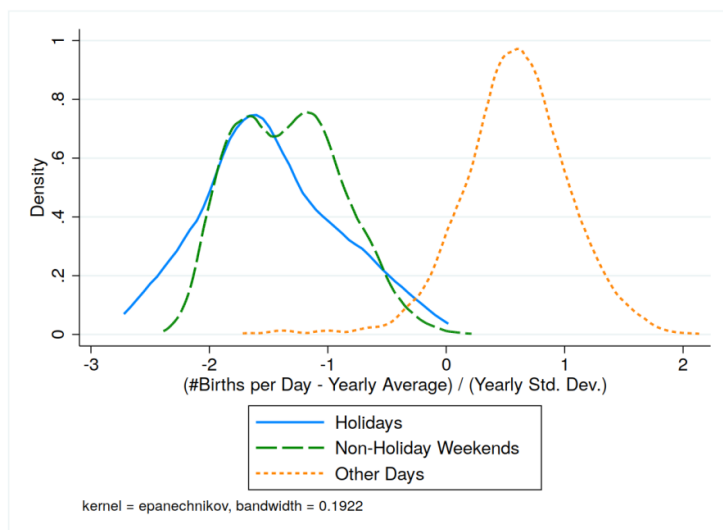
As the rate of medical interventions in the delivery process is much higher than it was decades ago, worries about the possible adverse effects of intervening linger. These quasi-experimental results, using holidays as a natural experiment, can be informative in that regard. As, in aggregate, the delivery mode is not impacted, these results are most useful for guiding decisions on when to intervene (and not on whether to intervene).

Even absent our extrapolation to understanding the effects of timing delivery interventions, this work is beneficial for hospitals deciding on holiday policies regarding staffing and the use of medical interventions such as C-sections, inductions, and stimulations. Our results suggest that the current shifting of births to accommodate the holiday

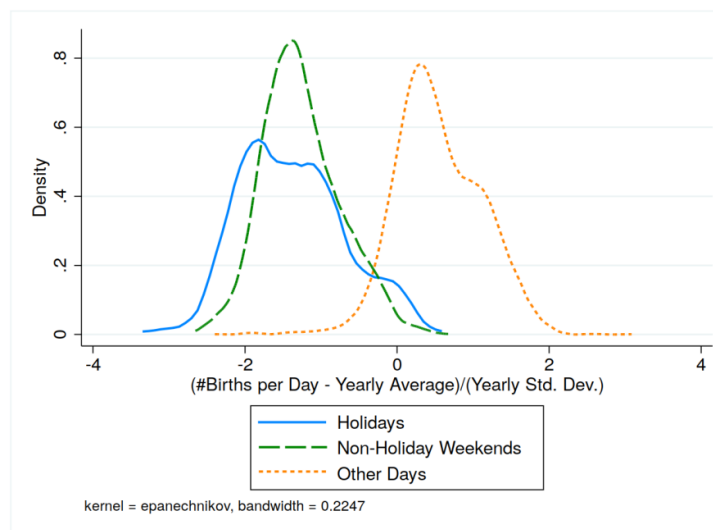
plans of both providers and patients does not have large adverse health consequences for either newborns or their moms at delivery.

Figure 2.1: Distribution of the Number of Births Across Different Days

A. California, 2000-2016



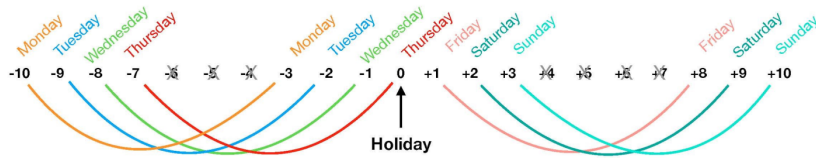
B. United States, 1969-1988



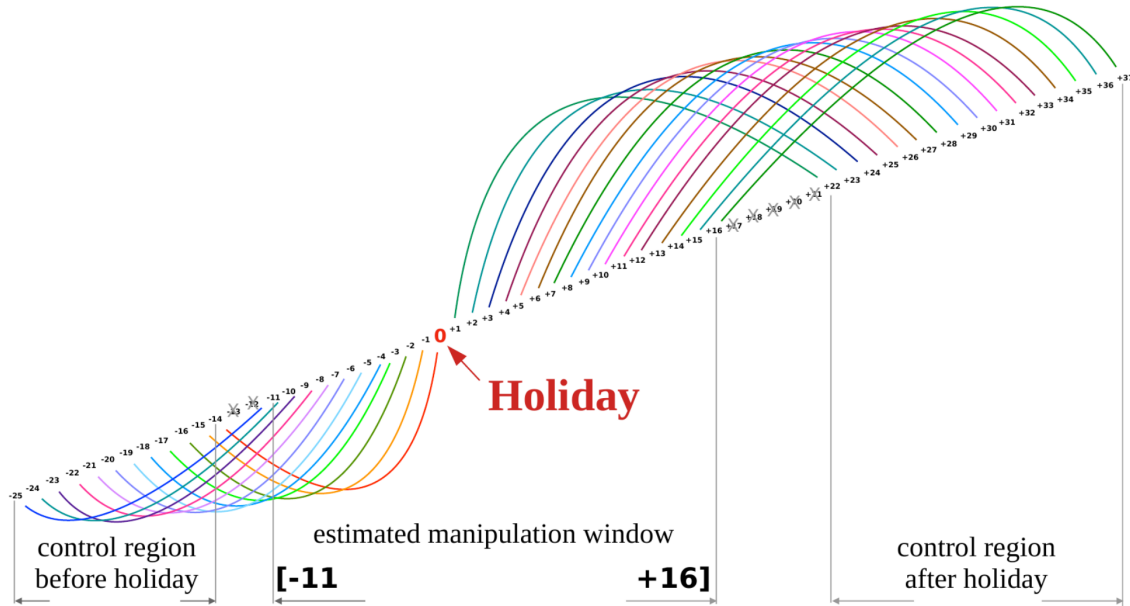
These figures show the distribution of daily births (as represented in Z-scores) for three types of days: holidays, non-holiday weekends, and other days. Panel A is for California and Panel B is for the entire United States. The set of holidays contains New Year's Day, Presidents' Day, Memorial Day, Independence Day, Labor Day, Thanksgiving, and Christmas Day. Starting in 1971, 11 states (DE, GA, IA, IN, KS, KY, LA, NC, NM, RI, WI) stopped celebrating Presidents' Day. Panel B therefore includes data of all states prior to 1971, and data of all states celebrating Presidents' Day starting in 1971. For states and years for which the data are based on a 50% sample of births, observations were multiplied by two.

Figure 2.2: Graphical Representation of Estimation Strategy

A. Illustrative Example with a Hypothetical Manipulation Window

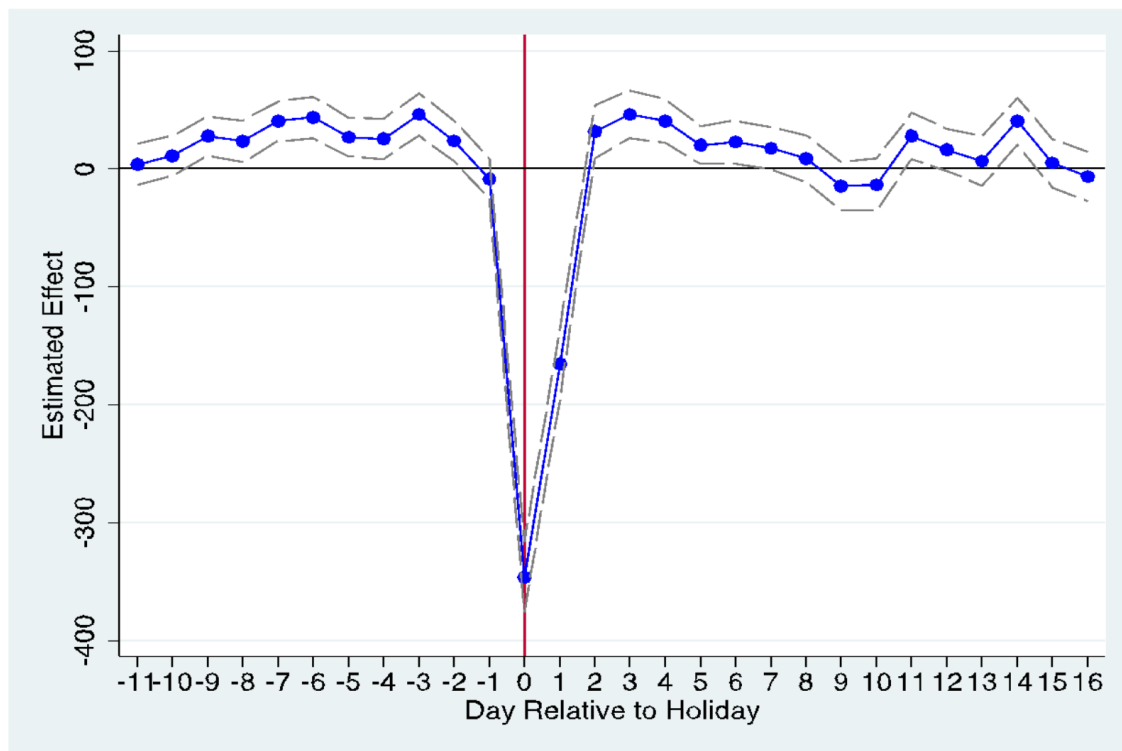


B. Estimated Manipulation Window



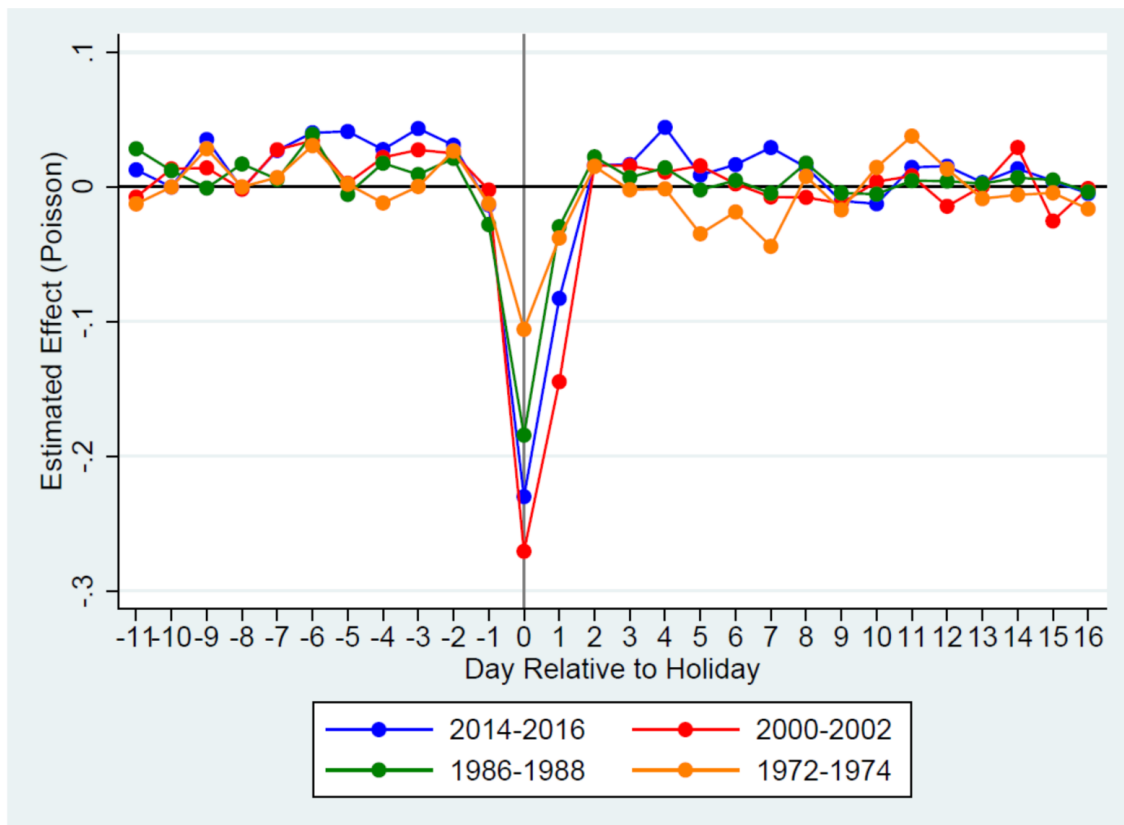
This figure illustrates how days in the manipulation window are matched with days of the same day of the week in the control region. Panel (A) illustrates this for a hypothetical example for which the estimated window is $[-3,+3]$ and all holidays fall on a Thursday. Since the control region before and after the holiday must include days that fall on the same day of the week, the control region for this hypothetical example contains days 7-10 before and days 8-10 after a holiday. Panel (B) illustrates the matching procedure for the actual estimated manipulation window, i.e., $[-11,+16]$. As all days are matched with the closest days that fall on the same day of the week in the control region, the control group therefore spans 14-25 days before a holiday and 22-37 days after a holiday.

Figure 2.3: The Effect of a Holiday on the Daily Number of Births in California: 2000-2016



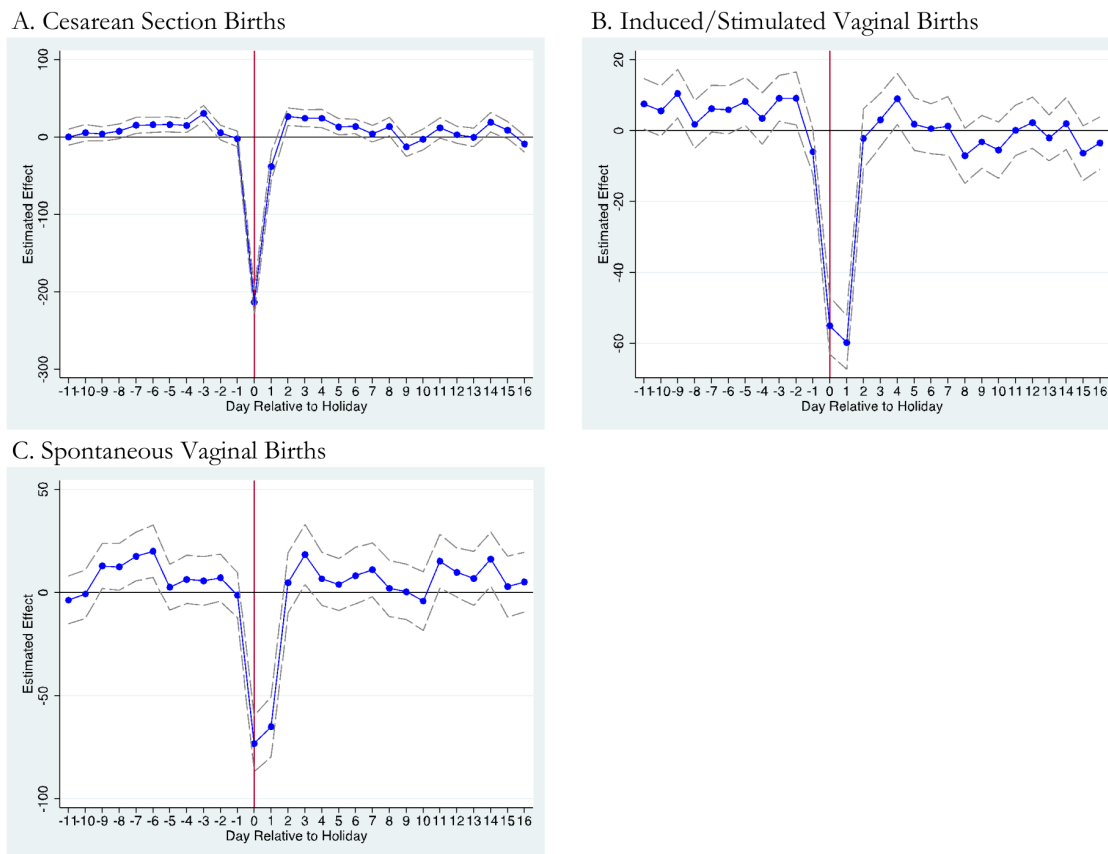
This figure shows the effect of a holiday on the daily number of births. Plotted are regression estimates from equation (2) with daily births as the dependent variable. On the x-axis is the day relative to the holiday (-1=day before holiday, 1=day after holiday, etc.). On the y-axis are estimates of the day relative to holiday dummy coefficients from equation (2) along with the 95% confidence interval.

Figure 2.4: The Effect of a Holiday on the Daily Number of Births in California over Time



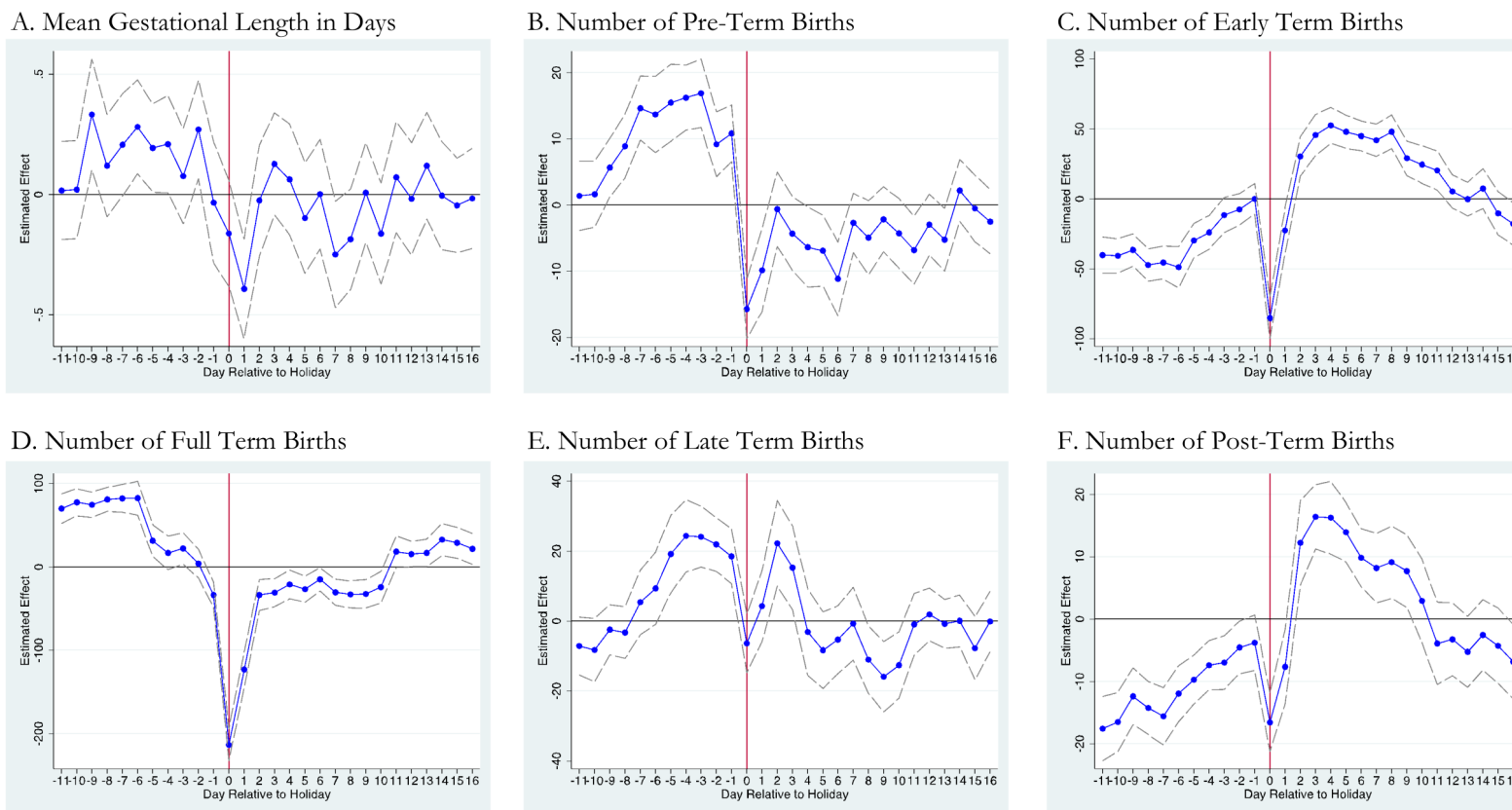
This figure shows the effect of a holiday on the daily number of births for different time periods (1972-1974, 1986-1988, 2000-2002, 2014-2016). Plotted are Poisson estimates of equation (2). On the x-axis is the day relative to the holiday (-1=day before holiday, 1=day after holiday, etc.). On the y-axis are estimates of the day relative to holiday dummy coefficients from equation (2). Since the publicly available data is based on a 50% sample 1972-1974, these observations were multiplied by two before conducting the analysis.

Figure 2.5: Shift in the Number of Births due to a Holiday by Delivery Type in California: 2000-2016



This figure shows the effect of a holiday on the daily number of cesarean section births (Panel A), induced/stimulated vaginal births (Panel B), and spontaneous vaginal births (Panel C). Plotted are regression estimates from equation (2). On the x-axis is the day relative to the holiday (-1=day before holiday, 1=day after holiday, etc.). On the y-axis are estimates of the day relative to holiday dummy coefficients from equation (2) along with the 95% confidence interval.

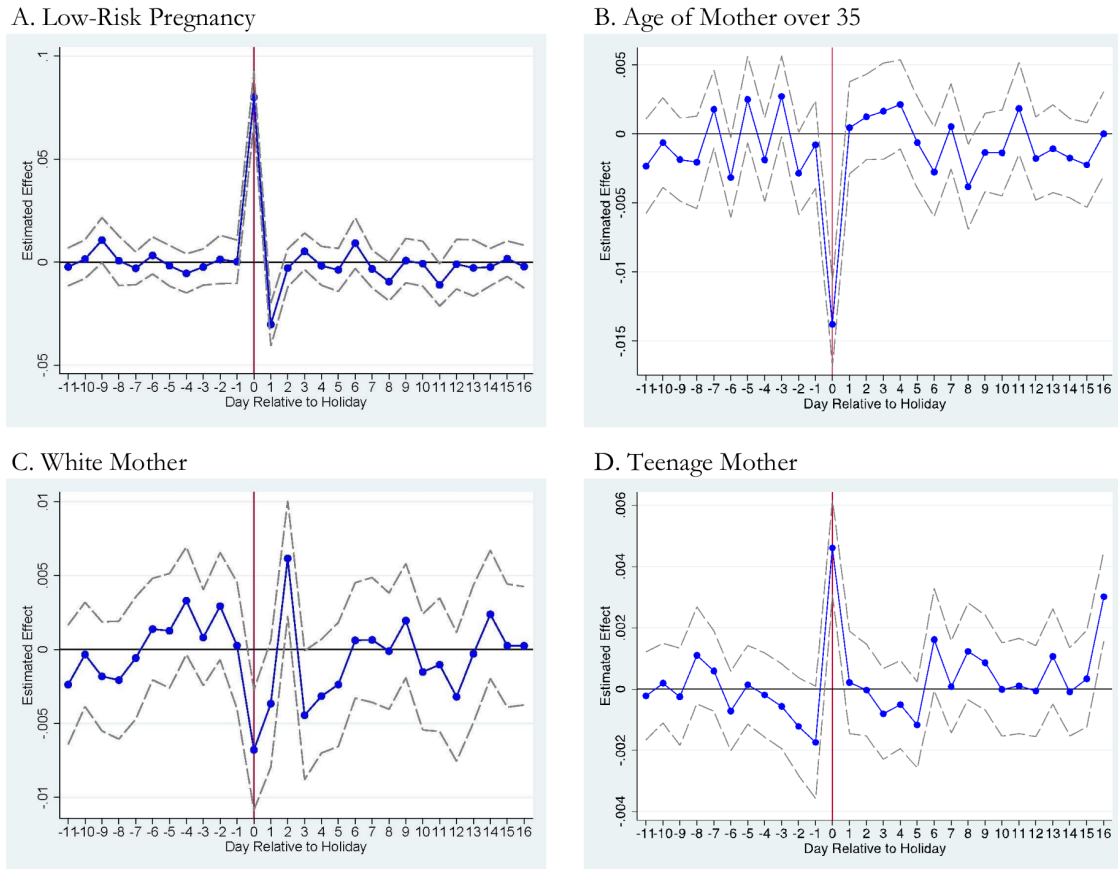
Figure 2.6: Shift in Births due to a Holiday by Gestational Length in California: 2000-2016



16

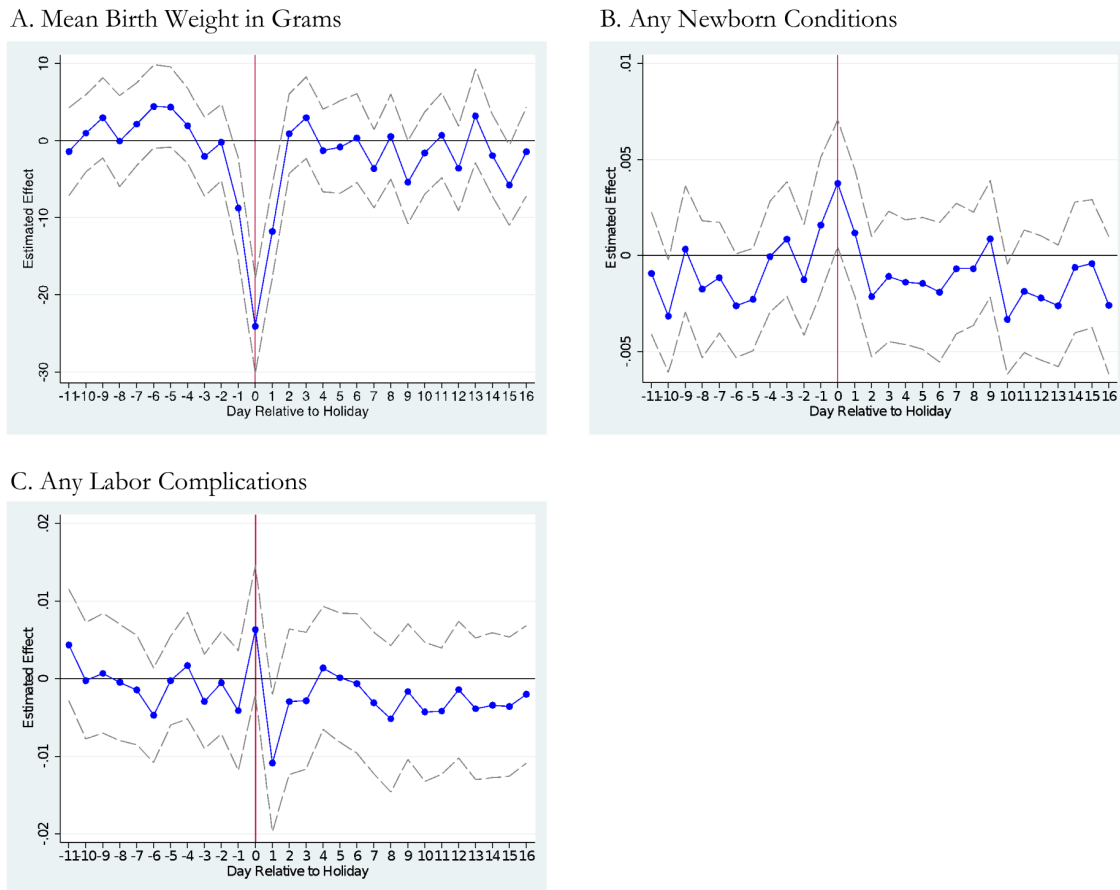
This figure shows the effect of a holiday on the mean gestation length (Panel A) as well as the number of daily births by term length category: pre-term meaning before 37 weeks (Panel B); early term or 37 0/7 weeks to 38 6/7 weeks (Panel C); full term or 39 0/7 weeks to 40 6/7 weeks (Panel D), late term or 41 0/7 weeks to 41 6/7 weeks (Panel E); and post-term or 42 weeks and later (Panel F). Plotted are regression estimates from equation (2). On the x-axis is the day relative to the holiday (-1=day before holiday, 1=day after holiday, etc.). On the y-axis are estimates of the day relative to holiday dummy coefficients from equation (2) along with the 95% confidence interval.

Figure 2.7: Shift in the Fraction of Births due to a Holiday by Maternal Characteristics in California: 2000-2016



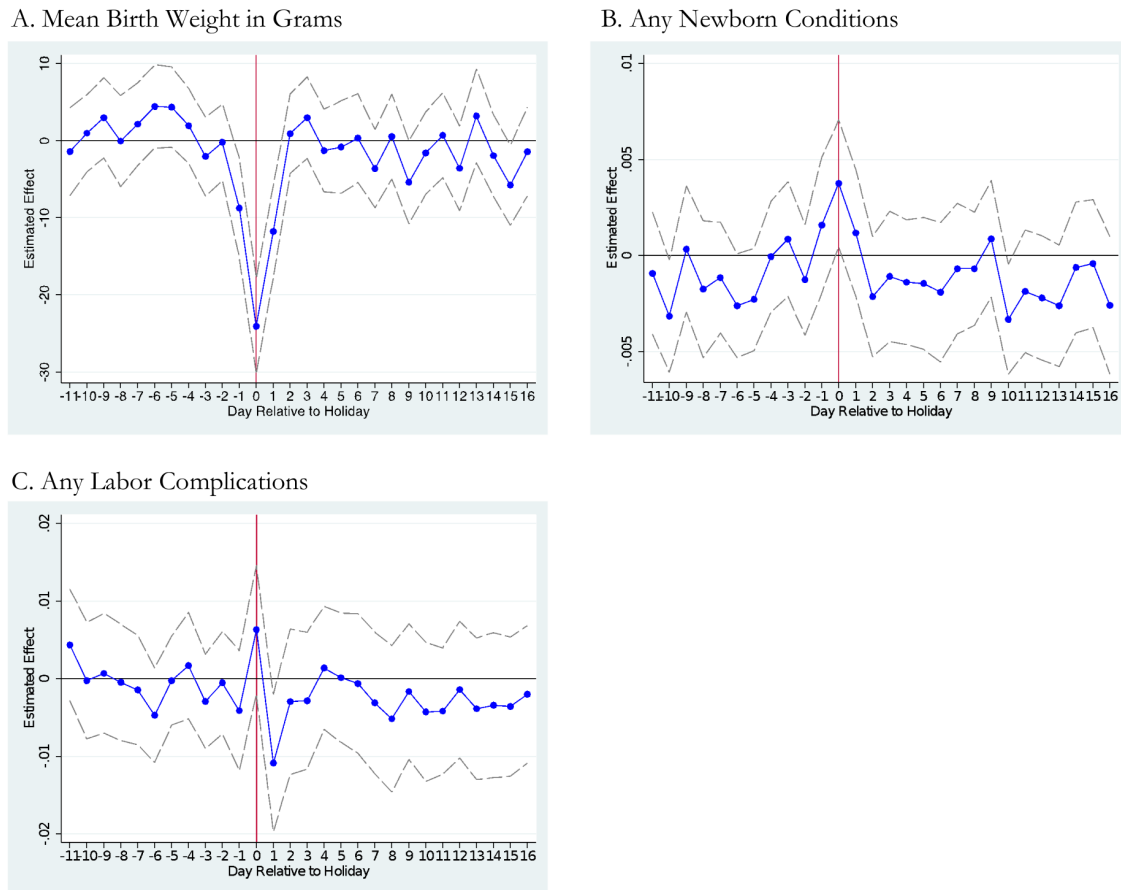
This figure shows the effect of a holiday on the fraction of births that are low risk (Panel A), fraction of births born to a mother over 35 (Panel B), fraction of births to a white mother (Panel C), and fraction of births to a teenage mom (Panel D). Plotted are regression estimates from equation (2). On the x-axis is the day relative to the holiday (-1=day before holiday, 1=day after holiday, etc.). On the y-axis are estimates of the day relative to holiday dummy coefficients from equation (2) along with the 95% confidence interval.

Figure 2.8: Birth Outcomes Contrasted with Counterfactual by Day Relative to Holiday in California: 2000-2016



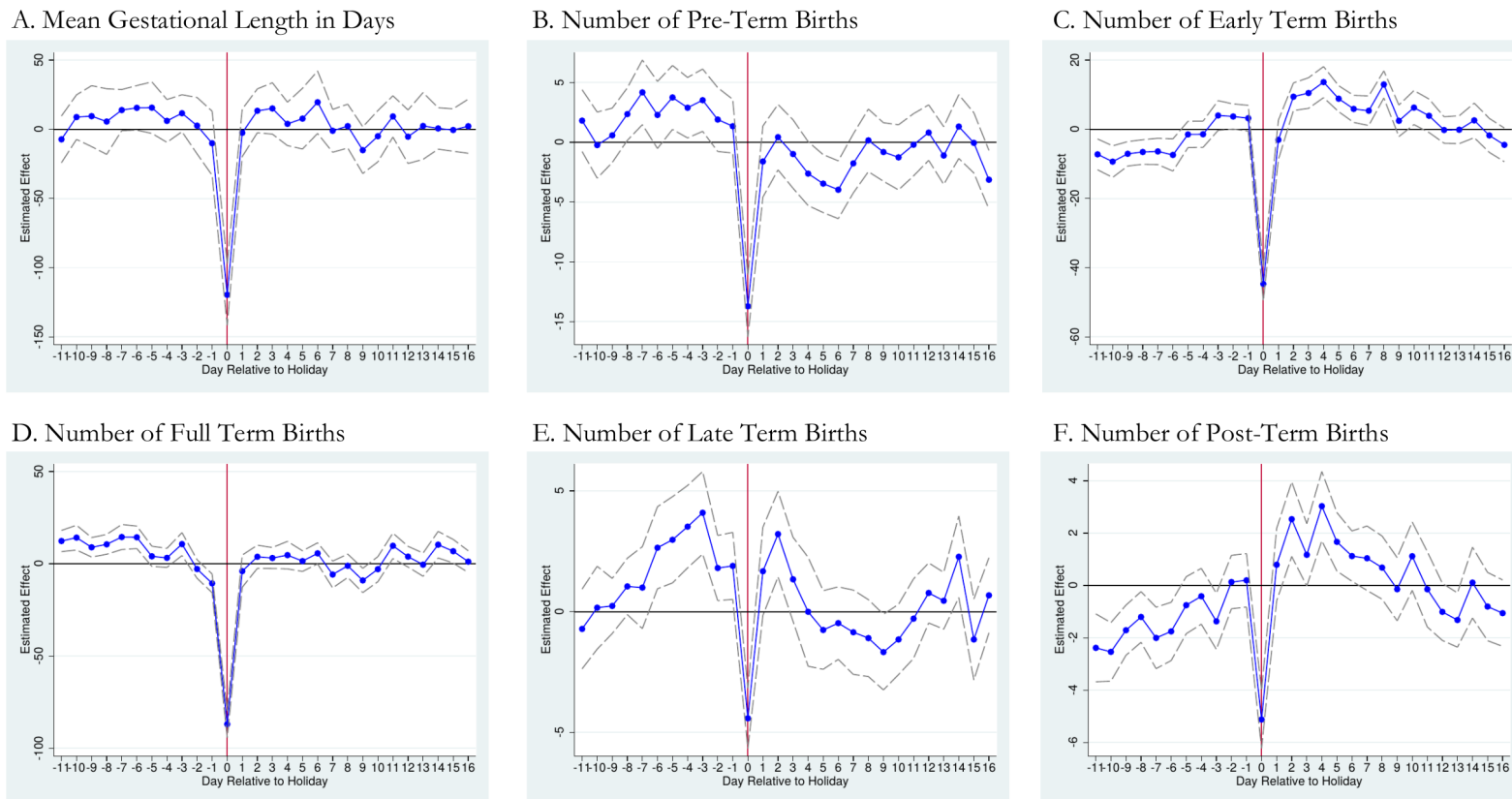
This figure shows the effect of a holiday on mean birth weight (Panel A), fraction of births with a noted newborn condition (Panel B) and the fraction of births with a noted labor complication (Panel C). Plotted are regression estimates from equation (2). On the x-axis is the day relative to the holiday (-1=day before holiday, 1=day after holiday, etc.). On the y-axis are estimates of the day relative to holiday dummy coefficients from equation (2) along with the 95% confidence interval.

Figure 2.9: Shift in the Number of Births due to a Holiday for High-Risk Pregnancies in California: 2000-2016



This figure shows the effect of a holiday for high-risk pregnancies on the daily number of births (Panel A) and the number of cesarean section deliveries (Panel B). Plotted are regression estimates from equation (2). On the x-axis is the day relative to the holiday (-1=day before holiday, 1=day after holiday, etc.). On the y-axis are estimates of the day relative to holiday dummies from equation (2) along with the 95% confidence interval.

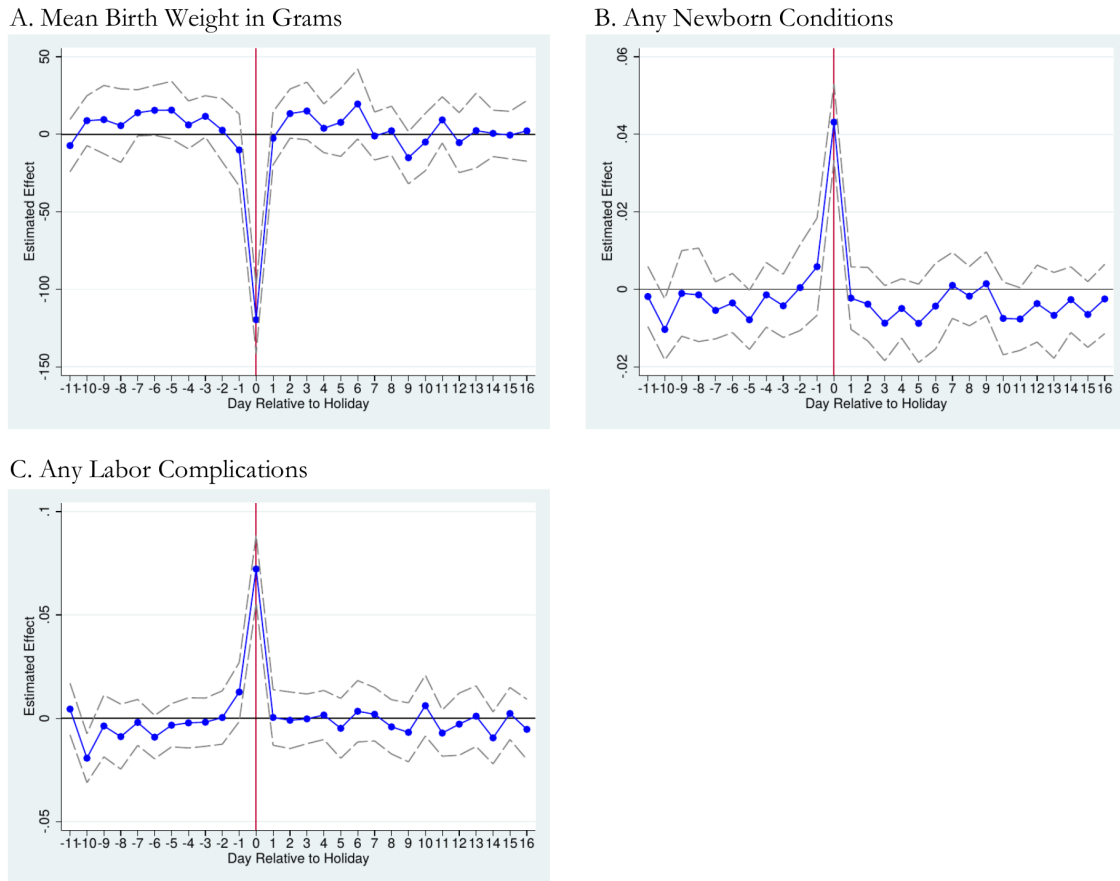
Figure 2.10: Shift in the Number of Births due to a Holiday by Gestational Length for High-Risk Pregnancies in California: 2000-2016



95

This figure shows the effect of a holiday on the mean gestation length (Panel A) as well as the number of daily births by term length category: pre-term meaning before 37 weeks (Panel B); early term or 37 0/7 weeks to 38 6/7 weeks (Panel C); full term or 39 0/7 weeks to 40 6/7 weeks (Panel D), late term or 41 0/7 weeks to 41 6/7 weeks (Panel E); and post-term or 42 weeks and later (Panel F) for high-risk pregnancies. Plotted are regression estimates from equation (2). On the x-axis is the day relative to the holiday (-1=day before holiday, 1=day after holiday, etc.). On the y-axis are estimates of the day relative to holiday dummy coefficients from equation (2) along with the 95% confidence interval.

Figure 2.11: Birth Outcomes Contrasted with Counterfactual by Day Relative to Holiday for High-Risk Births in California: 2000-2016



This figure shows the effect of a holiday on mean birth weight (Panel A), fraction of births with a noted newborn condition (Panel B), and the fraction of births with noted labor complication (Panel C) for high-risk pregnancies. Plotted are regression estimates from equation (2). On the x-axis is the day relative to the holiday (-1=day before holiday, 1=day after holiday, etc.). On the y-axis are estimates of the day relative to holiday dummy coefficients from equation (2) along with the 95% confidence interval.

Table 2.1: Mean Daily Births Overall and by Delivery Mode in California: 2000-2016

Type of Days	Obs (Number of Days)	Total Births	Cesarean Section	Spontaneous Vaginal	Induced/ Stimulated Vaginal
All	6210	1442 (229)	442 (130)	707 (126)	293 (63)
Holiday	119	1118 (126)	272 (62)	617 (101)	229 (53)
Weekend	1761	1150 (106)	267 (39)	642 (96)	241 (49)
Other	4330	1570 (128)	518 (67)	736 (127)	316 (54)
7-Holiday Analytic Sample	4599	1444 (233)	442 (132)	709 (127)	293 (64)
Holiday	119	1118 (126)	272 (62)	617 (101)	229 (53)
Weekend	1300	1153 (106)	267 (40)	644 (96)	241 (49)
Other	3180	1575 (131)	520 (68)	738 (128)	316 (55)
5-Holiday Analytic Sample	4256	1447 (233)	443 (132)	709 (127)	294 (64)
Holiday	85	1164 (112)	291 (59)	633 (102)	240 (51)
Weekend	1220	1155 (107)	268 (40)	645 (96)	242 (49)
Other	2951	1576 (139)	520 (72)	738 (128)	317 (55)

Standard deviations are in parentheses. All days uses the full set of births in California from 2000-2016, and considers the same set as the 7-Holiday Analytic Sample as holidays. The analytic samples include the holiday interval (11 days before until 16 days after a holiday) as well as the control group (14-25 days before a holiday, and 22-37 days after a holiday). The 7-Holiday Analytic Sample considers New Year's, President's Day, Memorial Day, Independence Day, Labor Day, Thanksgiving, and Christmas Day as holidays. The 5-Holiday Analytic Sample considers President's Day, Memorial Day, Independence Day, Labor Day, and Thanksgiving as holidays. Other days are non-holiday, non-weekend days. Means for births classified as "delivery type unknown" are not shown here. The number of observations are the number of days by day type, multiplied by years in the sample (e.g., 7 holidays times 17 years yields 119 observations).

Table 2.2: Aggregate Effect of the Holiday Period on Births and Delivery Types

	Total Births	Cesarean Section Births	Spontaneous Vaginal Births	Induced/ Stimulated Vaginal Births
<i>Panel A: Without Christmas and New Year's</i>				
Holiday Interval	-4.07 (2.99)	-2.42 (1.72)	1.05 (1.84)	-2.71*** (0.98)
Daily Mean Births	1447	443	709	294
Number of Observations	4256	4256	4256	4256
<i>Panel B: Including Christmas and New Year's</i>				
Holiday interval	1.23 (2.86)	-1.55 (1.63)	4.36*** (1.63)	-1.60* (0.90)
Daily Mean Births	1444	442	709	293
Number of Observations	4599	4599	4599	4599

Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The holiday interval is the period 11 days prior to a major holiday and 16 days after a major holiday. Panel A considers the following as major holidays: Presidents' Day, Memorial Day, Independence Day, Labor Day and Thanksgiving. The 4256 observations correspond to approximately 250.3 days over 17 years, Panel B adds Christmas and New Year's Day to the holiday list. The 4599 observations correspond to approximately 270.5 days over 17 years.

Table 2.3: Aggregate Effect of the Holiday Period on Term Length

	Number of Births by Term Length					
	Mean Gestational Age	Pre-Term: before 37 weeks	Early Term: 37 0/7 - 38 6/7 weeks	Full Term: 39 0/7- 40 6/7 weeks	Late Term: 41 0/7- 41 6/7 weeks	Post-Term: 42 weeks or later
<i>Panel A: Without Christmas and New Year's</i>						
Holiday Interval	0.03 (0.03)	1.52** (0.68)	-5.62*** (1.82)	0.38 (2.76)	3.58*** (1.32)	-3.32*** (0.74)
Daily Mean of Outcome	275	141	360	679	138	80
Number of Observations	4256	4256	4256	4256	4256	4256
<i>Panel B: Including Christmas and New Year's</i>						
Holiday Interval	0.02 (0.02)	0.36 (0.62)	1.92 (1.75)	-1.54 (2.50)	1.60 (1.23)	-0.37 (0.68)
Daily Mean of Outcome	275	141	360	677	137	81
Number of Observations	4599	4599	4599	4599	4599	4599

Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The holiday interval is the period of time covering the period 11 days prior to a major holiday through 16 days after a major holiday. Panel A considers the following as major holidays: Presidents' Day, Memorial Day, Independence Day, Labor Day and Thanksgiving. The 4256 observations correspond to approximately 250.3 days over 17 years, Panel B adds Christmas and New Year's Day to the holiday list. The 4599 observations correspond to approximately 270.5 days over 17 years.

Table 2.4: Effects on Number of Births Classified by the Characteristics of Moms over the Holiday Interval

	Low-Risk Births	Low-Risk First Births	Moms over Age 35	Teenage Moms	White Moms	Moms with High School Degree or Less	Deliveries with Private Insurance Payment	Deliveries with Public Insurance Payment
<i>Panel A: Without Christmas and New Year's</i>								
Holiday Interval	-0.76 (0.65)	-0.48 (0.54)	-2.10*** (0.75)	0.23 (0.28)	-3.99 (2.57)	1.52 (2.30)	-8.49*** (1.81)	5.22*** (1.67)
Daily Mean of Outcome	121	93	205	39	1113	699	701	689
Number of Observations	2756	2756	4256	4256	4256	3502	4256	4256
<i>Panel B: Including Christmas and New Year's</i>								
Holiday Interval	-0.61 (0.60)	-0.47 (0.50)	-1.44** (0.70)	0.00 (0.25)	0.85 (2.43)	1.47 (2.08)	-2.13 (1.69)	4.30*** (1.58)
Daily Mean of Outcome	120	93	204	39	1111	700	699	688
Number of Observations	2974	2974	4599	4599	4599	3790	4599	4599

Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The holiday interval is the period of time covering the period 11 days prior to a major holiday through 16 days after a major holiday. Panel A considers the following as major holidays: Presidents' Day, Memorial Day, Independence Day, Labor Day and Thanksgiving. The 4256 observations correspond to approximately 250.3 days over 17 years, Panel B adds Christmas and New Year's Day to the holiday list. The 4599 observations correspond to approximately 270.5 days over 17 years. Other payment categories (self-paid and other) are excluded for brevity. This category comprises less than 3 percent of deliveries and is unchanged over the holiday period. The sample size for low-risk births, overall or of first births, is lower because these variables cannot be constructed prior to 2006, when characteristics such as intrauterine growth restrictions were not captured in the California birth data. Similarly, the sample size for moms with a high school degree or less is smaller as the education variable is not available 2003-2005.

Table 2.5: Aggregate Effect of the Holiday Period on Birth Outcomes

	Mean Birth Weight	Low Birth Weight	Any Newborn Conditions	Any Labor Complications	Low Apgar Score
<i>Panel A: Without Christmas and New Year's</i>					
Holiday Interval	-2.00*** (0.70)	0.40 (0.29)	-0.88** (0.40)	-1.56 (1.02)	-0.49 (0.35)
Daily Mean of Outcome	3309	0.07	0.09	0.46	0.06
Fraction of Births Manipulated	0.013	0.013	0.013	0.013	0.013
Implied IV Estimate	-159	0.032	-0.070	-0.124	-0.039
Number of Observations	4256	4256	4256	4256	2506
<i>Panel B: Including Christmas and New Year's</i>					
Holiday Interval	-0.66 (0.63)	0.17 (0.27)	-1.03*** (0.36)	-2.95*** (0.88)	-0.44 (0.31)
Daily Mean of Outcome	3309	0.07	0.09	0.46	0.06
Fraction of Births Manipulated	0.014	0.014	0.014	0.014	0.014
Implied IV Estimate	-48	0.012	-0.075	-0.215	-0.032
Number of Observations	4599	4599	4599	4599	2700

Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The holiday interval is the period of time covering the period 11 days prior to a major holiday through 16 days after a major holiday. Panel A considers the following as major holidays: Presidents' Day, Memorial Day, Independence Day, Labor Day and Thanksgiving.. The 4256 observations correspond to approximately 250.3 days over 17 years, Panel B adds Christmas and New Year's Day to the holiday list. The 4599 observations correspond to approximately 270.5 days over 17 years. The fraction of births manipulated is calculated as the effect of the holiday on births on the day of the holiday and the day after the holiday divided by the total number of births in the 28-day manipulation window period. The implied IV estimate is the ratio of the holiday interval effect in the table divided by the fraction of births manipulated, 0.0126. The sample size for the share of babies with a low Apgar score is lower, as this variable is not available prior to 2007. Estimates and standard errors are multiplied by 1000, except for the mean birth weight outcome.

Table 2.6: Aggregate Effect of the Holiday Period on Births and Delivery Type among High-Risk Pregnancies

	Total Births	Cesarean Section Births	Spontaneous Vaginal Births	Induced/ Stimulated Vaginal Births
Holiday Interval	-1.55 (1.27)	-1.28 (1.25)	-0.17 (0.11)	-0.10 (0.07)
Daily Mean of Outcome	259	248	8	3
Number of Observations	4256	4256	4256	4256

Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The holiday interval is the period 11 days prior to a major holiday and 16 days after a major holiday.

Table 2.7: Aggregate Effect of the Holiday Period on Term Length among High-Risk Pregnancies

	Number of Births by Term Length					
	Mean Gestational Age	Pre-Term: before 37 weeks	Early Term: 37 0/7 - 38 6/7 weeks	Full Term: 39 0/7- 40 6/7 weeks	Late Term: 41 0/7- 41 6/7 weeks	Post-Term: 42 weeks or later
Holiday Interval	0.06 (0.08)	-0.33 (0.34)	-1.09 (0.59)	-0.27 (0.90)	0.69** (0.22)	-0.46*** (0.16)
Daily Mean of Outcome	269	42	81	106	13	10
Number of Observations	4256	4256	4256	4256	4256	4256

Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The holiday interval is the period 11 days prior to a major holiday and 16 days after a major holiday.

Table 2.8: Characteristics of Moms over the Holiday Interval among Women with High-Risk Pregnancies

	Moms over Age 35	Teenage Moms	White Moms	Moms with High School Degree or Less	Deliveries with Private Insurance Payment	Deliveries with Public Insurance Payment
Holiday Interval	-0.56 (0.42)	0.01 (0.05)	-1.26 (1.016)	0.58 (0.76)	-1.96*** (0.71)	0.62 (0.73)
Daily Mean of Outcome	57	2	199	124	127	122
Number of Observations	4256	4256	4256	3502	4256	4256

Robust standard errors are in parentheses. The holiday interval is the period of time covering the period 11 days prior to a major holiday through 16 days after a major holiday. The sample size of moms with a high school degree or less is lower, as the education variable is not available in the data between 2003-2005.

Table 2.9: Aggregate Effect of the Holiday Period on Birth Outcomes among High-Risk Pregnancies

	Mean Birth Weight	Low Birth Weight	Any Newborn Conditions	Any Labor Complications	Low Apgar Score
Holiday Interval	-1.32 (2.34)	-0.05 (1.20)	-1.38 (1.21)	1.63 (1.71)	0.12 (1.02)
Daily Mean of Outcome	3187	0.14	0.15	0.45	0.07
Fraction of Births Manipulated	0.022	0.022	0.022	0.022	0.022
Implied IV Estimate	-60.03	-0.002	-0.063	0.074	0.005
Number of Observations	4256	4256	4256	4256	2506

Robust standard errors are in parentheses. The holiday interval is the period of time covering the period 11 days prior to a major holiday through 16 days after a major holiday. The 4256 observations correspond to approximately 250.3 days over 17 years. The sample size for the share of babies with a low Apgar score is lower, as this variable is not available prior to 2007. The fraction of births manipulated is calculated as the effect of the holiday on births on the day of the holiday, divided by the total number of births in the 28-day manipulation window period. The implied IV estimate is the ratio of the holiday interval effect in the table divided by the fraction of births manipulated. Estimates and standard errors are multiplied by 1000, except for the mean birth weight outcome.

Chapter 3

Knowing Me, Knowing You: An Experiment on Mutual Payoff Information and Strategic Uncertainty

3.1 Introduction

Game-theoretic models are typically motivated by the idea that players reason about the behavior of others and choose their strategies accordingly. This reasoning can be informed directly by observing the payoff structure of the game or indirectly by observing and learning from the actions of other players. The type of introspective reasoning supported by directly observing others' payoffs is often embedded in models of strategic decision making, such as higher-level reasoning in level- k models (Stahl and Wilson, 1994; Nagel, 1995), which assume mutual payoff information. Additionally, experiments using eye-tracking have found that subjects devote a sizable amount of attention to the payoffs

of other players (Knoepfle et al., 2009; Polonio and Coricelli, 2019), and it has been documented that subjects engage in higher-level reasoning when other players' payoffs can be observed (e.g., Kneeland, 2015). In contrast, in providing an interpretation for his seminal equilibrium concept, Nash (1950) makes it explicit that, "it is unnecessary to assume that the participants have full knowledge of the total structure of the game, or the ability and inclination to go through any complex reasoning processes." Similarly, theoretical models of learning explore how equilibria can be reached and selected through processes of learning, adaptation, and/or imitation rather than introspection (Fudenberg and Levine, 2009), and uncoupled learning models (e.g., Hart and Mas-Colell, 2006; Foster and Young, 2006; Young, 2009; Babichenko, 2010) describe how equilibria can be reached in the absence of information about other players' incentives or even their existence.

It is not well understood, however, how mutual payoff information—in addition to observing others' actions—affects players' behavior in strategic interactions. On the one hand, knowledge of payoffs can reveal opportunities to coordinate on Pareto-dominant outcomes; on the other hand, being aware of an opportunity to cooperate can increase the strategic tension of a game if the cooperative outcomes are associated with actions that are dominated for at least one player. Furthermore, it is not clear through which channels such an effect would operate. Mutual payoff information may affect players' initial perceptions of a game, the process through which they learn and gain experience, or both.

We present an experiment designed to study how mutual payoff information affects play in the Prisoner's Dilemma (PD) and the Stag Hunt (SH). Subjects are asked to play these games repeatedly with random re-matching of opponents each round. In our partial-information treatment, subjects observe their own payoffs and the action of their opponent after each round, but they never observe the other's payoffs. Comparing this

partial-information version to the full-information baseline treatment in which subjects additionally observe the payoffs of the other player allows us to detect differences in play that arise due to the presence or lack of mutual payoff information.

The appeal of contrasting the PD with the SH in our experiment is grounded in our conjecture that mutual payoff information affects behavior differently in these two games. While the SH exhibits a tension between a mutually desirable outcome and avoiding personal risk, the PD exhibits a tension between a socially optimal outcome and personal gain. In the SH, knowledge of the other's payoffs arguably reduces the strategic tension of the stage game by revealing that a mutually beneficial outcome exists. In the PD, on the other hand, mutual payoff information may increase the strategic tension of the stage game. Without knowing the other's payoffs, there is little reason not to choose the payoff-dominant action. Indeed in this circumstance there is arguably no "dilemma," as subjects are unaware of the socially optimal, albeit dominated, cooperative action pair.¹ By introducing mutual payoff information, however, players observe a socially optimal outcome which can only be reached at personal expense. In the presence of social preferences, either player might prefer such an outcome but still wonder whether the other player shares these preferences and will reciprocate, thereby potentially making one's own attempt to cooperate futile. This uncertainty may increase the strategic tension of the PD game.

Our main result is that the presence or absence of mutual payoff information has a strong effect on play in both games. The fraction of subjects who initially cooperate in the PD or who coordinate on the payoff-dominant equilibrium in the SH is substantially higher under full-information than under partial-information. In the SH, this effect is remarkably persistent: throughout all rounds of the game, the vast majority of sub-

¹For consistency with prior literature, we use the words "cooperate" and "defect" to describe the action choices in the PD with acknowledgement that they are only well-defined from the player's perspective in the presence of mutual payoff information.

jects choose the action consistent with the payoff-dominant equilibrium of the SH in the full-information treatment, while choosing the risk-dominant action under partial-information.² By contrast, in the PD play converges toward the unique NE of the game under both information treatments.

To investigate the channels driving this information treatment effect, we estimate a belief-learning model that is a special case of an Experience-Weighted Attraction (EWA) model (Camerer and Ho, 1999). In this model, players choose the action with the higher expected payoff (higher attraction) in each round, given their beliefs about the actions of other players. Beliefs are formed as weighted averages of the observed history of play and a prior belief. We estimate a set of four parameters for each game and information treatment: two parameters for the initial attractions (one for each strategy), one parameter related to how subjects weight past observations when updating beliefs, and one parameter governing how precisely subjects respond to their estimated beliefs (i.e. to the updated attractions they hold for each action).

To examine the importance of these parameters, we conduct simulations where we flip the estimated parameters at the information treatment level to isolate the effect associated with initial play versus learning. We find that doing so results in simulated play that sharply diverges from what we actually observe, suggesting that our information treatment creates significant differences both in initial play and also in learning across rounds.

We propose that our results may be explained, at least in part, by the impact mutual payoff information has on strategic uncertainty and structural uncertainty and the relationship these have to the strategic tensions inherent in each game. While strate-

²Strictly speaking, following Harsanyi and Selten (1988)'s canonical definition of risk dominance an equilibrium cannot be "risk-dominant" for a player who does not have access to full payoff information. For the sake of clarity and consistency, we will refer to the "risk-dominant" and "payoff-dominant" actions for both the full- and partial-information treatments in the SH game.

gic uncertainty reflects a player's uncertainty about which strategies the other players will choose, structural uncertainty is the uncertainty about the parameters of a game (Brandenburger, 1996). Removing mutual payoff information mechanically introduces structural uncertainty to both games, which we argue increases strategic uncertainty in the PD while decreasing it in the SH, and affects subject behavior in ways that are captured by changes in the parameters of our learning model.

First, for subjects who face greater strategic uncertainty, experimenting may be relatively more attractive. Indeed, we find that subjects respond less precisely in the SH partial-information treatment than in the full-information counterpart, while the opposite is true of the PD. Additionally, when faced with greater strategic uncertainty, subjects may place greater weight on recent observations as opposed to their initial perception of the game. This is reflected in our estimate of the belief updating parameter in the SH, which suggests that subjects place greater weight on recent observations under the partial-information treatment. While the effect is not significant in the PD, the direction of our estimate is consistent with this interpretation.

Our paper makes several contributions. First, we provide evidence that even in the absence of mutual payoff information most subjects eventually choose actions that correspond to Nash equilibria. Second, we show, to our knowledge for the first time, that mutual payoff information can affect equilibrium selection in the SH throughout all rounds and initial play in the PD. These differences in play are associated with greater coordination on the payoff-dominant equilibrium across all rounds of play for the SH and greater cooperation in the short-run for the PD. Finally, while a number of previous experiments (e.g., Mookherjee and Sopher, 1994; Cox et al., 2001; Feltovich and Oda, 2014) employ learning models in their analysis, the goal has been to evaluate such models in isolation or relative to one another. We take a different approach and instead interpret our learning model parameters as indicative of underlying changes in the way players perceive,

update, and react to the availability of opponent payoff information. We propose that these changes reflect how strategic and structural uncertainty vary across our information treatments.

Related experiments. A closely related collection of experiments studies how subjects behave in incomplete information environments similar to ours. We are not aware of an experiment, however, that employs the same information treatments we use to the PD and the SH game. Mookherjee and Sopher (1994) find that in a matching pennies game with fixed partners, choice frequencies tend towards the unique mixed strategy equilibrium whether or not subjects are presented with opponent payoff information. Oechssler and Schipper (2003) have subjects play incomplete information treatments in the SH and PD; however, our experiments are difficult to compare as their design creates incentives for experimentation in the initial rounds by paying subjects more later and by providing rewards for correct answers regarding the opponents' payoffs. Perhaps most notable for our experiment, subjects' play converges towards Nash equilibrium despite the inability of subjects to fully perceive the game structure.

Most similar to our design is Feltovich and Oda (2014) in which subjects play incomplete information versions of six games, including the SH and PD, with treatments for random re-matching and fixed matching. Their results suggest that the matching mechanism does matter in the incomplete information environment, with fixed-pair often leading to increased coordination on pure strategy equilibria, higher payoffs, and faster convergence, but no full-information treatments are run for comparison. McKelvey and Palfrey (2001) run an ambitious number of games and information treatments, including versions of the PD and SH, but do not report choice frequencies for direct comparison to our results.³

³Additional studies have altered the information structure in ways that make them less comparable to our environment. For example, Cox et al. (2001) and Danz et al. (2012) inform subjects about a set

Our paper also relates to a literature examining the impact of payoff information on the formation of cooperation. In both Friedman et al. (2015) and Huck et al. (2017), subjects play Cournot games that exhibit tension similar to our PD between competition and cooperation. Subjects do not have access to their own or others' payoff function but are told payoff functions are symmetric and time invariant, and they receive feedback on their own and others' payoffs and actions. Huck et al. (2017) introduce a comparison treatment where subjects are shown the possible payoffs they could have received based on their partners last action. Contrary to our results, in these games payoff information tends to lead to play that is more competitive, suggesting that payoff information hinders cooperation. Nax et al. (2016) study cooperation in a voluntary contribution game under different information structures, and find that contribution rates are similar when players have full information about the game and when they only get to observe their own payoffs.

Another strand of literature we contribute to addresses the question of which equilibrium play converges to in the SH. With one payoff-dominant and one risk-dominant equilibrium, the SH embodies a tradeoff between maximizing social efficiency and minimizing personal risk. Many experiments (e.g., Battalio et al., 2001; Schmidt et al., 2003; Dubois et al., 2012; Kendall, 2020) have been designed to better understand the conditions under which play converges to either equilibrium. A common feature of these studies is that subjects play SH games where payoffs are commonly known, and by varying these payoffs across games, diverse theoretical predictions can be disentangled. Our paper is related to this literature but takes a different approach; we keep payoffs constant across treatments but vary whether players get to observe the other's payoffs. Doing so allows

of payoffs from which the actual opponent payoffs may be drawn, Nicklisch (2011) introduce information asymmetries into the environment, and Nikolaychuk (2012) match subjects with a computer following a learning algorithm and let them observe their own earnings after each round, or the whole payoff matrix, in versions of the PD, SH, and Battle of the Sexes. Andreoni et al. (2007) vary the information that bidders have about their rivals' valuation in first- and second-price auctions, and document that subjects' behavior in response to this information is consistent with theoretical predictions.

us to identify mutual payoff information as an important factor for the payoff-dominant equilibrium to arise in the SH.

The remainder of this chapter is organized as follows. Section 3.2 provides details of our experimental design. Section 3.3 presents our main results in a descriptive manner. Section 3.4 introduces and estimates the belief learning model. Additionally, we present simulations to isolate the effect of initial play versus learning and to explore the robustness of our estimates. Section 3.5 provides a discussion and Section 3.6 concludes.

3.2 Experimental Design

Overview. In our experiment, subjects played SH and PD games in randomly re-matched pairs over many periods. We employed two information treatments per game: In full-information (“Full”) treatments, subjects were shown the complete payoff matrix, including own and opponent payoffs. In partial-information (“Partial”) treatments, on the other hand, we showed subjects only their own payoffs. In both treatments, players made choices simultaneously. At the end of each round we notified subjects of their opponent’s action, reminded them of their own action, and displayed their resulting payoff, but not their counterpart’s.

Games and information treatments. For each game and information treatment, Figure 3.1 depicts the payoff matrices, as well as the available payoff information from the row player’s perspective. In the SH, there are two pure-strategy equilibria. One is payoff-dominant (X, X) and the other one is risk-dominant (Y, Y) . In the PD, there is one equilibrium in strictly dominant strategies, (Y, Y) . Note that the Full and Partial treatments of each game have the same payoffs (though they are partially hidden from

subjects in the Partial treatment), thus keeping the best-response correspondences and equilibria constant across the two information treatments of a game.⁴

40 rounds with random re-matching. In each treatment, subjects played 40 rounds of a game and were randomly and anonymously re-matched with other subjects each round.⁵ The information treatment was the same for all subjects within a session. That is, in the Full treatment, it was common knowledge that subjects were being re-matched with other subjects who could observe the whole payoff matrix. Similarly, in the Partial treatment, it was common knowledge that subjects were being re-matched with other subjects who could only observe their own payoffs and that any opponent they would be matched with had the same payoffs as their previous opponents.

To credibly implement the Partial treatment, it was crucial that subjects were not able to infer that they were playing a symmetric game.⁶ We therefore employed a two-population matching mechanism, in which subjects were randomly divided into two groups (labeled A and B) at the beginning of the experiment and were exclusively matched with subjects of the opposite group throughout the session. Subjects were told that while they could not observe the payoffs of participants from the other group, all participants of the other group faced the same payoffs. Likewise, participants from the other group would always be matched with somebody of one's own group, facing the same payoffs as themselves. We used this two-population matching mechanism in both information treatments for consistency.

⁴This also means that the size of the basin of attraction for pure strategies X and Y —in our context the beliefs one would have to have that one's counterpart will select the opposite action in order for expected payoffs of one's own actions to be equal—are held constant between information treatments in the SH. Embrey et al. (2017) provide experimental evidence of a positive correlation between the size of the basin of attraction for a given strategy and the frequency with which subjects select this strategy.

⁵Ghidoni et al. (2019) find that cooperation rates in a PD game with 10 rounds are very similar when subjects are randomly re-matched in groups of 6 or with a new opponent each round.

⁶This would enable subjects to determine their opponent's payoffs, thereby undermining the information treatment.

Feedback after each round. After selecting an action in each round, subjects learned what action their opponent had chosen in that round and their resulting payoff. Throughout the 40 rounds of the game, we showed subjects a table on the left side of the screen with a list of their choices, their counterparts' choices, and their own payoffs for the current and previous rounds of the game. Figure 3.2 contains example screenshots depicting the information shown to subjects, before, during, and after selection of an action in the PD Partial treatment.⁷

Comprehension. The experimenters handed out written instructions, which they also read aloud to subjects. To participate, subjects had to correctly answer a comprehension quiz on paper.⁸ The instructions and the comprehension questions we used are provided in Appendices C.1.1 and C.1.2.

We explained to subjects how to read a payoff matrix based on an example of an asymmetric game with all payoffs visible. In sessions that included a Partial treatment, we then told subjects that in part of the experiment the other's payoffs would be covered. This was illustrated by using the same example game matrix, this time with gray squares covering the other's payoffs in each cell, hereby mimicking the interface used in the experiment. Since all subjects in a session were given the same instructions, it was implicitly communicated that all subjects would only see their own payoffs in the Partial treatment and thus their opponent (who would also only see their own payoffs) would not know the subject's payoffs.

⁷See Appendix C.1.3 for corresponding screenshots from experiments conducted under the SH Full treatment.

⁸Only 5 of 196 subjects failed to answer all quiz questions correctly the first time. We pointed out to these subjects what they did wrong and provided them with a second quiz version with different matrix entries; each of these 5 subject correctly answered all of these questions on the second attempt.

Organization of treatments, between versus within analysis. Each experimental session consisted of two blocks of 40 rounds each, one with a Full treatment and the other with a Partial treatment, for a total of 80 rounds of play. Before each session began we administered the relevant instructions followed by a comprehension quiz. We did not inform subjects of details of the second game until the first game was completed.

In sessions 1-6 subjects first played a Partial treatment of a game (SH in sessions 1-3, and PD in sessions 4-6), followed by the Full treatment of the same game. This design feature thus enables a within-subjects analysis. In sessions 7-12 subjects first played a Full treatment of one game (SH in sessions 7-9 and PD in sessions 10-12), followed by a Partial treatment of the other game (PD and SH, respectively). This enables a between-subjects analysis and addresses potential order effects that could occur if subjects only played the Full treatment of a game after first playing its Partial counterpart.

The allocation of treatments for each part of the 12 experimental sessions we conducted and the number of subjects per session are depicted in Table 3.1. Note that subjects never got to play the Full version of a game before the Partial version of the same game to avoid them inferring that the payoffs in the second game were the same as in the first game.

Our organization of treatments enables us to conduct both between-subjects and within-subjects analyses, as summarized in Table 3.2. We center our analysis on the pooled data that uses all available data from both the first and second games of all sessions, either for the SH or the PD. In results not reported, we find that separately analyzing the between-subjects or within-subjects data yields qualitatively similar results.⁹

⁹This is consistent with Duffy and Fehr (2018), who find that the frequency of playing the action that is associated with the Pareto-efficient outcome in the PD or the SH does not depend on the order in which these two games are played.

Experimental details. We programmed the experimental interface using Z-Tree (Fischbacher, 2007), and conducted all sessions at the University of California, Santa Barbara’s Experimental and Behavioral Economics Laboratory (EBEL) in April and September of 2018. We recruited subjects for the experiment from the EBEL subject pool, using the Online Recruitment System for Economic Experiments (ORSEE) tool (Greiner, 2015b). A total of 194 subjects participated in the twelve separate experimental sessions described in Table 3.1. Subjects were between 18 and 68 years old with a median age of 20, and 16% of them indicated Economics as their major or intended major.

Experimental sessions lasted between 45 and 55 minutes. Subjects were paid their payoff from a randomly selected round of the total of 80 rounds, plus an additional \$7.00 show-up fee. The average total payment was \$13.22, while the minimum payment was \$8.00 and the maximum was \$20.00.

3.3 Descriptive Results

In this section, we describe the main results of the experiment. We use the pooled sample of data as detailed in Section 3.2 but note that our results are qualitatively similar when using alternative samples.

We first look at the impact of the information treatment on choosing action X in aggregate terms. Recall that action X is associated with the Pareto-efficient outcome in both games. Figure 3.3 shows how the average rate of choosing action X evolves in each game and information treatment. For the SH, there is a large difference in play with Partial versus Full across all rounds. For the PD, there is initially a large difference in the rate of playing action X in the earlier rounds that deteriorates by the final rounds.

To test the significance of these results, we estimate regressions of the following form:

$$Y_{i,r,s} = \alpha + \beta \text{PartialInfo} + \gamma C_s + u_{i,r,s}, \quad (3.1)$$

where $Y_{i,r,s}$ is a binary indicator for subject i choosing action X in round r of session s . The vector C_s is a set of dummy variables that flexibly controls for session size.¹⁰ The variable *PartialInfo* is an indicator and our main variable of interest. It equals one if the action choice is made under the Partial treatment and zero otherwise. Thus, the estimated coefficient $\hat{\beta}$ can be interpreted as the percentage point difference in the probability of choosing action X under the Partial treatment compared to the Full treatment. This regression equation is used to compare the session-size-adjusted outcome means across treatments, not to fully explain behavior of subjects, which will be investigated more carefully in Section 3.4. We employ an ordinary least squares regression to estimate equation 3.1 and cluster standard errors at the subject-session-level. For our main results, we present estimates of equation 3.1 in Table 3.3, with the results for the SH in panel a) and the PD in panel b).

Initial play. For Full, the fraction of subjects who initially choose action X is substantially larger than for Partial in both games. In column (2) of panel a) in Table 3.3, we can see the treatment effect in the first 10 rounds of play for the SH is a 66.9 percentage point (pp) reduction in the probability of playing action X . The control group selects action X about 88.1% of the time, so the effect in percent terms is an 80.1% reduction relative the control mean. In column (2) of panel b), we can see that for the PD there is a treatment effect of -30.4pp in the probability of selecting action X in the first 10 rounds of play. This result is about -70.2% relative to the control mean given the control

¹⁰We do not use session fixed-effects in our main specification since the estimate $\hat{\beta}$ would only exploit the within-subjects data. Results are qualitatively similar when using the within-subjects data.

group plays action X about 43.3% of the time.

The very first round of each game is special as subjects in the Partial treatment could not observe any previous actions of their opponents and thus have absolutely no information about their incentives. We find that the pattern emerging in the first 10 rounds is qualitatively similar to the very first round: in the SH-Full treatment, 86.5% of subjects chose action X in the first round, but only 32.0% chose that action in the SH-Partial treatment. For the PD, the corresponding values are 64.3% and 17.0%.

Result 3.1. *Mutual payoff information has a large effect on initial play in both games. Under full-information, the fraction of subjects choosing action X (the action supporting Pareto-superior outcomes) is substantially higher.*

Equilibrium selection and convergence. Next, we analyze how play evolves across the 40 rounds of a game. In the SH, the initial effect is remarkably persistent throughout the game. As can be seen from panel a) of Table 3.3, the vast majority of subjects choose action X in the SH-Full treatment and action Y in the SH-Partial treatment throughout all rounds of the experiment. In column (1) of panel a), the average treatment effect across all rounds is -67.6pp, which equates to a treatment effect of -80.1% relative to the control mean of 84.4%.

These results directly impact equilibrium and efficiency in the respective treatments for the SH. While these results are partly driven by the random mechanism through which pairs of subjects were matched each round, we present them here to show the stark impacts mutual payoff information has for each game. For Partial, subjects tend to reach the risk-dominant equilibrium, while for Full, subjects tend to reach the payoff-dominant equilibrium, as can be seen in Figure 3.4.

We estimate equation 3.1 using a binary indicator for reaching a pure strategy Nash

equilibrium, which for the SH is (X, X) or (Y, Y) , and report the results in Table 3.4.¹¹ It is clear from panel a) that there is only a small difference in the rate of reaching an equilibrium due to treatment. For the average across all 40 rounds of play in column (1), we can see that the control group reaches an equilibrium about 81.1% of the time and the treatment group only does so 10.0pp less often (cluster p-value of 0.001). However, since the payoff-dominant equilibrium is more efficient than the risk-dominant one, partial information leads to lower efficiency. We again estimate equation 3.1 but this time use an efficiency ratio as the outcome.¹² In panel a) of Table 3.5, the subjects under Partial experience lower efficiency compared to the subjects under Full. The average across all 40 rounds is a reduction of -0.396 of the ratio, or -45.8% relative to the control ratio of 0.864.

Result 3.2. *Throughout all rounds of play in the SH, for the Full treatment, the vast majority of subjects select action X, which corresponds to reaching the payoff-dominant Nash equilibrium, while in the Partial treatment, most subjects choose action Y, which corresponds to reaching the risk-dominant Nash equilibrium.*

In the PD, on the other hand, play converges toward the unique Nash equilibrium of the game under both information treatments. We define convergence as occurring at a given round of play at which at least 80% of subjects (averaged across all sessions) play the equilibrium action, Y, and at least 80% play action Y for the remaining rounds of the game. For the PD-Full, this occurs at round 24, and for the PD-Partial, this occurs at round 3. As evidenced by Figure 3.3, by the end of the game both treatments converge

¹¹Our SH game also has a mixed strategy Nash equilibrium where subjects choose action X two-thirds of the time and choose action Y one third of the time. In Table C.1 of Appendix C.2 we show the share of subjects in each of four 10-round periods whose mix of actions are within 10pp of $p_X = 0.667$. Subjects exhibiting such patterns of play are in the minority in all treatments and periods but are more common during earlier rounds and in the full information treatment.

¹²The efficiency ratio is the total payoffs of both subjects in a given round of play divided by the total payoffs of the efficient outcome. Naturally, the random re-matching of subjects will induce some variation here.

toward the deviating action Y . Panel b) of Table 3.3 quantifies how the choice of action X evolves across rounds in the PD. By the last 10 rounds, the treatment effect degrades to a difference of only -8.6pp between the Full and Partial treatments (cluster p-value of 0.001), and the control group also decline to only selecting action X 11.3% of the time. Though this difference is still statistically significant by the last 10 rounds, it diminishes greatly and almost monotonically across all rounds of play. As panel b) of Table 3.4 shows, by the last 10 rounds of play, the treatment group is only slightly more likely to reach an equilibrium relative to the control group, whereas the difference is much larger in the initial rounds. The efficiency implications of this convergence can be seen in panel b) of Table 3.5. In the first 10 rounds of play, the treatment group is much less efficient than the control group (-22.6%), but by the last 10 rounds, this treatment effect is greatly attenuated (-6.8%).

Result 3.3. *In the PD, play in both treatments converges toward the unique Nash equilibrium of the game.*

In results not reported, we vary the sample (within-subjects or between-subjects samples) and the controls used (no controls for session size, session-fixed effects), and find that Results (1)-(3) are qualitatively similar regardless of sample or specification.

3.4 Estimating a Learning Model and Simulations

While our descriptive results show that mutual payoff information has substantial effects on initial play in both games and the long-run outcome in SH, it is not clear if these effects are driven by initial play, learning, or both. For example, it could be that initial play is all that is impacted and any long-run differences are due solely to a history dependence. Alternatively, opponent payoff information may impact the dynamic

learning process used by players, which has long-run implications for convergence and equilibrium selection.

In order to better understand the channels through which mutual payoff information operates, we estimate a learning model and perform simulations. Our purpose here is not to test models to determine which more accurately matches the data, but rather to apply one that helps distinguish between the impacts of initial play and learning dynamics. As such, we apply a model of fictitious play, specified as a special case of the Experience-Weighted Attraction (EWA) learning model (Camerer and Ho, 1999).¹³ In introducing this model, the authors comment that they “consider the scientific problem of figuring out how people choose their initial strategies as being fundamentally different than explaining how they learn.” The advantage of using the EWA model in our setting is that it allows us to investigate whether the effect of having mutual payoff information operates through initial play, ongoing learning, or both.

Model basics. Each round, players choose their actions (X or Y) based on the updated attractions of these two actions. Loosely speaking, attractions are players’ expected payoffs conditional on their beliefs, explained below in more detail. The attractions for X and Y depend on past observations of other players’ actions and a prior attraction that players bring into each particular game. We model subject behavior by a single representative agent, i.e., we assume that all subjects within a treatment take actions according to the same learning and decision-making mechanism, governed by the same parameter values. As different subjects may observe different histories of play, however, they may choose different actions as a result.

For each player i at the end of round t , action j has attraction $A_i^j(t)$, and these attrac-

¹³We choose this particular model to economize on the number of parameters to be estimated. Belief-learning models perform favorably compared with alternatives; for example, see Nyarko and Schotter (2003).

tions determine play in round $t + 1$ according to the following logistic choice probability function:

$$P_i^j(t + 1) = \frac{e^{\lambda \cdot A_i^j(t)}}{\sum_{k=1}^2 e^{\lambda \cdot A_i^k(t)}}, \quad (3.2)$$

where $P_i^j(t + 1)$ is the probability that action a_i^j is chosen by player i in round $t + 1$, and λ is the response sensitivity parameter, explained below in more detail. Attraction $A_i^j(t)$ is given as

$$A_i^j(t; \phi) = \frac{\phi^t \cdot A_i^j(0) + \sum_{m=0}^{t-1} \phi^m \cdot \pi_i(a_i^j, a_{-i}(t - m))}{\sum_{n=0}^t \phi^n}, \quad (3.3)$$

and can be interpreted as subject i 's expected payoff from action j after round t , given the subject's beliefs about the action chosen by other players, conditional on their own actions. That is, the underlying beliefs are defined over two states: the state that one's opponent plays X , conditional on oneself playing X , and the state that one's opponent plays X , conditional on oneself playing Y .¹⁴ Each belief is a weighted average of the history of play that has been observed and a prior – i.e., $A_i^j(0)$ – the initial attraction the representative agent brings into the particular game (e.g., SH Full treatment, PD Partial treatment). The parameter ϕ is the weighting decay rate, explained below in more detail. The functions $a_i(t)$ and $a_{-i}(t)$ are the chosen actions in round t of player i and of the opponent that the player faced in that round. Finally, $\pi_i(a_i^j, a_{-i}(t))$ is player i 's hypothetical payoff from choosing action j in round t , conditional on $a_{-i}(t)$, the actual actions of all other players in round t . Similarly, $\pi_i(a_i(t), a_{-i}(t))\pi_i(t)$ is the realized payoff for player i in round t .

¹⁴We do not require these probabilities to be the same. In other words, we allow for beliefs where the probability assigned to the other's action depends on a player's own action. We do this to allow for the possibility that subjects might perceive the occurrence of symmetric outcomes – i.e., (X, X) or (Y, Y) – as disproportionately likely.

Interpreting the parameters we estimate. For each treatment, we estimate a set of four parameters: two learning parameters (response sensitivity parameter λ and weighting decay rate ϕ) as well as the initial attractions ($A^X(0)$ and $A^Y(0)$).

The response-sensitivity parameter (also known as noise parameter) λ models players' sensitivity to differences between attractions (Camerer and Ho, 1999). This parameter can be thought of as a measure of the noise with which players respond to updated attraction values in each round. λ can take on values from 0 to $+\infty$. At one extreme, if $\lambda = 0$, subjects select from their action set in a uniformly random manner. At the other extreme, if λ approaches ∞ , subjects always best respond – i.e., they strictly choose the action with the largest attraction value in each round.

The weighting decay rate ϕ captures how much weight is put on the observations of previous rounds, relative to those of the most recent round. In particular, after having played t rounds, ϕ^r captures how much weight is put on the results of round $t - r$, relative to the results of round t . Parameter ϕ can take on values between 0 and $+\infty$.¹⁵ While values of ϕ approaching 0 would indicate that subjects only take their last opponent's action into account when forming beliefs, $\phi = 1$ would indicate that all observed actions are weighted equally, and values of ϕ approaching ∞ would indicate that subjects do not engage in learning at all but instead base their beliefs solely based on their initial attractions – i.e., they do not update or “learn.”

The initial attractions $A^X(0)$ and $A^Y(0)$ can be interpreted as subjects' expected payoffs from each respective action prior to beginning play in the first round, given their beliefs about the play of the other subject with whom they are matched, conditional on their own actions.¹⁶ Note that attractions are defined in terms of own payoffs, which

¹⁵Camerer and Ho (1999) comment that values for this decay rate are “presumably between zero and one.” We do not limit the value of ϕ from exceeding 1, but we do note that our results are consistent with estimated values of ϕ that are somewhat less than 1 in all treatments.

¹⁶Note that for a given expected probability of one's counterpart playing X in the first round, $P^X(1)$, there is a one-to-one correspondence between the expected value of X and the expected value of Y . For

means that $A^j(0)$ is constrained to take a value between the lowest and highest possible payoffs that can be derived from choosing action j . This is done mainly to maintain consistency with attractions in later rounds and to ease interpretation of these attractions.

Initial play is captured by the initial attractions as well as λ . First and foremost, $A_i^X(0)$ and $A_i^Y(0)$ represent the expected payoff of each respective action, conditional on beliefs. In addition, λ can affect initial play as lower values of λ reduce the probability that the action with the higher attraction is chosen. Learning, on the other hand, is captured by the learning parameters ϕ and λ , as well as the continually updated, history-dependent attractions for each subject, $A_i^X(t)$ and $A_i^Y(t)$.

Estimation and simulation details. The estimation is performed numerically, using maximum likelihood techniques. We employ the bootstrap procedure for estimating parameter sampling distributions, with $B = 2,000$ bootstrap samples per estimation, from which we then conduct inference using the Bias Corrected-accelerated (BCa) confidence interval method of inference pioneered by Efron and Tibshirani (1993).¹⁷

Consider a treatment with N subjects and $T = 40$ rounds. Then, the likelihood of observing subject i 's action history $\{a_i(1), a_i(2), \dots, a_i(T-1), a_i(T)\}$, given $(A^x(0), A^y(0), \phi, \lambda)$ is

$$\prod_{t=1}^T P_i^{a_i(t)}(t|A^x(0), A^y(0), \phi, \lambda), \quad (3.4)$$

example, if a player assesses in our SH game that $(X) = 8$ in the first round, this implies $P^X(1) = 0.7$ (note that $0.7 * 11 + (1 - 0.7) * 1 = 8$), which in turn implies that $(Y) = 0.7 * 9 + (1 - 0.7) * 5 = 7.8$. Thus, $(X) = 8 \Leftrightarrow (Y) = 7.8$. However, we allow both initial attractions to range independently of one another, between the minimum and maximum possible payoff for each respective action. As discussed above, this allows for "skewed beliefs," beliefs in which subjects assign too high (or too low) a probability on symmetric outcomes. For example, suppose subjects playing the SH Partial treatment believe that conditional on themselves playing X (Y), their opponent will play X with a probability of 0.7 (0.6) in the first round. Given this set of beliefs, $A^X(0)$ would be 8 ($= 0.7 * 11 + 0.3 * 1$), while $A^Y(0)$ would be 9.8 ($= 0.6 * 13 + 0.4 * 5$).

¹⁷We are unable to use more commonly used standard error techniques due to extreme levels of skew and kurtosis in the sample distribution of bootstrapped parameter estimates.

where $P_i^{a_i(t)}$ corresponds to the logistic probability function defined in equation (3.2). The joint likelihood function $\mathcal{L}(A^x(0), A^y(0), \phi, \lambda)$ of observing all subjects' action histories is given by

$$\mathcal{L}(A^x(0), A^y(0), \phi, \lambda) = \prod_i^N \{ \prod_{t=1}^T P_i^{a_i(t)}(t | A^x(0), A^y(0), \phi, \lambda) \}.$$

To test whether our parameter estimates lead to predicted behavior that is consistent with actual observed behavior, we conduct simulations of 1,000 sessions per treatment. Each one of these sessions has an even number of subjects between 14 and 20, chosen randomly.¹⁸

3.4.1 Results: Learning Model and Simulations

Parameter estimates. Table 3.6 reports estimated parameters of the learning model by game and information treatment. Estimates of response sensitivity parameter λ are higher for the SH Full treatment than for the SH Partial treatment, and lower for PD Full treatment than for the PD Partial treatment, indicating that mutual information increases the sensitivity of responses to attractions in the SH while decreasing it in the PD. As reported in Appendix C.2 Table C.3, these differences are significant at a 95% confidence level.

Similarly, estimates of weighting decay rate ϕ are higher for the SH Full treatment than for the SH Partial treatment, and are lower for the PD Partial treatment than for the PD Full treatment. As reported in Table C.4, however, these differences are only marginally significant for the SH ($p = 0.062$) and are not significant for the PD ($p = 0.301$).¹⁹

¹⁸The simulation results are robust to changes in the number of subjects in each session.

¹⁹Because ϕ enters the learning model with an exponent equal to the number of rounds since an observation has been made, small differences in the value of ϕ can lead to significantly different modeled behavior. For example, our estimates of ϕ for the SH imply that an observation made 3 rounds previously in the SH Full treatment would have a weight of 90% relative to the current round's observation,

Estimates of the initial attraction for the coordinating action (in the SH) and cooperative action (in the PD) both increase under mutual information. These differences are significant for the SH at a 95% confidence level and at a 90% confidence level for the PD. It is worth pointing out that the estimates of $A^X(0)$ and $A^Y(0)$ in the Partial treatment are reasonably close to the expected payoffs associated with either action if one's prior belief that the other player chooses X versus Y is 50 : 50.²⁰ Considering that subjects did not get to observe any actions before the first round, holding a prior of 0.5 seems very intuitive in the Partial treatment.

Simulations with flipped parameters, initial play versus learning. Next, we shed light on whether the treatment effect operates primarily through initial play or learning. To do so, we conduct simulations wherein we swap the value of each parameter with the corresponding parameter value of the other treatment in the same game. For example, to isolate the role that the initial attractions play for explaining behavior in the SH Full treatment, we perform simulations using the estimated parameters of the SH Full treatment, except we use the initial attractions parameters from SH Partial treatment. Note that while the parameter estimates would already allow us to hypothesize in which direction average behavior may change, conducting this simulation exercise with flipped parameters has the advantage of providing insights on how economically meaningful these changes are.

Figure 3.5 shows model simulations with our estimated parameters (in solid lines)

while a similarly aged observation in the SH partial treatment would have only 63% relative weighting. After 6 rounds, the weights drop to 81% and 40%, respectively. This suggests that learning under the SH Partial treatment is much more heavily weighted to recent observations than it is in the SH Full treatment, implying a learning process that is much more sensitive to potential changes in behavior of one's counterparts.

²⁰For example, if a subject in the SH believes that her opponent is equally likely to choose action X and action Y , she expects to earn 6 ($= 0.5 * 11 + 0.5 * 1$) by choosing X , and 7 ($= 0.5 * 9 + 0.5 * 5$) by choosing Y . Similarly, a subject in the PD with the same prior belief would expect to earn 6 by choosing X and 9 by choosing Y .

alongside simulations with flipped parameters (in dotted lines). This is done separately for the SH (left panels) and the PD (right panels). In panel (a) and (b), estimates for the initial attractions are flipped at the information treatment level. Likewise, in panel (c) and (d), estimates of λ are flipped. Finally, in panel (e) and (f), estimates of ϕ are flipped.

For the SH, switching the initial attraction estimates results in quite dramatic differences of the simulated data. For the PD, differences are initially notable but vanish by about round 15. In both games, flipping the initial attractions reduces the simulated fraction of subjects playing X in the Full treatment but increases that fraction in the Partial treatment. This echoes the estimates of $A^X(0)$ in Table 3.6, where the SH Full and PD Full treatments have bigger values of $A^X(0)$ than their Partial counterparts. When $A^X(0)$ ($A^Y(0)$) is higher, the simulated fraction playing X goes up (down), which is an intuitive result.

Next, we look at whether our estimated differences for λ have meaningful consequences on behavior. When simulating behavior in the SH Full treatment with the λ of the SH Partial treatment, λ is now smaller than in the original simulations, moving behavior towards the *smaller* attraction, which is $A^Y(t)$ for the SH Full treatment. Likewise, if we simulate behavior in the SH Partial treatment of the λ of the SH Full treatment, we face a higher λ , leading behavior to move towards the *bigger* attraction, which is $A^Y(t)$ for the SH Partial treatment. Consequently, as panel (c) of Figure 3.5 shows, flipping the estimated λ parameters results in a lower fraction of playing X in both treatments of the SH. The same logic applies to the PD. Note that for the PD Full treatment, the attraction is initially higher for X , but starting in about round 5, the attraction is higher for Y . This is why in panel (d) of Figure 3.5, the simulation of the PD Full treatment with a flipped ϕ is initially lying above the original simulation but then falls below it.

When flipping ϕ , note that our estimates of ϕ are not statistically distinguishable for the PD and are only marginally statistically distinguishable for the SH (see Table 3.6). While we find no effect for the SH Partial, PD Full, and PD Partial treatments, there is a substantial drop of the fraction playing X over time in the SH Full treatment when we flip the ϕ parameters. This suggests that coordinating on the payoff-dominant Nash equilibrium is highly sensitive even to small changes of the model parameters. Note that the lower ϕ is, the more rapidly the weight placed on observations of earlier rounds is depreciated. Therefore, low values of ϕ make play more sensitive to volatility in behavior, as players are placing higher weights on a smaller set of (recent) observations when forming beliefs.

Taken together, the simulations where we flip parameters at the treatment level indicate that each of the four parameters we estimate (with the exception of ϕ in the PD) play an important role in explaining the differences in behavior across treatments. In fact, for our SH game the payoff dominant equilibrium is only maintained through a combination of the estimated parameters for initial attractions, belief updating, and response sensitivity. However, in the PD the effect is less persistent. While subjects' initial perceptions of the game and their response sensitivity change under mutual information, the way they process observations does not appear to change. That is, the initial perception of the game retains similar importance for future play under both information treatments in the PD.

Result 3.4. *Simulations suggest that the information treatment effect can be attributed to differences in both initial play and learning in both games.*

Model validation. To test whether the model performs reasonably well in fitting the observed data, we conduct simulations of 1,000 sessions per treatment using the parameter values we estimate. Each simulated session has an even number of subjects between

14 and 20, chosen randomly with equal probabilities. Figure 3.6 compares the observed average behavior in our experiment with simulated data that we generate by using the estimated learning model parameters. For each game and information treatment, solid lines depict the average fraction of subjects that chose action X in a given round, and dashed lines depict the corresponding simulated data averaged across 1,000 simulations. As Figure 3.6 shows, the simulated data is qualitatively very similar to the actual observed average behavior of subjects. Appendix Table C.2 provides regression analyses showing further support that the simulated data is on average statistically indistinguishable from the observed behavior.

3.5 Discussion

As discussed in Section 3.4, the information treatment effect operates through both initial play and learning. In this section, we discuss potential channels behind this treatment effect.

Recall that the appeal of contrasting the PD and the SH in our experiment is based on the idea that the structural uncertainty which is implied by the absence of mutual information affects strategic uncertainty differently in these games: Observing the other's payoffs arguably reduces strategic uncertainty in the SH, as it facilitates coordinating on the payoff-dominant Nash equilibrium. In the PD, however, mutual payoff information makes the cooperative action more attractive for players with social preferences. At the same time, players face uncertainty about the social preferences of others, and consequently mutual payoff information may increase strategic uncertainty.

We propose that differences in strategic uncertainty are captured by differences in the values of the learning model parameters. Conditional on the attractions, more prior uncertainty makes experimentation more beneficial, reflected by lower values of λ . Low

values of λ are therefore consistent with more strategic uncertainty. While we cannot identify that low values of λ indeed imply more strategic uncertainty, one way to think of low values of λ would be that this is suggestive evidence that the game is perceived to be more strategically uncertain, all else equal. Similarly, we argue that higher levels of strategic uncertainty are consistent with lower values of ϕ : If subjects place more weight on results from the most recent rounds (relative to that of earlier rounds), their behavior is more sensitive to noise and to potential emerging trends in the patterns of play of their counterparts.

Indeed, our estimates of λ imply that there is more noise in play in the SH Partial treatment than in the SH Full treatment and in the PD Full treatment than in the PD Partial treatment, which is consistent with the idea that mutual payoff information increases strategic uncertainty in the PD while decreasing it in the SH. Furthermore, our estimates of the weighting depreciating parameter ϕ indicate that subjects place more weight on counterpart actions from recent rounds, and thus tend to update their attractions more rapidly in the SH Partial treatment than in the SH Full treatment and in the PD Full treatment than in the PD Partial treatment, which is again consistent with the idea that the strategic uncertainty fueled by mutual payoff information goes in different directions in these two games.

Aside from our learning model, what other indicators of strategic uncertainty could one use to test the idea that mutual payoff information affects strategic uncertainty differently in the SH and the PD? While the literature addressing strategic uncertainty describes the concept in largely qualitative terms, Calford and Oprea (2017) utilize the size of the Basin of Attraction (BOA) as a simple quantitative measure of strategic risk. Applied in our context, the BOA index of a subject's least cooperative strategy (Y) is the probability they must assign to their counterpart playing the most cooperative of their strategies (X) in order to be indifferent between playing one's own most and least

cooperative strategies. The BOA index depends only on one's own payoffs and thus cannot explain the information treatment effects we examine. Moreover, this measure cannot provide useful insights into the strategic uncertainty that may exist within a given one-shot PD game: By convention, the BOA index is set to a value of 1.0, regardless of the degree to which one's "temptation" payoff exceeds their "reward" payoff or to which their "sucker" payoff falls below their "punishment" payoff, etc. The BOA index's invariant values for a one-shot PD do not reflect the varying levels of cooperation that are generally observed in this game.²¹

What other features of a game change as mutual payoff information is removed? Observing other players' payoffs may potentially also affect signaling and reputation building (as noted by Feltovich and Oda, 2014), focal points, and/or the expression of social preferences. Given that our experimental design employs random re-matching, we do not think that signaling or reputation concerns notably contribute to the information treatment effect we document. As for focal points, Crawford et al. (2008) suggest that four features of a game can potentially generate them: labels, pre-play communication, precedents based on shared histories, and the payoff structure of a game. Labels are kept constant across treatments in our design, and there is no pre-play communication in our experiment. While precedents of shared histories could be an issue if play in the first game of a session affects the second game, we do not find that the order in which games are presented to subjects in the experiment affects the behavior we observe. Finally, the payoff structure is a strategic element and as such not separate from the channel of strategic uncertainty. One could argue that non-strategic use of payoff structure, such as awareness that the game is symmetric, might affect play. However, this cannot predict the direction of the treatment effect we observe.

²¹See for example Charness et al. (2016), who show that increasing the reward payoff monotonically increases rates of cooperation.

While our experiment is not designed to study the effects of counterpart payoff information on social preferences, it is worth noting that our results for initial round play are nevertheless consistent with models that incorporate social preferences.²² We observe higher rates of initial coordination on socially optimal action pair (X, X) in the Full treatments of both our SH and PD games.²³ This is consistent with a social preference model in which agents behave as follows: If the socially optimal outcome can be identified—e.g., through mutual payoff information—they select the action associated with this outcome. In absence of such information, they select the level-1 action—i.e., the expected payoff-maximizing action assuming a counterpart who uniformly randomizes from their action set. A robust exploration of the impacts of counterpart payoff information on the expression of social preferences is beyond the scope of this paper and is left to further research.

3.6 Conclusion

We conduct an experiment in which subjects play repeated versions of SH and PD games with random re-matching. In the full-information treatment subjects observe the entire payoff matrix, while in the partial-information treatment subjects observe only their own payoffs. In both treatments, subjects observe the chosen actions after each round.

We find that mutual payoff information has a strong effect on initial play in both games. In the SH, the vast majority of subjects selects the action consistent with

²²We also compare the first-round play observed in our experiment with several other contemporary models of one-shot and initial game play (e.g. QRE, level- k , Cognitive Hierarchy, etc.). Of these, only models of social preference are able to rationalize results for both our SH and PD games.

²³Notably, in our PD Full treatment we observe initial rates of coordination on socially-optimal action profile (X, X) of 64.3%, with a 95% confidence interval that excludes 50% (see the final row of Table 3.6). If we apply the QRE model to analyze the forces shaping initial moves, this fact implies that errors or randomizing alone cannot rationalize our high rates of initial cooperation in the PD Full treatment: The attraction of X must be higher than that of Y for the rate of selection of X to exceed 50%.

the payoff-dominant Nash equilibrium under full-information but the action consistent with the risk-dominant Nash equilibrium under partial-information. This effect persists through all 40 rounds of the game. In the PD, on the other hand, we initially observe a pronounced difference – subjects in the full-information treatment are much more likely to cooperate – but play converges toward the unique Nash equilibrium in later rounds of the game.

We estimate a belief-learning model to study our information treatment effect on both initial play and on the learning process. These results and related simulations suggest that mutual payoff information alters not only the way subjects initially perceive the game but also the way they update and respond to their beliefs. The values of our estimated learning parameters are important in explaining the observed behavior, suggesting that learning by observing is an important feature of behavior in the long run regardless of whether or not mutual payoff information is available.

We propose that these effects are evidence of the effect that mutual payoff information has on strategic uncertainty. In the SH strategic uncertainty decreases as players can reason about the mutually beneficial, Pareto-efficient outcome. However, in the PD mutual payoff information reveals the tension inherent in the game and thus increases strategic uncertainty. In both cases, social preferences may also be at play when opponent payoff information is revealed.

Figure 3.1: The Games and Information Treatments From the Row Player’s Perspective

		Full		Partial	
		X	Y	X	Y
Stag Hunt (SH)	X	11, 11	1, 9	11, ■	1, ■
	Y	9, 1	5, 5	9, ■	5, ■
		X	Y	X	Y
Prisoner’s Dilemma (PD)	X	11, 11	1, 13	11, ■	1, ■
	Y	13, 1	5, 5	13, ■	5, ■

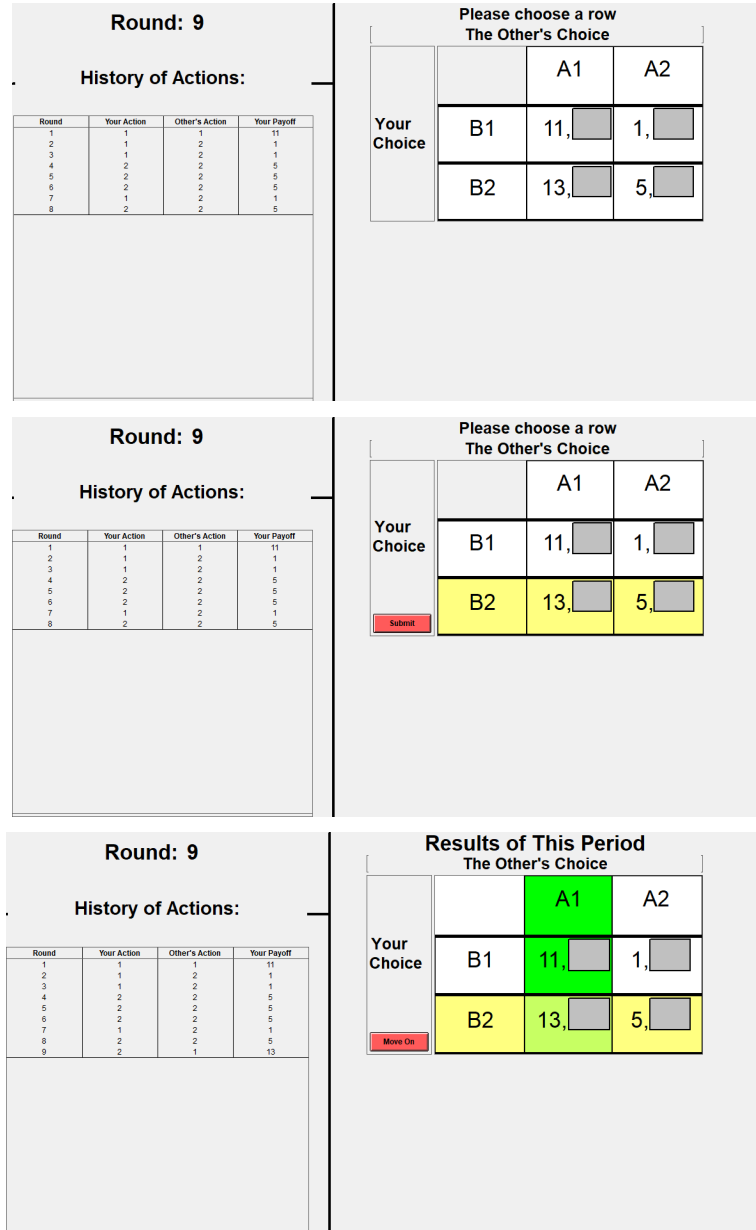
Table 3.1: Treatments by Session

Sessions	Part 1	Part 2	# Subjects per Session
1-3	SH - Partial	SH - Full	16, 16, 20
4-6	PD - Partial	PD - Full	16, 16, 18
7-9	SH - Full	PD - Partial	16, 14, 14
10-12	PD - Full	SH - Partial	16, 18, 14

Table 3.2: Between- vs. Within-Subjects Analysis

Analysis	Game	Data
Between-subjects	SH	first part sessions 1-3, first part sessions 7-9
Between-subjects	PD	first part sessions 4-6, first part sessions 10-12
Within-subjects	SH	sessions 1-3 (first and second part)
Within-subjects	PD	sessions 4-6 (first and second part)

Figure 3.2: Screenshots of Experimental Interface, PD Partial Treatment



This figure shows screenshots of the experimental interface of the PD Partial treatment for a subject that was assigned to Group B. The first panel displays the interface at the beginning of the ninth round; the subject sees the history of the first eight rounds but has not chosen their next action yet. The second panel depicts the same interface after the subject has selected – but not yet committed to – Action B2. Finally, the third panel shows the feedback the subject receives at the end of Round 9; they can see that the person from Group A they were matched with in Round 9 chose Action A1. Note in this panel that the History of Actions table has been updated with the results from Round 9.

Figure 3.3: Mean Action Rates by Treatment

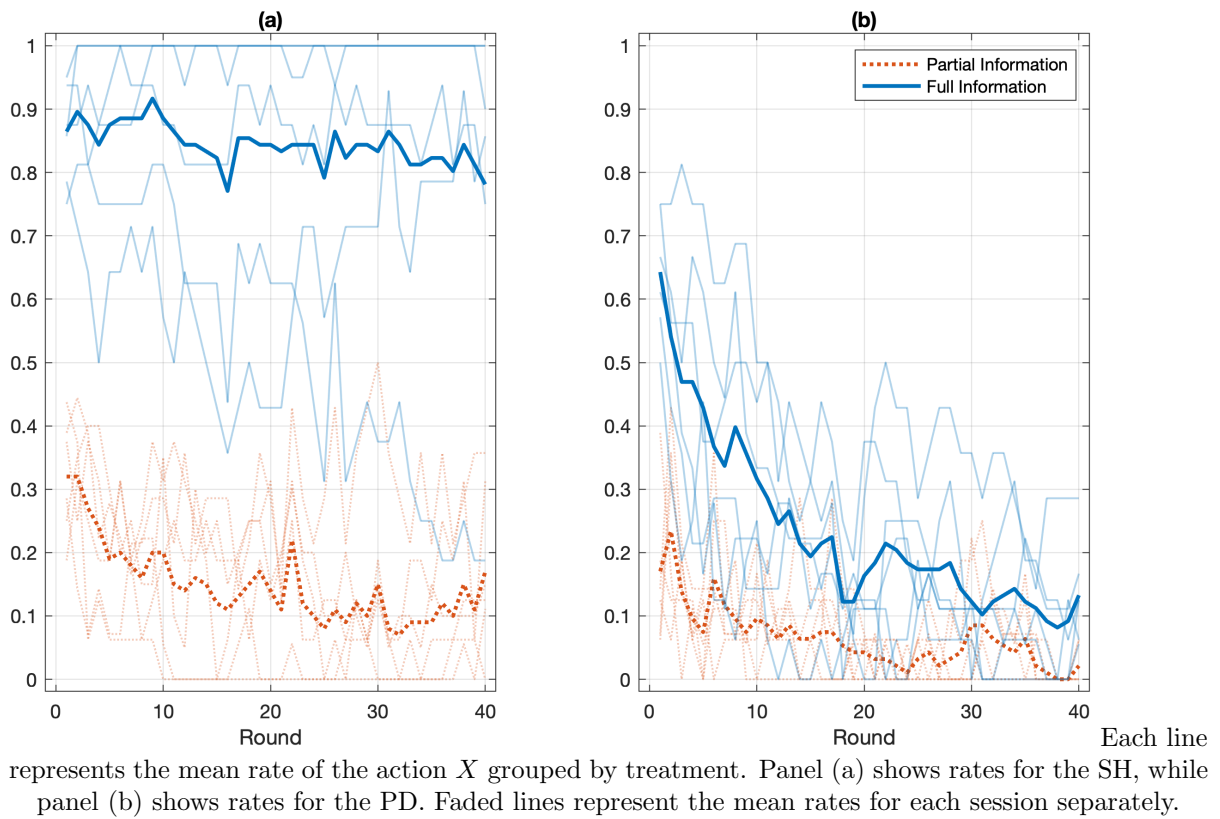
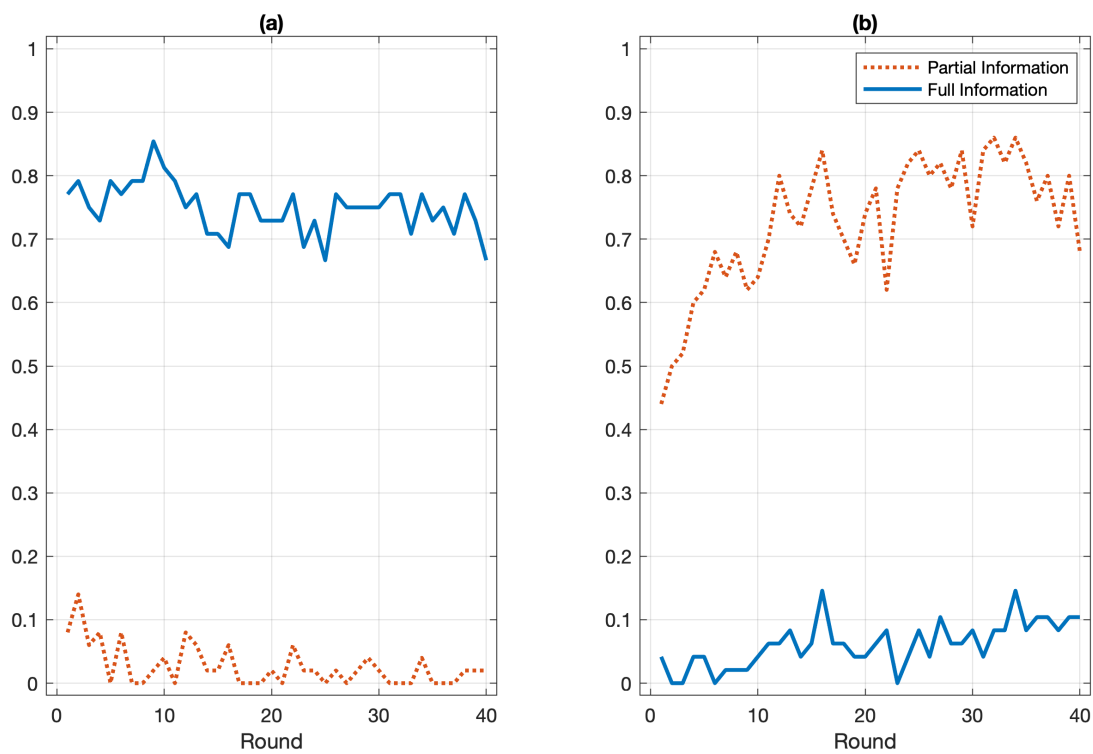
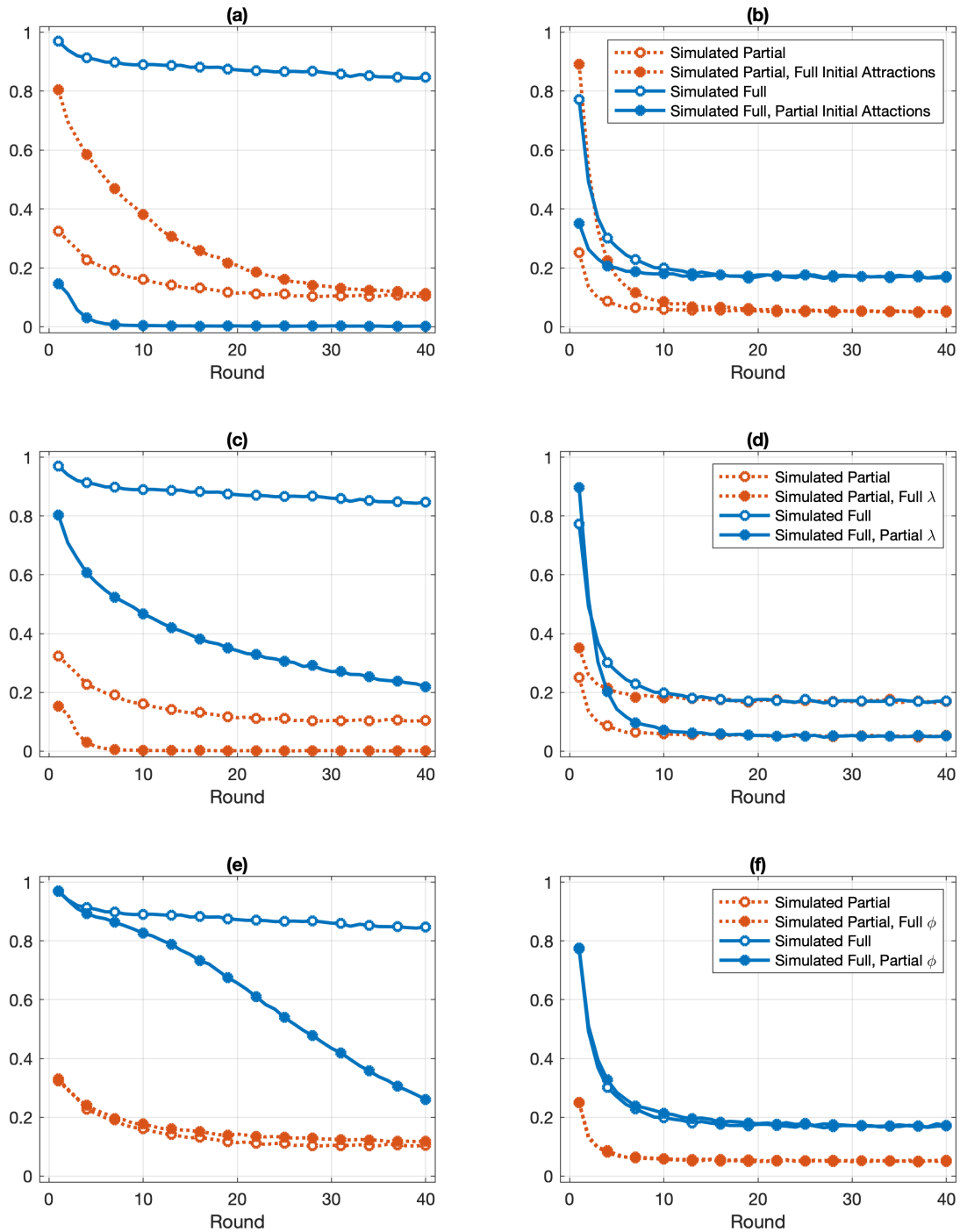


Figure 3.4: Equilibrium Rates for SH Sessions



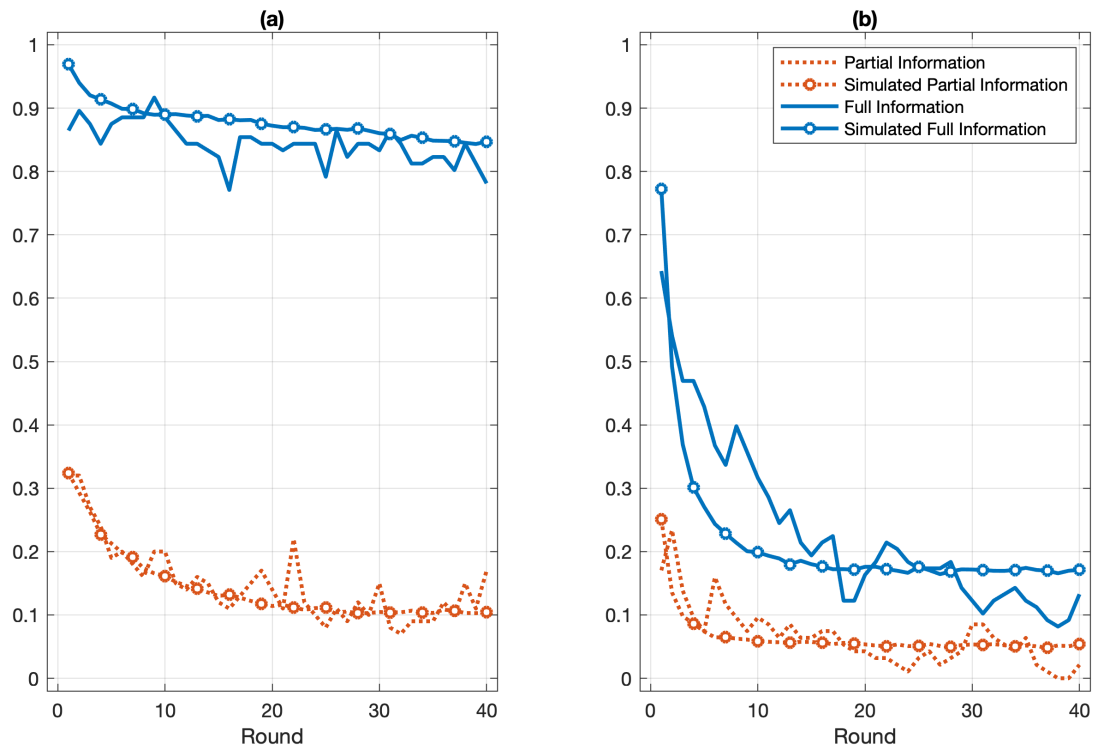
Panel (a) shows rates of the payoff-dominant equilibrium. Panel (b) shows rates of the risk-dominant equilibrium. Note that the sum of the solid blue lines adds up to less than one and the sum of the dotted red lines adds up to less than one. The remaining fraction is accounted for by the fact that a pure strategy Nash equilibrium is not always achieved in a given round.

Figure 3.5: Mean Action X Rates in Simulated Data of Six Sessions per Treatment, versus Same Rates in Simulated Data Using Parameters from Other Information Treatment of Same Game



Left hand Panels (a, c, e) show these rates for the SH, and right hand Panels (b, d, f) show them for the PD.

Figure 3.6: Mean Action X Rates in Pooled Data of Six Sessions per Treatment, versus Same Rates for 1000 Simulated Sessions



Panel (a) shows these rates for the SH, and Panel (b) shows them for the PD.

Table 3.3: Treatment Effect for Selection of Action X

	(1)	(2)	(3)	(4)	(5)
	Overall	1-10	11-20	21-30	31-40
a) Stag Hunt					
Treatment effect	-0.676	-0.669	-0.683	-0.677	-0.673
	(0.034)	(0.030)	(0.041)	(0.047)	(0.040)
Cluster p-value	0.000	0.000	0.000	0.000	0.000
Control mean	0.844	0.881	0.836	0.822	0.838
Number of clusters	144	144	144	144	144
N	7,840	1,960	1,960	1,960	1,960
b) Prisoner's Dilemma					
Treatment effect	-0.173	-0.304	-0.147	-0.156	-0.086
	(0.022)	(0.037)	(0.028)	(0.031)	(0.026)
Cluster p-value	0.000	0.000	0.000	0.000	0.001
Control mean	0.232	0.433	0.205	0.176	0.113
Number of clusters	142	142	142	142	142
N	7,680	1,920	1,920	1,920	1,920

The sample uses the pooled data. The regressions include controls for session size. Standard errors presented in parentheses are calculated using the cluster-robust method allowing for correlation between observations within a cluster. Clustering is at the session-subject level. Cluster p-value indicates the p-value from a two-sided t-test of the null hypothesis that the treatment effect is zero using the cluster-robust standard error.

Table 3.4: Treatment Effect for Reaching an Equilibrium Outcome

	(1)	(2)	(3)	(4)	(5)
	Overall	1-10	11-20	21-30	31-40
a) Stag Hunt					
Treatment effect	-0.100	-0.174	-0.050	-0.076	-0.099
	(0.028)	(0.036)	(0.036)	(0.033)	(0.033)
Cluster p-value	0.001	0.000	0.167	0.023	0.004
Control mean	0.811	0.808	0.798	0.831	0.808
Number of clusters	144	144	144	144	144
N	7,840	1,960	1,960	1,960	1,960
b) Prisoner's Dilemma					
Treatment effect	0.266	0.405	0.241	0.258	0.158
	(0.019)	(0.032)	(0.030)	(0.033)	(0.027)
Cluster p-value	0.000	0.000	0.000	0.000	0.000
Control mean	0.618	0.347	0.643	0.696	0.788
Number of clusters	142	142	142	142	142
N	7,680	1,920	1,920	1,920	1,920

The sample uses the pooled data. The regressions include controls for session size. Standard errors presented in parentheses are calculated using the cluster-robust method allowing for correlation between observations within a cluster. Clustering is at the session-subject level. Cluster p-value indicates the p-value from a two-sided t-test of the null hypothesis that the treatment effect is zero using the cluster-robust standard error.

Table 3.5: Treatment Effect for Efficiency Ratio

	(1) Overall	(2) 1-10	(3) 11-20	(4) 21-30	(5) 31-40
a) Stag Hunt					
Treatment effect	-0.396 (0.018)	-0.412 (0.015)	-0.386 (0.021)	-0.390 (0.026)	-0.394 (0.021)
Cluster p-value	0.000	0.000	0.000	0.000	0.000
Control mean	0.864	0.883	0.856	0.857	0.859
Number of clusters	144	144	144	144	144
N	7,840	1,960	1,960	1,960	1,960
b) Prisoner's Dilemma					
Treatment effect	-0.078 (0.011)	-0.147 (0.020)	-0.063 (0.014)	-0.066 (0.015)	-0.034 (0.015)
Cluster p-value	0.000	0.000	0.000	0.000	0.025
Control mean	0.554	0.650	0.539	0.527	0.498
Number of clusters	142	142	142	142	142
N	7,680	1,920	1,920	1,920	1,920

The sample uses the pooled data. The regressions include controls for session size. Standard errors presented in parentheses are calculated using the cluster-robust method allowing for correlation between observations within a cluster. Clustering is at the session-subject level. Cluster p-value indicates the p-value from a two-sided t-test of the null hypothesis that the treatment effect is zero using the cluster-robust standard error.

Table 3.6: Learning Model Parameter Estimates and 95% Confidence Intervals

Parameter	SH–Full	SH–Partial	PD–Full	PD–Partial
λ	1.5831 (1.2159 – 1.8776)	0.6516 (0.5184 – 0.7673)	0.4297 (0.3347 – 0.5242)	0.7426 (0.6615 – 0.9146)
ϕ	0.9649 (0.9011 – 1.0223)	0.8290 (0.7112 – 0.9911)	0.8011 (0.3009 – 0.9398)	0.8769 (0.6019 – 85.013)
$A^X(0)$	8.5765 (7.8948 – 9.7541)	5.9218 (5.1815 – 6.4317)	9.4480 (8.4351 – 10.5367)	7.1258 (6.9829 – 8.5193)
$A^Y(0)$	6.4400 (5.8775 – 6.6354)	7.0199 (6.8786 – 7.2157)	6.4995 (5.5125 – 7.3160)	8.5930 (8.5141 – 12.9087)
$L(\lambda)$	–1227.06	–1542.83	–1934.22	–882.90
n	3840	4000	3920	3760
$P^X(1)$ (learning model)	0.9672 (0.8002 – 0.9999)	0.3284 (0.1636 – 0.5069)	0.7802 (0.1991 – 0.9334)	0.2517 (0.0614 – 0.3999)
$p_X(1)$ (binomial model)	0.8646 (0.8038 – 0.9358)	0.3200 (0.2537 – 0.4366)	0.6429 (0.5641 – 0.7486)	0.1702 (0.1231 – 0.2825)

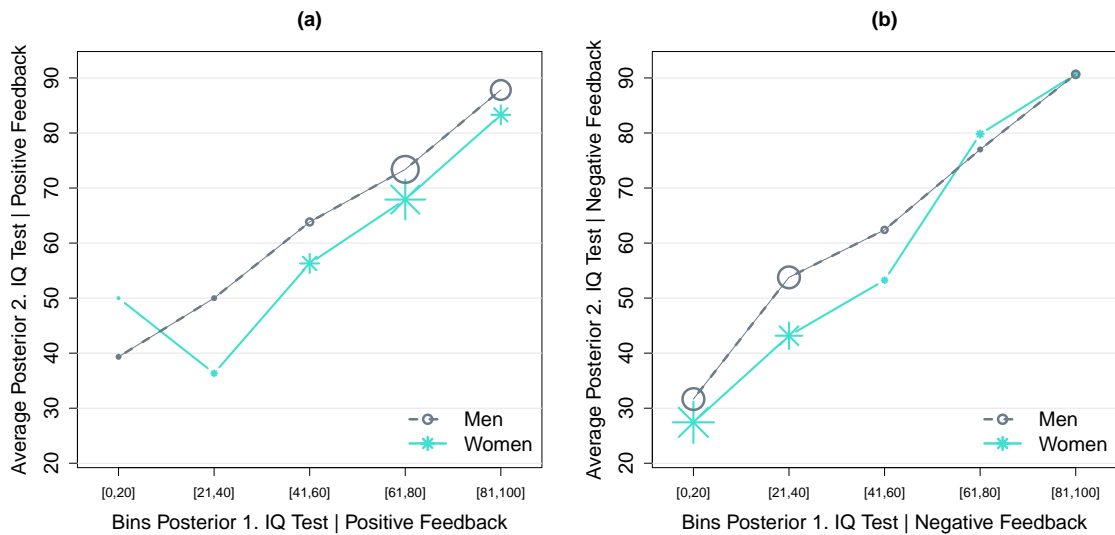
Results of tests of significance of the information treatment effect on parameters estimates for λ and ϕ in SH and PD are reported in Appendix C.2 Tables C.3 and C.4, respectively. Differences in values for λ are significant for both SH and PD ($p < 0.05$), while difference in values of ϕ are not significant for PD, and are only weakly significant for SH. For estimates of $p_X(1)$ we employ Agresti-Coull binomial confidence intervals; see Agresti and Coull (1998) and Brown et al. (2001) for mathematical definition and motivation for their use over the more commonly used Wald confidence interval approach.

Appendix A

Appendix to Chapter 1

A.1 Additional Figures and Tables

Figure A.1: Gender Differences in Posterior Beliefs About the Future, Given Beliefs About the Past.



This figure plots gender differences in posterior beliefs about passing the 2. IQ test, given posterior beliefs about the 1. IQ test. The size of the points represents the relative share of observations in a given bin category of prior beliefs about the 1. IQ test. Panel (a) shows this relationship conditional on having received positive, while panel (b) shows this relationship conditional on having received negative feedback. On average, men are more optimistic than women about passing the future IQ test, given their beliefs about having passed the first IQ test.

Table A.1: OLS Estimates of the Probability to Continue

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

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This table is an extension of Table 1.2, displaying only estimates that are relevant to the *Baseline* treatment. Robust standard errors in parentheses. Constants not displayed. Additional controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity). The mean of the reference group shows the average probability of continuing for all men (column 1), men who received positive feedback (columns 2, 4, and 5), and men who received positive feedback but failed the first IQ test (column 3) in the *Baseline*.

Table A.2: OLS Estimates of Prior and Posterior Beliefs.

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

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Constants not displayed. Additional controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity). Data from *Baseline* and *AlwaysInfo* combined. 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 A.3: OLS Estimates of Log-Likelihood Bayesian Updating.

	First Test		Future Test	
	(1)	(2)	(3)	(4)
α	0.834*** (0.0648)	0.842*** (0.122)	0.922*** (0.0624)	0.888*** (0.114)
β_p	1.227*** (0.156)	1.104*** (0.259)	0.839*** (0.140)	0.807*** (0.207)
β_n	1.672*** (0.159)	1.711*** (0.257)	1.104*** (0.145)	0.887*** (0.260)
α * Female		-0.00537 (0.135)		0.0594 (0.127)
β_p * Female		0.260 (0.317)		0.127 (0.276)
β_n * Female		-0.0708 (0.332)		0.376 (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

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Variants of equation 1.2 are estimated. Columns (1)-(2) estimate belief updating on the first IQ test, where $\phi = \frac{2}{3}$ by design. Columns (3)-(4) 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 A.4: OLS Estimates of the Probability to Continue.

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

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. This table only displays estimates that are relevant to the *Baseline* treatment, but uses data from all treatments. Robust standard errors in parentheses.

Constants not displayed. Additional controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity). Column (1) in this table corresponds to Column (1) of Table 1.2. CRRA and CARA risk parameters refer to the means of the risk parameter intervals computed under the assumption of narrow framing with a base wealth of 0. The number of observations in Columns (3) and (4) are lower as the risk parameters are not well-defined for all subjects. The mean of the reference group shows the average probability of continuing for men who received positive feedback in the *Baseline*. For columns (3) and (4), this average refers to the subset of subjects for which the risk parameters are well defined.

Table A.5: OLS Estimates of Risk Parameters

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

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Constants not displayed.

Additional controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity). The mean of the reference group refers to men's average estimated risk parameters. The number of observations refers to the number of subjects for which a respective risk parameter was well-defined.

Table A.6: Summary Statistics: AlwaysInfo Treatment Relative to Baseline Treatment

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

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

Table A.7: Probability of Continuing by 1. IQ Test Performance

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

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Constant not displayed. Additional controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity). This table is the extension version of Table 1.5.

Table A.8: OLS Estimates of Probability of Continuing, Baseline Treatment, Qualitative Controls.

	All					Typ. Occup.	Dad Works More	Cons. Attitudes
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	-0.103** (0.0422)	-0.0876* (0.0458)	-0.108** (0.0435)	-0.0945** (0.0450)	-0.0915* (0.0505)	-0.0907 (0.0556)	-0.122** (0.0567)	-0.166*** (0.0590)
Neg. Feedback	-0.106*** (0.0281)	-0.103*** (0.0295)	-0.105*** (0.0293)	-0.100*** (0.0277)	-0.103*** (0.0302)	-0.0879** (0.0425)	-0.0597 (0.0462)	-0.101** (0.0390)
Z-Score 1. IQ Test	0.0601*** (0.0151)	0.0691*** (0.0141)	0.0610*** (0.0154)	0.0594*** (0.0159)	0.0694*** (0.0150)	0.0561*** (0.0201)	0.0719*** (0.0246)	0.0811*** (0.0193)
Parents Typ. Occup. FEs	-	✓	-	-	✓	-	-	-
Parents Hours Worked	-	-	✓	-	✓	-	-	-
Own Attitudes FEs	-	-	-	✓	✓	-	-	-
Additional Controls	✓	✓	✓	✓	✓	✓	✓	✓
Mean Reference Group	0.68	0.68	0.68	0.68	0.68	0.64	0.67	0.68
Observations Baseline	94	94	94	94	94	56	49	53
Observations Total	205	201	191	205	191	119	104	112

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Constants not displayed. Only estimates relevant to the *Baseline* treatment are shown. Column (6) shows the sub-sample of subjects that did not disagree / strongly disagree that both their mother's and father's occupation was "typical for a woman/man of her/his generation." Column (7) shows the sub-sample of subjects that reported a strictly higher "hours worked for pay" for their father than mother in a "typical week" when they were a child. Column (8) shows the sub-sample of subjects that either disagreed or strongly disagreed that "women should pay their own way on dates," or that did not strongly disagree that "a wife with a family has no time for outside employment." Observation numbers in columns (1)-(5) differ as not all subjects answered the respective questions. Parental occupation fixed effects include a fixed effect for subjects' subjective assessment of whether their mother's/father's occupation is considered as typical for their generation. Parents' hours worked are the reported hours worked for pay in a typical week, separately for fathers and mothers. Own attitude fixed effect refer to subjects' agreement/disagreement with the statements captured in questions 13-16 in the end survey, see Appendix A.3. The mean of the reference group refers to the average continuation probability for men who received positive feedback in the *Baseline*. Additional controls: Zoom vs. in-person sessions and self-reported characteristics (US citizenship, English as a first language, GPA, major or intended major, race/ethnicity).

A.2 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 A.2 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 A.6.

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.¹ A number of visual and interactive features made the BDM especially intuitive to use.²

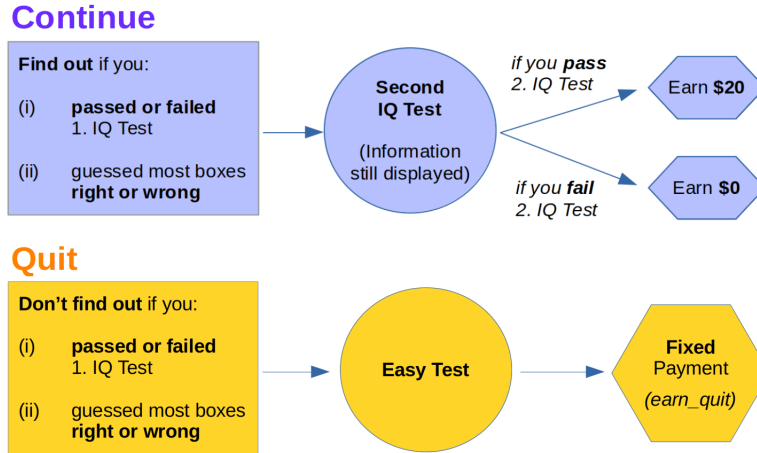
Guessing Game at the Beginning. Before the main part of the experiment began, a trivial “*Guessing Game*” was conducted. This game is not meaningful in the *Baseline* or the *AlwaysInfo* treatment. The reason for including it was to keep things consistent with a third treatment for which the data may be collected in the future.³

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.

¹The 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.

²The 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.

³In the “*Guessing Game*”, subjects had to guess which 3 out of 6 closed boxes contain a ball, see Figure ???. Correct guesses 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.



Q#		Option A		Option B	
1	Would you rather...	quit with $earn_quit = \$22$	or	continue	?
2	Would you rather...	quit with $earn_quit = \$21$	or	continue	?
3	Would you rather...	quit with $earn_quit = \$20$	or	continue	?
4	Would you rather...	quit with $earn_quit = \$19$	or	continue	?
.	.	.		.	
.	.	.		.	
.	.	.		.	
20	Would you rather...	quit with $earn_quit = \$3$	or	continue	?
21	Would you rather...	quit with $earn_quit = \$2$	or	continue	?
22	Would you rather...	quit with $earn_quit = \$1$	or	continue	?
23	Would you rather...	quit with $earn_quit = \$0$	or	continue	?

Your switch point: \$7

This means:

- You choose to **quit** if $earn_quit$ is \$7 or more.
- You choose to **continue** if $earn_quit$ is less than \$7.



If you move on, you finalize your **switch point** to be **\$7**.

Figure A.2: Screenshot of the BDM decision interface in the *Baseline* treatment with an example switch point of \$7.

A.3 Instructions and Experimental Interface

A.3.1 Instructions

Instructions were displayed on the screen and read out loud by the experimenter. Numbers next to the text on the slides indicate the order in which the text was displayed to subjects. The screenshots below display the last step on a given slide, with all information displayed.

Instructions at the very beginning of the experiment

Instructions

1. *You are about to participate in an experiment in the economics of decision-making. If you follow these instructions carefully and make good decisions, you can earn a **CONSIDERABLE AMOUNT OF MONEY**, which will be **PAID TO YOU IN CASH** at the end of the experiment. In addition to your other earnings, you will receive a show-up fee of **\$5**.*
2. *Your computer screen will display useful information. Remember that the information on your computer screen is **PRIVATE**. To ensure the best results for yourself, and accurate data for the experimenters, please **DO NOT COMMUNICATE** with the other participants at any point during the experiment. Please **turn your cell phone off** and avoid opening any other browsers or programs on your computer.*
3. *Economics experiments have a strict **policy against deception**. If we do anything deceptive, or don't pay you cash as described, then you can contact the **campus Human Subjects Committee** and we would be in serious trouble. Our interest is in seeing how people with an **accurate understanding** of how their decisions influence their outcomes and earnings make economic decisions.*
4. *In the following instructions, we will give you some important information about the experiment. These instructions are meant to clarify how the experiment actually works and how you can earn money. If you have any **questions** at any stage of the experiment, or need assistance of any kind, please **raise your hand** and the experimenter will come to you.*

Guessing Game

Before we begin with the main part of the experiment, there will be a guessing game.

You will see **six boxes** on the screen.



The computer randomly picked which **three of these boxes contain a ball**. The other three are empty.



Your task will be to **guess which boxes contain a ball**.

Whether or not you guess correctly entirely depends on **chance**, and will **not** impact your earnings in this experiment.

Just two more things

2. In the experiment, you will answer questions and make choices on the computer. At the bottom right of each page, there will be a little **blue arrow sign** that you can **click to move on**.



Once you click on that arrow sign, you **finalize the choices and answers** you made on a page. You **cannot go back** and change your answers and choices later.

It is therefore important that you always assure yourself that all your answers and choices are the way you want them to be **before** clicking on the blue arrow sign to move on.

Just two more things

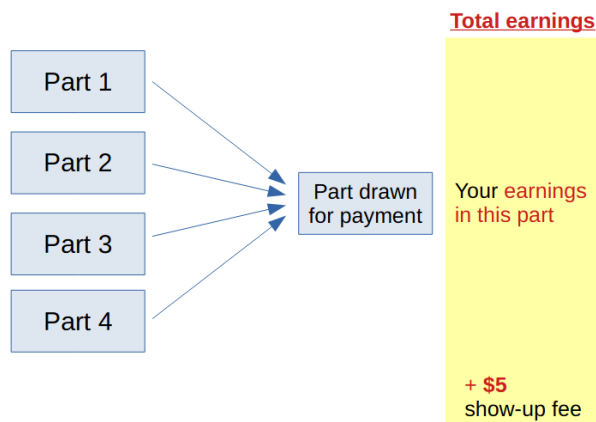
1. You will sometimes see the following message:

Please wait for further instructions.

After giving you further instructions, the experimenter will tell you the **password** to move on.

Instructions before first IQ test

Overview: 4 Parts



The remainder of the experiment consists of **4 parts**.

At the end, the computer will **randomly draw** one **part that counts for payment**, and each part is equally likely to be drawn. You won't be told which part was drawn for payment.

Your **total earnings** in this experiment will be:

- Your earnings in the part that is drawn for payment,

PLUS

- A show-up fee of \$5.

Part 1: IQ test

1. In a few minutes, you will be asked to take an **IQ test**. This test is frequently used to **measure intelligence**.

The IQ test will consist of **7 questions**.
 - You **pass** this test if you **answer at least 5** of these questions **correctly**.
 - **Otherwise, you fail**.
2. You will have **90 seconds to answer one question**, and a timer will indicate how much time is left. After 90 seconds have passed, your answers will be submitted automatically.

Unanswered questions will be counted as **wrong**.
3. Whether you pass or fail the test will **not** depend on how other participants perform on the test. All that matters is how many questions **you** get right.
4. Here's how you get paid in this part of the experiment:
 - If you **pass the IQ test**, you get paid **\$20**.
 - If you **fail the IQ test**, you get paid **\$0**.

Instructions before eliciting prior beliefs


Part 2: Assessment tasks

1. In the next part of the experiment, you will be asked to **make assessments of how likely you think** a given situation is. For example, we might ask you to assess the following situation:
2. There will be a **bar on the screen** that you can move with your mouse from numbers from 0 (at the very left) to 100 (at the very right). To answer the question, **move the bar** to the position that represents your **true assessment**.

How likely (out of 100) do you think it is that **it will be raining outside at the end of the experiment?**

0 100

Move the bar to make assessment.



3. **For example, suppose you think** that the chance that it will be raining outside at the end of the experiment is **27%**. In this case, you **should move the bar** to the **position** where the number **equals 27**, as shown on the screenshot.
4. It is **important** that you always indicate your **true assessment** of how likely a given situation is. As we will explain to you shortly, this will **maximize your chance of winning money** in this part of the experiment.

5. We will now explain to you how exactly the payment scheme for the assessment tasks works.

If you find the details hard to follow, **all you have to remember** is that we will pay you in a way that **guarantees** that you **maximize** your chance of **winning money** if you **always report your true assessment**.

6. Here is how exactly we pay you in this part of the experiment:

Let's call the **number you report in your assessment** with the slider bar **X**. After you submit your assessment, the **computer will draw a number between 0 and 100**, and each number is equally likely to be drawn. Let's call this number **Y**.

The numbers **X**, **Y**, and whether or not the situation in the assessment question occurs will determine if we pay you **\$20** or **\$0** for this part of the experiment.

If $Y \geq X$, we pay you **\$20** with a chance of **Y%**, and **\$0** with a chance of **(100-Y)%**.

If $Y < X$, we pay you **\$20** if the situation in the questions occurs, and **\$0** otherwise.
(In the example from before, this means that you would get paid \$20 if it was indeed raining at the end of the experiment.)

This payment scheme **guarantees** that you **maximize your chance of getting paid \$20** if you always report your **true assessment** of how likely a given situation is with the slider bar.

7. There will be a total of **4 assessment tasks**. The computer will **randomly pick one assessment task** that counts **for payment**.
(Depending on your later choices in the experiment, not all four assessment tasks might be eligible for payment.)

8. In the **first assessment task**, you will be asked **how likely** you think it is that you **passed the IQ test**. Some of the other assessment tasks refer to hypothetical scenarios about a future IQ test.

Make sure to always **read the questions carefully**, so that you understand which situation you are assessing in each task. If you have any questions, raise your hand and we will come and clarify.

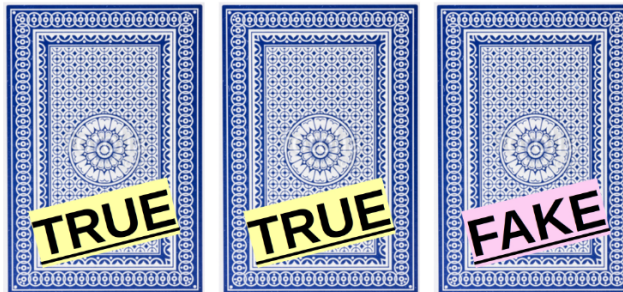
Instructions before feedback (cards)

Information on your test performance

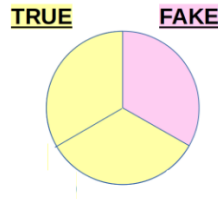
Before we ask you to answer the remaining assessment questions, you will receive information about how you performed on the first IQ test.

Cards

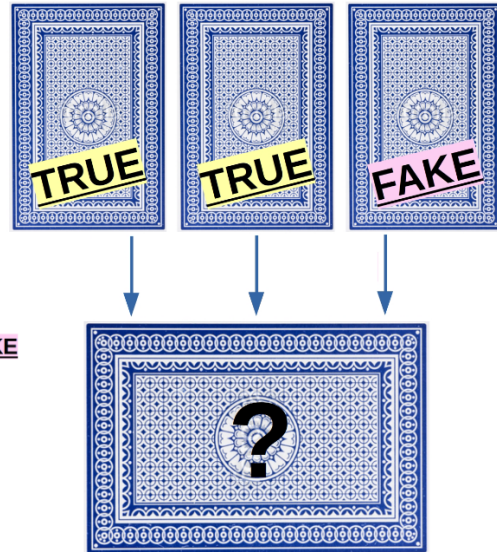
1. The computer will generate three blue cards. If we flip these cards over, they will **display information on whether you passed or failed** the IQ test.
2. Importantly, **two of these cards** say the **truth**, but **one card** is **fake** and says the contrary. This means that two cards will tell you truthfully if you passed or failed, but one card will always lie.



3. After generating the three cards, the **computer will randomly draw one card** and show you the information that is written on it. **Each card is equally likely to be drawn.**
4. The card that is drawn **will either say that you passed, or that you failed the IQ test.** But since you don't know which card was drawn, you **don't know if the information on the card is true or false.**
5. You do know, however, that the computer randomly picked one card of the three, two of which say the truth, and one of which is fake.



Importantly, this means that the **card you will see is twice as likely to tell the truth, than to be fake.**



6. The experiment will proceed as follows: The computer will first generate 3 blue cards.
 - *If you passed the IQ test*, it will generate two cards saying that you passed, and one card saying that you failed the IQ test.
 - *If you failed the IQ test*, it will generate two cards saying that you failed, and one card saying that you passed the IQ test.
7. Out of these **3 cards**, the computer will **randomly pick one**, flip it over and **show it to you.**

Remember that it is **twice as likely** that the **card will say the truth**, than that the card will be fake.

Instructions before eliciting posterior beliefs

Assessment tasks 3 & 4

1. Now that you have seen your card, there will be **two more assessment tasks**, very similar to the ones you just completed.

You will now have the opportunity to **adjust your assessments**.
2. This time, the **slider bar** will **initially** be placed at the **position** of your **previous assessment** of a given situation.

For example, we asked you before how likely (out of 100) you think it is that you passed the first IQ test. If the value you reported before was 50, for instance, then the slider bar will automatically be placed at 50 at the beginning of the new corresponding assessment task.
3. But now that you have seen the card, you **can move the slider bar around** as you like, and **make adjustments** to your assessments.

Importantly, make sure that before you click on the blue arrow sign, you bring the slider bar in the position that represents your **new assessment** of a given situation.
4. You will **maximize your chance of winning \$20** by always reporting your **true assessment** with the bar.

Instructions before practice BDM

Two Options

1. In the remainder of the experiment, you will always have two options: **Option A** and **Option B**. We are then going to ask you a list of questions similar to this one:
2. In each question, you pick either **Option A** (take **Path A** and earn some amount called **Earn_A**), or **Option B** (take **Path B**). We will later explain to you what each of these mean.
3. After you answer all 23 questions, the computer will randomly draw one of them. Each question is equally likely to be drawn.
 - If you chose **Option A** in the question that is drawn, you will take **Path A** and earn the indicated amount, **Earn_A**.
 - If you chose **Option B** in that question, you will take **Path B**.
4. **Make sure to answer all questions truthfully**, so that you always end up with the option you like better, no matter which question gets drawn.

Q#		Option A		Option B
1	Would you rather take...	Path A with <i>Earn_A</i> =\$22	or	Path B ?
2	Would you rather take...	Path A with <i>Earn_A</i> =\$21	or	Path B ?
3	Would you rather take...	Path A with <i>Earn_A</i> =\$20	or	Path B ?
4	Would you rather take...	Path A with <i>Earn_A</i> =\$19	or	Path B ?
.	.	.		.
.	.	.		.
.	.	.		.
20	Would you rather take...	Path A with <i>Earn_A</i> =\$3	or	Path B ?
21	Would you rather take...	Path A with <i>Earn_A</i> =\$2	or	Path B ?
22	Would you rather take...	Path A with <i>Earn_A</i> =\$1	or	Path B ?
23	Would you rather take...	Path A with <i>Earn_A</i> =\$0	or	Path B ?

5. We assume you will choose **Option A** in the first few questions, but at some point will switch to choosing **Option B**. So, to save time, we will simply ask you at which dollar value you would like to switch.
6. The computer will then “fill out” your answers to all 23 questions based on your **switch point** (choosing **Option A** for all questions before or at your switch point, and **Option B** for all questions after your switch point).
7. To report your switch point, there will be a slider bar with numbers from 0 to 22, that will initially be placed at the value of \$0. To **report your switch point**, **move the slider around** with your mouse or arrow keys.

Q #		Option A		Option B
1	Would you rather take...	Path A with <i>Earn_A</i> =\$22	or	Path B ?
2	Would you rather take...	Path A with <i>Earn_A</i> =\$21	or	Path B ?
3	Would you rather take...	Path A with <i>Earn_A</i> =\$20	or	Path B ?
4	Would you rather take...	Path A with <i>Earn_A</i> =\$19	or	Path B ?
.	.	.		.
.	.	.		.
.	.	.		.
20	Would you rather take...	Path A with <i>Earn_A</i> =\$3	or	Path B ?
21	Would you rather take...	Path A with <i>Earn_A</i> =\$2	or	Path B ?
22	Would you rather take...	Path A with <i>Earn_A</i> =\$1	or	Path B ?
23	Would you rather take...	Path A with <i>Earn_A</i> =\$0	or	Path B ?

8. Here's an example:
Suppose you prefer to take **Path A**, but only if *Earn_A* is at least \$6. Otherwise, you would rather take **Path B**.

In this case, the switch point you report should be **\$6**, as shown on the screenshot.

Example.

Your switch point: \$6

This means:


- You choose **Option A** if *Earn_A* is \$6 or more.
- You choose **Option B** if *Earn_A* is less than \$6.

Q #		Option A		Option B
1	Would you rather take...	Path A with <i>Earn_A</i> =\$22	or	Path B ?
2	Would you rather take...	Path A with <i>Earn_A</i> =\$21	or	Path B ?
3	Would you rather take...	Path A with <i>Earn_A</i> =\$20	or	Path B ?
4	Would you rather take...	Path A with <i>Earn_A</i> =\$19	or	Path B ?
.	.	.		.
.	.	.		.
.	.	.		.
20	Would you rather take...	Path A with <i>Earn_A</i> =\$3	or	Path B ?
21	Would you rather take...	Path A with <i>Earn_A</i> =\$2	or	Path B ?
22	Would you rather take...	Path A with <i>Earn_A</i> =\$1	or	Path B ?
23	Would you rather take...	Path A with <i>Earn_A</i> =\$0	or	Path B ?

8. *Here's an example:*
 Suppose you prefer to take **Path A**, but only if **Earn_A** is at least \$6. Otherwise, you would rather take **Path B**.

In this case, the switch point you report should be **\$6**, as shown on the screenshot.

Example.
 Your switch point: \$6

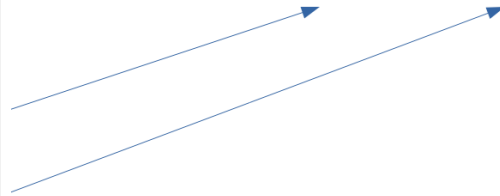


This means:


- You choose **Option A** if **Earn_A** is \$6 or more.
- You choose **Option B** if **Earn_A** is less than \$6.

Q#	Option A	or	Option B
1	Path A with Earn_A=\$22	or	Path B ?
2	Path A with Earn_A=\$21	or	Path B ?
3	Path A with Earn_A=\$20	or	Path B ?
4	Path A with Earn_A=\$19	or	Path B ?
5	Path A with Earn_A=\$18	or	Path B ?
6	Path A with Earn_A=\$17	or	Path B ?
7	Path A with Earn_A=\$16	or	Path B ?
8	Path A with Earn_A=\$15	or	Path B ?
9	Path A with Earn_A=\$14	or	Path B ?
10	Path A with Earn_A=\$13	or	Path B ?
11	Path A with Earn_A=\$12	or	Path B ?
12	Path A with Earn_A=\$11	or	Path B ?
13	Path A with Earn_A=\$10	or	Path B ?
14	Path A with Earn_A=\$9	or	Path B ?
15	Path A with Earn_A=\$8	or	Path B ?
16	Path A with Earn_A=\$7	or	Path B ?
17	Path A with Earn_A=\$6	or	Path B ?
18	Path A with Earn_A=\$5	or	Path B ?
19	Path A with Earn_A=\$4	or	Path B ?
20	Path A with Earn_A=\$3	or	Path B ?
21	Path A with Earn_A=\$2	or	Path B ?
22	Path A with Earn_A=\$1	or	Path B ?
23	Path A with Earn_A=\$0	or	Path B ?

9. To see what this means, let's take a look at the expanded table.
- (1) If a question with an **Earn_A** of \$6 or higher gets drawn, you **get Option A**: You take Path A and get the indicated fixed payment (which is at least \$6).
- (2) If a question with an **Earn_A** less than \$6 gets drawn, you **get Option B**: You take Path B and get no fixed payment.



Example.
 Your switch point: \$6



This means:

- You choose **Option A** if **Earn_A** is \$6 or more.
- You choose **Option B** if **Earn_A** is less than \$6.

Q#	Option A	or	Option B
1	Path A with Earn_A=\$22	or	Path B ?
2	Path A with Earn_A=\$21	or	Path B ?
3	Path A with Earn_A=\$20	or	Path B ?
4	Path A with Earn_A=\$19	or	Path B ?
5	Path A with Earn_A=\$18	or	Path B ?
6	Path A with Earn_A=\$17	or	Path B ?
7	Path A with Earn_A=\$16	or	Path B ?
8	Path A with Earn_A=\$15	or	Path B ?
9	Path A with Earn_A=\$14	or	Path B ?
10	Path A with Earn_A=\$13	or	Path B ?
11	Path A with Earn_A=\$12	or	Path B ?
12	Path A with Earn_A=\$11	or	Path B ?
13	Path A with Earn_A=\$10	or	Path B ?
14	Path A with Earn_A=\$9	or	Path B ?
15	Path A with Earn_A=\$8	or	Path B ?
16	Path A with Earn_A=\$7	or	Path B ?
17	Path A with Earn_A=\$6	or	Path B ?
18	Path A with Earn_A=\$5	or	Path B ?
19	Path A with Earn_A=\$4	or	Path B ?
20	Path A with Earn_A=\$3	or	Path B ?
21	Path A with Earn_A=\$2	or	Path B ?
22	Path A with Earn_A=\$1	or	Path B ?
23	Path A with Earn_A=\$0	or	Path B ?

10. You will have noticed that the following:
- The **lower** your switch point, the more likely it is you get **Option A**.
 - The **higher** your switch point, the more likely it is that you get **Option B**.
11. Now ask yourself: If you report a switch point of \$6, for example, does this guarantee you that you get paid at least \$6?
 The answer is **no**: It also matters which question gets drawn.
- If a question with an **Earn_A** of \$6 or more is drawn, you indeed get a fixed payment of at least \$6.
 - If a question with an **Earn_A** of less than \$6 is drawn, however, you take **Path B** and don't get a fixed payment.

12. You will now have the opportunity to familiarize yourself with how the slider works.

Instructions before main decision (continue/quit) - *Baseline* treatment

Part 3

In this part of the experiment, you have the following two options:

*You can either **continue** or **quit**.*

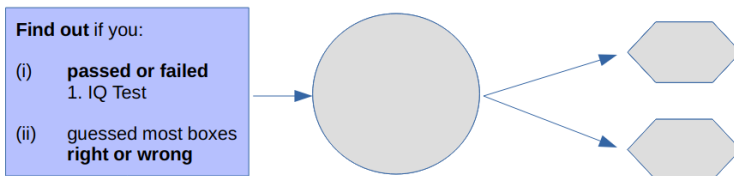
Each option consists of three steps, which we will now explain in detail.

We will now highlight what is different if you **continue** or **quit**.

No matter which path you take, you will spend the **same time** in the experiment.

Quitting does **not** mean that you finish earlier.

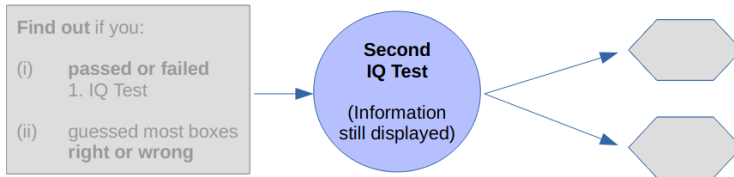
Continue



Step 1:

If you continue, you will find out whether you passed or failed the first IQ test, and whether you guessed most boxes right or wrong in the guessing game.

Continue

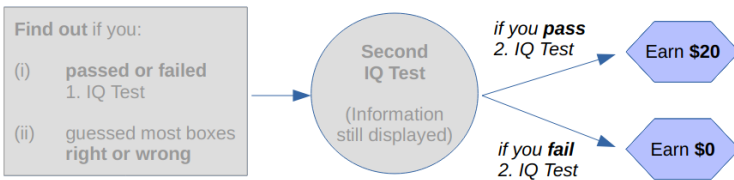


Step 2:

You will then be asked to take a second IQ test.

While you take the second IQ test, the information from Step 1 will be displayed next to each question.

Continue

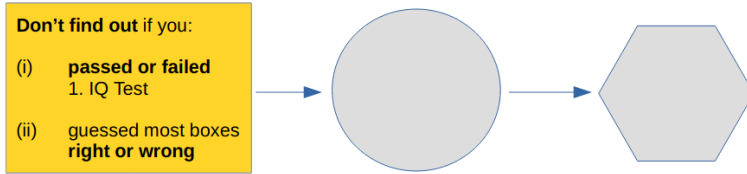


Step 3:

If you pass the second IQ test, you earn \$20.

If you fail the second IQ test, you earn \$0.

Quit



Step 1:

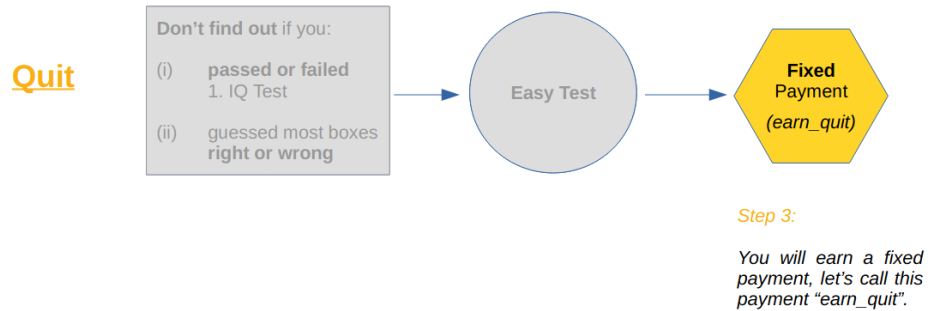
*If you quit, you **don't** find out whether you passed or failed the IQ test, and whether you guessed most boxes right or wrong.*

Quit



Step 2:

You will then be asked to take a very easy test.



Difference 1: The test you will take

Quit

Easy Test

If you *quit*, you will **take an easy test**. This test will be of a similar style, but **much easier** than the IQ test you took at the beginning.

- There will be 7 questions, and you will have 90 seconds to answer one question.
- It is very likely that you **can solve all questions** of this easy test.
- You cannot "pass" or "fail" the easy test, and you won't get any direct feedback on how well you did.

Continue

Second IQ Test

If you *continue*, you will take a **second IQ test**. This test will be of a similar style, and the **level of difficulty will be similar** to the IQ test you took at the beginning.

- There will be 7 questions, and you will have 90 seconds to answer one question.
- You will **pass** this IQ test if you **solve at least 5 questions correctly**. Otherwise, you fail.
- Whether you pass or fail does not depend on the performance of other participants.
- You will not get any direct feedback on whether you passed or failed the second test.

Difference 2: Information on your performance

Don't find out if you:

- (i) **passed or failed** 1. IQ Test
- (ii) guessed most boxes **right or wrong**

Quit

If you *quit*, nobody will tell you if you really passed or failed the first IQ test, and nobody will tell you if you guessed most boxes right or wrong.

We will just never mention your past performance again.

Find out if you:

- (i) **passed or failed** 1. IQ Test
- (ii) guessed most boxes **right or wrong**

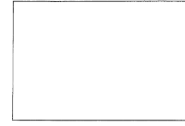
Continue

If you *continue*, you will find out

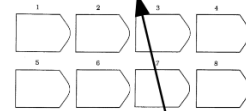
- (i) if you **passed or failed** the first IQ test, and
- (ii) if you guessed most boxes **right or wrong**

before you take the second IQ test.

In addition, **while** you take the second IQ test, a **reminder** with this information will be displayed at the center of every question.



- Reminder:
- **passed** or **failed** 1. IQ test,
 - guessed most boxes **right/wrong**

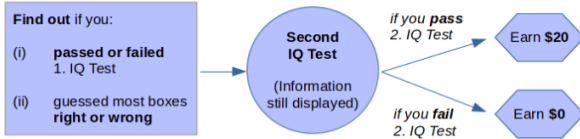


What's next

1. In what is next, we will show you the list of questions, and will ask you to report your switch point - the dollar value after which you would like to switch from *Option A* to *Option B*.
2. Given your switch point, the computer knows for which dollar values of *earn_quit* you prefer to *quit*, and for which values of *earn_quit* you prefer to *continue*.
3. It is important that you report your **true switch point**, so that you end up with the option you like more in each case.

Q#		Option A		Option B	
1	Would you rather...	quit with <i>earn_quit</i> =\$22	or	continue	?
2	Would you rather...	quit with <i>earn_quit</i> =\$21	or	continue	?
3	Would you rather...	quit with <i>earn_quit</i> =\$20	or	continue	?
4	Would you rather...	quit with <i>earn_quit</i> =\$19	or	continue	?
.
.
.
20	Would you rather...	quit with <i>earn_quit</i> =\$3	or	continue	?
21	Would you rather...	quit with <i>earn_quit</i> =\$2	or	continue	?
22	Would you rather...	quit with <i>earn_quit</i> =\$1	or	continue	?
23	Would you rather...	quit with <i>earn_quit</i> =\$0	or	continue	?

Continue



Quit



Q#	Option A		Option B
1	Would you rather... quit with $earn_quit = \$22$	or	continue ?
2	Would you rather... quit with $earn_quit = \$21$	or	continue ?
3	Would you rather... quit with $earn_quit = \$20$	or	continue ?
4	Would you rather... quit with $earn_quit = \$19$	or	continue ?
.	.		.
.	.		.
20	Would you rather... quit with $earn_quit = \$3$	or	continue ?
21	Would you rather... quit with $earn_quit = \$2$	or	continue ?
22	Would you rather... quit with $earn_quit = \$1$	or	continue ?
23	Would you rather... quit with $earn_quit = \$0$	or	continue ?

Here is an example:

- Suppose you would rather quit, but only if the fixed payment you get for quitting, $earn_quit$, is at least \$11. For lower values of $earn_quit$, you would rather continue.
- In this case, you should report a switch point of \$11. This means that:
 - If a question with an $earn_quit$ of \$11 or higher is drawn, the computer will make sure you quit.
 - If a question with an $earn_quit$ of less than \$11 is drawn, the computer will make sure you continue.

Example.

Your switch point: \$11

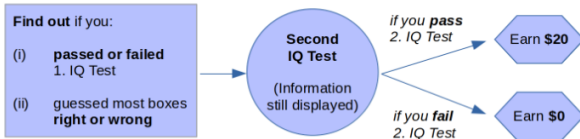
This means:

- You choose to quit if $earn_quit$ is \$11 or more.
- You choose to continue if $earn_quit$ is less than \$11.

Remember that reporting a switch point of \$11, for example, does not guarantee that you get paid at least \$11 in this part of the experiment. It also depends on which question gets drawn!



Continue



Quit



Q#	Option A		Option B
1	Would you rather... quit with $earn_quit = \$22$	or	continue ?
2	Would you rather... quit with $earn_quit = \$21$	or	continue ?
3	Would you rather... quit with $earn_quit = \$20$	or	continue ?
4	Would you rather... quit with $earn_quit = \$19$	or	continue ?
.	.		.
.	.		.
20	Would you rather... quit with $earn_quit = \$3$	or	continue ?
21	Would you rather... quit with $earn_quit = \$2$	or	continue ?
22	Would you rather... quit with $earn_quit = \$1$	or	continue ?
23	Would you rather... quit with $earn_quit = \$0$	or	continue ?

Make sure the slider is in the correct position before you move on.

Here are two more examples of how your slider and switch point could look like.

Example.

Your switch point: \$19

This means:

- You choose to quit if $earn_quit$ is \$19 or more.
- You choose to continue if $earn_quit$ is less than \$19.



Example.

Your switch point: \$3

This means:

- You choose to quit if $earn_quit$ is \$3 or more.
- You choose to continue if $earn_quit$ is less than \$3.



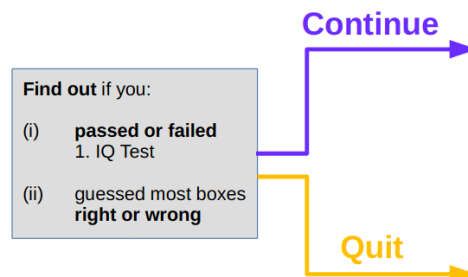
Instructions before main decision (continue/quit) - *AlwaysInfo* treatment

(Only instructions that differ across treatments are displayed here.)

In this part of the experiment, you have the following two options:

*You can either **continue** or **quit**.*

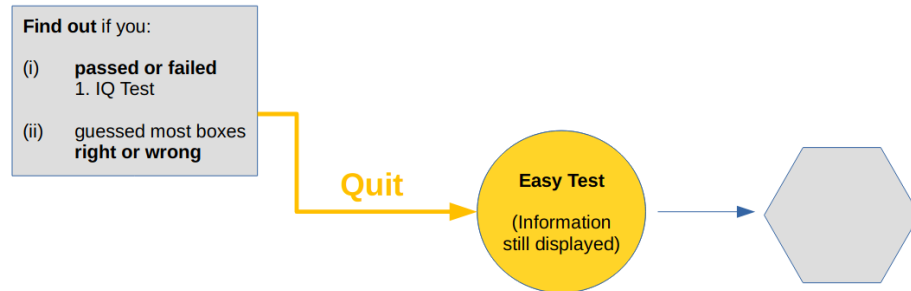
Each option consists of three steps, which we will now explain in detail. Step 1 does not depend on whether you continue or quit.



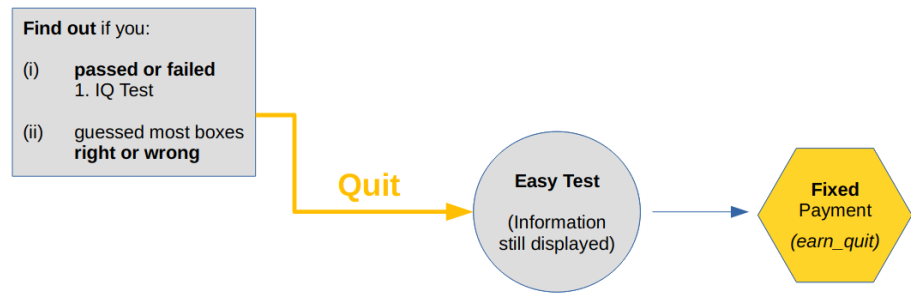
Step 1:

You will find out whether you passed or failed the first IQ test, and whether you guessed most boxes right or wrong in the guessing game.

What happens in Step 2 and 3 will depend on whether you continue or quit.



Step 2:
 If you quit, you will be asked to take a very easy test.
 While you take the easy test, the information from Step 1 will be displayed next to each question.

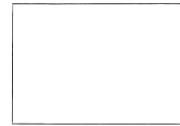


Step 3:
 You will earn a fixed payment, let's call this payment "earn_quit".

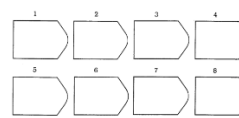
Remember that the information from Step 1 – whether you passed or failed the first IQ test, and whether you guessed most boxes right or wrong – will be displayed while you take the next test.

If you *continue*, this information will be displayed next to each question of the second IQ test.

If you *quit*, this information will be displayed next to each question of the easy test.



- Reminder:
- passed or failed 1. IQ test,
 - guessed most boxes right/wrong



Instructions before risk task

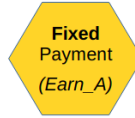
Part 4

Two options

You have two options in Part 4.

Option A:

Option A means that you get a fixed payment, called $Earn_A$.

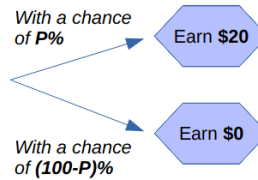


Option B:

Option B means that you get the following lottery:

- > With a chance of $P\%$, you earn \$20.
- > But with a chance of $(100-P)\%$, you earn \$0.

P will be replaced with an actual number in the experiment.



Important:

P is **not** just some random draw between 0 and 100. P is a **fixed number**, and you will see what P is before you report your switch point.

This means that you know exactly with which probability the lottery pays you \$20 and \$0 before you decide between Option A and Option B.

Q#	Option A	Option B
1	Would you rather get... a fixed payment of $Earn_A=\$22$	or the lottery ?
2	Would you rather get... a fixed payment of $Earn_A=\$21$	or the lottery ?
3	Would you rather get... a fixed payment of $Earn_A=\$20$	or the lottery ?
4	Would you rather get... a fixed payment of $Earn_A=\$19$	or the lottery ?
.	.	.
20	Would you rather get... a fixed payment of $Earn_A=\$3$	or the lottery ?
21	Would you rather get... a fixed payment of $Earn_A=\$2$	or the lottery ?
22	Would you rather get... a fixed payment of $Earn_A=\$1$	or the lottery ?
23	Would you rather get... a fixed payment of $Earn_A=\$0$	or the lottery ?

Just as before, we will show you the list of questions, and we will ask you to report at which dollar value you would like to switch from Option A to Option B.

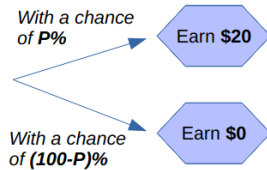
Given your switch point, the computer can "fill out" your answers to all questions.

- If you chose Option A in the question that is drawn for payment, you get a fixed payment of some $Earn_A$.
- If you chose Option B in the question that is drawn for payment, you get a lottery where you earn \$20 with a chance of $P\%$, and \$0 with a chance of $(100-P)\%$.

Option A



Option B



Finally, make sure the slider is in the correct position before you move on. Here are some examples of how your slider and switch point could look like.

The image shows three vertically stacked example screenshots of the experimental interface. Each screenshot contains the following text:

- Example.**
- Your switch point: \$19 (top), \$11 (middle), and \$3 (bottom).
- This means:
- You choose the **fixed payment** if *Earn_A* is \$19 or more. (top), \$11 or more. (middle), and \$3 or more. (bottom).
- You choose the **lottery** if *Earn_A* is less than \$19. (top), \$11. (middle), and \$3. (bottom).

Each screenshot also features a vertical slider on the right side, with a blue bar at the bottom and a yellow bar at the top. The slider's position corresponds to the switch point value: approximately 80% yellow for \$19, 30% yellow for \$11, and 10% yellow for \$3.

A.3.2 Experimental Interface

This section displays example screenshots of the *Baseline* treatment of the experiment. Horizontal lines indicate page breaks in the interface. Whenever subjects encountered a screen saying “Please wait for further instructions,” a new set of instructions was displayed on the screen and read out loud by the experimenter, see Section A.3.1. After that, subjects received the password to move on. Raven’s matrices were taken from test number 844 of the *Advanced Progressive Matrices Set II* (Raven, 1973).

Please wait for further instructions.



(Instructions at the very beginning of the experiment here.)

Quick comprehension quiz

Which statements are correct? (Click on all that apply.)

I will get a bonus payment if I can guess all three boxes correctly.

How well I guess will not impact my earnings in this experiment.

The guessing game is just about luck, not about skill.

There will be three boxes in this guessing game.



Guess which **three boxes** contain a ball.



Please wait for further instructions.



(Instructions before first IQ test here.)

PART 1/4
of the experiment begins now.



IQ Test.

Remember:

- This test is frequently used to **measure intelligence**.
- You **pass** if you can solve **at least 5 out of 7 questions correctly**.
- **Otherwise**, you **fail**.

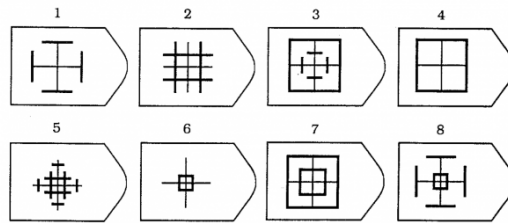
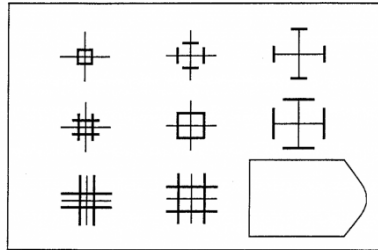
If you are ready to take the IQ test, click on the blue arrow sign below.



IQ Test.

01 11

Question 1/7



Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

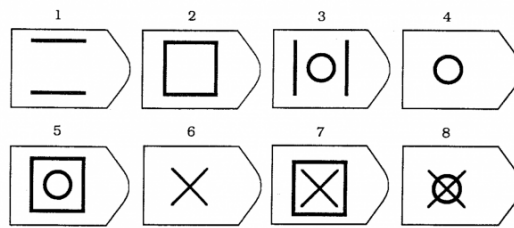
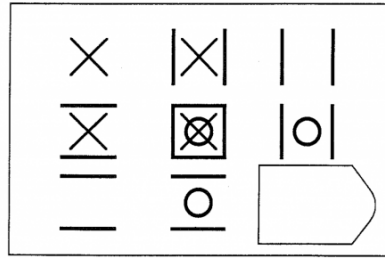
01 11



IQ Test.

0109

Question 2/7



Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

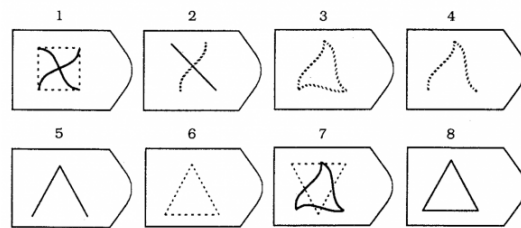
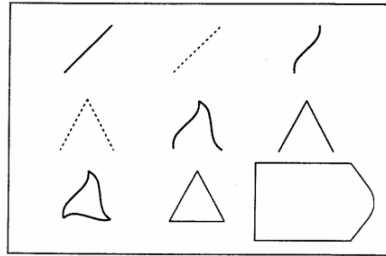
0109



IQ Test.

0104

Question 3/7



Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

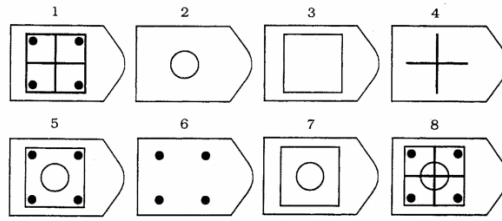
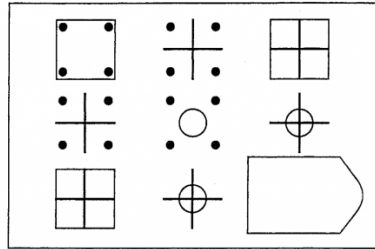
0104



IQ Test.

0121

Question 4/7



Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

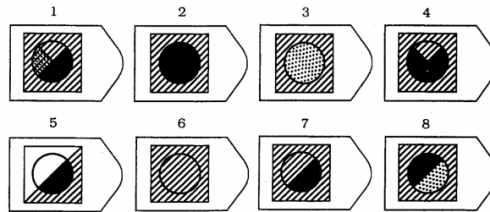
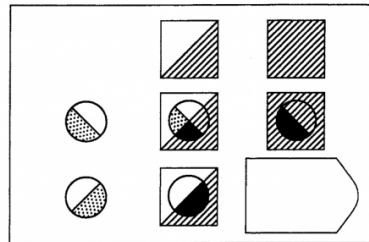
0121



IQ Test.

0119

Question 5/7



Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

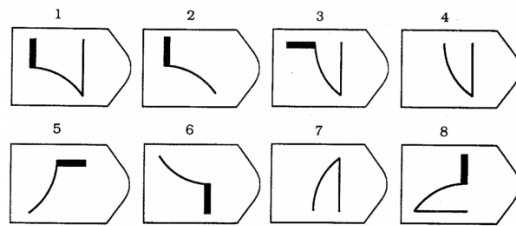
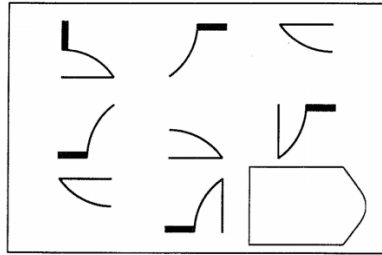
0119



IQ Test.

0121

Question 6/7



Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

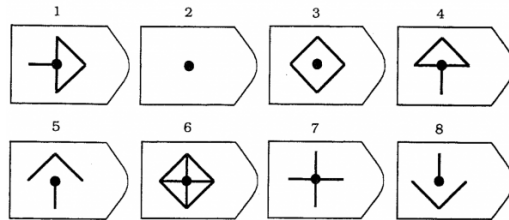
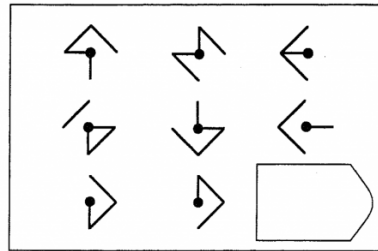
0121



IQ Test.

0108

Question 7/7



Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

0108



Please wait for further instructions.



(Instructions before prior beliefs here.)

Assessment task 1/4

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

(You passed if you answered at least 5/7 questions correctly.)

0

100

Move the bar to make your assessment.



Relevant information

Later in this experiment, you might be asked to take another IQ test.

Let's call this the **future IQ test**.

- It would consist of similar questions, and have a **similar level of difficulty**.
- You would again have **90 seconds** to answer one question.
- You would again **pass** if **you** can solve at least **5 out of 7 questions correctly**.

Before taking the **future IQ test**, you would see if you passed or failed the first IQ test, and if you guessed most boxes right or wrong. This information would still be visible *while* you take the test.



Assessment task 2/4

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

(You pass if you answer at least 5/7 questions correctly.)

0

100

Move the bar to make your assessment.



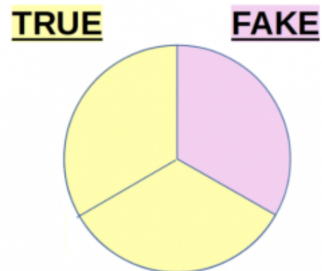
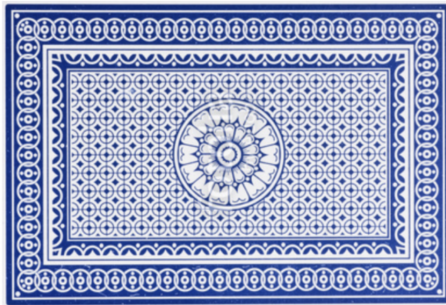
Please wait for further instructions.

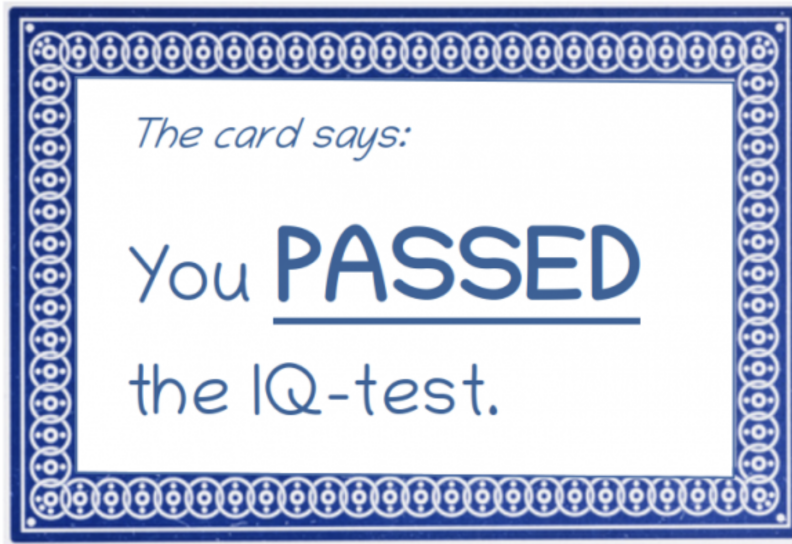


(Instructions before feedback here.)

Cards

The computer has drawn your card. Remember that it is **twice as likely** that the **card will say the truth**, than that the card will be fake. Click the blue arrow symbol to **flip the card over**.





Which statement is correct?

The card said I PASSED the IQ-test.

The card said I FAILED the IQ-test.

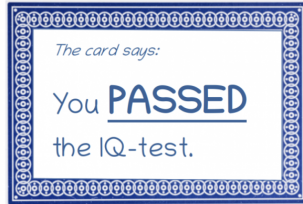


Now that you have seen what your cards says, you can adjust your assessments from before.



Assessment task 3/4

Now that you have seen your card:



How likely (out of 100) do you think it is that you **passed** the **IQ-test**?
(You passed if you answered at least 5/7 questions correctly.)

0

100

Move the bar to make your assessment.



Relevant information

As you know, later in this experiment, you might be asked to take another IQ test. Let's call this the **future IQ test**.

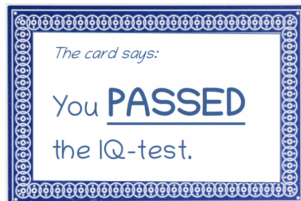
- It would consist of similar questions, and have a **similar level of difficulty**.
- You would again have **90 seconds** to answer one question.
- You would again **pass** if **you** can solve at least **5 out of 7 questions correctly**.

Before taking the **future IQ test**, you would see if you passed or failed the first IQ test, and if you guessed most boxes right or wrong. This information would still be visible *while* you take the test.



Assessment task 4/4

Now that you have seen your card:



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

(You pass if you answer at least 5/7 questions correctly.)

0

100

Move the bar to make your assessment.



Please wait for further instructions.



(Instructions before practice BDM here.)

This is just a test.

Your decision below does **not** count yet.

Q#		Option A		Option B	
1	Would you rather take...	Path A with $Earn_A = \$22$	or	Path B	?
2	Would you rather take...	Path A with $Earn_A = \$21$	or	Path B	?
3	Would you rather take...	Path A with $Earn_A = \$20$	or	Path B	?
4	Would you rather take...	Path A with $Earn_A = \$19$	or	Path B	?
.	.	.		.	
.	.	.		.	
.	.	.		.	
20	Would you rather take...	Path A with $Earn_A = \$3$	or	Path B	?
21	Would you rather take...	Path A with $Earn_A = \$2$	or	Path B	?
22	Would you rather take...	Path A with $Earn_A = \$1$	or	Path B	?
23	Would you rather take...	Path A with $Earn_A = \$0$	or	Path B	?

Move the slider bar around to familiarize yourself with it.

Test: Your switch point: \$8

This means:

- You choose to take **Path A** if $Earn_A$ is \$8 or more.
- You choose to take **Path B** if $Earn_A$ is less than \$8.



Once you are familiar with it, move on.



Please wait for further instructions.



(Instructions before main decision (continue/quit) here.)

PART 3/4
of the experiment begins now.



Quick comprehension quiz**Which statement is correct?**

If I continue, I earn \$20, but only if I passed the first IQ test.

If I continue and fail the second test, I don't earn anything.

If I continue, I earn \$20 if I passed at least one IQ test - either the first or the second test.

Which statement is correct?

If I quit, I will finish earlier.

If I quit, I won't find out if I really passed or failed the first IQ test.

If I quit, I get a fixed payment, but only if I pass the easy test.

Which statement is correct?

No matter if I continue or quit,

... I will have to take another IQ test.

... I will learn if I passed or failed the first IQ test.

... I will take a test with 7 more questions.

Which statement is correct?

If I quit, how much I earn will also depend on which question from the list is drawn.

The switch point I report is just hypothetical and will not affect my payments.

If I learn that I guessed most boxes correctly, this will increase my earnings.

Which statement is **correct**?

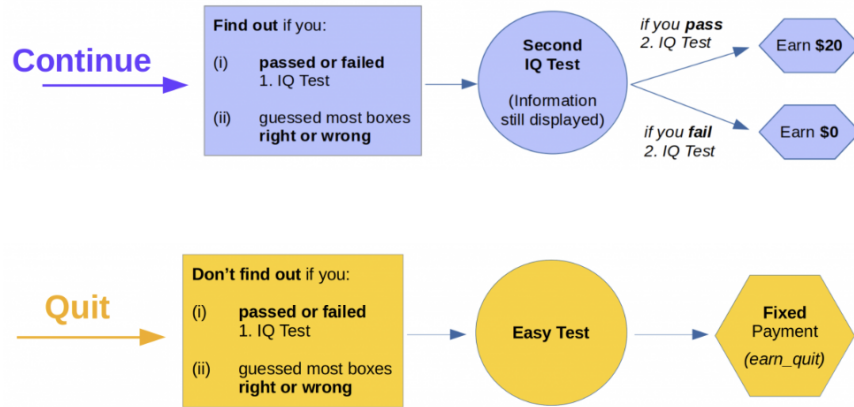
The lower the switch point of `earn_quit` that I report, the more likely it is that I continue.

If I report a switch point of `$X`, I will earn at least `$X`.

If I report a switch point of `$X`, I will earn at least `$X` if a question of an `earn_quit` of `X` or more is drawn. Otherwise, I continue, with no fixed payment.



Q#		Option A		Option B	
1	Would you rather...	quit with <code>earn_quit=\$22</code>	or	continue	?
2	Would you rather...	quit with <code>earn_quit=\$21</code>	or	continue	?
3	Would you rather...	quit with <code>earn_quit=\$20</code>	or	continue	?
4	Would you rather...	quit with <code>earn_quit=\$19</code>	or	continue	?
.	.	.		.	
.	.	.		.	
.	.	.		.	
20	Would you rather...	quit with <code>earn_quit=\$3</code>	or	continue	?
21	Would you rather...	quit with <code>earn_quit=\$2</code>	or	continue	?
22	Would you rather...	quit with <code>earn_quit=\$1</code>	or	continue	?
23	Would you rather...	quit with <code>earn_quit=\$0</code>	or	continue	?



Move the slider to report your **switch point**.

Your switch point: \$17

This means:

- You choose to **quit** if *earn_quit* is \$17 or more.
- You choose to **continue** if *earn_quit* is less than \$17.



If you move on, you finalize your **switch point** to be **\$17**.



Continue.

Given your reported switch point, and the question that was drawn,
you **continue**.



This is your **actual, true performance** on the first IQ-test:

You failed the first IQ-test.

This is your **actual, true performance** on the guessing game:

You guessed most boxes wrong.

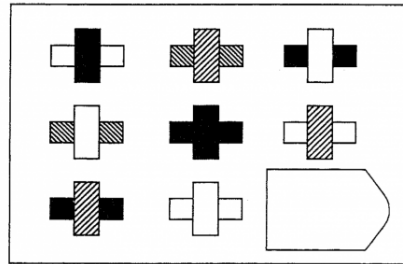
You will now take the second IQ test.



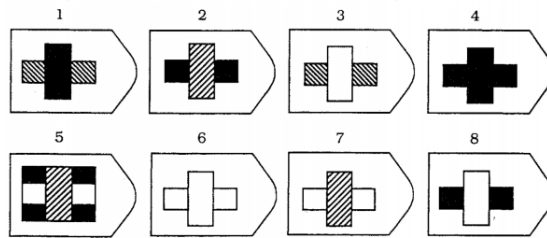
IQ Test.

0114

Question 1/7



You failed the first IQ-test.
You guessed most boxes wrong.



Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

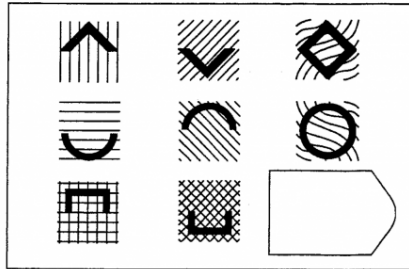
0114



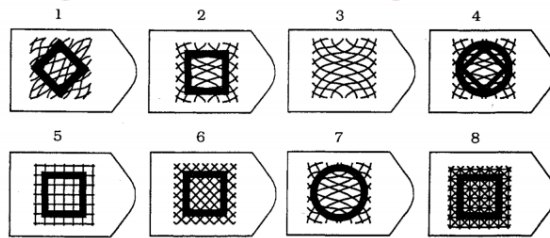
IQ Test.

0107

Question 2/7



You failed the first IQ-test.
You guessed most boxes wrong.



Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

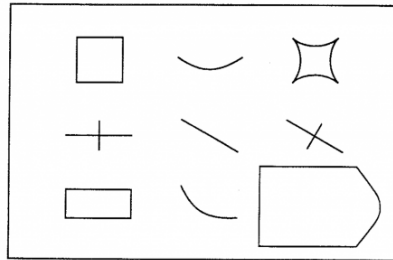
0121



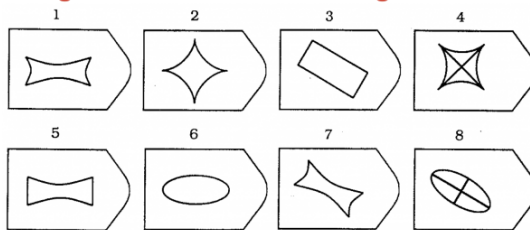
IQ Test.

0112

Question 3/7



**You failed the first IQ-test.
You guessed most boxes wrong.**



Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

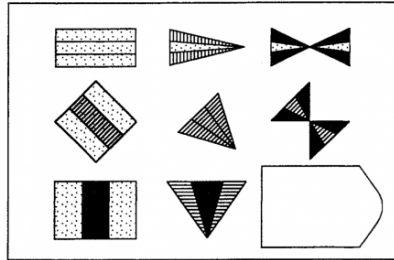
0112



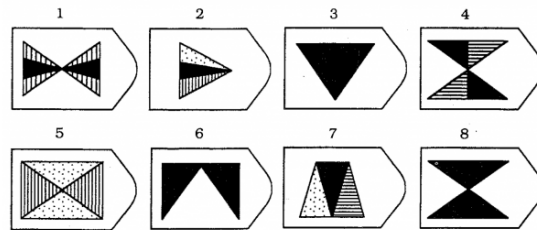
IQ Test.

0122

Question 4/7



You failed the first IQ-test.
You guessed most boxes wrong.



Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

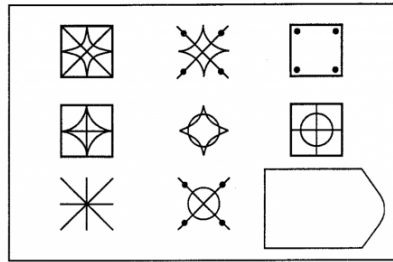
0122



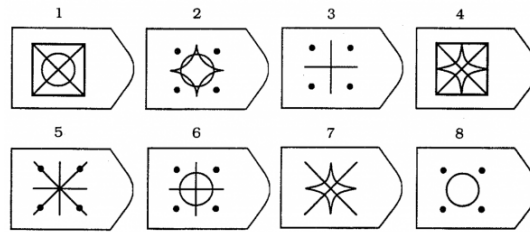
IQ Test.

0122

Question 5/7



You failed the first IQ-test.
You guessed most boxes wrong.



Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

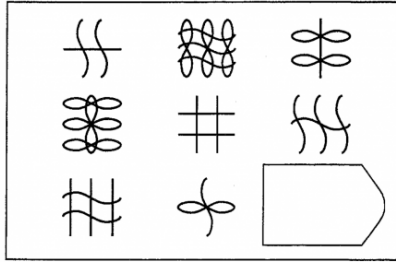
0122



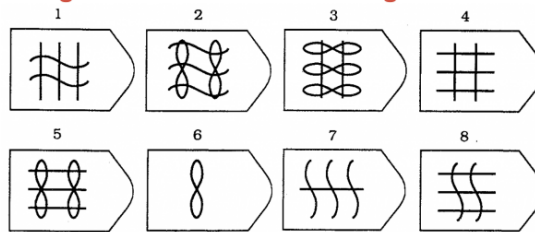
IQ Test.

0121

Question 6/7



You failed the first IQ-test.
You guessed most boxes wrong.



Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

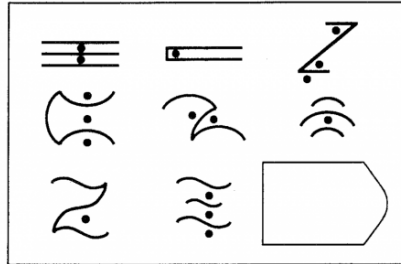
0121



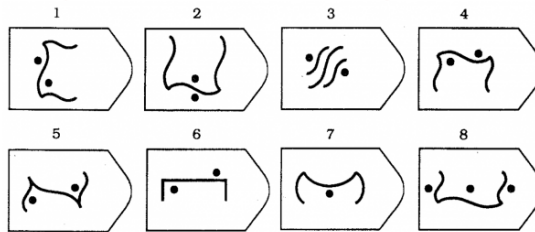
IQ Test.

0123

Question 7/7



You failed the first IQ-test.
You guessed most boxes wrong.



Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

0123



Quit.

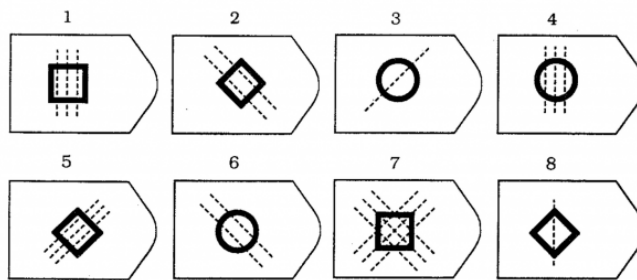
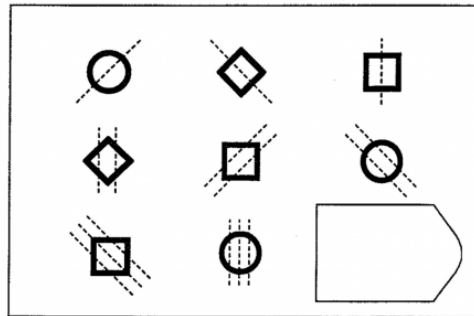
Given your reported switch point, and the question that was drawn,

you **quit**.



0112

Question 1/7

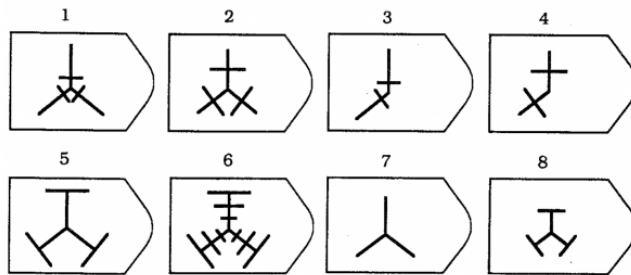
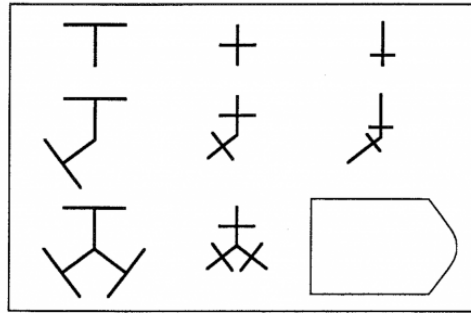


Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

0118

Question 2/7

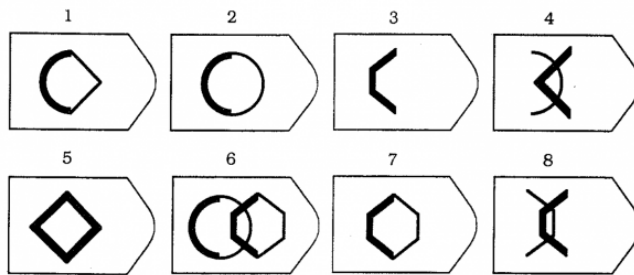
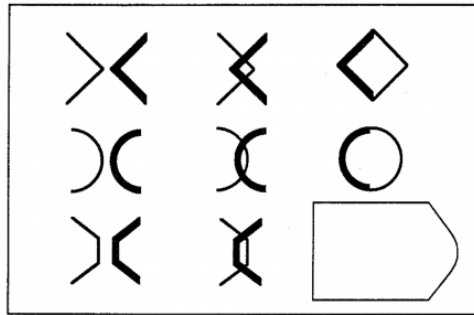


Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

0122

Question 3/7

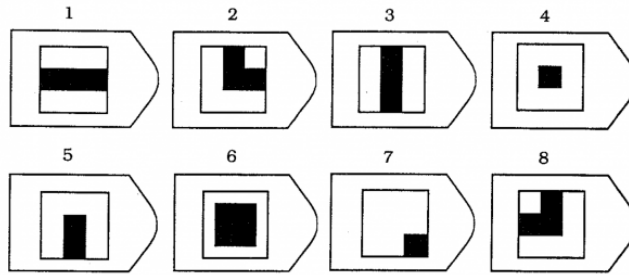
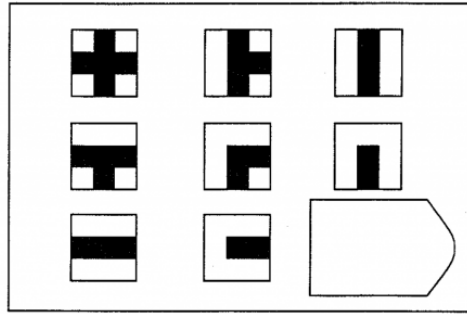


Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

0119

Question 4/7

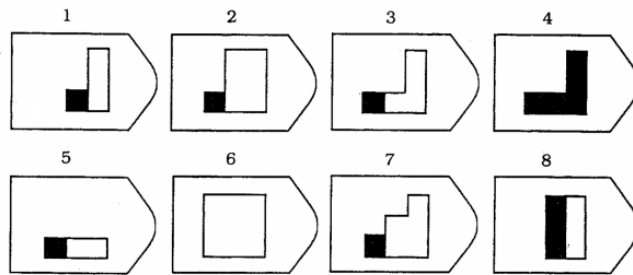
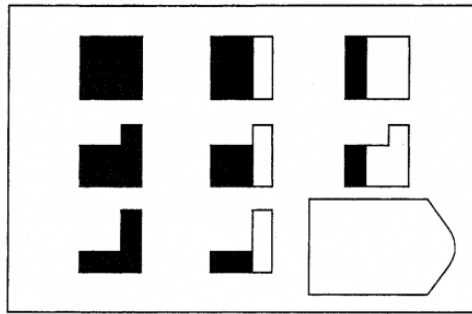


Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

0114

Question 5/7

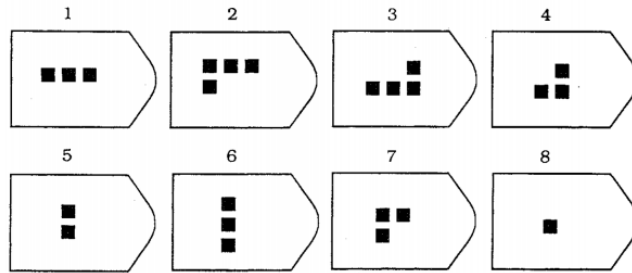
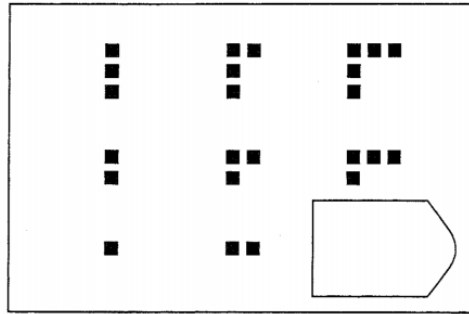


Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

0110

Question 6/7

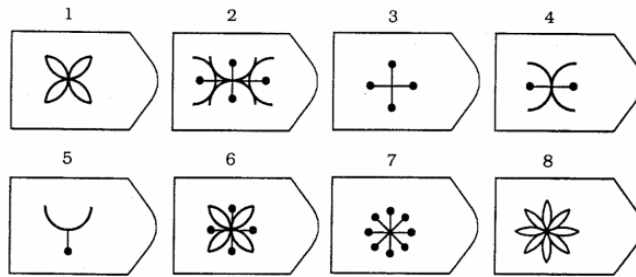
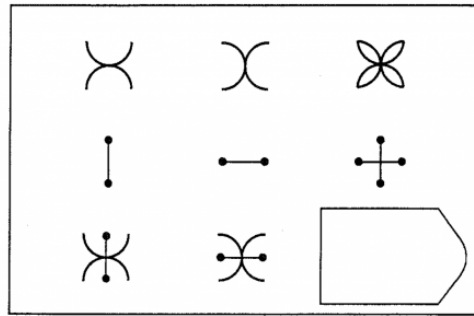


Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

0123

Question 7/7



Which piece completes the pattern?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

Please wait for further instructions.



(Instructions before risk task here.)

PART 4/4
of the experiment begins now.



Quick comprehension quiz**Which statement is correct?**

If I take the lottery, I will earn either \$20 or nothing.

If I take the fixed payment, I will earn \$20 for sure.

Which statement is correct?

If I take the lottery, I will earn \$20 with a chance of P , where P is some random draw between 0 and 100.

If I take the lottery, I will earn \$20 with a chance of P , where P is a fixed number. I will see what P is **before** I report my switch point.

Which statement is correct?

If I report a switch point of $\$V$, I am guaranteed a payment of at least $\$V$ in this part of the experiment, no matter which question gets drawn.

If I report a switch point of $\$V$, I am guaranteed a payment of at least $\$V$, **but only if** a question with a dollar amount of V or higher is chosen. Otherwise, I get the lottery.



Q#		Option A		Option B
1	Would you rather get...	a fixed payment of $Earn_A=\$22$	or	the lottery ?
2	Would you rather get...	a fixed payment of $Earn_A=\$21$	or	the lottery ?
3	Would you rather get...	a fixed payment of $Earn_A=\$20$	or	the lottery ?
4	Would you rather get...	a fixed payment of $Earn_A=\$19$	or	the lottery ?
.	.	.		.
.	.	.		.
.	.	.		.
20	Would you rather get...	a fixed payment of $Earn_A=\$3$	or	the lottery ?
21	Would you rather get...	a fixed payment of $Earn_A=\$2$	or	the lottery ?
22	Would you rather get...	a fixed payment of $Earn_A=\$1$	or	the lottery ?
23	Would you rather get...	a fixed payment of $Earn_A=\$0$	or	the lottery ?

Option A:

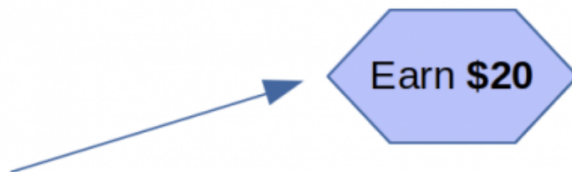
You get a **fixed payment**, $Earn_A$.

**Option B:**

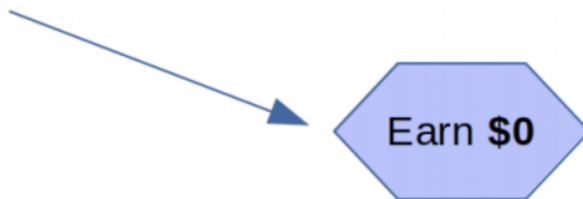
$P = 71\%$

You get the following **lottery**:

A **71%** chance of earning **\$20**,



and a **29%** chance of earning **nothing**.



Move the slider bar to report your **switch point**.

Your switch point: \$6

This means:

- You choose the **fixed payment** if *Earn_A* is \$6 or more.
- You choose the **lottery** if *Earn_A* is less than \$6.



If you move on, you finalize your **switch point** to be **\$6** .



You're almost done!

Please answer the **survey questions** starting at the next page. Then, remain seated until we call your computer ID number. We will pay you after that.

In part 3 of the experiment, why was your **switch point to continue \$17?**

Characters remaining: 500



In part 4 of the experiment, why was your **switch point to take the lottery \$6?**

Characters remaining: 500



What is your gender identity?

- Male
- Female
- Other



Which of the following is closest to your (anticipated) major?

Art, music, literature, dance

Economics, accounting

English, languages, communication

Engineering

Environmental studies, biology

History, geography, global studies

Humanities

Mathematics

Science

Was this major your intended major, or were you initially working towards a different major?

This major has been my initial choice.

I was initially aiming for a different major.



(The following two questions were only displayed if a subject indicated they were initially aiming for a different major in the previous question.)

Which major would have been your initial choice?

Why did you **not** pursue the major that would have been your initial choice?

Characters

remaining: 500

What is your nationality?

United States of America

Other



What is your GPA?



What is your ethnicity?

Black or African-American

Asian or Asian-American

Hispanic

White

Other



Which of the following do you think is true about the IQ test used in this experiment?

On average, **international students performed better** than domestic students.

On average, **international students performed worse** than domestic students.

I **don't think that there is a difference** between the performance of international and domestic students.



Which of the following do you think is true about the IQ test used in this experiment?

On average, **female students performed better** than male students.

On average, **female students performed worse** than male students.

I don't think that there is a difference between the performance of female and male students.



To what extent do you agree or disagree with the following statement?

"My mom's occupation is very typical for women of her generation."

Strongly Agree.

Agree.

Neither Agree nor Disagree.

Disagree.

Strongly Disagree.



To what extent do you agree or disagree with the following statement?

"My dad's occupation is very typical for men of his generation."

Strongly Agree.

Agree.

Neither Agree nor Disagree.

Disagree.

Strongly Disagree.



To what extent do you agree or disagree with the following statement?

"Women should pay their own way on dates."

Strongly Agree.

Agree.

Neither Agree nor Disagree.

Disagree.

Strongly Disagree.



To what extent do you agree or disagree with the following statement?

"A wife with a family has no time for outside employment."

Strongly Agree.

Agree.

Neither Agree nor Disagree.

Disagree.

Strongly Disagree.



When you were a child, how many hours did your **father work for pay** in a typical **week** (approximately)?

When you were a child, how many hours did your **mother work for pay** in a typical **week** (approximately)?



Did you understand the instructions in this experiment?

Yes, everything was clear.

No, I did not understand everything.



(The following question was only displayed if a subject answered “No, ...” to the previous question.)

What was not clear?

Characters

remaining: 500



Is English your first language?

yes

no



Do you have any comments about the experiment? Please let us know.



A.4 Classroom Field Study

With the aim of testing the outside validity of the belief formation patterns discovered in the laboratory, a classroom field study was conducted with Econ 1 students at UC Santa Barbara in the fall quarter of 2021. Econ 1 is usually the first economics class that students take at UCSB. More than half of all students enrolled in Econ 1 are freshmen students, and approximately 25 – 30% of students that complete this course end up majoring in economics. Roughly 45% of Econ 1 students at UCSB are women. All students enrolled in Econ 1 in the 2021 fall quarter were invited to participate in a “short research survey.” An email announcing this study as well as reminder emails were sent out by the course instructor. Students were informed that the purpose of this study was to investigate people’s beliefs about future success. For completing this study (which took students slightly less than 4 minutes on average), they earned 0.5 bonus points that counted towards their final grade in Econ 1, which accounted for roughly 12.5% of the point gap between two letter grades.⁴ In addition, students who completed the survey could earn a \$50 prize by making accurate assessments.⁵ To comply with the human subjects protocol, students were given the option to complete a “research alternative task” to earn the same 0.5 bonus points, which took roughly the same time to complete, and consisted of ten slider tasks. It was pointed out to students in both the announcement emails and the instructions that their Econ 1 instructor and TA were not involved as researchers in this study.

⁴The maximum score students could achieve in this class was 100. There were four midterm exams, each worth up to 15 points, and the three best scores accounted for 45% of a student’s final grade. The point gap between most letter grades in Econ 1 was 4 points, and thus the 0.5 bonus points accounted for roughly 12.5% of the gap between grades.

⁵To award these prizes, the same crossover mechanism as in the main experiment was used (see Section 1.2), however in the interest of keeping the time to participate in the survey as short as possible (and thus increase compliance), the details of this mechanism were not explained to participants. Subjects were informed that they could email the researcher if they had questions about the compensation mechanism, but no inquiries were made.

The classroom study was conducted on October 15, 2021 in the hours following first Econ 1 midterm exam. After finishing the first exam, students received an email with a link to the research survey. Upon clicking on this link, they could opt for either the research survey or the alternative task. Students knew they could complete this survey within a pre-announced time window of a few hours following the first midterm exam, but before learning their exam score. Students opting for the research study had to answer the following two questions, and were reminded that reporting accurate assessments increased their chance of winning a \$50 prize.

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

Note that these questions were kept as similar as possible to the elicitation subjects' beliefs with regard to the first and the future IQ test in the experiment. The survey was conducted after the first midterm quiz so that students had not received any previous performance feedback in the form of midterm quizzes in Econ 1. To mimic the binary pass/fail event of the IQ test, a cutoff of 12 was chosen, approximately matching the average score of previous quarters. As students participated in the survey before learning their actual exam scores, this setting is most similar to the prior beliefs elicited in the experiment. In addition, students were asked to report their race identity and gender identity.

A.5 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

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

Equation A.1 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.⁶ In what follows, risk parameters are computed as the mean of that interval, separately under the assumption of CRRA and CARA utility functions.

⁶Under 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, $U(x, r) = \frac{e^{-rx}}{r}$.

A.6 Individual Returns to Continuing versus Quitting

Did subjects with a higher ex-ante probability of continuing financially benefit from continuing (relative to quitting), and are there gender differences therein? Computing whether continuing paid off at the individual level requires estimating counterfactual outcomes: How much would have subjects who continued earned, had they quit? Recall that conditional on reporting the same switch point in Part 3 of the experiment, it is random who continues and who quits. In what follows, suppose that Part 3 of the experiment is drawn for payment. For subjects who continued and reported a switch point s , by construction of the BDM their expected bonus earnings of quitting are $\frac{s+22}{2}$. Their actual bonus earnings of continuing, on the other hand, 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 expected earnings, had they quit.

Figure A.3 shows that subjects who continued in the *Baseline* treatment 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.⁷ The average premium of continuing is computed as the difference between a quintile's average earnings for continuing and a quintile's average expected earnings for quitting.

Figure A.3 illustrates that for women, the average premium of continuing tends to increase with their ex-ante probability of continuing, i.e., women who were ex-ante more likely to continue were indeed more likely to pass the second IQ test, and thus on

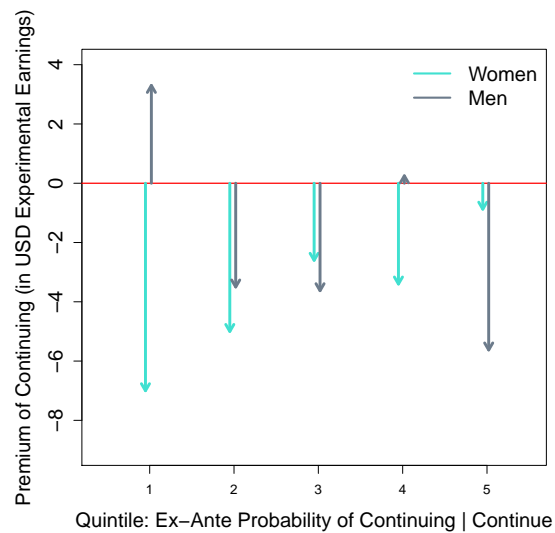
⁷That is, after ranking all subjects that continued by their probability of continuing (separately by gender), Quintile 1 captures the 20% of subjects with the lowest probability of continuing, etc.

average benefited more from continuing than women with a lower ex-ante probability of continuing. That being said, on average their expected earnings from quitting would have exceeded their realized earnings from continuing across the distribution. In other words, on average women would have had higher expected earnings in the experiment by quitting more often. More specifically, women who continued on average lose between \$1 – \$7 in experimental earnings relative to their expected earnings for quitting, as the downward-facing arrows in Figure A.3 demonstrate.

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 that are grouped in Quintile 5 in Figure A.3.

Figure A.3: 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.

Appendix B

Appendix to Chapter 2

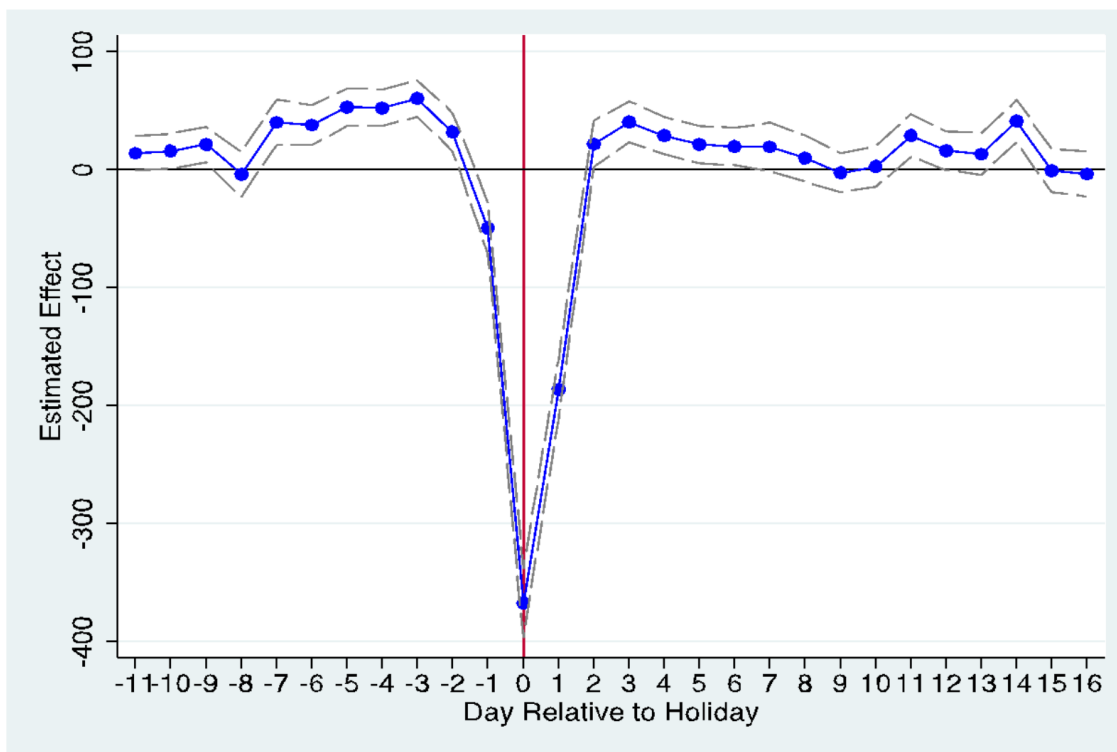
B.1 Additional Figures and Tables

Figure B.1: Excess Births - Missing Births for Varying Windows around a Holiday

		Days after holiday																		
		21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3
Days before holiday	21	1021	943.1	888.1	810.2	797.2	746.0	638.8	561.1	527.7	477.9	457.9	384.3	349.4	328.2	279.6	325.6	313.2	296.0	220.1
	20	1040	977.0	909.7	852.2	774.9	771.8	703.4	436.2	375.7	313.0	261.8	189.2	124.9	110.9	268.0	291.5	283.3	246.1	194.4
	19	963.3	919.1	892.9	849.3	802.8	806.0	749.2	418.4	355.7	294.0	246.1	174.7	116.4	106.0	293.6	304.1	289.2	254.7	197.2
	18	870.5	860.6	843.8	825.3	792.0	794.0	732.6	359.0	299.2	240.6	194.3	125.5	75.7	69.9	272.1	280.6	266.0	225.6	166.9
	17	814.6	810.3	805.2	774.8	750.3	738.7	682.0	299.4	246.4	191.3	151.4	93.2	52.6	52.8	258.7	272.5	255.1	217.2	157.4
	16	750.8	746.8	740.0	717.7	697.5	668.3	613.4	257.3	206.9	155.2	114.2	63.0	29.5	39.5	264.4	273.4	253.0	207.8	142.3
	15	686.7	683.0	673.7	641.9	616.9	587.7	540.3	211.2	166.5	117.0	73.6	29.5	4.6	16.2	245.7	253.1	228.2	179.6	112.7
	14	638.8	627.7	616.0	577.1	550.6	520.8	481.3	171.8	126.8	79.5	41.3	2.4	-21.1	-7.0	220.7	223.1	198.2	152.0	90.6
	13	399.0	329.4	263.2	211.8	164.6	134.8	119.3	26.8	-7.5	-47.1	-94.7	-133.7	-172.0	-174.8	-71.2	-62.2	-57.2	-88.5	-106.4
	12	327.3	259.5	192.9	141.5	95.6	68.0	56.5	-63.8	-96.3	-130.1	-173.0	-200.5	-231.0	-225.8	-114.2	-103.9	-99.9	-129.4	-143.3
	11	250.7	178.5	112.0	63.4	24.4	2.2	-3.4	-112.7	-132.7	-153.3	-186.6	-199.1	-222.0	-213.5	-154.2	-136.8	-131.8	-156.6	-166.4
	10	185.8	110.4	41.9	-2.4	-34.0	-50.4	-48.6	-144.9	-158.9	-169.2	-192.6	-202.2	-222.9	-218.1	-160.8	-140.3	-136.2	-157.4	-168.3
	9	119.4	40.9	-22.0	-62.4	-86.7	-96.9	-89.6	-164.6	-176.6	-183.8	-202.5	-214.5	-233.9	-228.0	-170.9	-149.5	-148.3	-169.0	-176.9
	8	19.0	-59.3	-120.4	-152.4	-171.3	-172.5	-162.5	-217.8	-223.5	-224.3	-232.2	-247.3	-267.3	-259.9	-199.6	-175.0	-174.9	-193.6	-207.5
	7	-58.5	-136.3	-189.9	-216.3	-230.2	-226.8	-217.6	-258.3	-256.3	-250.1	-254.9	-268.1	-290.0	-278.5	-214.8	-190.4	-190.1	-210.4	-226.5
	6	-108.6	-189.1	-223.0	-255.4	-261.0	-270.7	-294.4	-409.1	-414.1	-420.0	-426.0	-440.5	-439.4	-418.2	-304.0	-273.4	-266.6	-281.5	-300.4
	5	-187.9	-276.8	-313.5	-344.1	-350.3	-358.1	-374.0	-462.6	-462.8	-463.5	-464.1	-467.6	-463.4	-440.1	-323.4	-292.8	-285.8	-301.3	-319.0
	4	-247.1	-334.3	-375.4	-405.5	-415.5	-427.7	-439.4	-516.1	-512.5	-502.4	-495.9	-488.6	-483.6	-461.2	-348.6	-317.9	-311.8	-327.4	-345.0
	3	-305.8	-391.7	-436.7	-466.8	-483.4	-494.6	-503.8	-563.5	-559.2	-539.3	-525.5	-515.9	-508.4	-485.8	-374.4	-344.5	-339.0	-355.2	-373.4

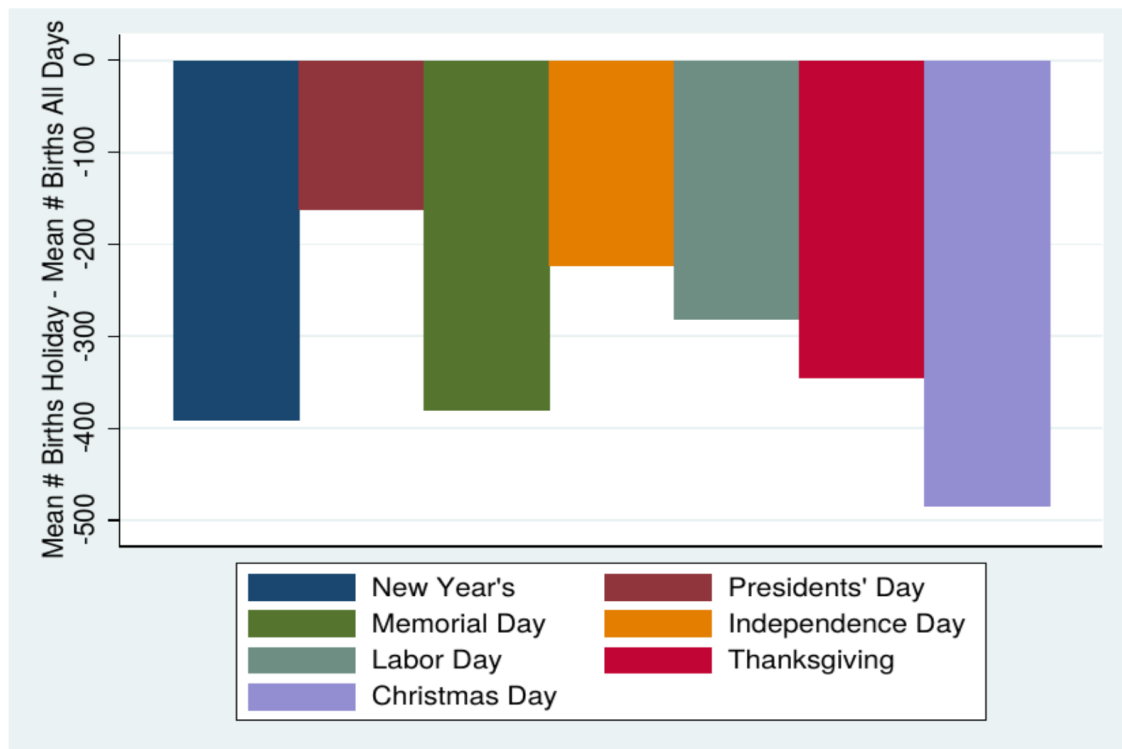
This figure shows the excess minus the missing number of births for varying windows around a holiday. We choose as the optimal window the one that minimizes the absolute value of this criterion. For our case, the optimal window is the range of days from 11 days before a holiday through 16 days after a holiday.

Figure B.2: Shift in the Number of Births due to a Holiday including Christmas and New Years in California: 2000-2016



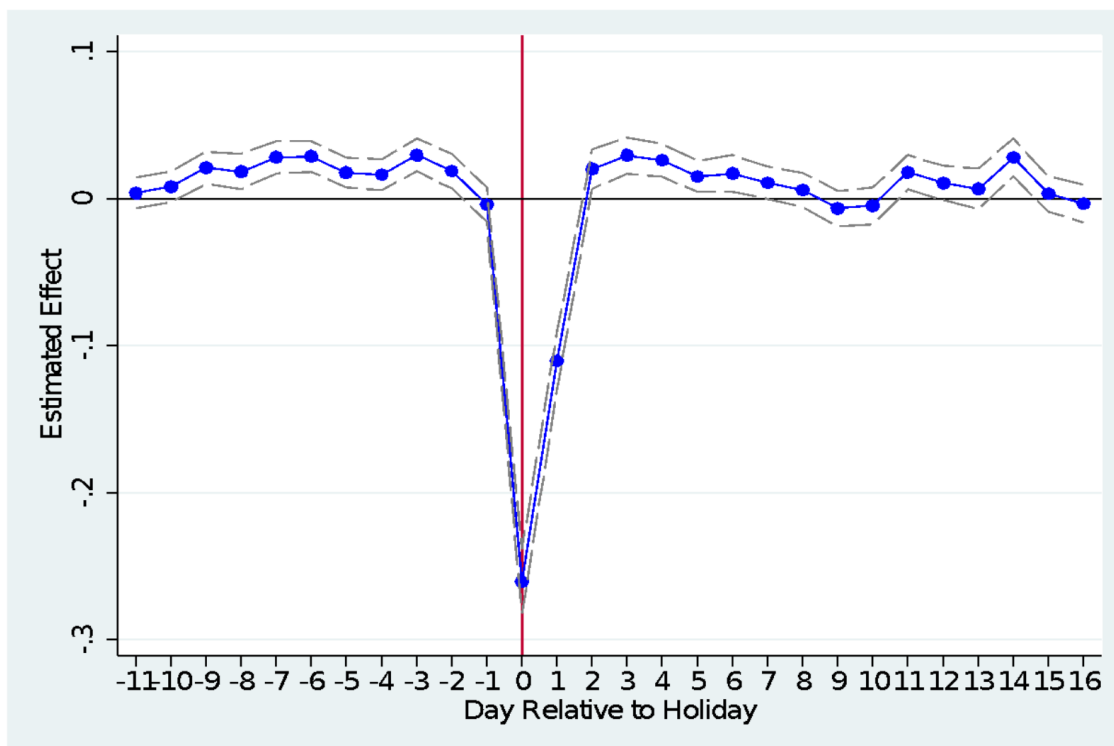
This figure shows the effect of a holiday on births. In this graph, the set of holidays considered is New Year's Day, Presidents' Day, Memorial Day, Independence Day, Labor Day, Thanksgiving, and Christmas. Plotted are regression estimates from equation (2). On the x-axis is the day relative to the holiday (-1=day before holiday, 1=day after holiday, etc.). On the y-axis are estimates of the day relative to holiday dummy coefficients from equation (2) along with the 95% confidence interval.

Figure B.3: The Holiday Decline by Specific Holiday



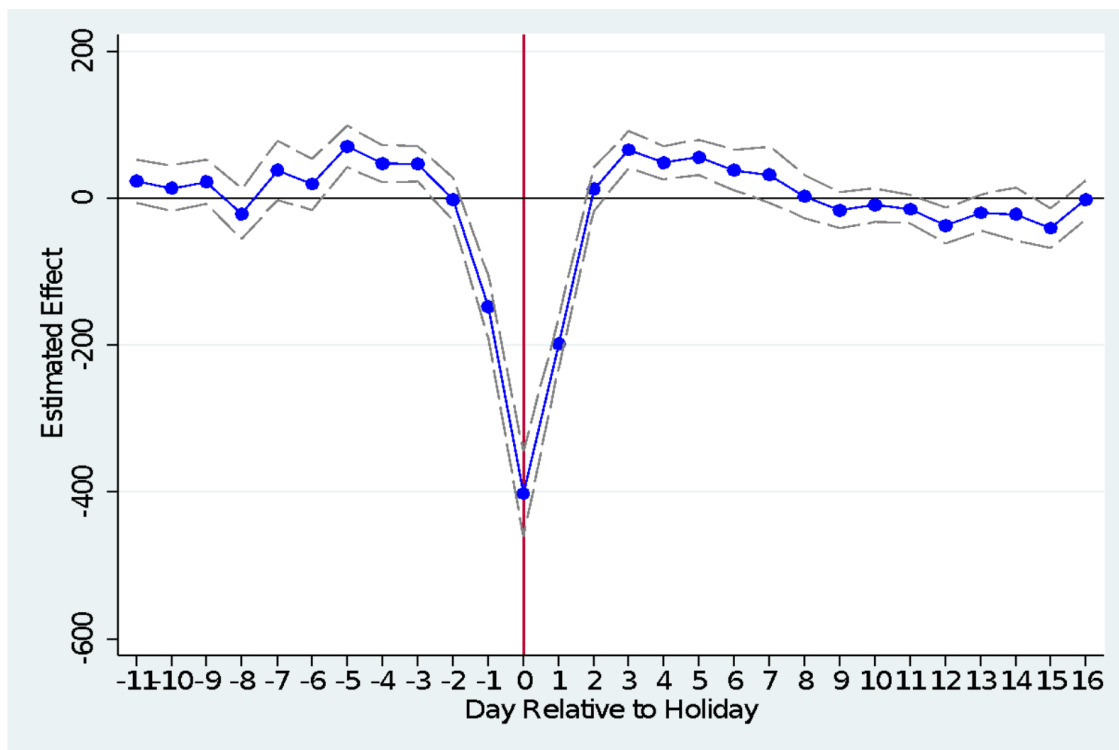
This figure shows, for each holiday, the mean number of births on that holiday minus the mean of the daily number of births for California from 2000-2016.

Figure B.4: Poisson Model Estimates of the Shift in the Number of Births due to a Holiday in California: 2000-2016



This figure shows the effect of a holiday on the daily number of births. Estimates are derived from a Poisson model. On the x-axis is the day relative to the holiday (-1=day before holiday, 1=day after holiday, etc.). On the y-axis are estimates of the day relative to holiday dummy coefficients from equation (2) along with the 95% confidence interval.

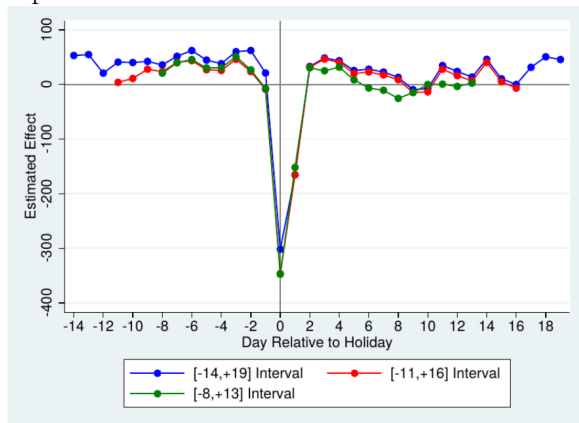
Figure B.5: Shift in the Number of Births due to a Holiday using Holidays that Rotate Days of the Week in California: 2000-2016



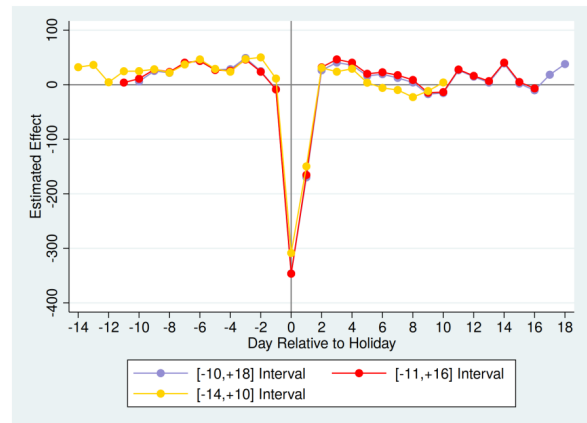
This figure shows the effect of a holiday on births. The holidays considered are those that do not occur on the same day of the week each year - New Year's Day, Independence Day, and Christmas. Plotted are regression estimates from equation (2). On the x-axis is the day relative to the holiday (-1=day before holiday, 1=day after holiday, etc.). On the y-axis are estimates of the day relative to holiday dummy coefficients from equation (2) along with the 95% confidence interval.

Figure B.6: Shift in the Number of Births due to a Holiday in California Varying the Holiday Window: 2000-2016

A. Windows that are 3 Days Wider / Narrower Than Optimal Window

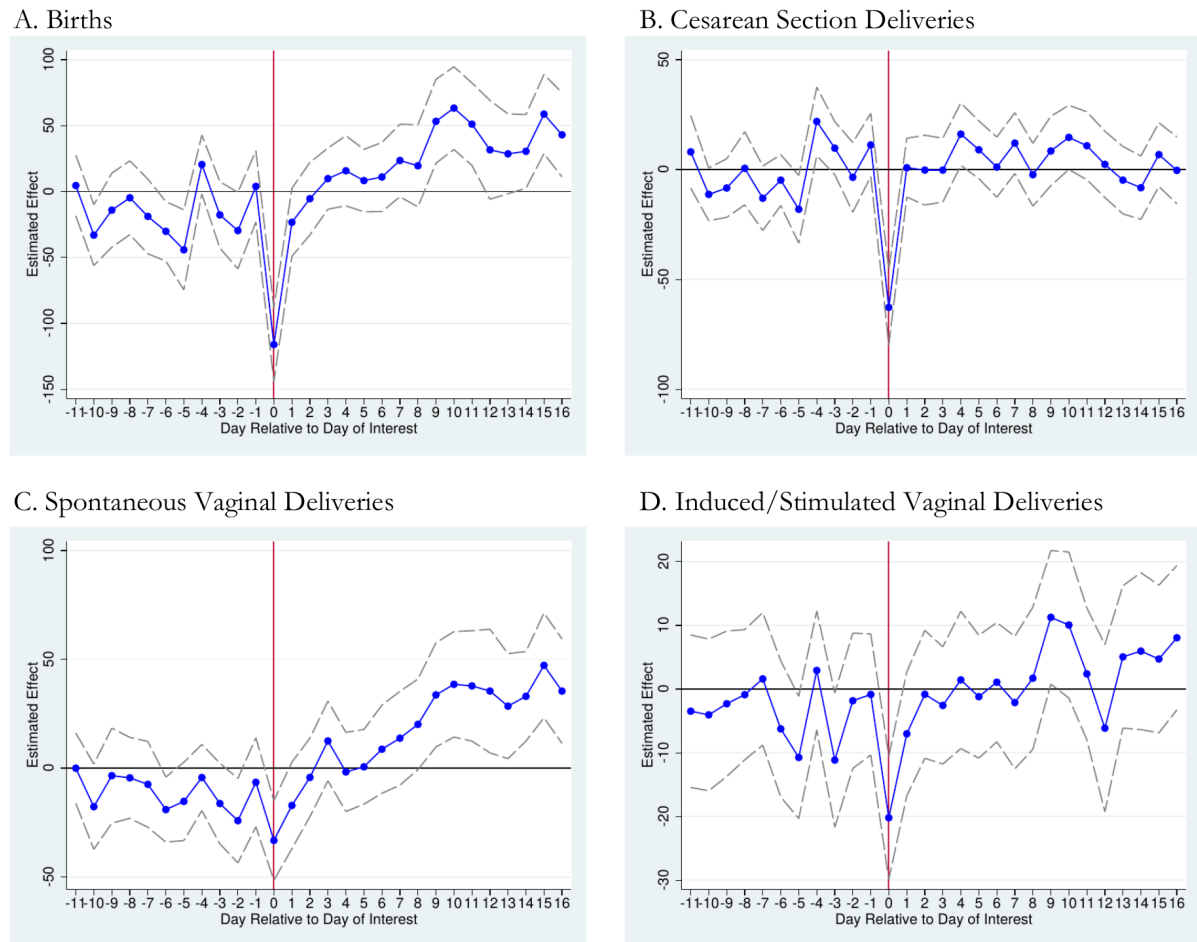


B. "Second-best" Optimal Windows, According to Grid Search



This figure shows the effect of a holiday on the daily number of births using different window sizes. Plotted are regression estimates from equation (2) with daily births as the dependent variable. On the x-axis is the day relative to the holiday (-1=day before holiday, 1=day after holiday, etc.). On the y-axis are estimates of the day relative to holiday dummy coefficients from equation (2). Panel A displays both a narrower and a wider window around each holiday. The [-8,+13] interval uses the days between 8 days before each holiday to 13 days after each holiday. The [-14,+19] interval uses the days between 14 days before each holiday to 19 days after each holiday. The [-11,+16] interval (our "optimal" window) uses the days between 11 days before and 16 days after. Panel B displays two windows that are "second-best" according to our grid search, as the absolute sum of the excess and missing numbers of births, as estimated in equation (1), are second-closest to zero.

Figure B.7: Shift in the Number of Births due to September 11th (2001-2016) and Friday the 13th (2000-2016) in California



This figure shows the effect of September 11 (after 2001) and Friday the 13 on the number of births (Panel A), the number of births delivered via cesarean section (Panel B), spontaneous vaginal deliveries (Panel C), and induced/stimulated deliveries (Panel D). Plotted are regression estimates from equation (2). On the x-axis is the day relative to the day of interest, i.e. September 11 or Friday the 13, (-1=day before day of interest, 1=day after day of interest, etc.). On the y-axis are estimates of the day relative to day-of-interest-dummy coefficients from equation (2) along with the 95% confidence interval.

Table B.1: Aggregate Effect of the Holiday Period on Birth Outcomes, Controlling for Moms over Age 35, Private and Public Delivery Payment

	Mean Birth Weight	Low Birth Weight	Any Newborn Conditions	Any Labor Complications	Low Apgar Score
Holiday Interval	-1.32* (0.69)	0.31 (0.29)	-1.06*** (0.39)	-2.33** (0.95)	-0.52 (0.35)
Mean Dependent Variable	3309	0.07	0.09	0.46	0.06
Fraction of Births Manipulated	0.013	0.013	0.013	0.013	0.013
Implied IV Estimate	-105	0.025	-0.084	-0.185	-0.041
Observations	4256	4256	4256	4256	2506

Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The holiday interval is the period of time covering the period 11 days prior to a major holiday through 16 days after a major holiday. The 4256 observations correspond to approximately 250.3 days over 17 years. The sample size for the share of babies with a low Apgar score is lower, as this variable is not available prior to 2007. The share of moms over age 35, the share of private insurance delivery payments, as well as the share of public insurance delivery payments are included as additional controls. Estimates and standard errors are multiplied by 1000, except for the mean birth weight outcome.

Table B.2: Aggregate Effect of the Holiday Period on Other Birth Outcomes

	Very Low Birth Weight	Newborn was in Neo-natal Intensive Care Unit (NICU)	Newborn had Assisted Ventilation	Mean Number of Newborn Conditions	Mean Number of Labor Complications
<i>Panel A: Without Christmas and New Year's</i>					
Holiday Interval	0.08 (0.13)	-0.10 (0.30)	-0.14 (0.14)	-1.55** (0.68)	-2.20 (2.01)
Daily Mean of Outcome	0.01	0.05	0.01	0.14	0.72
Fraction of Births Manipulated	0.013	0.013	0.013	0.013	0.013
Implied IV Estimate	0.006	-0.008	-0.011	-0.123	-0.175
Number of Observations	4256	4256	4256	4256	4256
<i>Panel B: Including Christmas and New Year's</i>					
Holiday Interval	0.07 (0.12)	-0.41 (0.27)	-0.18 (0.13)	-1.73*** (0.60)	-5.14*** (1.73)
Daily Mean of Outcome	0.01	0.05	0.01	0.14	0.72
Fraction of Births Manipulated	0.014	0.014	0.014	0.014	0.014
Implied IV Estimate	0.005	-0.030	-0.013	-0.126	-0.375
Number of Observations	4599	4599	4599	4599	4599

Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The holiday interval is the period of time covering the period 11 days prior to a major holiday through 16 days after a major holiday. The set of holidays in Panel A includes Presidents' Day, Memorial Day, Independence Day, Labor Day and Thanksgiving. The 4256 observations correspond to approximately 250.3 days over 17 years, Panel B adds Christmas and New Year's Day to the holiday list. The 4599 observations correspond to approximately 270.5 days over 17 years. The fraction of births manipulated is calculated as the effect of the holiday on births on the day of the holiday and the day after the holiday after divided by the total number of births in the 28-day manipulation window period. The implied IV estimate is the ratio of the holiday interval effect in the table divided by the fraction of births manipulated. Estimates and standard errors are multiplied by 1000.

Table B.3: Aggregate Effect of the Holiday Period on Birth Outcomes for High-Risk Pregnancies, Controlling for Private Delivery Payment

	Mean Birth Weight	Low Birth Weight	Any Newborn Conditions	Any Labor Complications	Low Apgar Score
Holiday Interval	-1.40 (2.34)	0.13 (1.20)	-1.42 (1.23)	1.44 (1.71)	0.09 (1.02)
Mean Dependent Variable	3309	0.14	0.15	0.45	0.07
Fraction of Births Manipulated	0.022	0.022	0.022	0.022	0.022
Implied IV Estimate	-63.78	0.006	-0.065	0.066	0.004
Observations	4256	4256	4256	4256	2506

Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The holiday interval is the period of time covering the period 11 days prior to a major holiday through 16 days after a major holiday. The 4256 observations correspond to approximately 250.3 days over 17 years. The sample size for the share of babies with a low Apgar score is lower, as this variable is not available prior to 2007. The share of private insurance delivery payments, as well as the share of public insurance delivery payments, are included as additional controls. Estimates and standard errors are multiplied by 1000, except for the mean birth weight outcome.

Table B.4: Poisson Model Estimates of Aggregate Effect of the Holiday Period on Births and Delivery Types

	Total Births	Cesarean Section Births	Spontaneous Vaginal Births	Induced/Stimulated Vaginal Births
Holiday Interval	-0.02 (0.02)	-0.04 (0.04)	0.01 (0.02)	-0.09*** (0.03)
Daily Mean Births	1447	443	709	294
Number of Observations	4256	4256	4256	4256

Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The holiday interval is the period of time covering the period 11 days prior to a major holiday through 16 days after a major holiday.

Table B.5: Aggregate Effect of the Holiday Period on Births and Delivery Types using Only Holidays that Vary Day of the Week

	Total Births	Cesarean Section Births	Spontaneous Vaginal Births	Induced/Stimulated Vaginal Births
Holiday Interval	-12.78*** (4.00)	-2.33 (2.24)	-4.43** (2.02)	-6.02*** (1.18)
Daily Mean Births	1447	443	710	294
Number of Observations	4256	4256	4256	4256

Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The holiday interval is the period of time covering the period 11 days prior to a major holiday through 16 days after a major holiday. The holidays considered are Independence Day, Christmas, and New Year's Day. The 4256 observations correspond to approximately 250.3 days over 17 years.

Table B.6: Aggregate Effect of the Holiday Period on Births and Delivery Types with Varying Window Size

	Total Births	Cesarean Section Births	Spontaneous Vaginal Births	Induced/ Stimulated Vaginal Births
<i>Panel A: Optimal Window of [-11, +16]</i>				
Holiday Interval	-4.07 (2.99)	-2.42 (1.72)	1.05 (1.84)	-2.71*** (0.98)
Daily Mean Births	1447	443	709	294
Number of Observations	4256	4256	4256	4256
<i>Panel B: Holiday Window [-8, +13]</i>				
Holiday Interval	-9.74*** (3.11)	-2.02 (1.75)	-3.72* (2.01)	-4.00*** (1.03)
Daily Mean Births	1445	442	711	293
Number of Observations	3519	3519	3519	3519
<i>Panel C: Holiday Window [-14, +19]</i>				
Holiday Interval	8.54*** (2.26)	3.83*** (1.24)	3.47*** (1.08)	1.24** (0.63)
Daily Mean Births	1450	446	710	295
Number of Observations	4943	4943	4943	4943
<i>Panel D: Holiday Window [-10, +18]</i>				
Holiday Interval	-4.26 (3.07)	-2.76 (1.74)	1.60 (1.88)	-3.11*** (0.99)
Daily Mean Births	1449	445	709	293
Number of Observations	4368	4368	4368	4368
<i>Panel E: Holiday Window [-14, +10]</i>				
Holiday Interval	-4.74 (2.78)	-0.92 (1.72)	-1.21 (1.99)	-2.60*** (1.00)
Daily Mean Births	1445	448	711	296
Number of Observations	3816	3816	3816	3816

Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Panel A provides our main birth displacement results from Table 2. Panel B shows results using a window that is reduced by 3 days on each side. Panel C shows results using a window that is increased by 3 days on either side. Panels D and E consider the “second-best” optimal windows, according to our grid search.

Table B.7: Aggregate Effect of the Holiday Period on Maternal Characteristics with Varying Window Sizes

	Low-Risk Births	Low-Risk First Births	Moms over Age 35	Teenage Moms	White Moms	Moms with High School Degree or Less	Deliveries with Private Insurance Payment	Deliveries with Public Insurance Payment
<i>Panel A: Optimal Window of [-11, +16]</i>								
Holiday Interval	-0.76 (0.65)	-0.48 (0.54)	-2.10*** (0.75)	0.23 (0.28)	-3.99 (2.57)	1.52 (2.30)	-8.49*** (1.81)	5.22*** (1.67)
Mean Dep Var	121	93	205	39	1113	699	701	689
Observations	2756	2756	4256	4256	4256	3502	4256	4256
<i>Panel B: Holiday Window [-8, +13]</i>								
Holiday Interval	-1.24* (0.67)	-0.81 (0.57)	-1.90** (0.77)	-0.45 (0.30)	-7.91*** (2.72)	-2.30 (2.50)	-7.24*** (1.88)	-2.06 (1.77)
Mean Dep Var	120	93	204	39	1112	699	700	689
Observations	2280	2280	3519	3519	3519	2901	3519	3519
<i>Panel C: Holiday Window [-14, +19]</i>								
Holiday Interval	0.71 (0.45)	0.34 (0.36)	-0.77 (0.53)	0.11 (0.17)	6.92*** (1.87)	4.46*** (1.42)	3.41*** (1.29)	5.50*** (1.21)
Mean Dep Var	121	93	205	39	1116	700	702	691
Observations	3203	3203	4943	4943	4943	4067	4943	4943
<i>Panel D: Holiday Window [-10, +18]</i>								
Holiday Interval	-0.79 (0.66)	-0.56 (0.55)	-2.26*** (0.75)	0.24 (0.29)	-3.92 (2.63)	3.38 (2.33)	-8.27*** (1.84)	4.67*** (1.74)
Mean Dep Var	121	93	205	39	1115	700	702	690
Observations	2830	2830	4368	4368	4368	3595	4368	4368
<i>Panel E: Holiday Window [-14, +10]</i>								
Holiday Interval	-0.35 0.64	-0.27 (0.52)	-1.02 (0.74)	-0.40 (0.30)	-4.07 (2.63)	-2.10 (2.45)	-5.24*** (1.85)	1.24 (1.67)
Mean Dep Var	121	94	206	39	1120	703	705	693
Observations	2472	2472	3816	3816	3816	3144	3816	3816

Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Panel A provides our main birth displacement results. Panel B (C) shows results using a window that is reduced (increased) by 3 days on each side. Panels D and E consider our “second-best” optimal windows. Variations in the number of observations reflect the different window sizes, as well as the unavailability of variables in some years.

Table B.8: Aggregate Effect of the Holiday Period on Births Outcomes with Varying Window Sizes

	Mean Birth Weight	Low Birth Weight	Any Newborn Conditions	Any Labor Complications	Low Apgar Score
<i>Panel A: Optimal Window of [-11, +16]</i>					
Holiday Interval	-2.00*** (0.70)	0.40 (0.29)	-0.88** (0.40)	-1.56 (1.02)	-0.49 (0.35)
Mean Dependent Variable	3309	0.07	0.09	0.46	0.06
Fraction of Births Manipulated	0.013	0.013	0.013	0.013	0.013
Implied IV Estimate	-159	0.032	-0.070	-0.124	-0.039
Observations	4256	4256	4256	4256	2506
<i>Panel B: Holiday Window [-8, +13]</i>					
Holiday Interval	-1.17 (0.75)	0.59* (0.32)	-0.40 (0.43)	-0.74 (1.11)	-0.02 (0.36)
Mean Dependent Variable	3309	0.07	0.09	0.46	0.06
Fraction of Births Manipulated	0.016	0.016	0.016	0.016	0.016
Implied IV Estimate	-71.07	0.036	-0.024	-0.045	-0.001
Observations	3519	3519	3519	3519	2074
<i>Panel C: Holiday Window [-14, +19]</i>					
Holiday Interval	1.64*** (0.45)	-0.41** (0.20)	-0.11 (0.23)	0.20 (0.41)	-0.11 (0.23)
Mean Dependent Variable	3309	0.07	0.09	0.46	0.06
Fraction of Births Manipulated	0.010	0.010	0.010	0.010	0.010
Implied IV Estimate	168.02	-0.042	-0.011	0.020	-0.011
Observations	4943	4943	4943	4943	2911
<i>Panel D: Holiday Window [-10, +18]</i>					
Holiday Interval	-1.97*** (0.71)	0.37 (0.30)	-1.07** (0.42)	-1.85* (1.04)	-0.50 (0.35)
Mean Dependent Variable	3309	0.07	0.09	0.46	0.06
Fraction of Births Manipulated	0.013	0.013	0.013	0.013	0.013
Implied IV Estimate	-155	0.029	-0.084	-0.145	-0.039
Observations	4368	4368	4368	4368	2572
<i>Panel E: Holiday Window [-14, +10]</i>					
Holiday Interval	-1.38* (0.72)	0.42 (0.29)	-0.38 (0.43)	-0.43 (1.14)	0.078 (0.36)
Mean Dependent Variable	3310	0.07	0.09	0.46	0.06
Fraction of Births Manipulated	0.013	0.013	0.013	0.013	0.013
Implied IV Estimate	-104.72	0.031	-0.029	-0.033	0.006
Observations	3816	3816	3816	3816	2248

Robust standard errors are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Panel A provides our main birth displacement results from Table 5. Panel B (C) shows results using a window that is reduced (increased) by 3 days on each side. Panels D and E consider our “second-best” optimal windows, according to our grid search. Variations in the number of observations reflect the different window sizes. The sample size for the low Apgar score outcome is lower, as this variable is not available prior to 2007. Estimates and standard errors are multiplied by 1000, except for the mean birth weight outcome.

Table B.9: Sub-group Analysis of Manipulated Births as a Percent of Total Births in the Window

Subgroup	N	Mean # Daily Births	# Births in Window	Lower Bound # of Manipulated Births	Percent of Total Births in the Window
Total Sample	4256	1447.0	40516	511.9	1.26
Teenage Mom	4256	39.0	1093	7.6	0.70
Older Mom	4256	204.5	5726	88.8	1.55
High-Risk Birth	4256	258.8	7246	159.2	2.20
Low-Risk First Birth	2756	93.0	2603	38.7	1.49
Low-Risk Birth	2756	120.5	3374	54.6	1.62
White Non-Hispanic Mom	4256	419.8	11754	179.1	1.52
Black Non-Hispanic Mom	4256	83.1	2327	26.1	1.12
Asian Non-Hispanic Mom	4256	114.7	3212	33.3	1.04
Hispanic Mom	4256	722.5	20230	239.3	1.18
Mom has Some College or More Education	3502	687.8	19258	253.5	1.32
Mom has High School or Less Education	3502	698.8	19566	237.3	1.21
Public Insurance Delivery	4256	688.9	19289	236.6	1.23
Private Insurance Delivery	4256	700.8	19622	262.1	1.34
Self-pay Delivery	4256	39.3	1101	8.1	0.73
Kaiser	4256	187.7	5256	34.0	0.65

Each row represents the analysis from a separate sub-group. The 4256 observations correspond to approximately 250.3 days over 17 years. The sample size for low-risk births, overall or of first births, is lower because these variables cannot be constructed prior to 2006, when characteristics such as intrauterine growth restrictions were not captured in the California birth data. Similarly, the sample size is smaller for education outcomes, as the education variable is not available in the data between 2003-2005.

Appendix C

Appendix to Chapter 3

C.1 Material Presented to Subjects during the Experiment

C.1.1 Instructions

Welcome!

You are about to participate in an experiment on decision-making. In this experiment, you can earn a considerable amount of **money**, which will be paid to you in cash, privately, at the end of the experiment. How much you earn will depend on your decisions, the decisions of other participants, and chance.

Please **do not communicate** with the other participants at any point during the experiment. Make sure that your phone is turned off now.

To make sure that everybody understands the tasks in this experiment, we will begin with some basic instructions. **If you have any questions**, or need assistance of any kind, raise your hand and the experimenter will come and help you.

Basic Instructions

At the beginning of the experiment, the computer will randomly assign each participant to one of **two groups: *Group A* and *Group B***. Once assigned, participants will remain in the same group throughout the entire experiment. There will be the same number of participants in each group.

Participants assigned to the **same group** will each see the **same payoff table and other information** on their computer screen at the beginning of the experiment.

The experiment consists of **2 parts**. For now, we will explain to you what is happening in the first part of the experiment. Once the first part is completed, we will explain to you what is happening in the second part.

The first part of the experiment consists of **40 rounds**. In each round, you will be **randomly paired** with a participant from the other group. You will not know who of the other participants is assigned to which group, and you will also not know with whom you are randomly paired in any given round. In each round, it is equally likely that you will be paired with any of the participants from the other group. You will never be paired with somebody from your own group.

In each of the 40 rounds, you and the person you are currently paired with will be asked to make a decision on the computer. In what follows, we will explain to you how you can make these decisions.

The Decision Tasks

In each of the 40 rounds, you will be able to **choose one of two actions**. The participant you are paired with will also be able to choose one out of two actions. In each round, everybody will have to **choose an action before seeing the action that the other participant has chosen**.

Below, we show you an example of how a decision task could look like on the computer.

In the experiment, you will see a similar table on your computer screen, but with different numbers.

Example of a Payoff Table

The **table below shows the payoffs** associated with each combination of your choice and the choice of the participant you are paired with. This is an example of how a decision task could look like on the computer; please note that the actual numbers you will see in the experiment will be different from those shown in this example. We will now explain to you how you can interpret the numbers in the table.

The **first entry** in each cell (i.e. the number before the comma) represents **your payoff**. The **second entry** in each cell (i.e. the number after the comma) represents the **payoff of the person you are paired with**.

Please choose a row			
The Other's Choice			
		B1	B2
Your Choice	A1	5, 7	4, 3
	A2	2, 4	10, 6

All **cell entries** of the table show the **payoffs that are associated with each combination of your choice and the other participant's choice**:

- For example, if you select “A1” and the other participant selects “B1”, you earn 5 Dollars and the other participant earns 7 Dollars.

- As another example, if you select “A2” and the other participant selects “B1”, you earn 2 Dollars and the other participant earns 4 Dollars.
- Another example: if you select “A1” and the other participant selects “B2”, you earn 4 Dollars and the other participant earns 3 Dollars.
- Another example: if you select “A2” and the other participant selects “B2”, you earn 10 Dollars and the other participant earns 6 Dollars.

How to Make Decisions

Suppose that the computer assigned you to *Group A*. In this example, you will be asked to choose either “A1” or “A2”. Remember that if you are in *Group A*, then in each round, you will be paired with somebody from *Group B*, and they will be asked to choose either “B1” or “B2”. If you should get assigned to *Group B*, however, then **you** will be asked to choose between “B1” and “B2”, and the **other participant** will be asked to choose between “A1” and “A2”, as they would be assigned to *Group A*.

In the experiment, you will see a table similar to the example above on your computer screen. **To make a choice**, you will **click on one of the rows in the table**.

Once you select a row, it will change color and a red *SUBMIT* button will appear. Your choice will be finalized once you click on the *SUBMIT* button. After submitting your choice, you will need to wait until the other participant you are paired with has also made their choice. Once you and the participant you are paired with have made your choices, those choices will be highlighted and your payoff for the round will appear. **Remember that you will only see the choice of the other participant once you have submitted your own choice.** After each of the 40 rounds, you will see an overview of the choice you made, the choice the other participant made, and your payoffs of the round.

Example of How a Payoff Table Will Look Like in the First Part of the Experiment

In the example above, we showed you an example of a decision task. In each cell of the table above, you could see both your own payoff and the other's payoff for each combination of your and the other participant's choices. In the actual experiment, however, **you will only see your own payoffs**. The payoffs of the other person will be covered.

Here is an example of what a table in the experiment could actually look like:

Please choose a row			
		The Other's Choice	
		B1	B2
Your Choice	A1	5, <input type="text"/>	4, <input type="text"/>
	A2	2, <input type="text"/>	10, <input type="text"/>

This **table shows your payoffs** associated with each combination of your choice and the choice of the participant you are paired with. It does not show, however, the payoffs of the person you are paired with. The **first entry in each cell** (i.e. the number before the comma sign) **represents your payoff**. For the first part of the experiment, you will never know how much the person you are paired with earns in each combination of your and their choice.

- For example, if you select “A1” and the other participant selects “B1”, you earn 5 Dollars, but you don't know how much the other participant earns.
- As another example, if you select “A2” and the other participant selects “B1”,

you earn 2 Dollars, but you don't know how much the other participant earns.

- Another example: if you select "A1" and the other participant selects "B2", you earn 4 Dollars, but you don't know how much the other participant earns.
- Another example: if you select "A2" and the other participant selects "B2", you earn 10 Dollars, but you don't know how much the other participant earns.

Summary: What It Means to Not See the Other's Payoffs

1. You only **know your own payoffs**, and you know that everybody in your group has the same payoffs.
2. You **do not know** the payoffs of the person you are paired with.
3. This means that you do not know what payoffs participants in the other group are getting. You know, however, that every participant in the other group is getting the same payoffs.

How much will you get paid in the end?

At the end of the experiment, for each participant the computer will randomly select a number between 1 and 80, corresponding to each of the rounds of the experiment. Every participant will get paid, in US Dollars, the amount of their payoff in that particular round, PLUS the show-up fee of 7 dollars. Before that, a short questionnaire will appear on your screen.

Summary

- There are a total of 80 rounds in the experiment, divided into two parts of 40 rounds each.

- You will make a decision in each of these 80 rounds.
- **At the beginning of the experiment**, half of all participants will be randomly assigned to *Group A*, and the other half will be assigned to *Group B*.
- Participants stay assigned to the same group throughout the experiment.
- All *Group A* participants have the same payoffs, and all *Group B* participants have the same payoffs. These payoffs remain the same throughout all 40 rounds of that part of the experiment.
- **In each round**, you will be randomly paired with someone from the other group.
- This means that before each decision round, a new random pair will be formed.

COMPREHENSION QUIZ

FIRST GAME (40 ROUNDS)

PART II

(Handed out to subjects after they completed the first 40 rounds of the experiment.)

The second part of the experiment has a similar setup to the first part. You will again be presented with a payoff table and will be asked to make choices by clicking on the rows of the table. This part of the experiment will consist of another **40 rounds**. As before, in each round, you will be **randomly paired** with a participant from the other group.

The table you see in this part of the experiment **shows both your payoff and the payoff of the person you are paired with**.

Example of a Payoff Table

The **first entry** in each cell (i.e. the number before the comma) represents **your payoff**. The **second entry** in each cell (i.e. the number after the comma) represents the **payoff of the person you are paired with**.

Please choose a row			
The Other's Choice			
		B1	B2
Your Choice	A1	5, 7	4, 3
	A2	2, 4	10, 6

Before we begin, let me briefly remind you of the following:

- Once you select a row, you need to click on the red *SUBMIT* button to confirm your choice.
- After that, please don't forget to press the *MOVE ON* button so that the next round of the experiment can begin.
- Remember that **all participants have to make their choice before they can observe the choice of the person they are paired with**.
- After completing all 40 rounds, a short questionnaire will appear. You will get paid after that.

C.1.2 Comprehension Quiz

PARTIAL INFORMATION - VERSION 1

Comprehension Quiz

To make sure that you understand the instructions of this experiment, please answer the questions below.

Below, you see an example of a decision task, similar to the one that you might encounter in the experiment. In this example, you got assigned to Group A, and the person you are randomly paired with got assigned to Group B.

Please choose a row			
The Other's Choice			
		B1	B2
Your Choice	A1	2, <input type="text"/>	3, <input type="text"/>
	A2	7, <input type="text"/>	5, <input type="text"/>

Please answer the following questions. If you don't know the answer for sure, please insert a question mark (?) into the blank space.

1. If you choose "A2" and the other chooses "B2", what is the other's payoff?
2. If you choose "A2" and the other chooses "B1", what is your payoff?
3. If you choose "A1" and the other chooses "B2", what is your payoff?
4. If you choose "A2" and the other chooses "B1", what is the other's payoff?
5. In this example, which combination of your action and the other's action needs to happen so that you get a payoff of 5?

6. In this example, which combination of your action and the other's action needs to happen so that you get a payoff of 7?

FULL INFORMATION - VERSION 1

Comprehension Quiz

To make sure that you understand the instructions of this experiment, please answer the questions below.

Below, you see an example of a decision task, similar to the one that you might encounter in the experiment. In this example, you got assigned to Group A, and the person you are randomly paired with got assigned to Group B.

Please choose a row			
The Other's Choice			
		B1	B2
Your Choice	A1	2, 0	3, 3
	A2	7, 11	5, 3

Please answer the following questions. If you don't know the answer for sure, please insert a question mark (?) into the blank space.

1. If you choose "A2" and the other chooses "B2", what is the other's payoff?
2. If you choose "A2" and the other chooses "B1", what is your payoff?
3. If you choose "A1" and the other chooses "B2", what is your payoff?

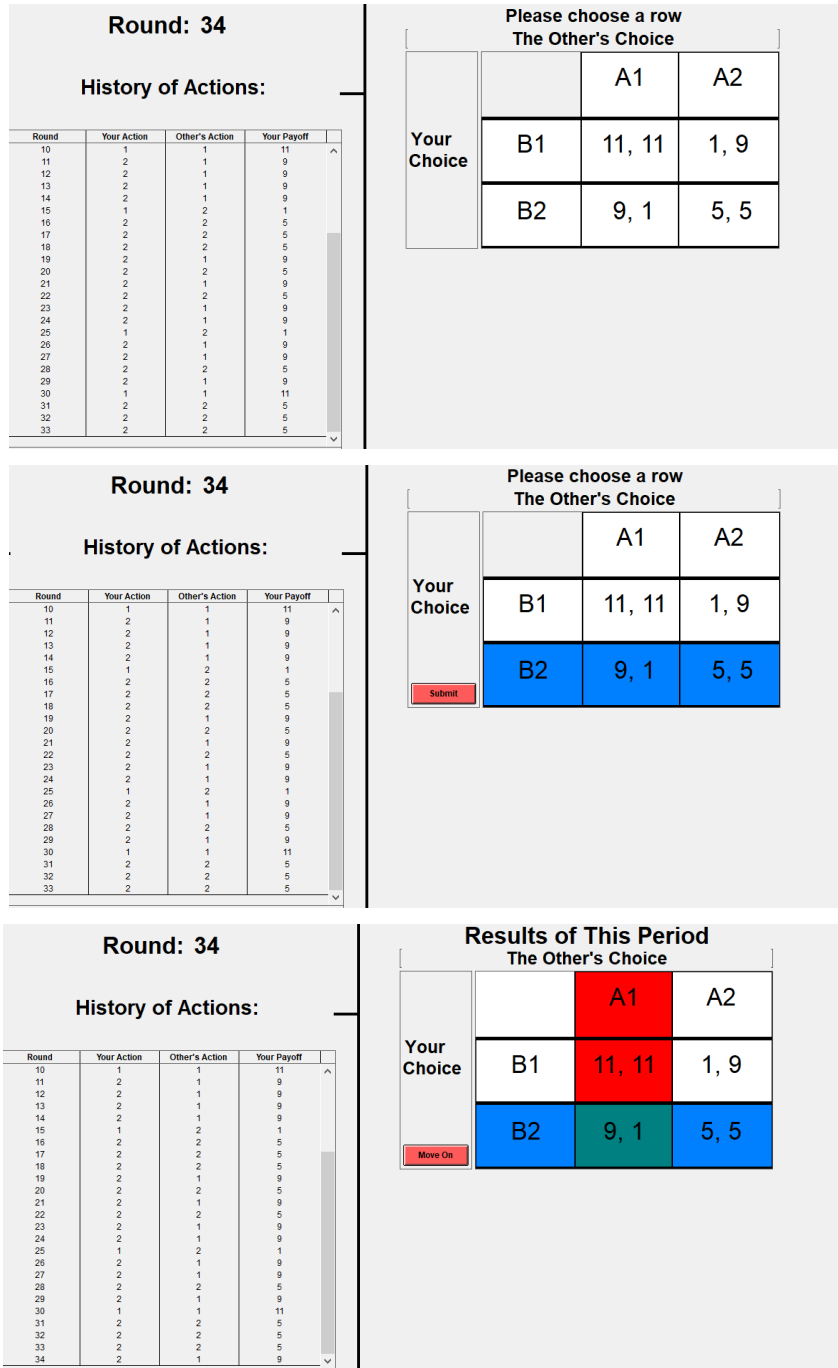
4. If you choose “A2” and the other chooses “B1”, what is the other’s payoff?
5. In this example, which combination of your action and the other’s action needs to happen so that you get a payoff of 5?
6. In this example, which combination of your action and the other’s action needs to happen so that the other gets a payoff of 3?

C.1.3 Experimental Interface

In Figure 3.2 we presented screenshots of the experimental interface before, during, and after subjects selected their actions in the PD-Partial treatment. For completion’s sake, Figure C.1 presents the corresponding screenshots for the SH-Full treatment:

C.2 Additional Tables

Figure C.1: Screenshots of Experimental Interface, SH-Full Treatment



We used a different color scheme for each of the two games in each experiment. We did this to help subjects understand and recall, when they are playing the second game of the experiment, that the current game they are playing is distinct from the first game they have already completed.

Table C.1: Share of subjects in the Stag Hunt choosing a mix of actions within 10 percentage points of the mixed strategy Nash equilibrium

	Rounds			
	1-10	11-20	21-30	31-40
a) Partial-information				
Fraction playing within .10 of MSNE	0.27	0.16	0.14	0.07
b) Full-information				
Fraction playing within .10 of MSNE	0.38	0.20	0.18	0.08

Given that the mixed strategy Nash equilibrium is to play action X 66.67% of the time, the fraction of subjects playing within 10 percentage points of the mixed strategy Nash equilibrium (MSNE) equals the total number of subjects playing action X between 56.67% and 76.67% of the time divided by the total number of subjects exposed to that information treatment.

Table C.2: Comparison of Observed and Simulated Data

height	(1)	(2)	(3)	(4)	(5)
	Overall	1-10	11-20	21-30	31-40
a) Stag Hunt, Partial					
Simulated	-0.007 (0.028)	-0.007 (0.021)	-0.010 (0.040)	-0.008 (0.039)	-0.002 (0.037)
Cluster p-value	0.805	0.755	0.798	0.832	0.946
Bootstrap p-value	0.820	0.778	0.810	0.854	0.961
Outcome mean	0.149	0.228	0.142	0.120	0.107
Number of clusters	1,006	1,006	1,006	1,006	1,006
N	686,400	171,600	171,600	171,600	171,600
b) Stag Hunt, Full					
Simulated	0.032 (0.073)	0.031 (0.045)	0.045 (0.079)	0.028 (0.083)	0.024 (0.108)
Cluster p-value	0.663	0.496	0.574	0.735	0.825
Bootstrap p-value	0.714	0.573	0.597	0.789	0.916
Outcome mean	0.844	0.881	0.838	0.836	0.822
Number of clusters	1,006	1,006	1,006	1,006	1,006
N	683,120	170,780	170,780	170,780	170,780
c) Prisoner's Dilemma, Partial					
Simulated	-0.002 (0.006)	-0.029 (0.011)	-0.010 (0.012)	0.017 (0.010)	0.016 (0.010)
Cluster p-value	0.814	0.006	0.432	0.083	0.108
Bootstrap p-value	0.820	0.069	0.588	0.146	0.194
Outcome mean	0.065	0.126	0.065	0.035	0.036
Number of clusters	1,006	1,006	1,006	1,006	1,006
N	687,360	171,840	171,840	171,840	171,840
d) Prisoner's Dilemma, Full					
Simulated	-0.018 (0.028)	-0.102 (0.064)	-0.025 (0.036)	-0.003 (0.040)	0.058 (0.034)
Cluster p-value	0.522	0.114	0.496	0.931	0.084
Bootstrap p-value	0.577	0.225	0.539	0.960	0.148
Outcome mean	0.232	0.433	0.205	0.176	0.113
Number of clusters	1,006	1,006	1,006	1,006	1,006
N	686,080	171,520	171,520	171,520	171,520

Standard errors presented in parentheses are calculated using the cluster-robust method allowing for correlation between observations within a cluster. The level of clustering is at the session. Cluster p-value indicates the p-value from a two-sided t-test of the null hypothesis that the different between the simulated data and observed data is zero using the cluster-robust standard error. Bootstrap p-value indicates the p-value from the empirical sampling distribution found with the bootstrapping method.

Table C.3: Comparison of Estimates of λ by Information Treatment

Estimate	λ -Stag Hunt	λ -Prisoner's Dilemma
Full-Information Treatment	1.5831 (1.2159 – 1.8776)	0.4297 (0.3347 – 0.5242)
Partial-Information Treatment	0.6516 (0.5184 – 0.7673)	0.7426 (0.6615 – 0.9146)
$H_0 : \lambda_{partial} \geq \lambda_{full}$ BCa interval test p-val	0.0014	
$H_0 : \lambda_{full} \geq \lambda_{partial}$ BCa interval test p-val		0.0368
Mann-Whitney U test	0.0018	0.0018
Kolmogorov-Smirnov test	0.0000	0.0000

One-sided BCa interval tests conducted using $B = 10,000$ bootstrap iterations, with bootstrapping being performed separately for full- and partial-information observations

Table C.4: Comparison of Estimates of ϕ by Information Treatment

Estimate	ϕ -Stag Hunt	ϕ -Prisoner's Dilemma
Full-Information Treatment	0.9649 (0.9011 – 1.0223)	0.8011 (0.3009 – 0.9398)
Partial-Information Treatment	0.8290 (0.7112 – 0.9911)	0.8769 (0.6019 – 85.013)
$H_0 : \phi_{partial} \geq \phi_{full}$ BCa interval test p-val	0.0624	
$H_0 : \phi_{full} \geq \phi_{partial}$ BCa interval test p-val		0.3005
Mann-Whitney U test	0.0017	0.0014
Kolmogorov-Smirnov test	0.0000	0.0000

One-sided BCa interval tests conducted using $B = 10,000$ bootstrap iterations, with bootstrapping being performed separately for full- and partial-information observations

Bibliography

- Agresti, A. and B. A. Coull (1998). Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician* 52(2), 119–126.
- Alan, S. and S. Ertac (2019). Mitigating the gender gap in the willingness to compete: Evidence from a randomized field experiment. *Journal of the European Economic Association* 17(4), 1147–1185.
- Almond, D., K. Y. Chay, and D. S. Lee (2005). The costs of low birth weight. *The Quarterly Journal of Economics* 120(3), 1031–1083.
- Almond, D., C. P. Chee, M. M. Sviatschi, and N. Zhong (2015). Auspicious birth dates among chinese in california. *Economics & Human Biology* 18, 153–159.
- Andreoni, J., Y.-K. Che, and J. Kim (2007). Asymmetric information about rivals’ types in standard auctions: An experiment. *Games and Economic Behavior* 59(2), 240–259.
- 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.
- Babichenko, Y. (2010). Uncoupled automata and pure Nash equilibria. *International Journal of Game Theory* 39(3), 483–502.
- Bailit, J. L., W. Grobman, Y. Zhao, R. J. Wapner, U. M. Reddy, M. W. Varner, K. J. Leveno, S. N. Caritis, J. D. Iams, A. T. Tita, et al. (2015). Nonmedically indicated induction vs expectant treatment in term nulliparous women. *American Journal of Obstetrics and Gynecology* 212(1), 103.e1–103.37.
- Barreca, A. I., J. M. Lindo, and G. R. Waddell (2016). Heaping-induced bias in regression-discontinuity designs. *Economic Inquiry* 54(1), 268–293.
- Battalio, R., L. Samuelson, and J. Van Huyck (2001). Optimization incentives and coordination failure in laboratory Stag Hunt games. *Econometrica* 69(3), 749–764.
- Bauer, T. K., S. Bender, J. Heining, and C. M. Schmidt (2013). The lunar cycle, sunspots and the frequency of births in germany, 1920–1989. *Economics & Human Biology* 11(4), 545–550.

- Becker, G. M., M. H. DeGroot, and J. Marschak (1964). Measuring Utility by a Single-response Sequential Method. *Behavioral Science* 9(3), 226–232.
- Berghella, V. (2018). Cesarean delivery: Preoperative planning and patient preparation.
- 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.
- Black, M., S. Bhattacharya, S. Philip, J. E. Norman, and D. J. McLernon (2015). Planned cesarean delivery at term and adverse outcomes in childhood health. *Journal of the American Medical Association* 314(21), 2271–2279.
- Black, S. E., P. J. Devereux, and K. G. Salvanes (2007). From the cradle to the labor market? the effect of birth weight on adult outcomes. *The Quarterly Journal of Economics* 122(1), 409–439.
- Bordalo, P., K. B. Coffman, N. Gennaioli, and A. Shleifer (2019). Beliefs about Gender. *American Economic Review* 109(3), 739–773.
- Borra, C., L. González, and A. Sevilla (2016). Birth timing and neonatal health. *American Economic Review* 106(5), 329–32.
- Borst, L. B. and M. Osley (1975). Holiday effects upon natality. *American Journal of Obstetrics and Gynecology* 122(7), 902–903.
- Boulvain, M., C. Stan, and O. Irion (2003). Membrane sweeping for induction of labour. *Cochrane Database of Systematic Reviews* CD000451.
- Brandenburger, A. (1996). Strategic and structural uncertainty in games. In R. Zeckhauser, R. Keeney, and J. Sebenius (Eds.), *Wise Choices: Decisions, Games, and Negotiations*, Chapter 13, pp. 221–232. Boston: Harvard Business School Press.
- Brown, L. D., T. Cai, and A. DasGupta (2001). Interval estimation for a binomial proportion. *Institute of Mathematical Statistics* 16(2).
- Brunner, B. and A. Kuhn (2014). Announcement effects of health policy reforms: evidence from the abolition of Austria’s baby bonus. *The European Journal of Health Economics* 15(4), 373–388.

- 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.
- Calford, E. and R. Oprea (2017). Continuity, inertia, and strategic uncertainty: A test of the theory of continuous time games. *Econometrica* 85(3).
- Camerer, C. and T.-H. Ho (1999). Experience-weighted attraction learning in normal form games. *Econometrica* 67(4), 827–874.
- Card, D., A. Fenizia, and D. Silver (2018). The health effects of cesarean delivery for low-risk first births.
- Charness, G., L. Rigotti, and A. Rustichini (2016). Social surplus determines cooperation rates in the one-shot Prisoner’s Dilemma. *Games and Economic Behavior* 100, 113–124.
- Chetty, R., J. N. Friedman, T. Olsen, and L. Pistaferri (2011). Adjustment costs, firm responses, and micro vs. macro labor supply elasticities: Evidence from danish tax records. *The Quarterly Journal of Economics* 126(2), 749–804.
- Coffman, K., M. Collis, and L. Kulkarni (2019). Stereotypes and Belief Updating. *Working Paper*.
- Coffman, K. B. (2014). Evidence on Self-stereotyping and the Contribution of Ideas. *The Quarterly Journal of Economics* 129(4), 1625–1660.
- Coffman, K. B., P. U. Araya, and B. Zafar (2021). A (dynamic) investigation of stereotypes, belief-updating, and behavior. Technical report, National Bureau of Economic Research.
- Cohen, J., P. Cohen, S. West, and L. Aiken (1983). Applied multiple regression/correlation analysis for behaviour science.
- Coutts, A. (2018). Good News and Bad News are Still News: Experimental Evidence on Belief Updating. *Experimental Economics* 22, 369–395.
- Cox, J. C., J. Shachat, and M. Walker (2001). An experiment to evaluate Bayesian learning of Nash Equilibrium play. *Games and Economic Behavior* 33, 11–33.

- Crawford, V. P., U. Gneezy, and Y. Rottenstreich (2008). The power of focal points is limited: Even minute payoff asymmetry may yield large coordination failures. *American Economic Review* 98(4), 1443–58.
- Croson, R. and U. Gneezy (2009). Gender Differences in Preferences. *Journal of Economic Literature* 47(2), 448–74.
- Danz, D. N., D. Fehr, and D. Kübler (2012). Information and beliefs in a repeated normal-form game. *Experimental Economics* 15(4), 622–640.
- 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.
- Dee, T. S., W. Dobbie, B. A. Jacob, and J. Rockoff (2019). The causes and consequences of test score manipulation: Evidence from the new york regents examinations. *American Economic Journal: Applied Economics* 11(3), 382–423.
- Diamond, R. and P. Persson (2016). The long-term consequences of teacher discretion in grading of high-stakes tests.
- Dickert-Conlin, S. and A. Chandra (1999). Taxes and the timing of births. *Journal of Political Economy* 107(1), 161–177.
- Dubois, D., M. Willinger, and P. Van Nguyen (2012). Optimization incentive and relative riskiness in experimental Stag-Hunt games. *International Journal of Game Theory* 41(2), 369–380.
- Duffy, J. and D. Fehr (2018). Equilibrium selection in similar repeated games: experimental evidence on the role of precedents. *Experimental Economics* 21(3), 573–600.
- Eckel, C. C. and P. J. Grossman (2008). Men, Women and Risk Aversion: Experimental Evidence. *Handbook of Experimental Economics Results* 1, 1061–1073.
- Edmonds, K. (2011). *Dewhurst’s textbook of obstetrics and gynaecology*. Hoboken, NJ: Wiley-Blackwell.
- Efron, B. and R. Tibshirani (1993). *An Introduction to the Bootstrap*. Chapman and Hall.
- 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*.

- Embrey, M., G. R. Fréchet, and S. Yuksel (2017). Cooperation in the finitely repeated Prisoner's Dilemma. *The Quarterly Journal of Economics* 133(1), 509–551.
- 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.
- Feltovich, N. and S. H. Oda (2014). Effect of matching mechanism on learning in games played under limited information. *Pacific Economic Review* 3, 260–277.
- Figlio, D., J. Guryan, K. Karbownik, and J. Roth (2014). The effects of poor neonatal health on children's cognitive development. *American Economic Review* 104(12), 3921–55.
- Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* 10, 171–178.
- Foster, D. and H. Young (2006). Regret testing leads to Nash Equilibrium. *Theoretical Economics*.
- Franco, C. (2018). How does Relative Performance Feedback affect Beliefs and Academic Decisions? Evidence from a Field Experiment. *Working Paper*.
- Fretts, R. C. (2018). Effects of advanced maternal age on pregnancy.
- Friedman, D., S. Huck, R. Oprea, and S. Weidenholzer (2015). From imitation to collusion: Long-run learning in a low-information environment. *Journal of Economic Theory* 155, 185–205.
- Fudenberg, D. and D. K. Levine (2009). Self-confirming equilibrium and the Lucas critique. *Journal of Economic Theory* 144(6), 2354–2371.
- Galal, M., I. Symonds, H. Murray, F. Petraglia, and R. Smith (2012). Postterm pregnancy. *Facts, Views & Vision in ObGyn* 4(3), 175.
- Gans, J. S. and A. Leigh (2009). Born on the first of july: An (un) natural experiment in birth timing. *Journal of Public Economics* 93(1-2), 246–263.
- Gans, J. S., A. Leigh, and E. Varganova (2007). Minding the shop: The case of obstetrics conferences. *Social Science & Medicine* 65(7), 1458–1465.
- Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin (2013). *Bayesian data analysis*. Texts in Statistical Science no. 3. Boca Raton, FL:CRC.

- Ghidoni, R., B. L. Cleave, and S. Suetens (2019). Perfect and imperfect strangers in social dilemmas. *European Economic Review* 116, 148–159.
- Golman, R., D. Hagmann, and G. Loewenstein (2017). Information Avoidance. *Journal of Economic Literature* 55(1), 96–135.
- Goodman, M. J., W. W. Nelson, and M. V. Maciosek (2005). Births by day of week: A historical perspective. *Journal of Midwifery & Women's Health* 50(1), 39–43.
- Gould, J. B., C. Qin, A. R. Marks, and G. Chavez (2003). Neonatal mortality in weekend vs weekday births. *Jama* 289(22), 2958–2962.
- Greiner, B. (2015a). Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE. *Journal of the Economic Science Association* 1(1), 114–125.
- Greiner, B. (2015b). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association* 1, 114–125.
- Grobman, W. A., M. M. Rice, U. M. Reddy, A. T. Tita, R. M. Silver, G. Mallett, K. Hill, E. A. Thom, Y. Y. El-Sayed, A. Perez-Delboy, et al. (2018). Labor induction versus expectant management in low-risk nulliparous women. *New England Journal of Medicine* 379(6), 513–523.
- Hamilton, P. and E. Restrepo (2006). Sociodemographic factors associated with weekend birth and increased risk of neonatal mortality. *Journal of Obstetric, Gynecologic & Neonatal Nursing* 35(2), 208–214.
- Hannah, M. E., W. J. Hannah, J. Hellmann, S. Hewson, R. Milner, A. Willan, and C. M. P. term Pregnancy Trial Group* (1992). Induction of labor as compared with serial antenatal monitoring in post-term pregnancy: a randomized controlled trial. *New England Journal of Medicine* 326(24), 1587–1592.
- Harsanyi, J. and R. Selten (1988). *A General Theory of Equilibrium Selection in Games*. MIT Press.
- Hart, S. and A. Mas-Colell (2006). Stochastic uncoupled dynamics and Nash Equilibrium. *Games and Economic Behavior* 57(2), 286–303.
- Hawe, E., A. MacFarlane, and J. Bithell (2001). Daily and seasonal variation in live births, stillbirths and infant mortality in England and Wales, 1979–96. *Health Statistics Quarterly* (9), 5–15.
- Healy, P. J. (2020). Explaining the BDM - or any random Binary Choice Elicitation Mechanism - to Subjects. *Working Paper*.
- Hendry, R. A. (1981). The weekend—a dangerous time to be born? *BJOG: An International Journal of Obstetrics & Gynaecology* 88(12), 1200–1203.

- Hong, J., H. Kang, S.-W. Yi, Y. Han, C. Nam, B. Gombojav, and H. Ohrr (2006). A comparison of perinatal mortality in Korea on holidays and working days. *British Journal of Obstetrics & Gynaecology* 113(11), 1235–1238.
- Huck, S., J. Leutgeb, and R. Oprea (2017). Payoff information hampers the evolution of cooperation. *Nature Communications* 8, 1–5.
- Jürges, H. (2017). Financial incentives, timing of births, and infant health: a closer look into the delivery room. *The European Journal of Health Economics* 18(2), 195–208.
- 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.
- Kavanagh, J., A. Kelly, and J. Thomas (2005). Breast stimulation for cervical ripening and induction of labour. *The Cochrane database of systematic reviews* (3), CD003392.
- Kendall, R. (2020). Decomposing coordination failure in Stag Hunt games. Working paper.
- Kleven, H. J. (2016). Bunching. *Annual Review of Economics* 8, 435–464.
- Kleven, H. J. and M. Waseem (2013). Using notches to uncover optimization frictions and structural elasticities: Theory and evidence from Pakistan. *The Quarterly Journal of Economics* 128(2), 669–723.
- Kneeland, T. (2015). Identifying higher-order rationality. *Econometrica* 83(5), 2065–2079.
- Knoepfle, D. T., J. T.-y. Wang, and C. F. Camerer (2009). Studying learning in games using eye-tracking. *Journal of the European Economic Association* 7(2-3), 388–398.
- 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.
- LaLumia, S., J. M. Sallee, and N. Turner (2015). New evidence on taxes and the timing of birth. *American Economic Journal: Economic Policy* 7(2), 258–93.
- Levy, B. R., P. H. Chung, and M. D. Slade (2011). Influence of valentine’s day and halloween on birth timing. *Social science & medicine* 73(8), 1246–1248.
- Lin, H.-C., S. Xirasagar, and Y.-C. Tung (2006). Impact of a cultural belief about ghost month on delivery mode in Taiwan. *Journal of Epidemiology & Community Health* 60(6), 522–526.

- Lo, J. (2003). Auspicious time and cesarean section. *Taiwan Journal of Public Health* 22, 134–140.
- 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. Punčochař (1994). Highly Confident but Wrong: Gender Differences and Similarities in Confidence Judgments. *Journal of Educational Psychology* 86(1), 114.
- Luthy, D. A., J. A. Malmgren, and R. W. Zingheim (2004). Cesarean delivery after elective induction in nulliparous women: the physician effect. *American Journal of Obstetrics and Gynecology* 191(5), 1511–1515.
- Macfarlane, A. (1978). Variations in number of births and perinatal mortality by day of week in england and wales. *British Medical Journal* 2(6153), 1670–1673.
- Mangold, W. D. (1981). Neonatal mortality by the day of the week in the 1974-75 arkansas live birth cohort. *American Journal of Public Health* 71(6), 601–605.
- Martin, P., M. Cortina-Borja, M. Newburn, G. Harper, R. Gibson, M. Dodwell, N. Dattani, and A. Macfarlane (2018). Timing of singleton births by onset of labour and mode of birth in nhs maternity units in england, 2005–2014: A study of linked birth registration, birth notification, and hospital episode data. *PloS ONE* 13(6), e0198183.
- Mathers, C. (1983). Births and perinatal deaths in australia: variations by day of week. *Journal of Epidemiology & Community Health* 37(1), 57–62.
- McKelvey, R. D. and T. R. Palfrey (2001). Playing in the dark: Information, learning, and coordination in repeated games. Working paper.
- 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*.
- Modlock, J., B. B. Nielsen, and N. Uldbjerg (2010). Acupuncture for the induction of labour: a double-blind randomised controlled study. *British Journal of Obstetrics & Gynaecology* 117(10), 1255–1261.
- Mookherjee, D. and B. Sopher (1994). Learning behavior in an experimental Matching Pennies game. *Games and Economic Behavior* 7(1), 62–91.
- Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review* 85(5), 1313–1326.

- Nash, J. (1950). Non-cooperative games. *Ph.D. Dissertation, Princeton University*.
- Nax, H. H., M. N. Burton-Chellew, S. A. West, and H. P. Young (2016). Learning in a black box. *Journal of Economic Behavior & Organization* 127, 1–15.
- Neugart, M. and H. Ohlsson (2013). Economic incentives and the timing of births: evidence from the german parental benefit reform of 2007. *Journal of Population Economics* 26(1), 87–108.
- Nicklisch, A. (2011). Learning strategic environments: an experimental study of strategy formation and transfer. *Theory and Decision* 71(4), 539–558.
- 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*.
- Nielsen, P. E., B. C. Howard, C. C. Hill, P. L. Larson, R. H. Holland, and P. N. Smith (2005). Comparison of elective induction of labor with favorable bishop scores versus expectant management: a randomized clinical trial. *The Journal of Maternal-Fetal & Neonatal Medicine* 18(1), 59–64.
- Nikolaychuk, O. (2012). Does it pay to know more in games of incomplete information? Working paper.
- Nyarko, Y. and A. Schotter (2003, December). An experimental study of belief learning using elicited beliefs. *Econometrica* 70(3), 971–1005.
- Oechssler, J. and B. Schipper (2003). Can you guess the game you are playing? *Games and Economic Behavior* 43, 137–152.
- of Obstetricians, A. A. C. and Gynecologists). Acog committee opinion no. 560: Medically indicated late-preterm and early-term deliveries. *Obstetrics and Gynecology* 121(4), 908–910.
- Oprea, R. and S. Yuksel (2021). Social Exchange of Motivated Beliefs. *Journal of the European Economic Association*.
- Oster, E. (2018). Expert behavior change in response to best practice changes: Evidence from obstetrics.
- Osterman, M. J. and J. A. Martin (2014). Recent declines in induction of labor by gestational age. *NCHS Data Brief, National Center for Health Statistics* (155), 1–8.

- Pasupathy, D., A. M. Wood, J. P. Pell, M. Fleming, and G. C. Smith (2010). Time of birth and risk of neonatal death at term: retrospective cohort study. *British Medical Journal* 341.
- 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*.
- Polonio, L. and G. Coricelli (2019). Testing the level of consistency between choices and beliefs in games using eye-tracking. *Games and Economic Behavior* 113, 566–586.
- Rask, K. and J. Tiefenthaler (2008). The role of grade sensitivity in explaining the gender imbalance in undergraduate economics. *Economics of Education Review* 27(6), 676–687.
- Raven, J. C. J. C. (1973). *Advanced Progressive Matrices*. London: H.K. Lewis.
- Restrepo, E., P. Hamilton, F. Liu, and P. Mancuso (2018). Relationships among neonatal mortality, hospital volume, weekday demand, and weekend birth. *Canadian Journal of Nursing Research* 50(2), 64–71.
- Rindfuss, R. R., J. L. Ladinsky, E. Coppock, V. W. Marshall, and A. Macpherson (1979). Convenience and the occurrence of births induction of labor in the united states and canada. *International Journal of Health Services* 9(3), 439–460.
- Royer, H. (2009). Separated at girth: Us twin estimates of the effects of birth weight. *American Economic Journal: Applied Economics* 1(1), 49–85.
- Saccone, G. and V. Berghella (2015). Omega-3 supplementation to prevent recurrent preterm birth: a systematic review and metaanalysis of randomized controlled trials. *American Journal of Obstetrics and Gynecology* 213(2), 135–140.
- Saez, E. (2010). Do taxpayers bunch at kink points? *American economic Journal: Economic Policy* 2(3), 180–212.
- Schmidt, D., R. Shupp, J. M. Walker, and E. Ostrom (2003). Playing safe in coordination games: the roles of risk dominance, payoff dominance, and history of play. *Games and Economic Behavior* 42(2), 281–299.
- Schulkind, L. and T. M. Shapiro (2014). What a difference a day makes: Quantifying the effects of birth timing manipulation on infant health. *Journal of Health Economics* 33, 139–158.
- Smith, C. A., C. A. Crowther, and S. J. Grant (2013). Acupuncture for induction of labour. *Cochrane Database of Systematic Reviews* (8).

- Spong, C. Y., B. M. Mercer, M. D'Alton, S. Kilpatrick, S. Blackwell, and G. Saade (2011). Timing of indicated late-preterm and early-term birth. *Obstetrics and Gynecology* 118(2 Pt 1), 323.
- Stahl, D. and P. Wilson (1994). Experimental evidence on players? Models of other players. *Journal of Economic Behavior & Organization* 25(3), 309–327.
- Stephansson, O., P. W. Dickman, A. L. Johansson, H. Kieler, and S. Cnattingius (2003). Time of birth and risk of intrapartum and early neonatal death. *Epidemiology* 14(2), 218–222.
- Tamm, M. (2013). The impact of a large parental leave benefit reform on the timing of birth around the day of implementation. *Oxford Bulletin of Economics and Statistics* 75(4), 585–601.
- Thomsen, D. M. (2018). Gender differences in candidate reemergence. *Working Paper*.
- Tita, A. T., M. B. Landon, C. Y. Spong, Y. Lai, K. J. Leveno, M. W. Varner, A. H. Moawad, S. N. Caritis, P. J. Meis, R. J. Wapner, et al. (2009). Timing of elective repeat cesarean delivery at term and neonatal outcomes. *New England Journal of Medicine* 360(2), 111–120.
- Walker, K. F., P. Wilson, G. J. Bugg, A. Dencker, and J. G. Thornton (2015). Childbirth experience questionnaire: validating its use in the united kingdom. *BMC Pregnancy and Childbirth* 15(1), 1–8.
- Wasserman, M. (2021). Gender Differences in Politician Persistence. *Review of Economics and Statistics*.
- Young, H. P. (2009). Learning by trial and error. *Games and Economic Behavior* 65(2), 626–643.
- Zhang, J., H. J. Landy, D. W. Branch, R. Burkman, S. Haberman, K. D. Gregory, C. G. Hatjis, M. M. Ramirez, J. L. Bailit, V. H. Gonzalez-Quintero, et al. (2010). Contemporary patterns of spontaneous labor with normal neonatal outcomes. *Obstetrics and gynecology* 116(6), 1281.
- Zimmermann, F. (2020). The Dynamics of Motivated Beliefs. *American Economic Review* 110(2), 337–361.