Identity Formation, Gender Differences, and the Perpetuation of Stereotypes

Erin Giffin∗

University of California, San Diego

CLICK FOR CURRENT DRAFT†

November 13, 2017

Abstract

Gender differences in economic decisions are well-documented and span a variety of important choices. Laboratory experiments have been used to identify potential mechanisms to explain these differences, and the results have most frequently been attributed to men and women having different preferences, especially when subjects’ choices are anonymous. In this paper, I propose a theoretical model that highlights that persistent gender differences can arise without differences in preferences. I show that if two groups are identical ex ante but there exists a stereotype about one of the groups, then groups will behave in ways consistent with this stereotype in equilibrium. Extending this to a multi-period model, I show that if individuals endogenously form group identities through habit formation, these differences will persist in the long-run, even after choices are no longer observed. The model thus depicts a mechanism through which external constraints are eventually internalized and captures how social norms can become self-enforced by individuals. Using multiple existing experimental datasets where gender data were collected but never analyzed, I find evidence consistent with my model’s predictions. I then conduct a new experiment to directly test the proposed mechanism. I show that by exposing subjects to external constraints in initial decisions, I mitigate gender differences in altruism. Moreover, this remains true even when those external constraints are removed. However, when subjects are not initially exposed to these constraints, women are significantly more generous than men.

∗I would like to thank Jim Andreoni, Isabel Trevino, and Joel Watson for their feedback throughout the course of this project. I am grateful to Kate Antonovics, Craig McKenzie, and Shanthi Manian for particularly helpful comments. I would also like to thank Jim Andreoni, B. Douglas Bernheim, and Justin Rao for making their data available. I am grateful to the Department of Economics and the Frontiers of Innovation Scholarship Program at UC San Diego for financial support.
†Or go to http://econweb.ucsd.edu/~egiffin/pdfs/Giffin_JMP.pdf
1 Introduction

Gender differences in economic decisions are well-documented and span many important dimensions of economic choices. Men and women differ in their consumption and savings behaviors (LIMRA, 2016), human capital investments (Ceci et al., 2014), choice of college major (National Center for Education Statistics, 2016), and occupational choice (Sapienza et al., 2009). At the household level, there are gender differences in the division of labor within the household (Bertrand et al., 2015) and expenditure on children (Thomas, 1990). Gender differences have also been documented in economic outcomes, with the greatest attention on the gender wage gap (e.g., Bertrand et al., 2010). While the results from observational data have established empirical facts, they frequently do not provide an explanation for these differences. However, understanding the mechanisms behind these differences is important both to interpret and understand results from observational data as well as to inform the optimal policy response. Experimentalists have studied gender differences in laboratory settings to examine potential mechanisms in a controlled environment. Experimental results that find that women are less competitive than men have been used to explain gender differences in occupational choice, namely why there are not more women in top executive positions (Saccardo et al., 2017). Results on how women are less likely to negotiate have been used to explain part of the gender gap in earnings (Babcock and Laschever, 2003). Results that women are more likely to accept requests have been used to explain why women are more likely than men to complete non-promotional tasks at work (Babcock et al., 2017).

There are two potential explanations for these results: either men and women face different constraints, or men and women differ in terms of fundamentals (i.e., men and women have different preferences or are different types). In anonymous laboratory settings, differences in observable choices are most often attributed to differences in preferences, because constraints that individuals face outside the laboratory should not apply (e.g., Croson and Gneezy, 2009).

In this paper, I propose a novel mechanism that can explain gender differences in anonymous lab settings without assuming different preferences. I propose that external costs, which are different for men and women, become internalized over time through habit formation. As a result, individuals will adhere to behaviors dictated by social norms even when no one is watching. In this paper, I focus on altruistic choices, as this is the focus of a large portion of experimental papers on gender differences (see, for example, Bolton and Katok, 1995; Eckel and Grossman, 1998; Andreoni and Vesterlund, 2001). I then test the model in two ways: I first conduct an empirical analysis to test the model’s predictions using existing experimental datasets where gender data were collected but never analyzed, and second I design and implement a new experiment as a direct test of the model’s mechanism. Using both of these methods, I find empirical support for the model’s validity.

My theoretical model begins with the assumption that men and women are identical in both their preferences (utility functions) and types. In the model, a decision-maker, who is either a man or a woman, chooses to act either selfishly or fairly, and this choice and their gender is observed. Individuals care about their own consumption as well as how others view them. Based on the decision-maker’s choice, observers make inferences about the decision maker’s character. I show
that if there is at least one observer who draws harsher inferences about a female decision-maker’s character if she chooses to act selfishly, then women will be more likely to behave generously in equilibrium. The model thus predicts that stereotypes will perpetuate: men and women behave differently ex post because of the stereotype, even though they were identical ex ante.

I then extend this to a multi-period model and allow individuals to endogenously form gender identities. Through habit formation, as decision-makers behave in a way that is stereotypical of their gender, the association between those behaviors and their own gender strengthens. Through identity formation, individuals internalize external constraints. I show that after gender identities are formed, gender differences in behaviors will persist, even after choices are no longer observed. So although initial group differences were driven by observers’ beliefs about men and women, these differences will be perpetuated in the long-run through identity formation.

I then conduct an empirical analysis using existing experimental data where gender data were collected but never analyzed and find evidence that is consistent with the model’s predictions. Specifically, I find that women are more generous than men when their decision is observed, even when given the opportunity to hide selfish actions, and that women are significantly more generous than men in an anonymous dictator game where they are asked to give a particular allocation. I also find an interesting secondary result: that although these datasets were not collected with the intention of examining gender differences, the results of the papers that originally used these datasets were partly or entirely driven by only one gender (men in one and women in the other). That is, I find that the results of the original papers were only statistically significant because either only men or only women were responsive to the experimental treatment and the result was strong enough to make the pooled result statistically significant.

I finally design and implement an experimental test of the model’s mechanism. In the experiment, subjects made a series of decisions on how to allocate $30 between themselves and their partner. To generate external constraints, in some decisions subjects’ choices were perfectly observed by others in the experiment, while in others, subjects had plausible deniability. For these decisions, there was a chance that subjects could not make a choice and an allocation where they kept everything was made for them. This offered subjects an opportunity to hide selfish actions, because if others in the experimental session saw that the subject was allocated everything, they could not be sure if the subject made this choice or if this choice was made for them.

Experimental treatments varied in the order subjects made decisions. Subjects’ first choice either offered no opportunity for plausible deniability (high external constraint) and this opportunity increased in subsequent decisions or their first choice offered the greatest opportunity for plausible deniability (low external constraint) and this opportunity decreased in subsequent decisions. I find evidence of persistence of behaviors, as I find that subjects’ decisions are relatively stable over the series of decisions. I also find that by imposing stricter constraints on subjects’ initial action, male subjects make more generous allocations and continue to behave similarly to women even in later

---

1I show that this is also true even if all observers draw identical inferences for both genders, but women anticipate that there is at least one observer who will draw harsher inferences against them.
decisions when these constraints are relaxed. Specifically, I find that when subjects’ first decision has low external constraints, at every level of nature intervening, women are more likely to choose equal allocations than men. However, by simply changing the order of decisions so early decisions have higher external constraints, I mitigate gender differences, as men and women are equally likely to choose a 50-50 split of the pie in this treatment.

This paper provides two important contributions to the gender differences literature. I first introduce a novel mechanism for understanding gender differences. This is the first model (to the best of my knowledge) that shows persistent gender differences without assuming differences in fundamentals. I also provide additional evidence of gender differences, even in contexts where the researchers were not looking for them. This suggests gender differences, and more largely adherence to social norms, may be more prevalent than we realize.

This paper additionally contributes to the literature on social norms and social prescriptions. These two are largely viewed as distinct, with the former relating to behaviors that are externally punished if not followed (e.g., Akerlof, 1976; Kandori, 1992; Cole et al., 1992) and the latter relating to behaviors that are self-enforced (e.g., Akerlof and Kranton, 2000; Huang and Wu, 1994). I connect these two literatures, as I propose a mechanism where one is generated by internalizing the other. This also suggests a powerful way in which social norms can perpetuate, as eventually external enforcement is no longer necessary for an individual to continue adhering to the norm.

Relatedly, this paper contributes to the literature on identity economics. There is a well-established literature (beginning with Akerlof and Kranton (2000)) on identity economics—the idea that individuals have an identity and derive disutility from taking an action inconsistent with that identity. While this mechanism makes good predictions for many behaviors, it does not address how these identities may form. It assumes that an individual is endowed with both a group membership and an identity with that group and does not want to deviate from the behavioral norms associated with that group. My model, in contrast, takes a step back. It does not assume that identities are endowed, but rather are endogenously formed. I assume that group membership is randomly assigned, but then individuals are incentivized to behave in ways consistent with the norms of their group membership. As agents continue taking actions consistent with their group, the association between themselves and the behaviors associated with their group strengthens through habit formation. Then, over time, gender identities are solidified. After this point, agents will continue to act in accordance with the prescriptions of their group membership, even if actions are not observed (so there are no external incentives for adhering to the social norm).

The most closely related paper in the theoretical literature is Coate and Loury (1993). Coate and Loury determine that even if two identifiable groups are identical ex ante, an affirmative action policy can create a situation in which employers correctly perceive the groups to be unequally productive ex post. This relates to my model in that both Coate and Loury (1993) and I are able to generate differences in observable behaviors without assuming differences in fundamentals about the groups. The most important distinction between their model and my own is that in their model, differences only persist as long as observers (in their model, employers) are able to observe
an individual’s group membership. If employers were not able to observe a potential employee’s group membership, groups would behave identically. In my model, because of the addition of habit formation, I show that group differences can persist even when observers cannot observe an individual’s group membership.

With respect to the experimental literature, my empirical analysis is most closely related to Andreoni and Vesterlund (2001) and DellaVigna et al. (2013). Both of these papers re-analyze an existing dataset and test for gender differences. Andreoni and Vesterlund find that men are more sensitive to the price of giving, while women appear more egalitarian, even when giving is expensive. DellaVigna et al. find that men and women are equally generous in a door-to-door solicitation, but that women become less generous when it is easy to avoid the solicitor. While both papers report significant gender differences, each of these papers re-analyzes only one dataset. In this paper I analyze multiple datasets, which allows me to come to different conclusions than one of these papers. Notably, DellaVigna et al. conclude from their analysis that women are more likely to be on the margin of giving, and are therefore more sensitive to experimental treatments. Using a larger number of datasets, I do not find support for this claim, as I find that men were more sensitive to experimental treatments in one of datasets I analyze as well as in my own experiments.

The paper proceeds as follows: In Section 2 I construct and analyze the model and develop a set of testable predictions. Section 3 presents the empirical analysis. Section 4 presents the experimental design, and the experimental results are presented in Section 5. Section 6 concludes. All proofs appear in the Appendix.

2 Model

I develop a model to analyze an individual’s decision to make altruistic choices. An individual may behave altruistically either because they care about fairness or because they desire others to perceive them as fair. In the model, individuals make a choice—to act either selfishly or fairly—and this choice is observed. Observers, after seeing the individual’s choice make an inference about their character, which is unobservable. Individuals may then act fairly because they inherently care about fairness to varying degrees and because they care about the inferences others make about their character. Individuals’ gender is visible and observers may form different inferences based on the individual’s gender. These different inferences provide different constraints for men and women, causing them to behave differently. Throughout an individual’s lifetime, they continue to face these same types of choices. As they continue to do so, they begin to internalize these different constraints. Eventually, individuals begin to self-enforce these mechanisms as these constraints become internalized.

This model thus shows how gender differences can be perpetuated, as I show that even when members of the two groups are identical ex ante, if there exists a stereotype that influences observers’ beliefs about the groups, group members may behave in ways consistent with this stereotype in equilibrium. Then, due to habit formation, these group difference will persist, even after choices
are no longer observed.

## 2.1 Setup

Two players—a decision-maker (D) and a receiver (R)—split a prize normalized to have unit value. D transfers \( x \in [0, 1] \) to R and consumes \( c = 1 - x \). Decision-makers belong to one of two groups and have label \( L \in \{M, W\} \) that discloses group membership. L is visible, making D’s group membership public information. Decision-makers are differentiated by a parameter, \( t \), that indicates the importance D places on fairness; \( t \) is D’s private information. The distribution of \( t \) has full support over the interval \([0, \bar{t}]\). \( K \) denotes the CDF, and I define \( K_T \) as the CDF obtained from \( T \), conditioning on \( T \leq t \). Groups and types (\( t \)) are uncorrelated, so groups are identical ex ante.

D cares about his own prize (\( c \)) as well as his social image (\( s \)), as perceived by an Audience (A), which includes R. \( F(c, s) \) is a utility function of \( c \) and \( s \). It is unbounded in both arguments, twice continuously differentiable, strictly increasing, and strictly concave in \( c \). The decision-maker also cares about fairness, which is determined by the extent to which the outcome departs from the fair alternative, \( x^F \).\(^2\) D’s total payoff is:\(^3\)

\[
U(x, s, t) = F(1-x, s) + tG(x - x^F)
\]

\( G \) is twice continuously differentiable, strictly concave, and reaches a maximum at zero. D’s social image, \( s \), depends on A’s perception of D’s fairness. I normalize \( s \) so that if A is certain D’s type is \( \bar{t} \), then D’s social image is \( \bar{s} \). \( \Phi \) denotes the CDF that represents A’s belief about D’s type and \( S(\Phi) \) is the associated social image. A forms an inference \( \Phi \) about \( t \) after observing \( x \) and \( L \). \( S \) is continuous and satisfies \( S(\Phi') > S(\Phi'') \) if \( \Phi' \) first-order stochastically dominates (FOSD) \( \Phi'' \). One possible functional form that the social image may take is \( E_D[E_A(t)] \), so D’s social image is her expectation of A’s expectation of her type.

I allow audience members to be heterogeneous and for audience members and the decision-maker to hold non-common priors. The decision-maker does not observe the inference directly, but she knows that A will judge her based on \( x \), so she accounts for this effect of her choice on A’s inference.

I restrict attention to pure strategy equilibria.

## 2.2 One Period Model

I first analyze the model where the game lasts only one period. For simplification, I restrict the decision-maker’s choice to \( x \in \{0, x^F\} \). Since the decision-maker’s choice is binary but there is a continuum of types, this precludes perfect separation. The following lemma shows that there is

\(^2\)\( x^F \) is most commonly \( \frac{1}{2} \), but I allow it to be a free parameter for generality.

\(^3\)This utility function was originally introduced by Andreoni and Bernheim (2009). They were the first to propose that individuals may act generously because they care about being perceived as fair.
a threshold type, \( t^*_L \), and all types above this threshold will choose to transfer the fair allocation while all types below the threshold will choose to transfer zero.

**Lemma 1.** There exists \( t^*_L \) for \( L \in \{M, W\} \) such that \( \forall t \geq t^*_L, D \) chooses \( x = x^F \) and \( \forall t < t^*_L, D \) chooses \( x = 0 \).

I first examine the case where all audience members hold the correct belief that groups are identical and that the decision-maker knows that \( A \) holds these beliefs. In this case, the threshold type will be the same across groups, as the next result shows.

**Proposition 1.** Let \( t^*_W \) denote the threshold type for group \( W \) and \( t^*_M \) denote the threshold type for group \( M \). If all audience members believe \( K^M = K^W = K \) and \( \Phi(t; W, x) = \Phi(t; M, x) = \Phi \) and these beliefs are common knowledge, then \( t^*_W = t^*_M = t^* \).

When all audience members know the true distribution of types and the decision-maker knows that they hold correct beliefs, then there will be no differences in group behavior. However, if even one audience member holds an incorrect belief about the groups, this result may break down.

I define a type of belief where audience members, upon observing a decision-maker choose the selfish allocation, draw harsher inferences about the decision-maker if she comes from group \( W \).

**Definition 1.** Belief \( B1 \): Belief such that \( \Phi(t; M, x = 0) \text{ FOSD } \Phi(t; W, x = 0) \).

There are multiple conditions that would lead audience members to draw inferences consistent with \( B1 \). Sufficient conditions for \( B1 \) include:

1. Distorted prior beliefs: An audience member believes that \( t \) is drawn from two different distributions and while the distribution of \( t \) still has full support over \([0, \bar{t}]\) for each group, he believes \( K^W > K^M \) for \( 0 < t < t^* \) and \( K^W < K^M \) for \( t^* < t < \bar{t} \) (\( K^W = K^M \) for \( t \in \{0, t^*, \bar{t}\} \)). These distorted beliefs imply that the audience member believes that members of group \( W \) are concentrated at the tails of the distribution, so members of group \( W \) are more likely to either be very low types or very high types.

2. Biased inferences: An audience member is biased (implicitly or explicitly) against group \( W \) and so upon observing \( x = 0 \) and that \( D \) is a member of group \( W \), he over-updates and arrives at different inferences about the decision-maker based on group membership.

Holding belief \( B1 \) means that that the audience member holds incorrect beliefs about the decision-maker’s type. I also allow for members of the audience to hold incorrect beliefs about the decision-maker’s preferences. In this case, audience members misspecify the decision-maker’s utility function. Specifically, audience members believe decision-makers from the two groups care about social image to different degrees and place different weights (\( \alpha \)) on the social image in their utility.

**Definition 2.** Belief \( B2 \): Belief that \( U_L = F(1-x, \alpha_L s) + tG(x-x^F), \alpha > 0 \) and \( \alpha_W > \alpha_M \).
If an audience member holds this belief, then he believes that members of group \( W \) care more about social image, and thus have a stronger preference for being perceived as fair, than members of group \( M \).\(^4\) If any audience members hold incorrect beliefs about either the decision-maker’s type or the decision-maker’s preferences, then group differences will arise, as the next result shows.

**Proposition 2.** If there exists at least one member of \( A \) who holds belief \( B_1 \) or belief \( B_2 \) and \( D \) knows this, then \( t^*_W < t^*_M \).

The above result illustrates that it is sufficient for just one member of the audience to hold incorrect beliefs to result in group differences. The next result shows that even if all audience members hold correct beliefs, this is not sufficient to guarantee no differences between groups. As the next result shows, even if all audience members hold correct beliefs about both decision-makers’ types and preferences, but the decision-maker believes that at least one audience member holds misspecified beliefs, then group differences will arise.

**Proposition 3.** If all members of \( A \) believe \( \Phi(t; W, x) = \Phi(t; M, x) = \Phi \) and \( \alpha_W = \alpha_M = \alpha \), but \( D \) believes there exists at least one audience member who holds belief \( B_1 \) or belief \( B_2 \), then \( t^*_W < t^*_M \).

This section examined an equilibrium where actions are perfectly observed. We can easily imagine scenarios where this is not the case. The next section examines the case where the observation of the decision-maker’s choice is noisy.

### 2.3 One Period Model with Noisy Signal

I now consider what happens if the signal, \( x \), is noisy. Suppose now that nature intervenes with probability \( p \in (0, 1) \). If nature intervenes, \( x = 0 \) is transferred regardless of the decision-maker’s choice. \( p \) is common knowledge, but \( R \) and \( A \) cannot observe if nature intervened.

The following result demonstrates that introducing some plausible deniability decreases the threshold type. This implies that a lower fraction of decision-makers will choose \( x = x_F \), resulting in greater pooling at the bottom.

**Lemma 2.** For \( L \in \{M, W\} \), \( t^*_L \) is increasing in \( p \).

Although the threshold type for both groups falls when decision-makers can “hide” behind nature, unless the decision-maker and all audience members hold correct beliefs, at each level of \( p \), a larger fraction of group \( W \) will still select the fair allocation, as demonstrated by the next result.

**Proposition 4.** Let \( t^*_{p,L} \) denote the threshold \( t \) for group \( L \) when the probability of intervention is \( p \). If there exists at least one audience member who holds belief \( B_1 \) or \( B_2 \) or \( D \) believes there exists at least one audience member who holds belief \( B_1 \) or \( B_2 \), then \( t^*_{p,W} < t^*_{p,M} \) for any \( p \in (0, 1) \).

\(^4\) Assume that if \( x = 0 \), then \( s < 0 \) and if \( x = x_F \), \( s > 0 \). Then, under belief \( B_1 \), a member of group \( W \) gets greater disutility from a low social image and greater utility from a high social image.
The above result shows that even with a noisy signal, group differences will still exist and that members of group $W$ will behave more generously even when there is an opportunity for plausible deniability. Although the fraction of both groups voluntarily giving $x = 0$ grows, at every level of $p$ this fraction will be smaller for group $W$ than for group $M$.

2.4 Multi-period Model with Habit Formation

I show above that when there are stereotype-based ideas about groups and these ideas influence beliefs about the groups, then individuals will behave consistently with this stereotype in equilibrium. That is, even when the two groups are ex ante identical, expectations can result in group differences. Now I want to determine if these differences can persist in the long-run even in contexts where social image is not a concern (for example, because the decision-maker’s choice is not observed in some period).

$D$ participates in a sequence of dictator games, getting rematched with a different receiver and audience in each game. Each game is denoted by $g \in [1, \bar{g}]$. The sequence consists of two phases: in the first phase ($g \in [1, \hat{g}]$) actions are observed, and in the second phase ($g \in [\hat{g} + 1, \bar{g}]$) actions are not observed. I assume that the decision-maker has habit formation, so the more times he has taken an action in the past, the more likely he is to take this action in the current period. Let $x_g$ denote $D$’s transfer in game $g$ and $s_g$ denote $D$’s social image in game $g$ (there is no transfer of social image between games because the audience is different in each game). $r \in [0, \overline{r}]$ is the weight $D$ places on habit formation. The decision-maker places more weight on more recent actions, so past actions are time-discounted by a factor $\delta \in (0, 1)$. $D$’s utility function for each game can be written as the following:

Phase 1:

$$U = F(1 - x_g, s_g) + tG(x_g - x^F) + rH\left(\sum_{j=1}^{g-1} \delta^j \mathbb{1}\{x_{g-j} = x_g\}\right)$$

Phase 2:

$$U = F(1 - x_g) + tG(x_g - x^F) + rH\left(\sum_{j=1}^{g-1} \delta^j \mathbb{1}\{x_{g-j} = x_g\}\right)$$

I assume $H$ is twice continuously differentiable, strictly increasing, non-negative, and $H(0) = 0$.

Note that the above utility functions differ in that $s$ does not enter the utility function in Phase 2. Since actions are not observed in this phase, the audience cannot draw inferences about $D$’s type and thus social image is not a concern.

The following result demonstrates that although initial group differences are due to contexts where social image is relevant, habit formation can eventually make these differences permanent, so members of the two groups behave differently even when choices are anonymous.

**Proposition 5.** For $r > 0$, if there exists at least one audience member who holds the same belief $B1$ or belief $B2$ or $D$ believes there exists at least one audience member who holds the same belief $B1$ or belief $B2 \forall g \in [1, \bar{g}]$, then $\exists g^\ast$ such that $\forall g > g^\ast$, $D$ chooses $x_g = x_{g-j}$ with probability 1.
This result illustrates that my model gives rise to behavioral differences between groups that persist in the long-run, even in contexts where choices are anonymous, despite the assumption that groups were identical ex ante. Group differences are initially driven by the difference in inferences, but these initial differences will eventually become permanent due to habit formation.

2.5 Discussion

This model proposes a mechanism by which individuals internalize external constraints. While the external constraints were initially necessary for group differences to arise, eventually these external constraints become internalized. Individuals then self-enforce social norms and consequently adhere to the norm even when no one is watching.

This model also allows for gender identities to form endogenously. Previous papers on identity assume either that identities are exogenously endowed ex ante (Akerlof and Kranton, 2000) or that another exogenous event, for example puberty, causes individuals to form gender identities (Bharadwaj and Cullen, 2017). My model does not require either of these, as in the model identities are formed entirely through habit formation. Thus, simply behaving in a way that is consistent with the norms of a particular group over time causes individuals to identify with that group.

2.6 Testable Predictions

The model generates two key testable predictions:

1. Women will be more generous when their choice is observable, even when they are offered opportunities for plausible deniability.

2. If women have sufficient life experience, women will be more generous even when no one is watching.

3 Empirical Analysis

I conduct an empirical analysis to test for evidence of the model’s predictions. I use existing data on dictator games where gender data were collected but never analyzed. Compared to using published results from the gender differences literature, this is a cleaner test of the model’s findings, as I was not aware of if there were gender differences in the data before conducting the analysis.

The empirical analysis uses data from two previous experiments that involve dictator games. I do not rely only on the results of these papers, but I use their raw data to perform new analysis. These datasets are Andreoni and Bernheim (2009) “Social Image and the 50-50 Norm: A theoretical and experimental analysis of audience effects” and Andreoni and Rao (2011) “The Power of Asking: How communication affects selfishness, empathy, and altruism”. Going forward, these studies will be referred to as AB and AR. Dataset AB allows me to test the model’s first prediction that women will be more generous than men when offered plausible deniability and dataset AR allows me to test the model’s second prediction that women will be more generous even in anonymous settings.
For each of these datasets, I first discuss the key features of the experimental design as well as the original paper’s main result for comparison to my new analysis. Then, I present my new analysis using gender data.

3.1 AB

AB examines preferences for fairness versus preferences for being perceived as fair. The experimental design allowed individuals to “hide” their selfish actions by giving them plausible deniability. At the beginning of the experiment, subjects were divided into pairs, and partners were seated opposite one another, so all subjects knew with whom they were paired. Allocators needed to decide how to split $20 between themselves and their partner. For 9 separate dictator games, there was a probability that nature intervened, which varied between 0, 0.25, 0.5, and 0.75. If nature intervened, the allocator could not choose the allocation, and instead a predetermined amount ($x_0$ or $20 - x_0$) was transferred. There were two treatments, one where $x_0 = 0$ and one where $x_0 = 1$.\(^5\) At the end of the experiment, one of the decisions was randomly selected and the outcome for each pair was made public.\(^6\) The experiment involved 120 subjects (60 men and 60 women), all undergraduates at the University of Wisconsin–Madison.\(^7\)

3.1.1 Original Results

Figures 1 and 2 show the distributions of dictators’ voluntary choices in the two conditions ($x_0 = 0$ and $x_0 = 1$, respectively). Values of $x$ are grouped into five categories: $x = 0$, $x = 1$, $2 \leq x \leq 9$, $x = 10$, and $x > 10$. Looking at Figure 1, when $p = 0$, 57 percent of allocators transfer half the prize. As $p$ increases, this fraction steadily declines, and when $p = 75$, only 28 percent of subjects split the prize equally. As $p$ increases, the fraction of subjects transferring nothing grows, starting at 30 percent when $p = 0$ and ending at 70 percent when $p = 75$.

Looking at Figure 2, a large fraction of subjects choose to split the prize evenly when $p = 0$ (69 percent) and, like in the previous condition, this fraction declines at $p$ increases, shrinking to 34 percent when $p = 75$. Conversely, the fraction of subjects transferring 1 to their partner grows substantially as $p$ increases, beginning at only 3 percent when $p = 0$ and growing to 48 percent when $p = 75$.

Table 1 reports the results of two linear probability models. Looking at the first column of Table 1, the probability of choosing $x = x_0$ increases by approximately 27 percentage points when $p$ increases from 0 to 0.25, and increases by approximately 15 percentage points when $p$ increases from 0.25 to 0.5. This suggests that there is a significant increase in pooling at $x_0$ at these increases in $p$ but not when $p$ rises from 0.5 to 0.75. Looking at the second column, the coefficients imply that there is a significant decrease in pooling at $x = 10$ when $p$ increases from 0 to 0.25, as the

\(^5\)Each subject participated in only one of the treatments.
\(^6\)The experimenter wrote the final allocation on the board at the front of the room. This decision sheet was also used to determine payments.
\(^7\)One pair in condition $x_0 = 1$ did not complete the experiment, so only 118 subjects are included in analysis.

11
Figure 1: Distribution of amounts allocated to partners, condition $x_0 = 0$

Figure 2: Distribution of amounts allocated to partners, condition $x_0 = 1$

probability of choosing $x = 10$ decreases by nearly 24 percentage points, but there is no significant decline when $p$ increases from 0.25 to 0.5 or 0.5 to 0.75. Similar results hold when I separate by condition (estimates reported in Table 2).

### 3.1.2 Gender Analysis

This dataset allows me to test the model’s first prediction that women will be more generous than men even when offered an opportunity to hide a selfish action behind a noisy signal.

Figures 3 and 4 show the distributions of dictators’ voluntary choices in the two conditions ($x_0 = 0$ and $x_0 = 1$, respectively) separately for men and women. The differences in these distributions is particularly striking in Figure 3. Nearly 40 percent of men transfer nothing when $p = 0$ and this increases to over 80 percent when $p = 75$. By contrast, only half as many women (approximately 20 percent) choose $x = 0$ when $p = 0$ and this fraction increases to 57 percent when $p = 75$. Looking at even-splits, 56 percent of men transfer half the prize when $p = 0$ and this shrinks to 12 percent
Table 1: Linear Probability Models

<table>
<thead>
<tr>
<th></th>
<th>Probability of Choosing $x = x_0$</th>
<th>Probability of Choosing $x = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p \geq 25$</td>
<td>0.271***</td>
<td>-0.237***</td>
</tr>
<tr>
<td></td>
<td>(0.0786)</td>
<td>(0.0761)</td>
</tr>
<tr>
<td>$p \geq 50$</td>
<td>0.153***</td>
<td>-0.0678</td>
</tr>
<tr>
<td></td>
<td>(0.0549)</td>
<td>(0.0476)</td>
</tr>
<tr>
<td>$p = 75$</td>
<td>0.000</td>
<td>-0.0169</td>
</tr>
<tr>
<td></td>
<td>(0.0487)</td>
<td>(0.0525)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.169***</td>
<td>0.627***</td>
</tr>
<tr>
<td></td>
<td>(0.0549)</td>
<td>(0.0514)</td>
</tr>
</tbody>
</table>

Observations 236 236

Standard errors (clustered at the individual level) in parentheses. All specifications include individual fixed effects. *** p<0.01, ** p<0.05, * p<0.1

Table 2: Linear Probability Models by Condition

<table>
<thead>
<tr>
<th></th>
<th>Probability of choosing $x = x_0$</th>
<th>Probability of choosing $x = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$x_0 = 0$</td>
<td>$x_0 = 1$</td>
</tr>
<tr>
<td>$p \geq 25$</td>
<td>0.233**</td>
<td>0.310**</td>
</tr>
<tr>
<td></td>
<td>(0.108)</td>
<td>(0.118)</td>
</tr>
<tr>
<td>$p \geq 50$</td>
<td>0.200**</td>
<td>0.103</td>
</tr>
<tr>
<td></td>
<td>(0.0869)</td>
<td>(0.0673)</td>
</tr>
<tr>
<td>$p = 75$</td>
<td>-0.0333</td>
<td>0.0345</td>
</tr>
<tr>
<td></td>
<td>(0.0683)</td>
<td>(0.0707)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.300***</td>
<td>0.0345</td>
</tr>
<tr>
<td></td>
<td>(0.0714)</td>
<td>(0.0860)</td>
</tr>
</tbody>
</table>

Observations 120 116 120 116

Standard errors (clustered at the individual level) in parentheses. All specifications include individual fixed effects. *** p<0.01, ** p<0.05, * p<0.1

when $p = 75$. This decrease is less substantial for women, as 57 percent transfer half the prize when $p = 0$ and this only shrinks to 43 percent when $p = 75$. This means in the decision with the highest level of plausible deniability, compared to men, over 3.5 times as many women are still opting to share the pie equally.

It is also interesting to note where these increases/decreases come from. When $p$ increases from 0 to 0.25, the same fraction of women give $x = 0$ and the fraction of women giving intermediate amounts ($x \in [1, 9]$) decreases to 0 when $p \geq 25$. For men, however, the fraction giving intermediate amounts stays relatively constant when $p$ increases from 0 to 0.25. This illustrates an interesting pattern in “switching” behavior. The increase in pooling at $x = 0$ for men when $p$ increases from 0 to 0.25 is driven by men switching from giving equal divisions to giving zero when there is an
opportunity for plausible deniability. The increase in pooling at \( x = 0 \) for women is driven by women who were giving intermediate amounts when choices were perfectly observable.

These results also suggest that women who switch from making equal divisions need a greater degree of plausible deniability before they are willing to change their behavior. While men changed their behavior from giving half to giving nothing at any positive level of plausible deniability, women needed this probability to be 0.5 in order for a majority fraction to choose \( x = 0 \). The willingness to take advantage of plausible deniability is clearly blunted for women compared to men.

Figure 3: Distribution of amounts allocated to partners by gender, condition \( x_0 = 0 \)

![Figure 3](image)

Figure 4: Distribution of amounts allocated to partners by gender, condition \( x_0 = 1 \)

![Figure 4](image)

Table 3 reports the results of linear probability models. Columns 1 and 3 report results for the \( x_0 = 0 \) condition and columns 2 and 4 report results for the \( x_0 = 1 \) condition. Looking at the first column, there is a statistically significant increase in pooling at \( x = 0 \) when \( p \) increases from 0 to 0.25 for women, but not for men. Conversely, there is a statistically significant increase when \( p \) increases from 0.25 to 0.5 for men but not for women. The coefficient for both is insignificant for \( p = 75 \). Looking at the third column, none of the coefficients are statistically significant for women. However, when \( p \) increases from 0 to 0.25, the probability that a man divides the prize
### Table 3: Linear Probability Models by Condition

<table>
<thead>
<tr>
<th>Probability of choosing $x = x_0$</th>
<th>Probability of choosing $x = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_0 = 0$</td>
<td>$x_0 = 1$</td>
</tr>
<tr>
<td>$p \geq 25$</td>
<td>0.187</td>
</tr>
<tr>
<td></td>
<td>(0.159)</td>
</tr>
<tr>
<td>$p \geq 50$</td>
<td>0.250*</td>
</tr>
<tr>
<td></td>
<td>(0.131)</td>
</tr>
<tr>
<td>$p = 75$</td>
<td>-0.000</td>
</tr>
<tr>
<td></td>
<td>(0.113)</td>
</tr>
<tr>
<td>$p \geq 25 \times$ Female</td>
<td>0.0982</td>
</tr>
<tr>
<td></td>
<td>(0.216)</td>
</tr>
<tr>
<td>$p \geq 50 \times$ Female</td>
<td>-0.107</td>
</tr>
<tr>
<td></td>
<td>(0.173)</td>
</tr>
<tr>
<td>$p = 75 \times$ Female</td>
<td>-0.0714</td>
</tr>
<tr>
<td></td>
<td>(0.148)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.300***</td>
</tr>
<tr>
<td></td>
<td>(0.0727)</td>
</tr>
<tr>
<td>$p \geq 25$ if Female</td>
<td>0.286*</td>
</tr>
<tr>
<td>$p \geq 50$ if Female</td>
<td>0.143</td>
</tr>
<tr>
<td>$p = 75$ if Female</td>
<td>-0.0714</td>
</tr>
</tbody>
</table>

Observations 120 116 120 116

Standard errors (clustered at the individual level) in parentheses. All specifications include individual fixed effects.

- *** $p < 0.01$
- ** $p < 0.05$
- * $p < 0.1$

...equally decreases by over 37 percentage points. This coefficient is statistically significant and over five times the magnitude of the coefficient for women.

Looking at the second and fourth columns ($x_0 = 1$ condition), gender differences are not as stark. The main notable difference is that there is a significant increase for men giving $x = 1$ when $p$ increases from 0 to 0.25. Similarly, there is a significant decrease in pooling at $x = 10$ as $p$ increases from 0 to 0.25 for women, but not for men.

Comparing condition $x_0 = 0$ to $x_0 = 1$, women behave relatively similarly between the two conditions. Even at the highest level of plausible deniability $p = 75$ in condition $x_0 = 1$, 40 percent of women chose an even split while another 40 percent chose to transfer $x_0$. This is similar to what happened in condition $x_0 = 0$, where these percentages were 57 and 43, respectively. On the other hand, 29 percent of men chose even splits and 48 percent chose $x = x_0$ when $p = 0.75$ while these percentages were 12 and 81, respectively in the $x_0 = 0$ condition.
3.1.3 Summary

There are clear differences in men’s and women’s behavior in the $x_0 = 0$ condition. A larger fraction of men, compared to women, chose to transfer nothing to their partner when choices were perfectly observable. While the pooling at $x = 0$ increased for both genders as subjects were able to “hide” their selfishness, at every level of plausible deniability, the fraction of men choosing to transfer zero was larger than it was for women. Conversely, while the fraction of men choosing to split the prize evenly sharply decreased as the probability of nature intervening increased, this decline didn’t begin until after $p$ was greater than 25 and the degree of decline was blunted compared to men.

Turning to condition $x_0 = 1$, the results were relatively similar between men and women. Women behaved relatively similarly between the two conditions. Thus, the lack of a real difference between the groups stems from men acting more similarly to women under this condition rather than women acting more similarly to men.

These results are in line with the model’s prediction that women will be more generous than men when provided opportunities to hide their selfishness behind noisy signal.

3.2 AR

AR examines the role of communication in giving decisions. The experiment involved an anonymous dictator game where they systematically varied who in the pair could speak. Pairs and roles were randomly assigned, and allocators decided how to split $10 between themselves and their partners. Pairs communicated via written messages that contained both a pass allocation (numerical request) and a free response message. There were five experimental treatments: Baseline (no communication), Ask (only the recipient sent a message), Explain (only the allocator sent a message), Ask-Explain (both sent a message, but the recipient sent his first), and Explain-Ask (both sent a message, but the allocator sent his first). Subjects made two allocations (with different partners) and participated in only one experimental treatment. The experiment involved 258 subjects (117 men and 141 women), all undergraduates at the University of California, San Diego.

3.2.1 Original Results

Andreoni and Rao find that anytime the recipient spoke, giving increased. In the baseline (no-communication) condition, subjects passed 15.3 MU on average. Giving was higher in the Ask condition, with subjects passing 23.25 MU on average, and this difference becomes statistically significant when only requests for an even division or less are considered (Wilcoxon rank-sum $z = 1.965, p < 0.049$). Giving was highest in the two-way communication conditions, and this difference is significantly different from Baseline (AE: $z = 3.29, p < 0.001$, EA: $z = 2.04, p < 0.041$). Figure 5 (left panel) presents mean pass values.

---

8Subjects divided 100 monetary units (MU) at an exchange rate of 1 MU = $0.10.
9The only restriction on messages was that they could not contain identifying information or promises outside the lab.
10This is a recreation of Figure 2 from Andreoni and Rao (2011).
(Left panel) Means of pass value by condition: Baseline (no communication), A (ask by recipients), E (explain by allocators), AE (ask then explain), EA (explain then ask). (Right panel) Fraction of equal divisions and pass=0 by condition. The allocator determined the final allocation of 100 MU between him/herself and an anonymous receiver. Bars give +/- 2 s.e.

Figure 5: Means of pass values and fraction of equal divisions and passes of zero by condition

3.2.2 Gender Analysis

This dataset allows me to test the second prediction that women will be more generous even when choices are anonymous. Figure 6 (left panel) presents mean pass values and the fraction of subjects who chose equal divisions and to pass zero (right panel) separately for men and women. When subjects did not communicate with one another (Baseline condition), men and women were equally generous on average (16.2 MU vs. 17.96 MU, respectively). However, compared to women, nearly three times as many men chose to allocate nothing to their partners (47.2 percent of men vs. 16.67 percent of women; Fisher’s Exact Test: $p = 0.026$).

Differences between men and women become stronger when receivers are allowed to speak. When only recipients send a message, women are approximately twice as generous as men on average, as men give 16.8 MU on average while women give up 31.1 MU on average–nearly one-third of the total pie (t-test: $t = -2.21, p = 0.033$). And again in this condition, women are substantially less likely to give nothing to their partners (50.0 percent of men vs. 16.67 percent of women; Fisher’s Exact test: $p = 0.046$). Comparing the distributions of allocations between men and women in this condition is even more striking. Figure 7 presents smoothed kernel densities of pass values for the Baseline and Ask conditions. The distributions for men and women in the Ask condition are both visibly and statistically significantly different (Wilcoxon rank-sum $z = -1.99, p = 0.046$; Kolmogorov-Smirnov $D = 0.42, p = 0.031$). Women were also more generous than men in two-way communication when allocators spoke first (Explain-Ask condition), as they were again significantly less likely to make zero allocations (43.8 percent of men vs. 13.6 percent of women; Fisher’s Exact test $p = 0.062$). Men and women were equally generous in the Ask-Explain condition, but this was
(Left panel) Means of pass value by condition: Baseline (no communication), A (ask by recipients), E (explain by allocators), AE (ask then explain), EA (explain then ask). (Right panel) Fraction of equal divisions and pass=0 by condition. The allocator determined the final allocation of 100 MU between him/herself and an anonymous receiver. Bars give +/- 2 s.e.

Figure 6: Means of pass values and fraction of equal divisions and passes of zero by condition and gender

due to men being more generous in this condition compared to the others. Namely, a much smaller fraction of men gave zero in this condition compared to all the others (18.8 percent in Ask-Explain compared to a minimum of 44 percent across the remaining conditions).

Figure 7: Smoothed kernel densities of pass values–Baseline and Ask conditions by gender

Men and women also respond differently to numerical pass requests. Looking at the difference between the recipient’s numerical request and the allocator’s pass value, again reveals large and significant gender differences. Women gave, on average, amounts closer to the request. In the Ask condition, the mean difference between the request men receive and what they give is more
than twice that for women (35 MU vs. 14.5 MU), and this difference is statistically significant ($z = 2.20, p = 0.028; D = 0.38, p = 0.07$). This size of this difference is heavily driven by a large number of men receiving requests of 50 MU (the modal request) and responding by giving nothing. This result is not due to men receiving higher pass requests (or conversely women receiving more “reasonable” requests), as the requests allocators’ received did not differ by gender in any condition.\footnote{Since partners were anonymous to one another, and receivers therefore didn’t know the gender of their partners, this is not surprising.}

This difference stays relatively stable for women between one- and two-way communication (Ask-Explain: 18.6 MU, Explain-Ask: 14.0 MU), but decreases for men (Ask-Explain: 20.4, Explain-Ask: 25.6). However, the decrease between Ask and Ask-Explain is marginally insignificant ($D = 0.31, p = 0.102$), while the decrease between Ask and Explain-Ask is not statistically significant.

\subsection{3.2.3 Summary}

These results are consistent with the model’s prediction that if given sufficient life experience, women will be more generous even in anonymous settings. Since the subjects in this experiment are college students, it is reasonable to believe that the women in the study have had enough experience to generate generous habits. Women were, in general, less likely to make perfectly selfish allocations. And when receivers were permitted to “speak,” women were substantially more generous than men. Women were responsive to this social norm even when their identity was unknown to all those involved in the study, including the beneficiary of their generosity.

Additionally, considering gender leads to a very different conclusion of the original results drawn from this dataset. Andreoni and Rao found that whenever the recipient spoke, giving increased. However, this conclusion is only true for women. For male subjects to give more generous allocations, there needed to be two-way communication and the recipient needed to speak first. This challenges their finding that giving was highest under two-way communication.

\subsection{3.3 Summary of Results}

In a public setting, women were less likely to exploit an opportunity to hide their selfishness when they were offered some degree of plausible deniability. Women were less likely to make perfectly selfish allocations and were substantially more generous in response to the presence of requests, even though choices were anonymous. These results provide empirical support consistent with the model’s predictions.

Another interesting finding from this analysis is that in both of the datasets in the empirical analysis, one gender was responsible for driving some or all of the published results. The measured average treatment effect was not representative of the sample. Instead, it was an average of two extremes—one group that was strongly affected by the treatment and another group that was either not affected at all or was affected to a significantly lesser degree. This analysis provides strong evidence that heterogenous treatment effects due to behavioral differences between men and women may be responsible for many experimental results. This suggests that even if an experimental
treatment was not designed with the intention of examining gender differences and even if it is not clear that the environment being studied should have differential effects on men and women, additional analysis to examine heterogenous treatment effects by gender should be performed.

4 Experimental Design

The empirical analysis provides evidence in support of the model’s predictions, but it does not test the model’s mechanism. In order to provide a more direct test of a key component of the model, I design and implement new experiments. While the model is designed to capture a complex process that takes place over an individual’s lifetime, I distill this down into a key feature that can be tested in a laboratory setting: early decisions can have persistent effects even when the constraints of those decisions change. In the experiment, subjects make a series of dictator game allocations. The games vary in the chance that nature intervenes and forces subjects to either keep everything or give everything to their partner with equal probability. If there is a chance that nature intervenes, this gives subjects an opportunity for plausible deniability if they choose to keep everything (because if others in the experiment observe an allocation where the subject keeps everything, they will be unable to determine if the subject made that choice or if nature forced that allocation). Experimental treatments vary in the order that subjects make decisions, so subjects’ first choice either (i) offers no opportunity for plausible deniability and this opportunity increases in subsequent decisions or (ii) offers the highest level of plausible deniability and this opportunity decreases in subsequent decisions. In the experiment, I examine if exposing subjects to high external constraints in initial decisions mitigates gender differences even when subjects can take advantage of plausible deniability in later decisions.

4.1 Procedures

All sessions were conducted at UCSD’s EconLab using undergraduate students recruited via email. Instructions were read aloud to subjects and they submitted all responses via experimental software. Subjects were divided into pairs, with partners and roles assigned randomly. Within each pair, one subject was designated as the decision-maker and the other pair the receiver. The decision-maker determined how the pair divided $30.

Each session proceeded as follows: Subjects were randomly divided into pairs, and partners were seated opposite one another. One-at-a-time, pairs stood up and greeted one another in order to identify themselves to their partner. Decision-makers made 27 decisions for how to split $30 between themselves and their partner. Decisions differed in the probability that they were forced to make a particular allocation. If a decision was “forced,” the decision-maker kept all $30 and transferred nothing to his partner or kept nothing and transferred all $30 to his partner with equal probability.\[12\] The probability that a decision was forced varied between 10 values (0, 0.01, 0.02, 0.03, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 1.00).

\[12\] This was done to make the ex ante outcome of being forced equal for both the decision-maker and the partner. This was to ensure that individuals did not try to maximize ex ante fairness by being more generous in decisions.
For each decision, decision-makers knew whether they were “forced” or free to make an allocation. This was to highlight for decision-makers that they knew whether their choice was forced but no one else did. After all subjects had submitted their decisions, one decision was selected at random to determine payments. At the end of the session, the outcome for all groups of this selected decision was written on the board at the front of the room. There were two treatment groups: one treatment where subjects made decisions in increasing order of being forced (starting with a zero probability of being forced and ending with 0.90) and another treatment where subjects made decisions in decreasing order of being forced (starting with 0.90 and ending with 0). I will refer to these as the Increasing treatment and the Decreasing treatment, respectively. This is a between subjects design (all subjects within a single session were in the same treatment and each subject participated in only one treatment).

At the end of the session, subjects were paid in cash. Sessions lasted approximately one hour, and subjects earned an average of $20, including a $5 show-up fee for their participation. 9 sessions (5 sessions of the Increasing treatment and 4 sessions of the Decreasing treatment) of 16-20 subjects per session were conducted, resulting in a total of 166 subjects (41 men and 42 women decision-makers).

5 Experimental Results

I seek to answer two questions: First, do individuals exhibit persistence in their choices—that is, is what individuals choose in each decision relatively stable even though the opportunity for plausible deniability varies? Second, does the order of the decisions matter? That is, if individuals are initially exposed to a low probability of nature intervening, are they more generous initially and does this generosity extend to later decisions where the opportunity for plausible deniability is high?

I formalize these questions into three hypotheses:

**Hypothesis 1:** Choices will be relatively stable even though the opportunity for plausible deniability varies. This means that as subjects move to the next decision in the series, they will not be significantly more likely to change their allocation.

**Hypothesis 2:** Women in the Decreasing treatment will be more generous than men in the Decreasing treatment.

\[ \Pr(\text{Pass} = 15|W, D) > \Pr(\text{Pass} = 15|M, D) \]

**Hypothesis 3:** Men and women in the Increasing treatment will be equally generous.

\[ \Pr(\text{Pass} = 15|W, I) = \Pr(\text{Pass} = 15|M, I) \]

where they were able to make an allocation in order to make up for forced decisions in which they were forced to make a selfish allocation.
When subjects are initially exposed to a high probability of intervention, I predict men will be more likely to take advantage of this plausible deniability. These differences will persist through the series of decisions, so even when there are low or no opportunities for plausible deniability, men will still be less likely than women to choose equal allocations. However, when subjects’ initial decisions have no probability of intervention, I predict that men will give equal allocations at approximately the same rate as women, and these initial generous actions will persist in later actions, even subjects are given the opportunity to hide a selfish action behind nature. These hypotheses mean that I predict that there will be gender differences in the Decreasing condition but these differences will be mitigated in the Increasing condition.

Looking first at Hypothesis 1, subjects’ behavior appears to exhibit persistence to a high degree. When regressing the probability of choosing to pass 15 (an even split of the pie) or pass zero on the probability that the choice was forced using linear probability models, only one coefficient is statistically significant. Looking at the first column of Table 4 (the outcome variable is the probability that the decision-maker passed 15), only one of the coefficients is statistically significant. This is the interaction term on the probability of forced being greater than $0.50 \times \text{Female}$. Although, choices seem to return back to their previous level, as the coefficient on $p \geq 75 \times \text{Female}$ is almost equal in magnitude but opposite in sign (it is not quite statistically significant). Moreover, the point estimates are very close to zero, with only two being greater than 0.10. Looking at the second column of this table (the outcome variable is the probability that the decision-maker passed zero), none of the coefficients are statistically significant. Similarly, the point estimates are very small in magnitude, with approximately one-third of them being approximately 0.03 or less in magnitude. Given that the opportunity for plausible deniability across choices varies greatly, the degree of stability of subject’s choices is surprising.

Turning to Hypothesis 2, large gender differences are apparent when comparing men and women in the Decreasing treatment. Figure 8 depicts the fraction of subjects who chose equal allocations (pass 15) in this treatment. These results are also available in Table 5. Note that in the figure, the order of decisions goes from right to left (starting with 0.90 and ending with 0). As evidenced in the figure, the faction of women who chose to split the pie equally is greater than the fraction of men who chose this allocation at every level of intervention. That is, women are always more likely than men to choose equal allocations, and these differences are significant. Looking at subjects’ first choice ($p = 0.90$), 56 percent of women chose to allocate 15 while only 21 percent of men did (two-sided Fisher’s Exact test: $p = 0.045$). Even in subjects’ last choice, where there no opportunity for plausible deniability, women were nearly twice as likely as men to choose to pass 15 (71 percent vs. 37 percent, two-sided Fisher’s Exact test: $p = 0.054$).

While the differences between men and women’s choices in the Decreasing condition are large, when subjects made decisions in the opposite order, gender differences were mitigated. Looking at Hypothesis 3, men and women’s behavior looks much more similar in the Increasing condition. Figure 9 depicts the fraction of subjects who chose 50-50 splits (pass 15) in this treatment. Note that in this figure the order of decisions goes from left to right (beginning with 0 and ending with
0.90). The fraction of men and women who chose equal divisions is not statistically different. In subjects' first choice, although a larger fraction of women choose to pass 15 to their partner–68 percent of women compared to 57 percent of men, this difference is not statistically significant (two-sided Fisher’s Exact test: \( p = 0.545 \)). Even when subjects are offered a large opportunity for plausible deniability in their last decision (\( p = 0.90 \)), men are still as likely as women to give equal allocations (38 percent of men vs. 41 percent of women; two-sided Fisher’s Exact test: \( p = 1.00 \))

The experimental results are in line with the hypotheses and illustrate that the the order of subjects’ decisions has a large influence on their behavior. The difference in the parameters of the initial decision not only changed subjects’ choices for that decision, but also their subsequent decisions. By initially exposing individuals to a high degree of plausible deniability, I relaxed the external constraints if subjects chose to act selfishly. This caused men to be less likely to give equal allocations in that decision, but this behavior persisted over the series of decisions, even
when there was no opportunity for plausible deniability. However, by exposing subjects to stricter external constraints on their first action, I mitigated gender differences, as men continued to behave generously even when they had ample opportunity to take advantage of plausible deniability. By simply changing the order in which subjects made decisions, I mitigated gender differences in subjects’ behaviors. The results of this experiment thus present evidence in support of the model’s mechanism.

6 Conclusion

I have proposed a theory of behavior that captures how the external constraints individuals face can eventually become internalized. This shows how social norms that are initially externally enforced later become self-enforced by the individual. This mechanism provides insight into gender differences in observed behaviors and provides an alternative explanation for behavioral differences between men and women.

This mechanism is important in the study of gender differences for two primary reasons. First, this mechanism provides a different interpretation for data on gender differences. Instead of differences in observables being due to differences in fundamentals (preference functions or types), differences in men and women’s choices could be indicative of men and women facing different constraints and these constraints have become internalized over time. This analysis also provides evidence for the power and prevalence of social norms. A collection of experiments that did not set out to study gender differences actually captured very strong gender differences, so much so that the significance of their pooled results relied on the treatment effect to only one gender. In these data, even when choices were anonymous, the power of an internalized social norm was present.

Second, the results of this paper suggest that there is room for policy intervention. The danger of attributing gender differences to differences in fundamental characteristics between men and women is that is suggests that policy will be ineffectual. There is no need to construct policy if different choices are because men and women “are” different. My mechanism suggests that policy can be effective, specifically policies that either target established beliefs about men and women in order to relax the constraints put on women’s behavior and policies that are targeted at habit breaking for women who have already learned to internalize social norms. There already exist a few policies that may be effective in achieving these ends. In July 2017, Britain’s advertising regulator, the Committee on Advertising Practice, announced that new rules would be developed to ban advertising that promotes gender stereotypes or mocks those who do not conform to them. For example, one of the types of ads the UK policy is targeting is advertisements involving cleaning products, which typically feature women using them, and thus subtly enforce the association between women and domestic labor. Another potential for policy would be habit-breaking for women who have already formed habits for particular behaviors. Within economics, a group of female economists formed the “I just can’t say no club” in order to address the frequent difficulty of women being able to say “no” to work requests that are often non-promotable in nature. Founding members
### Table 4: Linear Probability Models

<table>
<thead>
<tr>
<th>Probability of choosing Pass = 15</th>
<th>Probability of choosing Pass = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p \geq 1$</td>
<td>-0.0750</td>
</tr>
<tr>
<td></td>
<td>(0.0700)</td>
</tr>
<tr>
<td>$p \geq 2$</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>(0.0538)</td>
</tr>
<tr>
<td>$p \geq 3$</td>
<td>0.0250</td>
</tr>
<tr>
<td></td>
<td>(0.0464)</td>
</tr>
<tr>
<td>$p \geq 5$</td>
<td>0.0500</td>
</tr>
<tr>
<td></td>
<td>(0.0531)</td>
</tr>
<tr>
<td>$p \geq 10$</td>
<td>-0.100</td>
</tr>
<tr>
<td></td>
<td>(0.0742)</td>
</tr>
<tr>
<td>$p \geq 25$</td>
<td>0.0500</td>
</tr>
<tr>
<td></td>
<td>(0.0531)</td>
</tr>
<tr>
<td>$p \geq 50$</td>
<td>-0.0763</td>
</tr>
<tr>
<td></td>
<td>(0.0692)</td>
</tr>
<tr>
<td>$p \geq 75$</td>
<td>-0.0319</td>
</tr>
<tr>
<td></td>
<td>(0.0858)</td>
</tr>
<tr>
<td>$p = 90$</td>
<td>-0.0168</td>
</tr>
<tr>
<td></td>
<td>(0.0899)</td>
</tr>
<tr>
<td>$p \geq 1 \times \text{Female}$</td>
<td>-0.0917</td>
</tr>
<tr>
<td></td>
<td>(0.0935)</td>
</tr>
<tr>
<td>$p \geq 2 \times \text{Female}$</td>
<td>-0.0238</td>
</tr>
<tr>
<td></td>
<td>(0.0785)</td>
</tr>
<tr>
<td>$p \geq 3 \times \text{Female}$</td>
<td>-0.0488</td>
</tr>
<tr>
<td></td>
<td>(0.0736)</td>
</tr>
<tr>
<td>$p \geq 5 \times \text{Female}$</td>
<td>0.0452</td>
</tr>
<tr>
<td></td>
<td>(0.0807)</td>
</tr>
<tr>
<td>$p \geq 10 \times \text{Female}$</td>
<td>-0.0190</td>
</tr>
<tr>
<td></td>
<td>(0.0985)</td>
</tr>
<tr>
<td>$p \geq 25 \times \text{Female}$</td>
<td>-0.0500</td>
</tr>
<tr>
<td></td>
<td>(0.0738)</td>
</tr>
<tr>
<td>$p \geq 50 \times \text{Female}$</td>
<td>0.172*</td>
</tr>
<tr>
<td></td>
<td>(0.0921)</td>
</tr>
<tr>
<td>$p \geq 75 \times \text{Female}$</td>
<td>-0.139</td>
</tr>
<tr>
<td></td>
<td>(0.105)</td>
</tr>
<tr>
<td>$p = 90 \times \text{Female}$</td>
<td>0.0959</td>
</tr>
<tr>
<td></td>
<td>(0.112)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.583***</td>
</tr>
<tr>
<td></td>
<td>(0.0385)</td>
</tr>
</tbody>
</table>

**Observations:** 786 786

Standard errors (clustered at the individual level) in parentheses. All specifications include individual fixed effects. *** p<0.01, ** p<0.05, * p<0.1
<table>
<thead>
<tr>
<th>Probability of forced choice</th>
<th>Number of Observations</th>
<th>Increasing</th>
<th></th>
<th></th>
<th></th>
<th>Decreasing</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Pass = 15</td>
<td>Pass = 0</td>
<td>Pass = 15</td>
<td>Pass = 0</td>
<td>Pass = 15</td>
<td>Pass = 0</td>
<td>Pass = 15</td>
<td>Pass = 0</td>
</tr>
<tr>
<td></td>
<td>Men</td>
<td>Women</td>
<td>Men</td>
<td>Women</td>
<td>Men</td>
<td>Women</td>
<td>Men</td>
<td>Women</td>
<td>Men</td>
</tr>
</tbody>
</table>
include Linda Babcock and Lise Vesterlund, and the group has since spread to three national clubs. Educating women on how to effectively decline requests is a promising potential policy.

While the idea that external constraints become internalized has clear applications to gender differences research, it is also a mechanism that could apply to other social norms. From a general policy perspective, potential research could examine how we might encourage socially desirable or welfare-improving behaviors, and eventually these behaviors will become self-perpetuating through habit formation and self-enforcement. Further examining this mechanism and its application to the way economists think about how individuals make decisions is a promising area of future theoretical, experimental, and applied research.

References


LIMRA (2016). Men vs. Women: Who makes the financial decisions?

National Center for Education Statistics (2016). Bachelor’s, master’s, and doctor’s degrees conferred by postsecondary institutions, by sex of student and discipline division.


Appendix

Proof of Lemma 1. Let $s_x$ denote $D$’s social image upon choosing $x$. $D$ chooses $x = 0$ if and only if $U(0, s_0, t) > U(x^F, s_x^F, t)$, and $D$ chooses $x = x^F$ if and only if $U(0, s_0, t) < U(x^F, s_x^F, t)$. Then, $D$ is indifferent between the two when $U(0, s_0, t) = U(x^F, s_x^F, t)$. Define $t^*$ as $t$ such that $U(0, s_0, t^*) = U(x^F, s_x^F, t^*)$. Take any $\hat{t} > t^*$. Then, $U(0, s_0, \hat{t}) < U(0, s_0, t^*)$, since $F(1, s_0) + \hat{G}(-x^F) < F(1, s_0) + t^*G(-x^F)$, and $U(x^F, s_x^F, \hat{t}) > U(x^F, s_x^F, t^*)$, since $F(1-x^F, s_x^F) + \hat{G}(0) > F(1-x^F, s_x^F) + t^*G(0)$. Thus, $U(0, s_0, \hat{t}) < U(0, s_0, t^*) = U(x^F, s_x^F, t^*) < U(x^F, s_x^F, \hat{t})$. Then $D$ of type $t$ chooses $x = x^F$.

A parallel argument holds for $t < t^*$ and choosing $x = 0$. Since I only consider pure strategy equilibria, assume that if the decision-maker is indifferent, he breaks ties by choosing $x = x^F$. 

Proof of Proposition 1. As in Lemma 1, assume $t^*$ is the $t$ that satisfies $U(0, s_0, t^*) = U(x^F, s_x^F, t^*)$. If all members of A believe $K^W = K^M = K$ and $\Phi(t; W, x) = \Phi(t; M, x) = \Phi$, then $s_0, W = s_0, M$ and $s_x^W, W = s_x^W, M$. Then, for any $t$, $U(0, s_0, W, t) = U(0, s_0, M, t)$ and $U(x^F, s_x^W, W, t) = U(x^F, s_x^W, M, t)$. Therefore, $U(0, s_0, W, t^*) = U(0, s_0, M, t^*) = U(x^F, s_x^W, W, t^*) = U(x^F, s_x^W, M, t^*)$ and $t^*_W = t^*_M = t^*$. 

Proof of Proposition 2. First examine the case of one audience member holding belief B1. In this case, $\Phi(t; M, x = 0)$ FOSD $\Phi(t; W, x = 0)$. Then, $S(\Phi(t; W, x = 0)) < S(\Phi(t; M, x = 0)) \implies s_0, W < s_0, M$. Suppose that $t^*_W = t^*_M$. This would imply that $U(0, s_0, W, t^*_W) = U(x^F, s_x^W, W, t^*_W) = U(0, s_0, M, t^*_M) = U(x^F, s_x^W, M, t^*_M)$. But this cannot be true because $s_0, W \neq s_0, M$. Next, suppose $t^*_W > t^*_M$. This implies that $U(0, s_0, W, t^*_W) > U(0, s_0, M, t^*_W)$. But this can’t be true because $s_0, W < s_0, M$. Then it must be that $t^*_W < t^*_M$.

Next examine the case of one audience member holding belief B2. Consider any social image $s_0$ such that $s_0 = s_0, W = s_0, M$. If groups had utility functions $U_W$ and $U_M$, then for any fixed $t$, $U_W(0, s_0, t) < U_W(0, s_0, t)$, since $F(1, \alpha_W s_0) + tG(-x^F) < F(1, \alpha_M s_0) + tG(-x^F)$. Then, $t^*_W < t^*_M$. When an audience member holds belief B2, he believes this is the case, and thus upon observing $x = 0$, $\Phi(t; M, x = 0)$ FOSD $\Phi(t; W, x = 0)$ $\implies s_0, W < s_0, M$. Then, even when the utility function for both groups is $U$, $s_0, W < s_0, M$, then $t^*_W < t^*_M$. 

Proof of Proposition 3. If $A$ holds correct beliefs and makes correct inferences, then $\Phi(W) = \Phi(M) = \Phi$. But, if $D$ believes that at least one audience member holds belief B1 or B2, then $S_W(\Phi(t; W, x = 0)) < S_M(\Phi(t; M, x = 0)) \implies s_0, W < s_0, M$. Therefore, $t^*_W < t^*_M$.

Proof of Lemma 2. Take any $p_1, p_2 \in (0, 1)$ with $p_1 > p_2$. Define $t^*_p$ to be the type such that $U(0, s_0, p_i, t^*_i) = U(x^F, s_x^p, i, t^*_i)$. Since, $s_0, p_i > s_0, p_2$, $U(0, s_0, p_i, t^*_i) > U(0, s_0, p_2, t^*_i)$. Then, for $U(0, s_0, p_2, t^*_p) = U(x^F, s_x^p, p_2, t^*_p)$, $t^*_p < t^*_p$. 

Proof of Proposition 4. Suppose $t^*_p, W = t^*_p, M$. This would imply that $U(0, s_0, p, W, t^*_p) = U(x^F, s_x^p, p, W, t^*_p) = U(0, s_0, p, M, t^*_p) = U(x^F, s_x^p, p, M, t^*_p)$. But this cannot be true because $s_0, p, W \neq s_0, p, M$. Next suppose $t^*_p, W = t^*_p, M$. This implies $U(0, s_0, p, W, t^*_p) > U(0, s_0, p, M, t^*_p)$. But this cannot be true because $s_0, p, W < s_0, p, M$. Then it must be that $t^*_p, W < t^*_p, M$. 

29
Proof of Proposition 5. Without loss of generality, I focus on the actions of group $W$. Define $\tilde{t}$ to be the type such that $F(1-x^F) + \tilde{t}G(x^F - x^F) = F(1-0) + \tilde{t}G(0-x^F)$. By Lemma 1, $\forall t > \tilde{t}$, $D$ will choose $x = x^F$. Individuals of these types will give $x^F$ in phase 2 even without habit formation. In phase 1, members of group $W$ with type $t < t^*_W$ transfer 0, so they do not have any incentive to switch actions in phase 2. Then, restrict attention on decision-makers who are of are of types $t \in [t^*_W, \tilde{t}]$. These are types who would rather pick 0, but gave $x^F$ in phase 1 because actions were observable.

Looking at continuation payoffs, $D$ will choose $x = x^F$ in all $g$ iff the continuation payoff from giving $x^F$ is greater than or equal to the continuation payoff from giving $x = 0$. If $D$ transfers $x = x^F$, $D$’s utility is:

$$U = \sum_{g=g+1}^{g} [F(1-x^F) + tG(x^F - x^F) + rH(\sum_{j=1}^{g-1} \delta^j)] 
\tag{1}$$

If $D$ transfers $x = 0$, $D$’s utility is:

$$U = F(1-0) + tG(0-x^F) + rH(0) + \sum_{g=g+2}^{g} [F(1-0) + tG(0-x^F) + rH(\sum_{j=1}^{g-1} \delta^j)] 
\tag{2}$$

$D$ will choose $x^F$ in all periods iff $1 \geq 2$.

Simplifying (1), we obtain

$$U = [\bar{g} - (\bar{g} + 1)][F(1-x^F) + tG(x^F - x^F)] + \sum_{g=\bar{g}+1}^{\bar{g}} [rH(\sum_{j=1}^{g-1} \delta^j)]$$

Simplifying (2) yields

$$U = F(1-0) + tG(0-x^F) + [\bar{g} - (\bar{g} + 2)][F(1-0) + tG(0-x^F)] + \sum_{g=\bar{g}+2}^{\bar{g}} [rH(\sum_{j=1}^{g-1} \delta^j)]$$

$$= [\bar{g} - (\bar{g} + 1)][F(1-0) + tG(0-x^F)] + \sum_{g=\bar{g}+2}^{\bar{g}} [rH(\sum_{j=1}^{g-1} \delta^j)]$$

As $\bar{g}$ increases, the incentive to switch from 0 to $x^F$ decreases, because habit formation term for staying with $x^F$ increases and the number of periods to collect extra benefit of $F(1-0) + tG(0-x^F)$ decreases. So as $\bar{g}$ increases (approaches \bar{g}), (2) gets smaller and the second term of (1) gets larger. Then, if we make $\bar{g}$ arbitrarily large, there will be some $\hat{g}^*$ such that for $\bar{g} > \hat{g}^*$, $1 > (2)$. Then in games $g > \hat{g}$, $D$ will choose $x = x^F$. Thus, for for types $t \in [0, t^*_W)$, $D$ chooses $x = 0 \forall g \in [1, \bar{g}]$, for types $t \in [t^*_W, \tilde{t}]$, $D$ will choose $x = x^F \forall g \in [1, \bar{g}]$.

---

\[\text{For this proof, I assume no future discounting, as this is a stronger result. The result will obviously still hold if the decision-maker discounts future periods.}\]