The Female Sensitivity Hypothesis: Evidence From Experimental Economics

Felipe A. Araujo
Lehigh University

Neeraja Gupta
University of Richmond

Lise Vesterlund
University of Pittsburgh and NBER

Abstract: We assess the empirical evidence for the hypothesis that women more than men respond to changes in treatment. First, we examine whether the results of over two hundred experimental economics studies support the female sensitivity hypothesis. Second, using data from two studies (DellaVigna and Pope, 2022; Exley et al., 2025), we conduct over two hundred pairwise tests of the hypothesis. Both analyses show that gender is not predictive of responsiveness to treatment. We further examine how the hypothesis has been disseminated in the literature and find strong confirmation bias with the hypothesis predominantly being cited by studies that support it.

1. Introduction

The experimental literature on gender differences in economic behavior is extensive. Researchers have used both laboratory and field experiments to investigate gender differences across various domains, including social preferences (Andreoni and Vesterlund, 2001; Buchan et al., 2008; Eckel and Grossman, 2001), competition (Gneezy et al., 2003; Niederle and Vesterlund, 2007), negotiation (Exley et al., 2020; Leibbrandt & List, 2015; Small et al., 2007), risk and time preferences (Charness and Gneezy, 2012; Dohmen and Falk, 2011), self-perceptions (Coffman, 2014; Exley & Kessler, 2022), among others. A recurring assertion within the literature is that women more than men are responsive—or sensitive—to changes in experimental design or treatment conditions. We refer to this as the *female sensitivity hypothesis*.

Though the precise origin of the hypothesis is difficult to pin down within economics, early reference is made to leading studies in psychology. For example, Gilligan (1982) argues that women are more sensitive to social cues when determining appropriate behavior, suggesting that women may prioritize relational and emotional contexts in decision making. This view is further supported by evidence that women exhibit greater empathy (Eisenberg and Lennon, 1983) and heightened sensitivity to social information (Putrevu, 2001), and by numerous psychological studies arguing that women experience and express emotions more intensely than men (e.g., Barrett et al., 2000; Fujita et al., 1991; Kring and Gordon, 1998; Loewenstein et al., 2001). One of the most influential economics references of the female sensitivity hypothesis is the review paper by Croson and Gneezy (2009). In summarizing the literature on gender differences in social, risk, and competitive preferences, the authors argue that women's greater responsiveness serves as an organizing principle for observed gender differences in social preferences. They further note that women are more responsive to both the experimental context—features of the design unrelated to the primary treatment (e.g., pen-and-paper versus computerized experiments)—and to what they refer to as social cues, which include factors such as "the size of payoffs, price of altruism, or the repetition of the game, and psychological variables like the amount of anonymity […] and the way the situation is described" (Croson and Gneezy, 2009, p. 463).

The female sensitivity hypothesis, as presented in Croson and Gneezy (2009), has been remarkably influential in the experimental economics literature. Notably, researchers have invoked the hypothesis to explain not just gender differences in social preferences, but across diverse domains including

competition, negotiation, and time preferences.[1] Despite frequent references to it, a comprehensive assessment of the hypothesis' empirical support is lacking. Understanding whether the hypothesis holds is crucial for accurately interpreting gender differences in economic behaviors. If women are more responsive to experimental conditions, this would suggest that observed gender differences might be artifacts of experimental design rather than fundamental differences in preferences. Alternatively, if neither gender systematically is more responsive to experimental variations, it would suggest that gendered responses depend on the specific treatment variation.

The purpose of this paper is to provide the first comprehensive and thorough assessment of the female sensitivity hypothesis. While Croson and Gneezy establish support for the female sensitivity hypothesis from a select set of social preference comparisons, a statistical assessment of the hypothesis is only seen in single experiments with a few treatment variations (Boschini et al., 2012; Eriksson and Simpson, 2010; Lotz, 2014; Miller and Ubeda, 2012) or in response to performance pay (Bandiera et al., 2021).

Our empirical evaluation of the hypothesis has two parts. First, we provide an extensive review of the literature on gender differences in experimental economics and evaluate whether the results in each study align with the female sensitivity hypothesis. Second, we analyze data from two studies (DellaVigna & Pope, 2022; Exley et al., 2025), which combined include a large set of both real-effort and decision experiments with almost 20,000 participants. This data allows us to estimate gender differences in responsiveness, both between and within-subjects, in 203 pairwise treatment comparisons.

For our first analysis, we take the view of a scholar reading the literature and ask whether they would conclude that women are more likely than men to respond to treatment variation within a study. We start by examining gender differences in response to *any* changes in the experimental design or treatment. We then assess the results along the two dimensions outlined by Croson and Gneezy (2009) and evaluate the evidence for the female sensitivity hypothesis in: (a) studies focusing on social

[1] For example, in a study on gender differences in time preferences Dittrich and Leipold (2014, p. 413) write that "women have a higher context-sensitivity than men (Croson and Gneezy, 2009)." In a paper studying competitive preferences, Gill and Prowse (2014, p. 368) conclude that "[o]ur findings […] are consistent with the claim that women's behavior exhibits greater context sensitivity." In another paper on competitiveness Datta Gupta et al. (2013, p. 831) write that "[a] more plausible explanation of the higher competitiveness of women in this treatment is that women tend to be more sensitive to the social context in which they make choices." And in a paper about cheating, Ezquerra et al. (2018, p. 46) motivate their research questions by noting that "[…] males and female react differently to the context. Our aim is to see how these findings apply to cheating behavior […]".

preferences, which include those using dictator, ultimatum, trust, prisoner's dilemma, social dilemma, or public goods games, and (b) studies where the treatment variation involves changes to the experimental context or social cues, which we collectively refer to as experimental conditions.

Wanting to simultaneously assess the literary support for the female sensitivity hypothesis and how the hypothesis is referenced we focus on the literature after 2009. Further to present a comprehensive assessment of the literature we include studies independent of data availability and center our first evaluation on whether one gender and not the other responds to treatment. Our literature review consists of 257 experimental economics papers highlighting results by gender. Among these, 49 papers do not involve any treatment variation (e.g., risk elicitation without additional experimental manipulation) and are excluded from the analysis. Of the remaining 208 studies, one-quarter (52) provide evidence consistent with the female sensitivity hypothesis, with women responding significantly to treatment variation while men do not. Another quarter of studies (49) reach the opposite conclusion, namely that men significantly respond to treatment variation while women do not. In 80 papers, either both genders respond to treatment, or neither does, or findings vary across different sets of treatment comparisons in the study, indicating no consistent pattern of gender differences in responsiveness. Lastly, 27 studies present insufficient data to assess a gendered response to treatment. In summary, our comprehensive assessment of the literature reveals that 71% of studies present results inconsistent with women and not men responding to treatment variation.[2] This conclusion remains when we restrict the analysis to studies focusing on social preferences; on treatment variations related to experimental context or social cues; or on the intersection of the two.[3]

Second, to assess not only whether women are more likely than men to respond to treatment, but also whether the magnitude of the response varies by gender we leverage data from DellaVigna and Pope (2022) and Exley et al. (2025). Each study allows for a large number of pairwise comparisons of men's and women's responsiveness to variations in experimental design or treatment.

The results are in line with the findings from our literature review. Specifically, we find that women are neither more likely to respond significantly to treatment variation nor to exhibit larger treatment

[2] Our analysis of DellaVigna and Pope (2022) and Exley et al (2025) permits assessment of differences in magnitudes.
[3] The within-study assessment is consistent with the primary analysis in Croson and Gneezy (2009), where of the twelve examples used as evidence for the hypothesis nine are within-study. In presenting three between-study examples they note "Between-study comparisons of levels is always tricky, thus we are more careful in our interpretations here". To secure controlled comparisons we focus on within-study comparisons holding constant subject pools, experimental protocols, and research teams.

responses than men. Among the sixty between-subject pairwise comparisons from DellaVigna and Pope (2022), four show that women, but not men, significantly respond to treatment, while eight show the opposite. Support for the female sensitivity hypothesis weakens further when we consider the magnitude of the treatment response. No comparison shows a significantly larger response by women, whereas six show a significantly larger response by men. Depending on the analysis the female sensitivity hypothesis is rejected in 93% to 100% of the pairwise comparisons, and the results similarly reject the hypothesis when restricted to the domain of social preferences.[4] Finally, we examine within-subject responses using data from Exley et al. (2025). In this paper, participants make binary choices across a wide range of social preference treatments. We analyze 143 pairwise comparisons, examining the proportion of women and men who change their decisions (between a selfish and non-selfish option) across treatment pairs. Significant gender differences emerge in only 9 of the 143 comparisons. In five of those, women change their choices at significantly higher rates than men, while in the remaining four, men do so more often. Leaving the female sensitivity hypothesis rejected in 97% of the comparisons.

Having found no evidence in support of the female sensitivity hypothesis, we next try to understand why the hypothesis has prominence in the literature. To do this, we revisit the papers in our literature review and assess those that explicitly mention the hypothesis as stated by Croson and Gneezy (2009). We find that a full 18% of our sample references the female sensitivity hypothesis, and that they do so in a manner consistent with confirmation bias. Among the studies that explicitly cite the hypothesis, only 3% mention it unfavorably, arguing that their findings do not support it, 61% treat the hypothesis as an established fact or suggest that their findings align with it, while the remaining 34% studies acknowledge the hypothesis without offering any support or critique. This referencing pattern stands in stark contrast to the empirical findings from our broader assessment of the literature and from our analysis of the data from DellaVigna and Pope (2022) and Exley et al. (2025), all of which show no greater sensitivity to variations in treatments or experimental conditions among women compared to men. Consequently, a casual reader of the literature is likely to encounter a predominance of favorable references to the female sensitivity hypothesis, despite its rejection in systematic evaluation of existing research.

[4] We do not separately examine treatment variations and experimental conditions, as all the variations analyzed in DellaVigna and Pope (2022) and Exley et al. (2025) involve changes in experimental context or social cues, as defined by Croson and Gneezy (2009).

The remainder of the paper is organized as follows. Section 2 presents the results from our assessment of the experimental economics literature and section 3 analyzes data from both DellaVigna and Pope (2022) and Exley et al. (2025). Section 4 explores evidence on the dissemination of the female sensitivity hypothesis and concludes.

2. Analysis of the Literature

To review the evidence for the female sensitivity hypothesis, we assess the inference a scholar would draw from reviewing the experimental economics literature on gender differences. In providing a broad assessment of the literature our focus is on whether in a particular study women and not men significantly respond to treatment variation. In answering this question, we searched the literature for experimental economics papers published after 2009, where the research focus is on gender differences or where gender-related findings are highlighted.[5] We use the 2009 start date to mirror the year of publication of the review paper by Croson and Gneezy (2009), because it allows us to simultaneously evaluate the results and the extent to which the hypothesis is referenced.

We identified a total of 257 papers, of which 208 included at least one treatment variation, making them suitable for evaluating the female sensitivity hypothesis. After compiling the dataset, we classified each paper into one of four categories based on the response to treatment: (1) *Men:* studies where men respond significantly (at the 5% level) to treatment variation, but women do not. This classification was applied when the evidence was supported by either a single conclusive result or consistent gendered responses for all treatment variations in the study; similarly (2) *Women*: studies where women respond significantly (at the 5% level) to treatment variation, but men do not; (3) *Both/Neither*: studies where either both men and women respond significantly to treatment variations, or where neither gender responds significantly, or where there is no consistent pattern of gender differences in responsiveness across different treatment variations; and (4*) Insufficient Data*: studies

[5] See Online Appendix A for a detailed description of the review process and Online Appendix B for the complete list of included papers along with their classification.

where the data did not provide enough information to unequivocally assess a gendered response to treatment.[6]

To illustrate our categorization procedure, consider the following three papers from our dataset. Baldiga and Coffman (2018) examine whether the presence of a sponsor—someone with stakes in a participant's performance who can advocate for them—affects the gender gap in competitiveness. They find, across all payoff structures and levels of participant performance, that men are significantly more likely to enter competition relative to the baseline of no sponsorship, while no such treatment difference is observed for women. We classify this paper as *Men*. Heinz et al. (2012) investigate transfers in dictator games, comparing situations where the endowment is obtained via a windfall lottery win or through performance on a real-effort task. The average taking rate for women varies significantly by treatment (74.0% in the windfall condition versus 63.3% in the real-effort condition), while the taking rate for men is stable (73.8% in the windfall condition versus 75.4% in the real-effort condition). This study is classified as *Women*. Finally, Babcock et al. (2017) is classified as *Both/Neither*. The study explores gender differences in volunteering in response to the gender composition of the group. While women volunteer more than men in mixed-sex groups, the gender gap is eliminated in single-sex groups. Specifically, moving from mixed-sex to single-sex groups decreases the rate of volunteering for women by the same amount as it increases the rate for men.

We begin our analysis by examining all changes in experimental design and treatments and ask whether women more than men are responsive to such changes. We then narrow our focus to the two study categories noted by Croson and Gneezy (2009) as being more conducive to the female sensitivity hypothesis: (1) studies within the social-preferences domain and (2) studies where treatment variations pertain to the experimental context or social cues. Wherein the second category, experimental context refers to aspects of the experimental design that can vary while the primary treatment remains unchanged (e.g., face-to-face versus computerized interaction, or strategy versus game method elicitation). Social cues, on the other hand, encompass factors such as the incentives (monetary or otherwise), the size of payoffs, the price of altruism, game repetition, and psychological variables like the degree of anonymity between participants and experimenters, as well as how the situation is framed (Croson and Gneezy, 2009, p. 463). For simplicity, and because both context and

[6] The *Insufficient Data* classification includes papers that report a single coefficient to highlight gender difference in response. Without further information on standard errors by gender it is not possible to determine whether one gender and not the other significantly responds to treatment.

social cues are hypothesized to affect women's responses more than men's, we combine these two elements under the term *experimental conditions.* Finally, we also assess the intersection of the two. An advantage of our aggregate assessment of gender differences in responsiveness to treatment is that it avoids the challenges of classifying studies into specific categories.[7] Further, it encompasses the broad set of studies that reference the female sensitivity hypothesis, extending beyond those examining social preferences and changes in experimental conditions.

Table 1 presents the results of our literature review. Columns (1) and (2) consider all treatment variations, while columns (3) and (4) restrict attention to papers with changes in experimental conditions. Within each of those categories, we report our findings separately for papers in the social-preferences domain. For our main sample we see in column (1) that 24% (49 of 208) of studies find a significant treatment response for men but not for women, while 25% (52 of 208) of studies find a significant treatment response for women but not for men, and 38% of the studies (80 of 208) find no gender difference in responsiveness to treatment variations. Abstracting from the 13% of studies that we cannot classify, we have a full 71% (129 of 181) of the studies being inconsistent with women and not men responding to treatment. Column (2) reports the results for the 49 studies that explore gender differences in social preference domains, again showing no evidence of the female sensitivity hypothesis, with 18% finding that men and not women respond to treatment, 30% finding that women and not men respond, and 39% showing no gender difference in responsiveness. Abstracting from the studies that cannot be classified, 64% (28 of 43) of studies are inconsistent with women and not men responding to treatment variation.

Columns (3) and (4) restrict attention to papers where the treatment variation is classified as changes in experimental conditions, which includes changes in the experimental context or in social cues. Because of its broad definition, most of the papers in our dataset are classified as having changes in experimental conditions (183 out of 208). The results for the subset of studies classified as having

[7] While some domains clearly focus on social preferences, other-regarding preferences influence a wide range of experimental paradigms causing the category of social domain experiments to be somewhat vague. Further it is challenging to classify changes in experimental conditions. For example, Ariely et al. (2009) has a treatment comparison between subjects solving anagrams either privately or in front of a 10-person group. This treatment variation is easily classified as a change in experimental condition (shifting the private to public setting and thus altering the degree of anonymity). Flory et al. (2015), on the other hand, is an example of a harder-to-classify study. The paper examines gender differences in job-entry decisions using a natural field experiment, varying job advertisements, compensation schemes, and application procedures across male- and female-oriented job tasks. While the differences in payment schemes (fixed wage versus competition) lead to different expected payoffs and can be viewed as a change in social cues, as defined by Croson and Gneezy (2009), we consider this variation to be a primary treatment of interest rather than a peripheral aspect of the design. And so, we do not assign this study to the experimental conditions category.

changes to experimental conditions are similar to those in the larger set of papers. Considering all domains, we see in column (3) that the shares of papers that provide evidence that either only men or only women respond to treatment are 25 and 26%, respectively, while 38% of the studies do not find that one gender and not the other is responsive to treatment. Restricting the analysis to papers reporting changes in experimental conditions within the domain of social preferences does not alter the inference, where 64% (28 of 42) of the studies are inconsistent with women and not men responding to treatment variation. In summary, our assessment of the literature shows that there is no evidence that a significant response to treatment more often is seen for women than for men.

Table 1: Gender Differences in Treatment Responsiveness in the Experimental Economics Literature

	All Treatment Variations		Experimental Conditions	
	All Topics	Social Preferences	All Topics	Social Preferences
	(1)	(2)	(3)	(4)
Men	49	9	45	8
Women	52	15	47	15
Both/Neither	80	19	70	19
Insufficient Data	27	6	21	6
Total	208	49	183	48

Notes: The table summarizes the classification of papers from our literature review of experimental studies published on or after 2009. *Men*: a significant (at the 5% level) treatment response seen for men but not for women; *Women*: a significant (at the 5% level) treatment response seen for women but not for men. *Both*: both responsive or both unresponsive or results varied across multiple treatment comparisons. *Insufficient Data*: gender differences in responsiveness could not be inferred from the published results. This latter classification includes papers where a single coefficient shows gender difference, but where the information reported is not sufficient to separately identify the gender-specific response to treatment.

3. Analysis of Experimental Data from DellaVigna and Pope (2022) and Exley et al. (2025)

The female sensitivity hypothesis is too far-reaching to be refuted or supported by a single experimental study with a limited set of treatments. Hence our initial assessment asked whether the broader literature indicates that women more than men are likely to respond to treatment variation. However, to further evaluate the female sensitivity hypothesis, we want to determine whether the magnitude of the treatment response differs by gender. Well-suited for such an assessment are two recently published studies by DellaVigna and Pope (2022) and Exley et al. (2025). Each study includes an unusually large set of treatments which combined facilitate over 200 pairwise tests of the female sensitivity hypothesis, all while using the same experimental framework and population.

For consistency with our analysis of the broader literature, for each paper we first evaluate whether one gender and not the other significantly responds to treatment. Second, we use our access to the raw data to evaluate whether there are significant gender differences in the magnitude of the treatment response. Further adhering to our earlier analysis, we emphasize the domains thought to be conducive to the hypothesis. While all pairwise comparisons are over experimental conditions (corresponding to columns 3 and 4 in Table 1), comparisons in the social preference domain account for a subset of those in DellaVigna and Pope (2022) and for all comparisons in Exley et al. (2025).

3.1 DellaVigna and Pope (2022)

The DellaVigna and Pope (2022) study assesses the stability and predictability of behavioral economics results by exploring the behavior of over 18,000 participants in online experiments. One of two real effort tasks (typing the letters “a” and “b” or coding World War II conscription cards) was used to assess the behavioral response to one of sixteen treatments in one of five conditions. For instance, four of the treatments vary payment types (fixed payment versus piece rate), as well as the piece-rate amounts. Other treatments adopt a time preference incentive (payment in two weeks versus four weeks), a probabilistic incentive (50% chance of a small piece rate versus 1% chance of a large piece rate), or a charitable-giving incentive (low versus high piece rate for the Red Cross). There are also several treatments exploring psychological incentives, such as social comparisons, relative rankings, and task significance. Finally, the five different conditions present modifications to the study protocol that, while often observed in practice, are not directly related to the primary research

question. These include variations in the type of task, output measures, the presence of a consent form, and the demographic characteristics of participants.

To test the female sensitivity hypothesis, we use the between-subject design to assess how men and women on average respond to treatment variations and changes in experimental conditions. For example, we compare the changes in effort for both types of real effort tasks when participants are faced with a 1-cent versus a 4-cent piece rate, or when exposed to different types of psychological incentives, and so on. Examining the DellaVigna and Pope (2022) conditions that includes a gender identifier, we assess gendered treatment responses between the sixteen behavioral treatments within the two tasks (a/b typing and coding WWII conscription cards).[8]

The results for the 60 unique tests of gender differences in response to treatment are summarized in Table 2.[9] In columns (1) and (2), we use the same classification as for our literature assessment, evaluating whether one gender and not the other significantly responds to treatment. In columns (3) and (4), we classify each pairwise comparison based on whether the magnitude of the treatment response is greater for one gender. Specifically, we classify each pairwise comparison as *Men* (*Women*) if the treatment effect is larger for men (women) in absolute terms and the interaction coefficient in an OLS regression is significant at the 5% level. For all other cases, we classify the magnitude of the response as being indistinguishable between men and women (*Both*/*Neither*).

Table 2 columns (1) and (2) show that there are more instances where men and not women significantly respond to treatment, than there are cases where women and not men respond to treatment. For all treatment variations (column 1), 93% of comparisons are inconsistent with women and not men responding to treatment, and for the domain of social preferences (column 2) the share is 96%. The evidence of the female sensitivity hypothesis is equally weak when we look at the magnitude of the response. Columns (3) and (4) report the share of pairwise treatment comparisons where the magnitude of the response for men is larger, smaller, or no different, at the 5% level of significance, from that of women, for all tests and for tests in the social preference domain. The results

[8] Note that both tasks are elastic with respect to incentives, which is crucial for our purposes (Araujo et al., 2016). As alternative pairwise comparisons, one could consider comparisons within behavioral treatments and between task types (typing keys versus coding WWII conscription cards), though the difficulty here is that the tasks have different outcomes and, potentially, different levels of noise, which would challenge inference on gender responsiveness.

[9] See Online Appendix C for a detailed description of the treatments and the 60 pairwise comparisons.

show that, for the vast majority of pairwise comparisons there is no gender difference in responsiveness, and if anything, the magnitude of the response by men is greater than that by women.[10]

Table 2: Gender Differences in Treatment Responsiveness in DellaVigna and Pope (2022)

	Treatment Response		Magnitude of the Response	
	All Topics (1)	Social Preferences (2)	All Topics (3)	Social Preferences (4)
Men	8	4	6	0
Women	4	1	0	0
Both/Neither	48	23	54	28
Total	60	28	60	28

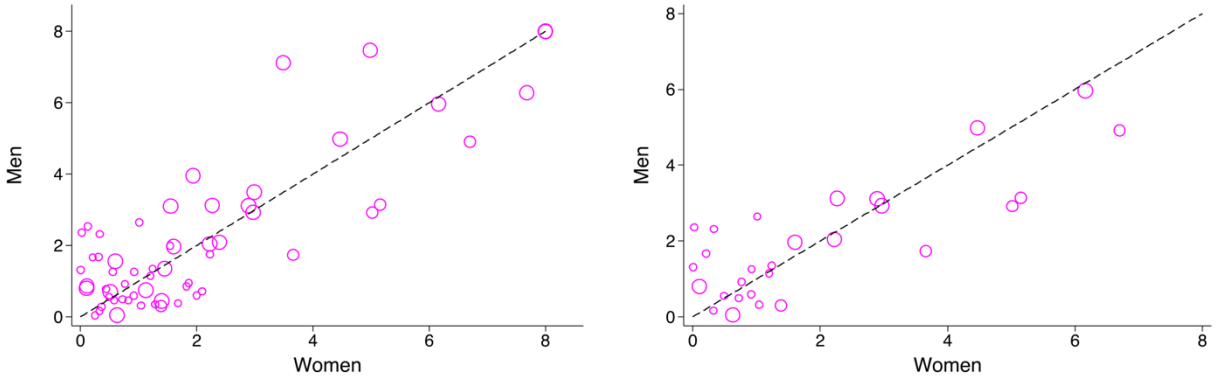
Notes: The table summarizes the classification of results on the female sensitivity hypothesis using data from DellaVigna and Pope (2022). We classify each pairwise comparison first using the method in the literature review (determining whether one gender and not the other has a directional treatment response) and second, by considering the magnitude of the response. Specifically, we classify each pairwise comparison as *Men* (*Women*) if the treatment effect is larger for men (women), in absolute terms, and the interaction coefficient in an OLS regression is significant at the 5% level. For all other cases, we classify the magnitude of the response as being indistinguishable between men and women (*Both*/*Neither*).

The gender differences in the magnitude of responses to treatments is further explored in Figure 1. Specifically, it illustrates changes in mean effort in response to treatment changes for men (y-axis) and women (x-axis), plotting in Panels a and b the absolute values of the test statistics (t-values) and in Panels c and d the absolute values of the effect sizes (Cohen's *d*). As before, we report results separately for all 60 comparisons and for the subset of 28 treatment changes in the social preference domain (e.g., donations to the Red Cross, social comparisons, etc.). In each graph, the size of the dots is proportional to the sample size for each pairwise comparison. If women were more sensitive to variations in treatment or in experimental conditions, we would expect to see most dots below the 45-degree line. However, as seen in Figure 1, the dots are roughly evenly distributed, with a noticeable concentration near the 45-degree line. This pattern suggests that the findings from DellaVigna and Pope (2022) align with the results from our assessment of the broader literature, showing no evidence that being female is predictive of greater responsiveness to treatment variation.

[10] While our main analysis uses a 5% cutoff for statistical significance the results are similar with a less conservative 10% cutoff. A 10% cutoff would change the numbers in column (3) from 6 to 8 for men and remain unchanged for women. In column (4) the numbers would change from 0 to 1 for men and remain unchanged for women.

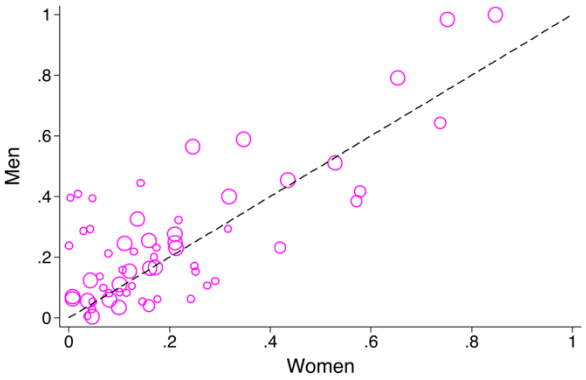
If anything, men appear slightly more sensitive to treatment than women, as the average effect size is 0.26 for male participants and 0.21 for female participants (p = 0.136, Mann-Whitney).

Figure 1: Scatter Plot of Test Statistics and Effect Sizes for Difference in Mean Effort by Gender Across Experimental Conditions in DellaVigna and Pope (2022)

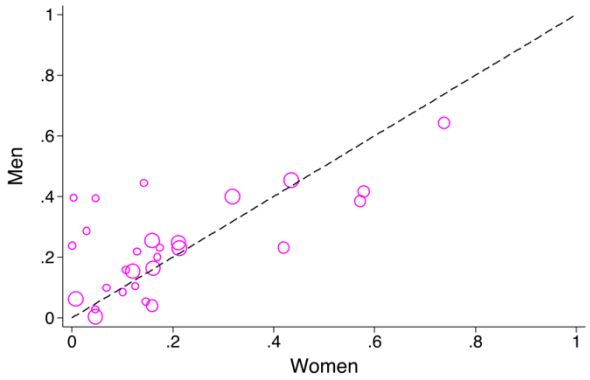


(a) t-Values: All Treatments

(b) t-Values: Social Preference Treatments



(c) Cohen's d: All Treatments



(d) Cohen's d: Social Preference Treatments

Notes: Each dot corresponds to a change in mean effort in response to a change in treatment or experimental condition. The axes show the absolute values of either the t-values or the Cohen's *d* effect sizes for women (x-axis) and men (y-axis). Panels (a) and (c) show results for all 60 treatment comparisons. Panels (b) and (d) show results for the 28 treatment comparisons in the social preference domain. Dot sizes are proportional to the combined sample sizes. Panel (a): t-values for pairwise comparisons across all treatment variations; values are top-coded at 8 for ease of visualization (see Online Appendix D for a version of the graph without top-coding). Panel (b): t-values for pairwise comparisons for treatment changes in social preference domains. Panel (c): Cohen's *d* values for pairwise comparisons across all treatment variations. Panel (d): Cohen's *d* values for pairwise comparisons for changes in social preference domains.

3.2 Exley et al. (2025)

The data from Exley et al. (2025) allows for further assessment of the female sensitivity hypothesis. While this study primarily focuses on beliefs about gender differences in social preferences, it includes experimental tasks that enable us to examine whether women are more responsive than men to treatment variations within the social preferences domain. Unlike DellaVigna and Pope (2022), who employ a between-subject design with continuous effort measures, Exley et al. (2025) use a within-subject design with binary choices (selfish versus non-selfish decisions). This complementary analysis allows us to assess the robustness of our findings across different experimental paradigms.

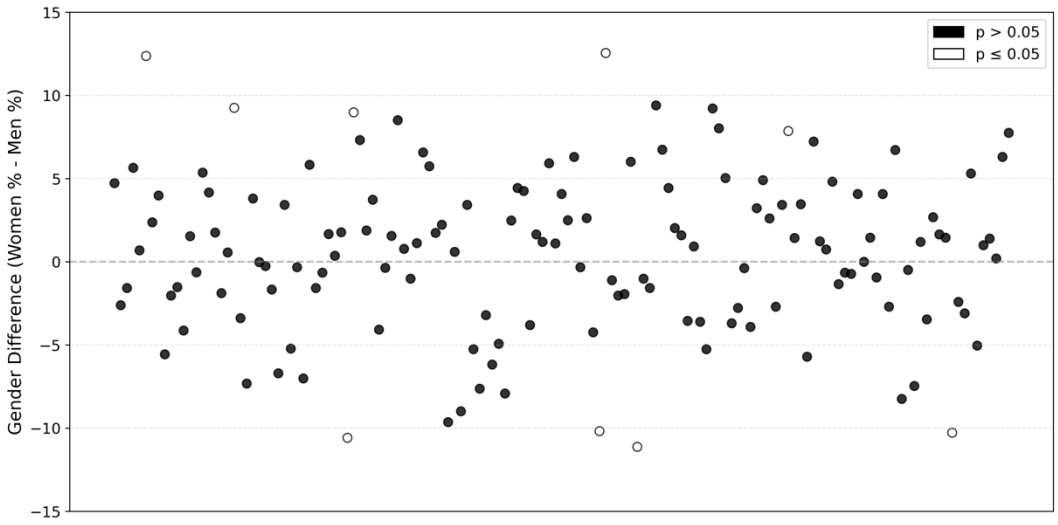
We assess the response to treatment for the three sets of experiments reported by Exley et al. (2025): (1) Economic Games with Undergraduate Students, (2) Economic Games with Online Participants, and (3) Economic Games with varying stakes. In both the undergraduate and online population studies, participants make decisions in seven different economic games: a classic dictator game, a dictator game where the selfish choice is more efficient, a dictator game with entitlement, an ultimatum game, a trust game, a prisoners' dilemma game, and a public goods game. Additionally, each game is played twice: once when outcomes affect only the participant, and once when outcomes affect others. In the stakes-vary experiments, participants make decisions across ten dictator games with varying stakes. Specifically, in each game, participants decide whether to keep 10 tokens for themselves or give 2, 4, 6, 8, 10, 20, 40, 60, 80 or 100 tokens to the other participant (where one token has the value of 10 cents).

Our analysis of the Exley et al. (2025) data focuses on within-subject changes in decisions across different experimental conditions. For each participant, we observe whether they switch between the selfish and non-selfish choice when moving from one experimental condition to another and compare the proportions of women and men who change their decisions between conditions.[11] We conduct a total of 143 pairwise comparisons. In both the undergraduate and online studies, we examine gender differences in responsiveness in two ways: (1) by comparing the share of women and men who change their decisions between the for-self and for-others versions of each of the seven games (14 comparisons) and (2) by comparing decisions between any two of the seven treatments separately for both self and others (84 comparisons). For the stakes-vary study, we compare the share of men and women who change their decisions between any two of the ten different treatments (45 comparisons).

[11] In using a within-subject design we can precisely assess the proportion of participants who respond to treatment. Our results are the same if we compare for each gender the proportion of selfish choices across treatments and instead assess the data as if it resulted from a between-study design (see Figure E.3 in Online Appendix E).

Figure 2 plots the difference, in percentage points, between the share of women and men that change their decisions (i.e., from selfish to non-selfish or vice-versa) between two treatment conditions, where positive values indicate cases where women are more likely to change their decisions than men.[12] Among the 143 pairwise comparisons, women are more likely to change their decisions in 79 cases, while men are more likely to do so in 64. However, a statistically significant gender difference, at the 5% level is only seen in 9 cases (6.3% of the pair-wise comparisons), where in five cases women change their decisions at significantly higher rates and in four cases men show significantly greater responsiveness. [13] That is 97% of the comparisons reject the female sensitivity hypothesis, revealing clearly that being female is not a reliable predictor of sensitivity to treatment variation.

Figure 2: Differences in the Share of Female and Male Participants that Changed Decisions in Pairwise Treatment Comparisons using Data from Exley et al. (2025)



Notes: Each dot corresponds to a pairwise treatment comparison and represents the difference, in percentage points, between the share of female participants that change their decisions (from selfish to non-selfish or vice versa) and the share of male participants that change their decisions. A positive value indicate that female participants change their decisions at higher rates than male participants.

[12] Using the within-subject design we also evaluate whether women more than men have a significant response to changes in treatment. Considering the share of women and men who change their decisions between treatments in each of the 143 pairwise treatment comparisons we find no observations are consistent with only women and not men responding significantly to treatment.

[13] Online Appendix E reports similar results using a 10% significance threshold. In that case, we observe 20 statistically significant differences: 10 larger for women and 10 larger for men.

In summary, our analysis of the Exley et al. (2025) data, complements and reinforces our findings from DellaVigna and Pope (2022). Across various experimental paradigms, participant populations, and outcome measures, we find no evidence that women are systematically more responsive than men to variations in experimental conditions.

4. Discussion and Conclusion

Our analysis challenges the frequently claimed female sensitivity hypothesis. By combining an extensive review of over 200 experimental economic studies with an empirical analysis of over 200 pairwise treatment comparisons from the data from DellaVigna and Pope (2022) and Exley et al. (2025), we provide strong evidence that being female is not a general predictor of sensitivity to treatment. Perhaps it should be expected that a blanket statement asserting greater female sensitivity to experimental treatments is unfounded. It would be surprising if women responded more than men, regardless of the nature of the treatment variation. We would instead expect men to be more responsive to changes in treatment that matter more to them, and men and women to be equally responsive in the many cases where they care similarly about the variations in treatment.

The lack of empirical support for the female sensitivity hypothesis has important implications for experimental economics, particularly in how gender differences in treatment responsiveness are understood. Our findings make the case for a more nuanced approach in designing and interpreting experiments, as well as a reconsideration of policies based on assumed gender differences in sensitivity. Rather than attributing variations in responsiveness to gender, researchers should focus on identifying the factors that drive differences in behavior.

To understand how the female sensitivity hypothesis has been disseminated and influenced researchers' interpretations of their findings, we examine studies from our literature review that explicitly reference the hypothesis. That is, we assess references to the hypothesis in the set of papers included in our literature review in section 2. We focus on papers that mention the female sensitivity hypothesis and classify each paper based on its stance. A study is classified as *neutral* if it merely acknowledges the existence of the hypothesis without endorsing it or assuming it as a scientific fact; as *favorable* if it either it presents empirical evidence in support of the hypothesis or explicitly assumes that women are more sensitive to treatment variations than men; and as *unfavorable* if it either presents the hypothesis while casting doubt on its

validity or argues that its own results contradict the hypothesis. By assessing how frequently the hypothesis is supported, refuted, or reported neutrally we can assess the current perception of the female sensitivity hypothesis.

Of the 38 studies that directly mention the female sensitivity hypothesis as stated in Croson and Gneezy (2009), 23 are classified as *favorable*, meaning they either conclude that their evidence supports the hypothesis or simply cite it as an established fact; 13 are classified as *neutral*, and only 2 studies reject the hypothesis and argue that their evidence contradicts it – and are thus classified as *unfavorable*. Both our literature review and our analysis of the data from DellaVigna and Pope (2022) and Exley et al. (2025) reveal a striking discrepancy between the pattern of referencing and the pattern of results supporting the female sensitivity hypothesis. While only 3% of citing studies are unfavorable towards the hypothesis, the empirical evidence reveals instead that the vast majority of treatment responses are inconsistent with the hypothesis. This is indicative of literary confirmation bias. Studies that find results in line with the hypothesis are more likely to cite it, while studies that do not find supportive evidence do not address it. Researchers may see a single result as insufficient for challenging the hypothesis and thus defer assessment for a more comprehensive review (similar to that presented here), while a comparable confirmatory result is more readily referenced to reinforce the hypothesis.

The consistent lack of empirical support for the female sensitivity hypothesis is noteworthy. Prior research presents no evidence that women are any more sensitive and fickler to treatment than men. Despite the lack of empirical support, the continued reference to the hypothesis fosters misperception and a trivialized understanding of gendered behavior. This form of selective referencing and confirmatory bias threatens our collective knowledge. Evidence-based assessments may be particularly important when the upheld hypotheses are based on stereotypical perceptions of human behavior.

References

Andreoni, J., & Vesterlund, L. (2001). Which is the fair sex? Gender differences in altruism. *The Quarterly Journal of Economics*, *116*(1), 293–312.

Araujo, F. A., Carbone, E., Conell-Price, L., Dunietz, M. W., Jaroszewicz, A., Landsman, R., Lamé, D., Vesterlund, L., Wang, S. W., & Wilson, A. J. (2016). The slider task: An example of restricted inference on incentive effects. *Journal of the Economic Science Association*, *2*(1), 1–12.

Ariely, D., Gneezy, U., Loewenstein, G., & Mazar, N. (2009). Large stakes and big mistakes. *The Review of Economic Studies*, *76*(2), 451–469.

Babcock, L., Recalde, M. P., Vesterlund, L., & Weingart, L. (2017). Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review*, *107*(3), 714–747.

Baldiga, N. R., & Coffman, K. B. (2018). Laboratory evidence on the effects of sponsorship on the competitive preferences of men and women. *Management Science*, *64*(2), 888–901.

Bandiera, O., Fischer, G., Prat, A., & Ytsma, E. (2021). Do women respond less to performance pay? Building evidence from multiple experiments. *American Economic Review: Insights*, *3*(4), 435–454.

Barrett, L. F., Lane, R. D., Sechrest, L., & Schwartz, G. E. (2000). Sex Differences in Emotional Awareness. *Personality and Social Psychology Bulletin*, *26*(9), 1027–1035. https://doi.org/10.1177/01461672002611001

Boschini, A., Muren, A., & Persson, M. (2012). Constructing gender differences in the economics lab. *Journal of Economic Behavior & Organization*, *84*(3), 741–752.

Buchan, N. R., Croson, R. T. A., & Solnick, S. (2008). Trust and gender: An examination of behavior and beliefs in the Investment Game. *Journal of Economic Behavior & Organization*, *68*(3–4), 466–476.

Charness, G., & Gneezy, U. (2012). Strong evidence for gender differences in risk taking. *Journal of Economic Behavior & Organization*, *83*(1), 50–58.

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

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, *47*(2), 448–474.

Datta Gupta, N., Poulsen, A., & Villeval, M. C. (2013). Gender matching and competitiveness: Experimental evidence. *Economic Inquiry*, *51*(1), 816–835.

DellaVigna, S., & Pope, D. (2022). Stability of Experimental Results: Forecasts and Evidence. *American Economic Journal: Microeconomics*, *14*(3), 889–925. https://doi.org/10.1257/mic.20200129

Dittrich, M., & Leipold, K. (2014). Gender differences in time preferences. *Economics Letters*, *122*(3), 413–415.

Dohmen, T., & Falk, A. (2011). Performance pay and multidimensional sorting: Productivity, preferences, and gender. *American Economic Review*, *101*(2), 556–590.

Eckel, C. C., & Grossman, P. J. (2001). Chivalry and solidarity in ultimatum games. *Economic Inquiry*, *39*(2), 171–188.

Eisenberg, N., & Lennon, R. (1983). Sex differences in empathy and related capacities. *Psychological Bulletin*, *94*(1), 100–131. https://doi.org/10.1037/0033-2909.94.1.100

Eriksson, K., & Simpson, B. (2010). Emotional reactions to losing explain gender differences in entering a risky lottery. *Judgment and Decision Making*, *5*(3), 159.

Exley, C. L., Hauser, O. P., Moore, M., & Pezzuto, J.-H. (2025). Believed Gender Differences in Social Preferences. *The Quarterly Journal of Economics*, *140*(1), 403–458. https://doi.org/10.1093/qje/qjae030

Exley, C. L., & Kessler, J. B. (2022). The gender gap in self-promotion. *The Quarterly Journal of Economics*, *137*(3), 1345–1381.

Exley, C. L., Niederle, M., & Vesterlund, L. (2020). Knowing when to ask: The cost of leaning in. *Journal of Political Economy*, *128*(3), 816–854.

Ezquerra, L., Kolev, G. I., & Rodriguez-Lara, I. (2018). Gender differences in cheating: Loss vs. Gain framing. *Economics Letters*, *163*, 46–49.

Flory, J. A., Leibbrandt, A., & List, J. A. (2015). Do competitive workplaces deter female workers? A large-scale natural field experiment on job entry decisions. *The Review of Economic Studies*, *82*(1), 122–155.

Fujita, F., Diener, E., & Sandvik, E. (1991). Gender differences in negative affect and well-being: The case for emotional intensity. *Journal of Personality and Social Psychology*, *61*(3), 427.

Gill, D., & Prowse, V. (2014). Gender differences and dynamics in competition: The role of luck. *Quantitative Economics*, *5*(2), 351–376.

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Harvard University Press.

Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, *118*(3), 1049–1074.

Heinz, M., Juranek, S., & Rau, H. A. (2012). Do women behave more reciprocally than men? Gender differences in real effort dictator games. *Journal of Economic Behavior & Organization*, *83*(1), 105–110.

Leibbrandt, A., & List, J. A. (2015). Do women avoid salary negotiations? Evidence from a large-scale natural field experiment. *Management Science*, *61*(9), 2016–2024.

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, *127*(2), 267.

Lotz, S. (2014). Is Women's Behavior More Context-Dependent than Men's? Gender Differences in Reluctant Altruism. *Gender Differences in Reluctant Altruism (December 18, 2014)*.

Miller, L., & Ubeda, P. (2012). Are women more sensitive to the decision-making context? *Journal of Economic Behavior & Organization*, *83*(1), 98–104.

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics*, *122*(3), 1067–1101.

Putrevu, S. (2001). Exploring the Origins and Information Processing Differences Between Men and Women: Implications for Advertisers. *Academy of Marketing Science Review*, *10*, 1–14.

Small, D. A., Gelfand, M., Babcock, L., & Gettman, H. (2007). Who goes to the bargaining table? The influence of gender and framing on the initiation of negotiation. *Journal of Personality and Social Psychology*, *93*(4), 600.