3-2021

To Perform or Not to Perform? Examining the Effects of Gender and Written Communication Style on Task Completion

Alvaro Plachejo
ajplachejo@stcloudstate.edu

To Perform or Not to Perform? Examining the Effects of Gender and Written Communication Style on Task Completion

By

Álvaro Plachejo

A Thesis

Submitted to the Graduate Faculty of

St. Cloud State University

in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in Applied Economics

March, 2021

Thesis Committee:
Mana Komai Molle, Chairperson
Patricia Hughes
Susan Parault Dowds

Abstract

I study the effect of two power categories, gender (male/female) and written communication style (strong language/weak language) on performance. To examine this relationship, these two attributes are considered in a request to perform a task using an experimental design on Amazon Mechanical Turk. Three variations of the experiment are performed: announcing an additional monetary reward for task performance, not announcing the reward, and explicitly warning that no reward is provided. I find significant differences in task performance caused by communication style such that weak language achieves 22.4 percent higher probability of task performance for male requesters while strong language has 11 percent more likelihood to achieve task performance for female requesters. Only the last experiment finds robust results and in two experiments no conclusions can be drawn due to the lack of variation in task performance. Omitted variable bias and lack of sample power might explain inconsistent results across experimental designs, and, while the lack of monetary reward in the third experiment does not allow for conclusions about task performance in hierarchical relationships, it does on the willingness of participants to perform a helpful act for the requester.

Keywords: Gender, Language style, Leadership, Performance

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Female leaders may face barriers in diverse settings by conscious or unconscious discrimination of followers based on gender roles expectations, influencing performance, and explaining the fact that females are less likely to experience mobility in the organizational hierarchy and, when they do, they face different treatment than their male counterparts. The purpose of this research is to further examine the relationship between power status and the ability to influence task completion. Specifically, this study seeks to measure the effect of two power categories, gender (male/female) and written communication style (strong language/weak language), to determine performance. An experimental design is used in which various groups will receive the same request using different frames to measure the impact of framing on task performance.

Exploring these issues is fundamental to the field of management because it can confirm performance biases and help to develop an understanding on mechanisms to reduce them, thus improving leaders' performance. Additionally, leaders can learn to navigate those biases during leadership transition periods until the capacity to change them is built. In this context, I provide contributions to three main areas of the literature. In the field of behavioral economics and framing, by continuing the analysis of framing across gender. Additionally, the experiment is of interest to further understand language style use, a topic that is unexplored in the literature. Finally, the experimental design contributes to further the understanding of differences in influence between gender roles. In this context, where task performance is of relevance to the productivity of organizations, awareness of gender differences in leadership may be key to overcome gender barriers and discrimination in management.

This project researches the relationship between two attributes of a request, gender and communication style, and the performance of a task by means of an experimental design. Four hypotheses guide this inquiry. First, (H.1.) the gender of the requester influences the decision to perform the task. In second place, (H.2.) the language of the request will influence the decision to perform the task. Next, I study congruity between the gender role of the requester and the language of the request, under the hypothesis that (H.3.) a male requester using strong language achieves different performance from followers than a female requester using weak language. The last hypothesis, following the literature on framing, is that (H.4.) these effects are different depending on the followers' gender.

An experimental design in Amazon Mechanical Turk tests these hypotheses by offering a sociodemographic survey that participants complete for a reward, followed by an additional task that participants choose whether or not to perform. The task request, however, is randomly framed with attributes of gender and language style of the requester. Therefore, the treatment is entirely exogenous. Furthermore, the experiment is repeated three times, with variations in the nature of the incentives: either announcing an additional monetary reward for task performance, not providing the reward, and explicitly announcing that no reward is provided.

I find significant differences in task performance caused by communication style and congruity between language and gender, but only in the last experimental design. The former treatment effect has a large magnitude of 11 to 24 percent, showing that weak language is more effective than strong language for male leaders and the opposite for female leaders. The latter effect is very small. Results are not consistent across specifications and the incentives of each experiment. I also find that the participants' gender shows no difference in framing effects.

These results must be qualified in the context of limitations that arise due to omitted variable bias that may explain inconsistent results across specifications. Furthermore, while the lack of monetary reward in the third experiment does not allow for conclusions about task performance in hierarchical relationships, it does on the willingness of participants to perform a helpful act for the requester. Limitations of this study are explored thoroughly later on.

In order to present this study, I first explore the conceptual framework that embodies the relevant literature. The third chapter focuses on describing the methodology used to test the stated hypotheses, including the experimental design and the econometric model. The fourth chapter presents and interprets the results of the experiments. Finally, the conclusions discuss the contributions and limitations of this study.

Chapter 2. Conceptual Framework

Social psychologists define power as control over another's resources and outcomes (Keltner et al., 2003) and, more specifically, as the capacity to recruit others in the service of one's agenda (Simon & Oakes, 2006). Within this context, being able to influence others to work towards your own vision seems to make a person powerful. Paradoxically, however, powerful people are the ones who are able to do the influencing in the first place (which in turn, makes them more powerful). In fact, Bruckmüller et al. (2012) found that high group social status (relative social prestige and prominence) is a determinant of normativity, with higher status identities (such as male gender, whiteness, and heterosexuality) becoming cultural default values and implicit norms against which to explain intergroup differences.

This research contributes to three main areas of the literature. First, that of behavioral economics and framing, by further studying the influence of framing across gender and using gender as an attribute frame in itself. Second, the literature on language style and its effect to achieve goals. Finally, by designing an experiment setting to study differences in influence between males and females to achieve task performance, the project contributes to the managerial literature by exploring gender discrimination, the so-called glass ceiling and glass cliff.

Pioneering research on framing in economics was focused on uncertainty and risky choices by studying the effects of the ways in which risk is "framed" or presented to agents required to choose alternatives or scenarios with the same expected utility. In this context, framing effects refer to shifts in behavior produced by presenting choices involving risk in different ways, which sometimes related to decisions yielding the same risk level but a different frame or to decisions with variations in the level of risk (Tversky & Kahneman, 1981). Levin et al. (1998) have advanced the knowledge in this field by developing a taxonomy of framing effects and the underlying mechanisms behind those effects. Three kinds of framing are identified: the aforementioned risky-choice framing, where framing is related to the probability associated with each outcome; attribute framing, which focuses on the characteristics of an alternative that incentivizes a specific behavior; and goal framing, where the goal of an action or behavior is framed by highlighting the gains or losses associated with it. More recent studies attempt to combine different types of frames (Peng et al., 2013). This project focuses on attribute framing by presenting four different treatments with changes in the attributes of gender of the requester and language style of the request.

While the literature on framing is reviewed in Levin et al. (1998), I describe below an influential paper for each type of framing as a means of illustration. First, risky choice framing is well known after Tversky and Kahneman's (1981) work on the hypothetical "Asian disease problem," where participants choose a program out of a set of differently framed alternatives that determines the probability of survival or death of the population. Attribute framing effects are usually related to willingness-to-pay and willingness-to-accept as influenced by variations in the characteristics of an object, such as the country of origin of automobiles (Levin et al., 1996). Finally, Kahneman et al. (1990) exemplify goal framing with the endowment effect theory, which states that individuals have different evaluations of gaining (preferred) and losing (aversion); in other words, willingness to accept exceeds willingness to pay.

Framing happens when a change in the context (frame) causes people to react differently to a particular piece of information or to an otherwise identical situation. One plausible explanatory mechanism of framing is that humans' reflective system does not do the work that would be required to check and see whether reframing the request would produce a different outcome (Thaler & Sunstein, 2009). Therefore, framing may be caused by somewhat mindless motivations behind actions when agents act as passive decision makers, but a complete theory on framing has not been developed yet (Payne et al., 1993; Hasseldine & Hite, 2003).

The gendered effects of frames is a topic that is frequently addressed in the framing literature. For example, Hasseldine and Hite (2003) find that females show larger tax compliance than males and both groups significantly respond better to positively framed messages that point out gains from compliance rather than penalties from non-compliance. Furthermore, these sex differences in framing are widely explored in many contexts (Huan and Wang, 2010), including medical decision-making (Peng et al., 2013), entrepreneurship and risky choice (Emami, 2017), human capital investment and borrowing (Bartholomae et al., 2019), performance and task difficulty (Jian-jun et al., 2011), and even the choosing of potential mates (Saad and Gill, 2014). In general, findings tend to conclude that females exhibit greater sensitivity to negative framing, with some exceptions (Croson and Gneezy, 2009).

The aforementioned studies focus on behavior as determined by the characteristics of the agent, but it is not less relevant to study the behavior of agents related to their perceptions of others' characteristics. Consider the resume experiment in Bertrand and Mullainathan (2004), which uses a field experiment to tests employers' willingness to respond to white-named resumes compared to black-named resumes. Here, a principal chooses to offer a job to a fictional candidate based on his/her resume, and these resumes are treated with names typically assigned to white or black people. The researchers find that employers are less likely to call back a job applicant who is perceived to be black. This is a form of attribute framing in which race is the changing attribute and, in a similar fashion, the effects of gender framing can be tested. For example, Miller et al. (1991) study the relationship between gender and frames in the context of causal explanations. The authors look at differences in the explanation of political behavior as influenced by the framing of gender in each situation, which requires participants to explain the behavior of different gender-based characterizations.

In managerial settings, the so-called “glass ceiling” embodies the idea that females face barriers in the organizational hierarchy, causing gaps and underrepresentation of women in high-earnings jobs (Pande and Ford, 2014; Guvenen et al., 2014). In top managerial positions, the decomposition of earnings shows that 75 percent of the wage gap between sexes is explained by the size of the firm and the roles of female executives; however, at least 5 percent of that wage differential remains unexplained (Bertrand and Hallock, 2001). Furthermore, Bertrand et al. (2010) have identified three factors that explain the gap in earnings between males and females in management positions: differences in training pre-graduation, career interruptions, and weekly hours of work, with the latter two being specifically associated to childcare.

Although the literature has reported mixed findings on sex differences in leadership (Van Engen and Willemsen, 2004), Adams and Funk (2012) found gender-based differences in risk attitudes and values across a sample of board of directors. However, their findings are somewhat contradictory, as female directors, as expected, seem to be more benevolent and less power oriented, but, at the same time, are less risk-averse than their male counterparts. Similarly, Grossman et al. (2015) find that females show more cooperative behavior and hesitation to lead. However, the researchers also conclude that followers behave the same way regardless of the gender of the leader, which appears to suggest that, at least in carefully designed games in a laboratory setting, followers show no bias related to leadership gender. That being said, other studies have identified a gap between male and female leaders in performance evaluations and rewards, where females tend to be evaluated worse than males and their rewards are fewer (Grossman et al., 2019; Joshi et al., 2015).

An important feature of leadership is mannerism and language. Sacavem et al. (2017) study the delivery style of leaders and the mood of followers, finding that dominant and immediate leaders achieve better performance and perception. Across renowned business blogs and media there has been a discussion about language style, using a typology that classifies language as “strong” or “weak.” According to Weissman (2011a, 2011b), strong language is definite, specific, and concrete. It provides the audience with as much certainty as possible by replacing conditional language with forward-looking statements, such as “I am confident,” “I am convinced,” “I am optimistic,” and “I expect.” Strong language uses positive statements (such as “What I am…” rather than “What I am not…”) and meaningful words stated in a declarative, assertive mood because it is more likely to produce meaningful actions.

Conversely, weak language employs conditional terms such as “I believe,” “I think,” and “I feel,” which casts doubt on the competence of the presenter. Finishing a sentence with “does that make sense?” reflects doubt about the ability of the audience to comprehend the message. Furthermore, using qualifying words such as “sort of,” “kind of,” “just,” “pretty much,” “basically,” and “really” lessen the importance and value of the nouns and verbs they accompany and reduce the credibility of the speaker. Additionally, the phrase “to be honest…” makes it seem as if the speaker was not being truthful before. Minimizing wording is another feature of weak language. For example, stating that one may not have as much expertise as others reduces the presenter's credibility. Similarly, using tag lines at the end of a sentence, such as "don't you think?" or "isn't it?" weakens the authority of the speaker because it shows that he/she is not completely confident and requires the reassurance of the audience (Marcus, 2011). Lastly, using negative statements (such as "What I am not…" rather than "What I am…") fails to provide information and sounds defensive.

This typology appears unexplored in the literature and it has not been a part of the experimental research on leadership, but studies on gender roles do explore the idea that people tend to seek congruity between their gender roles and their environment, and that incongruity will lead to worse evaluations (Eagly & Karau, 2002). A study by Bruckmüller & Abele (2010) found that members of normative groups (such as males) were perceived to be more "agentic" (competent, assertive, and independent) and less "communal" (warm, cooperative, and empathic). Similarly, Ellingsen et al. (2013) examined gender differences in social dilemmas across two different frames, community and stock market, and found that the difference in behavior between men and women was statistically significant in the former but not the latter. That is, women were significantly more cooperative than men in specific situations only.

Following the notion of glass ceiling and the suggestion that, once bypassed, women in positions of power face different treatment that men (Bruckmüller et al., 2014; Groeneveld et al., 2020), this project evaluates gender role congruency and language as barriers for female leaders to achieve performance. In this context, it is expected that results will show females face less responsiveness to achieve task performance from participants, thus representing a form of "glass cliff’ arising from gender roles, which females are expected to compensate for to become effective leaders.

In conclusion, the literature allows me to shape a series of hypotheses regarding the nature of performance as determined by the attributes of the request, such as the gender of the requester and the language style of the request. To explore gender roles and effectiveness of leadership in this context, the first expectation is that there will exist a different, and likely negative, result when females make a request. Secondly, by distinguishing between communication styles, it should be possible to assess responsiveness to communication and gender roles, while also testing differences that may arise from congruency between language and sex, as agents may expect males to be more assertive and determined. Finally, experimental literature would suggest that female participants will exhibit different performance than males because of their higher sensitivity to framing.

Chapter 3. Methodology

Experimental Design

This research develops three different experimental designs using Human Intelligence Tasks (HITs) in Amazon Mechanical Turk (MTurk). MTurk is an online labor market where requesters post jobs that are completed by MTurk workers. Jobs are presented in the form of HITs that offer a specific reward. Workers are free to choose which HITs to complete based on their title, reward amount, and a short description that can be accessed by clicking on the HIT. Once a worker decides to complete the HIT, he/she can do so by clicking on the “Accept & Work” button, which will lead to the survey screen where a link to Qualtrics will be available to access the survey.

There are many empirical studies that address the validity of experiments performed using Mechanical Turk. For example, Berinsky et al. (2012) show that respondents recruited via MTurk are more representative of the U.S. population than in-person convenience samples. Furthermore, Huff and Tingley (2015) compared participants of an MTurk survey against those of the Cooperative Congressional Election Survey (CCSE) and discovered that respondents in both samples were similar in terms of gender, race, geographical location, occupation, and political ideology. The one demographic difference they found was regarding age, since MTurk respondents were significantly younger than those of the CCSE, with the majority of respondents being under 45. This seems to make sense given the different platforms in which each of the surveys is administered.

The first experiment considers two monetary rewards. Initially, individuals are paid $0.10 for completing a demographic survey HIT. Questions are standard sociodemographic variables and it was developed following the American Community Survey questionnaire available at iPUMS (Ruggles et al., 2020). After the survey is finalized, a request is presented offering an additional $0.10 reward for completing another task. Hara et al. (2018) recorded 2,676 workers performing 3.8 million tasks on Amazon’s Mechanical Turk and showed that workers earned on average a median hourly wage of only $2.00 ($0.033 per minute). Therefore, offering $0.10 for a one-minute survey is attractive enough to recruit the participants required for the study.

The request to complete the second task includes two different randomized treatments, that is four different frames. These frames will vary by the gender of the requester (male/female) and the language of the request (strong language/weak language). The four frames will be: (1) Male requester using strong language, (2) male requester using weak language, (3) female requester using strong language, and (4) female requester using weak language. For validity reasons, treatments being randomized implies a 25% probability for each frame, and the actual data may have minor differences with respect to the true probability.

The first treatment regards the sex of the requester, which is inferred by the signed name on the request. If the requester is male, the request is signed by Nathan Johnson. If the requester is female, the request is signed by Michelle Johnson. To avoid the influence of ethnic perceptions on the decision to complete the additional task, ethnically neutral first names were purposefully chosen. Following Sisense’s (2018) analysis of the names of 372,534 babies born in New York City between 2011-2016, the most ethnically neutral male names were Richard, Marcus, and Nathan and the most ethnically neutral female names were Aria, Michelle, Chloe, and Isabelle. Their neutrality was determined based on the representativeness of each name across multiple ethnic groups (Asian, Black, Hispanic, and White). The last name for both requesters, Johnson, was chosen from a list of the most common last names in the United States, with the expectation that it will also have high representativeness across multiple ethnic groups.

The second treatment is determined by the request being written using strong language or weak language. Following Weissman (2011b), characteristics of strong language include certainty and avoiding conditional wording. Then, I determine that a prototypical statement will contain the positive statements previously identified: “I am confident,” “I am convinced,” “I am optimistic,” and “I expect.” Study participants being offered to complete the additional task by a male or a female using strong language receive the following message:

> Thank you for completing the demographic survey. This study is important to advance knowledge in the social sciences. I am convinced that your participation will make a difference. I want to offer you the opportunity to complete an additional task for a bonus payment of $0.10. My experience as a researcher makes me confident that your participation will impact my findings. If you choose not to do it, you will still receive compensation for the demographic survey. [Nathan Johnson/Michelle Johnson].
> Will you complete the additional task?

In the previous section, the conditional statements “I believe,” “I think,” and “I feel,” (Weissman, 2011a) where identified as characteristics of weak language. Furthermore, weak language ends communication with doubts -“does that make sense?”- and employs qualifiers such as “sort of,” “kind of,” “just,” “pretty much,” “basically,” and “really.” Combined with Marcus’ (2011) idea that minimizing wording is a feature of weak language, study participants being offered to complete the additional task by a male or a female using weak language will receive the following message:

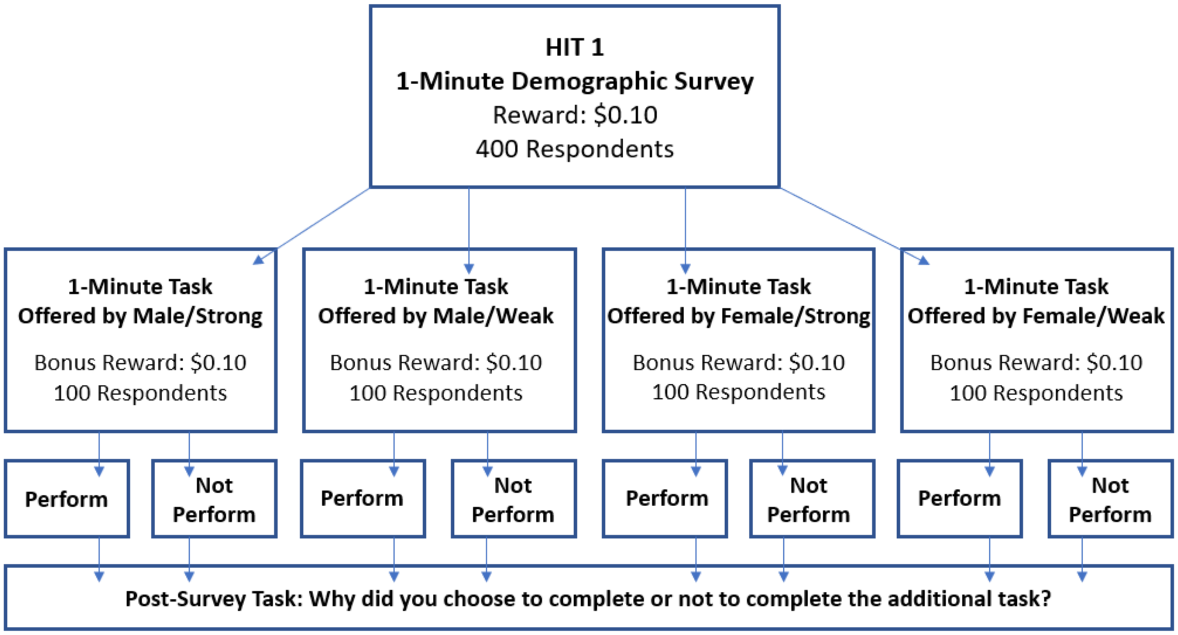
> Thank you for completing the demographic survey. I think this study may possibly advance knowledge in the social sciences. I think your participation will make a difference (let us hope so). I want to offer you the opportunity to complete just one additional task for a bonus payment of $0.10. As a researcher, I feel that your participation may somewhat impact my findings (keeping my fingers crossed!). If you

choose not to do it, you will still receive compensation for the demographic survey. I hope this makes sense. [Nathan Johnson/Michelle Johnson].
Will you complete the additional task?

A binary yes/no decision is offered. The HIT progression is depicted below:

Figure 1

Survey Algorithm



The additional task will be the same for all participants and it consists of responding to an open-ended question: "Why did you choose to complete or not to complete the additional task?" While the participant's answer may shed some light into the decision, the study is interested on the decision to complete the additional task based on the frame through which it was offered, rather than the participant's rationalization of such decision.

Because many HIT rewards in MTurk range around $0.10, the monetary incentive to perform the additional task is very appealing. As such, large compliance with the task may bias results due to the low variation in denying the task. The other two experimental designs attempt to solve this issue by excluding the second monetary reward (while still paying $0.10 for completing the initial demographic survey). In the second experiment, the frames are modified such that there is no mention of the monetary incentive. In the third experiment, however, the message explicitly warns participants that there is no monetary payment for completing the additional task.

Econometric Model

To answer the research questions, a Linear Probability Model is proposed to estimate the following equation:

(1) $Perform_i = \beta_0 + \beta_1(Female)_i + \beta_2(Weak)_i + \beta_3[(Female)_i \times (Weak)_i] + \beta_4 X_i + \epsilon_i$

The dependent variable is a binary variable of task performance that is 1 if the individual replied "yes" to the request to perform the additional task or zero otherwise. The first treatment is the gender of the requester, with male being the baseline. The second treatment is a binary variable identifying the language of the request, which is equal to 1 if it displays weak language and zero if it displays strong language. The independent variable X_i is a vector of control variables associated with the ith respondent: sex, education, race, income, marital status, age, employment status, political affiliation, and the type of setting the individual lives in (small city, large city, suburban, or rural). A full description of these variables is presented in Table 1.

The baseline is the mean task performance given a male requester using strong language. The estimator β_1 is interpreted as the percentage increase in average task performance, or the likelihood of performing, for a female requester among requesters using strong language. β_2 captures the difference in the conditional mean of task performance given that language is weak versus the baseline of strong language, among male requesters. Finally, β_3 is an interaction such that the combined coefficients $\beta_1 + \beta_2 + \beta_3$ represent the difference in the conditional mean of task performance between a female requester using weak language and the baseline, a male requester using strong language. Coefficients $\beta_2 + \beta_3$ allows to compare a female requester using strong language with a female requester using weak language.

Table 1

List of socioeconomic control variables

Variable	Type of Variable	Description
Female Respondent	Binary	1 Female Respondent 0 Male Respondent
Educational Attainment	Categorical	0 Less than high school 1 Regular high school diploma 2 GED or alternative credential 3 Some college credit, but less than one year of college credit 4 One or more years of college credit, no degree 5 Associate's degree 6 Bachelor's degree 7 Master's degree 8 Professional degree beyond a bachelor's degree 9 Doctorate degree
White	Binary	1 White 0 Minority (Hispanic, African American, Asian, others)
Income	Continuous	Total yearly income in dollars

Table 2 Continued

Married	Binary	1 Married or cohabitation 0 never married, widowed, divorced, separated
Age	Continuous	Age in years
Employed	Binary	0 Unemployed 1 Employed
Republican	Binary	0 Democrat or other 1 Republican
City	Categorical	0 Small urban 1 Large urban 2 Suburban 3 Rural

It follows that Hypothesis 1 analyzes the influence of the gender of the requester on the decision to complete the task. With the null hypotheses being $\beta_1 = 0$ and $\beta_1 + \beta_3 = 0,$ a female requester would be equally likely to achieve task performance than a male requester, regardless of the type of language used, thus rejecting Hypothesis 1. Hypothesis 2 regards the influence of the language of the request on the decision to perform and is explored through β_2, such that a positive estimator will imply that weak language has a positive effect in influencing agents to perform across male requesters, and $\beta_2 + \beta_3$ captures the difference between a female requester using strong language and a female requester using weak language. To study the congruity between gender roles and language stated in Hypothesis 3, I use the interaction estimator $\beta_1 + \beta_2 + \beta_3$ that compares the baseline of male requester using strong language to a request made by a female using weak language.

The fourth Hypothesis studies whether the respondent's sex influences his/her willingness to perform the additional task. To analyze this question, a triple interaction model is proposed, following the equation:

(2) $Perform_i = \gamma_0 + \gamma_1(Weak)_i + \gamma_2(Female)_i + \gamma_3(Female\_Res)_i + \gamma_4\,[(Weak)_i \times (Female)_i] + \gamma_5[(Weak)_i \times (Female\_Res)_i] + \gamma_6[(Female)_i \times (Female\_Res)_i] + \gamma_7[(Weak)_i \times (Female)_i \times (Female\_Res)_i] + \gamma_8 X_i + \mu_i$

In this model, $Female$ is the gender treatment, $Female\_Res$ represents the sex of the participant, and the baseline is the performance of a male participant with a congruent request from a male requester using strong language. Similarly, the difference between the baseline and a female participant facing a weak-language request from a female requester is measured by $\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 + \gamma_5 + \gamma_6 + \gamma_7$. If the combined estimators are positive, female participants are more likely than male participants to comply to a request that is congruent between gender roles and language style. A statistically significant result in this estimator, independent of its direction, would suggest that individuals of a particular sex are more influenced by congruity. If that were the case, the results of the experiment could be consistent with empirical findings that women are more sensitive to framing than men (Croson & Gneezy, 2009).

Following the notation above, the null and alternative hypotheses can be summarized as:

1.A. $H_0\colon \beta_1 = 0, H_A\colon \beta_1 \neq 0$ explores the difference in performance generated by a female requester and a male requester, both using strong language. Under the null hypothesis, both requesters achieve the same outcome. An individual significance t-test is used to test this hypothesis.

1.B. $H_0\colon \beta_1 + \beta_3 = 0, H_A\colon \beta_1 + \beta_3 \neq 0$ explores the difference in performance generated by a female requester and a male requester, both using weak language. Under the null hypothesis, both requesters achieve the same outcome. A joint significance F-test is used to test this hypothesis.

2.A $H_0\colon \beta_2 = 0, H_A\colon \beta_2 \neq 0$ explores the difference in performance generated by male requesters using weak language and strong language. Under the null hypothesis, both language styles achieve the same results. An individual significance t-test is used to test this hypothesis.

2.B. $H_0\colon \beta_2 + \beta_3 = 0, H_A\colon \beta_2 + \beta_3 \neq 0$ explores the difference in performance generated by female requesters using weak language and strong language. Under the null hypothesis, both language styles achieve the same results. A joint significance F-test is used to test this hypothesis.

3. $H_0\colon \beta_1 + \beta_2 + \beta_3 = 0, H_A\colon \beta_1 + \beta_2 + \beta_3 \neq 0$ explores the difference in performance generated by a request made by a female using weak language and a request made by a male using strong language. Under the null hypothesis, there is no difference in performance. A joint significance F-test is used to test this hypothesis.

4. $H_0\colon \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 + \gamma_5 + \gamma_6 + \gamma_7 = 0, H_A\colon \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 + \gamma_5 + \gamma_6 + \gamma_7 \neq 0$ explores the difference in the likelihood to perform between a female participant receiving a request from a female using weak language and a male participant receiving a request from a male using strong language. Under the null hypothesis, females and males show no difference in

their reactions to the congruency of the request. A joint significance F-test is used to test this hypothesis.

Hypotheses 1, 2, and 3 can be tested using both equation (1) and equation (2), because the beta coefficients in equation (1) are also represented in equation (2). Hypothesis 4, however, can only be tested using equation (2). The analysis will have to consider the consistency of the estimates between both equations and the results of joint significance tests. The next section attempts to evaluate these equations and interpret their results.

Chapter 4. Analysis

This chapter begins by presenting descriptive statistics for all three experiments. Next, I use non-parametric and regression analysis to analyze the findings from each experiment. The mean t-test analysis attempts to compare participants in all three experiments using their sociodemographic characteristics, to examine their similarities and to determine whether the three experimental designs achieve the goal of the experiment by reducing biases. Finally, regression analysis presents the main results and estimates of the treatment effects.

Summary statistics for experiment 1 are presented in Table 2. This experiment included an explicit reward to perform and 90.1% of 413 participants chose to perform the task. As seen below, 46.7% of participants were randomly assigned to the weak language treatment and 44.6% were assigned to the female requester treatment. Regarding socioeconomic variables: the mean age was 36.39 years old, with a minimum of 18 and a maximum of 79 years of age. Average income was $51,212 while 79.2% of the sample was employed the week before participating in the survey. Also, 67.6% of the sample is white, 46.0% of the participants are female, and 40.7% define themselves politically as Republicans.

Table 3

Summary statistics (experiment 1)

	(1)	(2)	(3)	(4)	(5)
VARIABLES	N	mean	sd	min	max
Age	413	36.39	11.06	18	79
Income	413	51,212	88,066	0	1.100e+06
Female	413	0.460	0.499	0	1
Employed	413	0.792	0.407	0	1
Married	413	0.608	0.489	0	1
Republican	413	0.407	0.492	0	1
White	413	0.676	0.469	0	1
Treatment weak	413	0.467	0.500	0	1
Treatment female	413	0.446	0.498	0	1
Perform	413	0.901	0.299	0	1

Table 3 presents summary statistics for the second experiment, where no monetary reward for participating was awarded, but it was not explicitly announced. In this scenario, there are 406 participants and 85.2% of them chose to perform the additional task. A weak language treatment was assigned to 49.3% of the participants and 52.7% received the female requester treatment. Mean age is 36.25 years old, average income is $44,291, and employment rate in the previous week was 81.8%. Furthermore, similar to the first experiment, 69.5% of the sample is white, 42.9% of the participants are female, and 43.3% identify as Republicans.

Table 4

Summary statistics (experiment 2)

	(1)	(2)	(3)	(4)	(5)
VARIABLES	N	mean	sd	min	max
Age	406	36.25	10.93	18	74
Income	406	44,291	41,937	0	360,000
Female	406	0.429	0.495	0	1
Employed	406	0.818	0.387	0	1
Married	406	0.623	0.485	0	1
Republican	406	0.433	0.496	0	1
White	406	0.695	0.461	0	1
Treatment weak	406	0.493	0.501	0	1
Treatment female	406	0.527	0.500	0	1
Perform	406	0.852	0.355	0	1

Summary statistics for the last experiment are presented in Table 4, in which, under explicit non-monetary reward, 64.8% of the 403 participants chose to perform the additional task. Weak language and female requester treatment were randomly allocated to 46.7% and 43.4% of participants, respectively. Furthermore, the mean participant is 37.26 years old and earns $49,048 per year. Summary statistics show that 43.7% of participants in this experiment are female, 77.7% were employed during the previous week, 58.8% are married, 40.9% identify politically as Republicans, and 73.9% are white.

The first experiment had the most decisions to perform the additional task, which suggests that the $0.10 reward is enough for agents to not be indifferent between performance and nonperformance. Table 5 shows the results of a two-sample t-test that compares the mean performance between any two experiments, under the null hypothesis that both samples have equal means ($\bar{\mu}_1 = \bar{\mu}_2$) and the alternative hypothesis that the means are not equal ($\bar{\mu}_1 \neq \bar{\mu}_2$). The results show that, when variance in responses across experiments is accounted for, there is no statistical difference at 1% between experiments 1 and 2, which means that both experiments offer a similar response rate to task performance (90% and 85%, respectively).

Table 5

Summary statistics (experiment 3)

	(1)	(2)	(3)	(4)	(5)
VARIABLES	N	mean	sd	min	max
Age	403	37.26	12.65	18	79
Income	403	49,048	68,102	0	960,000
Female	403	0.437	0.497	0	1
Employed	403	0.777	0.417	0	1
Married	403	0.588	0.493	0	1
Republican	403	0.409	0.492	0	1
White	403	0.739	0.439	0	1
Treatment weak	403	0.467	0.499	0	1
Treatment female	403	0.434	0.496	0	1
Perform	403	0.648	0.478	0	1

This result indicates that the monetary incentive might produce a biased coefficient, and that treatment effects will appear smaller due to the fact that the financial reward seems to compensate participants for their opportunity cost of performing. In the first experiment this financial reward is explicit and, as explained previously, significant for the context in which it is being offered. In the second experiment, while no reward was offered, the qualitative responses to the question “Why did you choose to complete or not to complete the additional task?” indicate that the absence of an explicit statement about the lack of reward made many participants assume that additional compensation would be offered, which produces a similar bias than the explicit offer of a reward. The third experiment, however, has a statistically different mean from both the first and the second experiments (measured in the last two columns, respectively), which can be attributed to explicitly stating that there was no additional reward associated with performance. This suggests that treatment effects should be more salient in this third experimental design.

Table 6

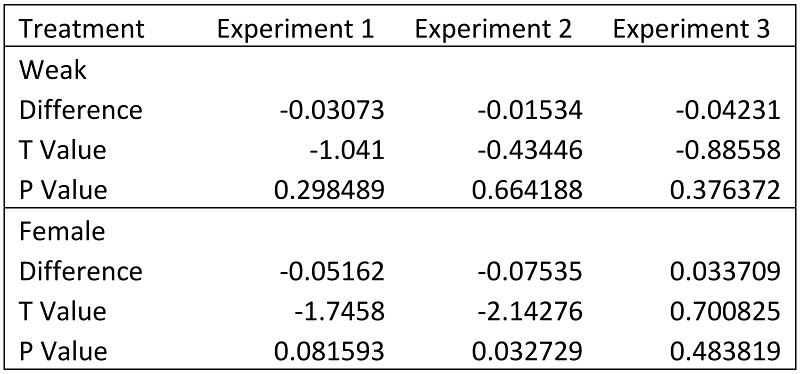
Two-Tailed Mean T-Test between Experiments

Variable	Experiment 1 (Mean)	Experiment 2 (Mean)	Test Statistic p-value	Experiment 3 (Mean)	Test Statistic p-value	Experiment 2 and 3 p-value
Perform	0.90	0.85	0.03	0.65	0.00	0.00
Female	0.46	0.43	0.37	0.44	0.50	0.82
White	0.68	0.69	0.56	0.74	0.04	0.16
Age	36.39	36.25	0.85	37.26	0.30	0.23
Income	51211.65	44290.85	0.15	49048.40	0.70	0.23
Employment	0.79	0.82	0.35	0.78	0.60	0.15
Married	0.61	0.62	0.65	0.59	0.57	0.31
Conservative	0.41	0.43	0.44	0.41	0.94	0.49
N	413	406		403		

Table 5 also reveals that there are no significant differences at p-values less than 1% between participants' sociodemographic characteristics, excluding factor variables. This means that differences in coefficients that could arise in regression analysis will be the result of different decisions made by participants and not of differences in the composition of the participant pools.

Table 7

Two Tailed Mean Test of Performance by Treatment



Treatment	Experiment 1	Experiment 2	Experiment 3
Weak			
Difference	-0.03073	-0.01534	-0.04231
T Value	-1.041	-0.43446	-0.88558
P Value	0.298489	0.664188	0.376372
Female			
Difference	-0.05162	-0.07535	0.033709
T Value	-1.7458	-2.14276	0.700825
P Value	0.081593	0.032729	0.483819

Table 6 performs non-parametric test of differences in means by each of the two treatments. For the weak language treatment, there are no significant differences in the means across all experiments. Meanwhile, the female requester has different results across experiments. In the first experiment, a female requester gets approximately five percentage points less performance than a male requester with a p-value less than 10 percent; the second experiment finds a negative difference of 7 percentage points that is significant at less than 5 percent.

Regression results for all three experiments are shown in Table 7, using the linear probability model to estimate equation (1). The first column provides estimates for the experiment with monetary incentive, while columns 2 and 3 show results for the experiments with no monetary incentive and explicit no monetary incentive, respectively. There is no evidence that participants are more likely to perform the task under the treatment of language or gender of the requester, as all coefficients remain statistically insignificant across experiments. That is to say, we cannot reject the null hypotheses that the gender of the requester has no influence on the decision to complete the task and that the language of the request has no effect on the decision to complete the task. These correspond to hypotheses 1.A. and 2.A., respectively.

Table 8

Regression analysis of double interaction

	(1)	(2)	(3)
VARIABLES	Experiment 1	Experiment 2	Experiment 3
Treatment weak = 1	-0.030	-0.007	0.099
	[0.045]	[0.059]	[0.064]
Treatment female = 1	-0.033	0.054	0.009
	[0.044]	[0.052]	[0.068]
Treatment weak x Treatment female	0.134**	0.047	-0.107
	[0.060]	[0.073]	[0.099]
Female	0.033	0.040	0.014
	[0.028]	[0.036]	[0.050]
White	0.033	-0.050	0.036
	[0.035]	[0.038]	[0.056]
Age	0.003**	0.002	0.005**
	[0.001]	[0.002]	[0.002]
Income	0.000	0.000	-0.000
	[0.000]	[0.000]	[0.000]
Employed	-0.014	-0.071*	-0.070
	[0.040]	[0.043]	[0.063]
Education = 2	0.437*	0.331	0.234
	[0.263]	[0.348]	[0.210]
Education = 3	0.216	0.217	0.466*
	[0.311]	[0.367]	[0.248]
Education = 4	0.426	0.257	0.275
	[0.269]	[0.351]	[0.218]
Education = 5	0.508*	0.382	0.383*
	[0.260]	[0.344]	[0.201]
Education = 6	0.431	0.172	0.166
	[0.264]	[0.353]	[0.215]
Education = 7	0.452*	0.249	0.360*
	[0.259]	[0.344]	[0.189]
Education = 8	0.463*	0.287	0.350*
	[0.262]	[0.345]	[0.196]
Education = 9	0.275	0.048	0.289
	[0.297]	[0.375]	[0.251]
Education = 10	0.487*	0.520	0.348
	[0.261]	[0.350]	[0.317]
Married	-0.013	0.051	0.026
	[0.036]	[0.039]	[0.052]

Table 9 Continued

Republican	0.001	0.030	-0.023
	[0.032]	[0.038]	[0.053]
City Setting = 2	-0.083**	-0.081*	-0.068
	[0.038]	[0.049]	[0.078]
City Setting = 3	-0.081**	-0.007	-0.045
	[0.035]	[0.046]	[0.079]
City Setting = 4	-0.077	-0.098*	0.010
	[0.053]	[0.058]	[0.087]
Constant	0.401	0.540	0.175
	[0.268]	[0.357]	[0.220]
Observations	413	406	403
R-squared	0.093	0.086	0.052
F-statistic	1.744	2.322	1.163
Female + Female x Weak	0.101	0.102	-0.0980
Weak + Female x Weak	0.105	0.0400	-0.00812
Weak x Female + Weak + Female	0.0715	0.0941	0.00112

Note. Robust standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

The corresponding effects for hypothesis 1.B. and 2.B. require to sum across coefficients, which I present at the bottom of Table 7. The sign of the summed coefficients is not consistent across experiments. The sum of Female with the Interaction, that is hypothesis 1.B., is positive in the first two experiments while negative in the third experiment. Table 8 shows the results for the F-test on the linear restriction and there is no evidence that females and males using weak language achieve different performance from participants. With respect to hypothesis 2.B., the summation of weak with the interaction is also jointly insignificant and I cannot find evidence that females achieve a different outcome by using weak or strong language.

Table 10

Linear restriction F-Test, Model 1 (p-values)

	Experiment 1	Experiment 2	Experiment 3
Female, Weak x Female	0.089447	0.954704	0.448786
Weak, Weak x Female	0.103913	0.661053	0.164549
Female, Weak, Weak x Female	0.228784	0.501692	0.210409

Congruity in gender and language is analyzed through adding the coefficients of the two main effects and the interaction. According to the results in Table 8, the effects of congruity are very small and jointly insignificant. Overall, this model suggests that there are no differences in performance between female requesters using weak language and male requesters using strong language. Furthermore, across all three models in Table 7 we have consistent results that show that the null hypotheses cannot be rejected.

Table 9 shows estimates from the linear probability model in equation 2, including a triple interaction between the two treatments and the participants' sex. Results are not consistent across all experimental designs and they differ from those discussed above. In the first and second experiments, no treatment is statistically significant, including the combined effects. Thus, I will focus on discussing the results of the third experiment.

When the lack of monetary incentive is explicitly mentioned in experiment 3, results differ from those achieved by the first regression model. I find no significant effect for the female requester treatment alone, corresponding to hypothesis 1.A. However, the summation of the female coefficient with the double interaction implies that a request made by a female using weak language is 22.6 percent less likely to influence performance than a request made by a male using weak language. This result was tested using an F-test presented in Table 10 and was found to be significant at less than 5% only in the third experiment.

Table 11

Regression analysis of second model

	(1)	(2)	(3)
VARIABLES	Experiment 1	Experiment 2	Experiment 3
Treatment weak = 1	0.008	-0.001	0.224***
	[0.064]	[0.087]	[0.085]
Treatment female = 1	0.007	0.091	0.108
	[0.063]	[0.077]	[0.094]
Treatment weak x Treatment female	0.092	0.078	-0.334**
	[0.084]	[0.105]	[0.129]
Female = 1	0.094*	0.109	0.127
	[0.052]	[0.083]	[0.089]
Treatment weak x Female	-0.086	-0.010	-0.277**
	[0.084]	[0.116]	[0.129]
Treatment female x Female	-0.090	-0.090	-0.217
	[0.082]	[0.103]	[0.136]
Treatment weak x Treatment female x Female	0.098	-0.068	0.528***
	[0.111]	[0.145]	[0.194]
White	0.030	-0.040	0.035
	[0.035]	[0.038]	[0.056]
Age	0.003**	0.002	0.004**
	[0.001]	[0.002]	[0.002]
Income	0.000	0.000	-0.000
	[0.000]	[0.000]	[0.000]
Employed	-0.013	-0.070*	-0.074
	[0.040]	[0.042]	[0.063]
Education = 2	0.434	0.355	0.265
	[0.264]	[0.358]	[0.202]
Education = 3	0.223	0.232	0.535**
	[0.313]	[0.379]	[0.235]
Education = 4	0.423	0.285	0.299
	[0.271]	[0.359]	[0.215]
Education = 5	0.507*	0.396	0.430**
	[0.262]	[0.353]	[0.192]
Education = 6	0.431	0.201	0.237
	[0.266]	[0.361]	[0.208]

Table 12 Continued

Education = 7	0.452*	0.274	0.398**
	[0.261]	[0.354]	[0.182]
Education = 8	0.463*	0.304	0.394**
	[0.263]	[0.355]	[0.189]
Education = 9	0.286	0.071	0.296
	[0.298]	[0.382]	[0.242]
Education = 10	0.485*	0.535	0.400
	[0.263]	[0.363]	[0.305]
Married	-0.015	0.058	0.027
	[0.036]	[0.040]	[0.053]
Republican	0.004	0.030	-0.013
	[0.032]	[0.039]	[0.052]
City Setting = 2	-0.084**	-0.077	-0.063
	[0.038]	[0.048]	[0.077]
City Setting = 3	-0.081**	0.000	-0.039
	[0.034]	[0.048]	[0.078]
City Setting = 4	-0.074	-0.096	0.008
	[0.053]	[0.059]	[0.086]
Constant	0.381	0.489	0.094
	[0.271]	[0.368]	[0.220]
Observations	413	406	403
R-squared	0.097	0.094	0.071
F-statistic	1.708	2.213	1.425
Female + Weak x Female	0.0983	0.169	-0.226
Weak + Weak x Female	0.0991	0.0767	-0.110
Weak + Female + Weak x Female	0.106	0.168	-0.00148
Full Interaction Effects	0.121	0.110	0.160

Note. Robust standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

The coefficient for weak language corresponds to hypothesis 2.A. and shows that a male requester using weak language is 22.4 percent more likely to achieve task performance from followers than a male requester using strong language, keeping everything else constant. This is a very large effect that allows me to reject the null hypothesis that both language styles achieve the same performance with a p-value of less than 1 percent. Furthermore, when evaluating hypothesis 2.B., I find that there is a 11 percent difference in performance between a female requester using weak language and a female requester using strong language, in favor of the latter, that is jointly significant at less than 1 percent according to Table 10. Combined, these findings are intriguing, as they suggest that a man can better influence the decision to perform by using weak language, while a woman benefits from using strong language. In other words, these results seem to indicate that incongruity between gender roles and language style could lead to increases in productivity for both male and female leaders.

Hypothesis 3 is tested by adding the two treatments and their interaction. There is a very small, yet significant, negative effect that shows that a male requester using strong language achieves higher performance than a female requester using weak language. Paired with the previous findings from hypothesis 2, this result suggests that, while incongruency seems to be more efficient than congruency in increasing productivity, men would benefit more than women from a leadership style that matches their perceived gender role. Finally, Hypothesis 4 is tested using the Full Interaction Effects, which is the sum of the treatment effects and the double and triple interactions $(\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 + \gamma_5 + \gamma_6 + \gamma_7)$. This effect shows that a female participant who receives a request from a woman using weak language is 16 percent more likely to perform than a male participant who receives a request from a man using strong language. However, this effect is not jointly significant and I cannot reject the null hypothesis, suggesting that, in this case, male and female followers do not react differently to the framing.

Table 13

Linear restriction F-Test, Model 2 (p-values)

	Experiment 1	Experiment 2	Experiment 3
Female, Weak x Female	0.542287	0.935354	0.033886
Weak, Weak x Female	0.546957	0.670689	0.004535
Female, Weak, Weak x Female	0.823441	0.428671	0.01143
Full Interaction Effects	0.205217	0.218003	0.075973

To test the robustness of the results presented above, Probit analysis is performed to measure the consistency of results across both models. To simplify, Table 11 shows the results of the analysis for the treatment effects corresponding to the first equation and Table 7, and omits the results for the control variables. The Probit estimates indicate that only the interaction term has a significant effect at less than 1%. These results are generally consistent with the Linear Probability Model that estimated equation 1, as it fails to identify any significant effect.

Table 14

Treatment effects for equation 1. Probit Analysis

	(1)	(2)	(3)
VARIABLES	Experiment 1	Experiment 2	Experiment 3
Treatment weak = 1	-0.211	-0.023	0.280
	[0.230]	[0.233]	[0.178]
Treatment female = 1	-0.232	0.238	0.020
	[0.237]	[0.228]	[0.181]
Treatment weak x Treatment female	1.373***	0.283	-0.293
	[0.467]	[0.335]	[0.268]
Observations	413	406	403

Note. Robust standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Table 12 shows the results of the Probit analysis for the second equation and Table 9, also omitting the results for the control variables. This estimation method seems to give different results for the interaction between female requester and weak language in experiment 1, which is now significant at 10%; however, the low significance does not suggest inconsistency with the linear probability model. Experiment 3 also has similar results in direction and significance of the coefficients. The magnitude, however, is hard to interpret in these cases and requires marginal effects. In general, Tables 11 and 12 are consistent with the Linear Probability Model.

Table 15

Treatment effects for equation 2. Probit Analysis

	(1)	(2)	(3)
VARIABLES	Experiment 1	Experiment 2	Experiment 3
Treatment weak = 1	-0.018	-0.060	0.657***
	[0.291]	[0.284]	[0.251]
Treatment female = 1	-0.036	0.312	0.291
	[0.319]	[0.286]	[0.249]
Treatment weak x Treatment female	0.919*	0.538	-0.940***
	[0.550]	[0.436]	[0.359]
Female = 1	0.641*	0.385	0.364
	[0.360]	[0.389]	[0.244]
Treatment weak x Female	-0.582	0.064	-0.822**
	[0.485]	[0.502]	[0.369]
Treatment female x Female	-0.581	-0.239	-0.611*
	[0.488]	[0.500]	[0.367]
Treatment weak x Treatment female x Female		-0.525	1.509***
		[0.684]	[0.547]
Observations	369	378	403

Note. Robust standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Several intuitions behind these results must be pointed out. First, the inconsistencies in results between estimates from model 1 and model 2 could be related to omitted variable bias. As the literature points out, framing varies across gender, so omitting the triple interaction could be misleading because it neglects to acknowledge that effect. In other words, the combined effects of the positive correlation between framing and respondents' sex and the negative correlation between framing and female performance could generate a downward bias. The F-test for joint significance, which sheds light on model selection by comparing a general model with a nested specification, does not provide any conclusive evidence, especially for the third experimental design, so that bias cannot be rejected. However, in both models we find no evidence of a significant difference in framing across genders.

Following the non-parametric test results, the small differences in performance across treatment groups cautions that the treatment effects might be too small given the sample sizes used in every experiment. Before this study, there was no evidence on the expected effect of the treatments, which made it difficult to estimate an effective sample size. However, after the experiments I can observe that the differences in mean performance range from 3 to 7.5 percentage points. For instance, in the third experiment, the mean performance for the control group is 0.68 while the standard deviation is approximately 0.46, which is the smallest standard deviation across experiments. Given that standard deviation and the largest possible effect, power analysis suggests that a 5% significance level with 90% power requires over 1,720 observations to identify treatment effects, while interactions may require an even larger sample size. Unfortunately, all three experiments in this study fail to have enough power. Additionally, due to the limited resources of this project, the need to change the experimental design twice to address the lack of variance in performance prevented any possible increase in the sample size of a given experiment.

Lastly, and related to the previous point, the consistent failure to obtain any results in experiments 1 and 2 is also related to the lack of variability that stems from the experimental design. In the context of the MTurk platform, an offer to perform a task for $0.10, regardless of its characteristics, seems to more than compensate for the participants' opportunity cost of finding alternative work. Consequently, when over ninety percent of the sample is choosing to perform, it cannot be accurately concluded that the treatment is the cause of such decision for the marginal respondent –the one that is indifferent to the reward and will be affected by the treatment. A different experimental design is required to properly measure the treatment effects in a setting with such a high rate of participation. Experiment 3 was an attempt to achieve that by explicitly removing the additional reward offered to perform. However, while this change reduces the positive participation bias caused by the monetary reward, it also changes the context of the analysis. Therefore, findings from experiment 3 may be more descriptive of altruistic behavior, where performance reflects the willingness to help the researcher via a selfless or kind act, which is often inconsistent with managerial settings.

Chapter 5. Conclusions

The main goal of this study was to analyze differences in task performance of participants produced by gender and communication style. Three experimental designs were posted in MTurk and two linear probability models are used to evaluate the research hypotheses. In addition, a Probit model was used to replicate the results in order to test for robustness. To control for heteroskedasticity, all results use robust standard errors, as it is traditional in the economic literature.

Also, non-parametric tests across experiments suggest that there is no significant difference in the performance response rate across the first and second experiments, which offer or are perceived to offer a monetary reward, while the response rate in the third experiment is indeed different. This result is very important because it is related to the biases produced by the experimental design. In particular, the high rates of task performance in the first two experiments bias coefficients downward and explain why there are no significant differences found in any of the treatments or interactions. Furthermore, in all three experiments the study finds no evidence of gender differences in the participants' response to the framing, contrary to results that are well established in the literature.

In regard to language, while results are not consistent across experiments and specifications, the study finds a robust but very small difference in performance between men using strong language and women using weak language. However, larger significant effects are found to support the notion that incongruency between gender roles and language style could increase productivity for both male and female leaders. Males using weak language achieve better performance by followers than males using strong language, while female leaders benefit more from using strong language. These results suggest that, while gender by itself does not explain differences in productivity, a contribution of this study is to provide initial evidence of the gendered effects of language in influencing behavior in the context of productivity and performance. These results are not entirely consistent with the framework proposed by Weissman (2011a, 2011b), who would expect weak language to achieve no results. In that sense, these findings suggest that further scholarly examination is needed to accurately qualify the interactions between gender roles and language style.

Since the third experiment explicitly warns participants that there is no reward for completing the additional task, the caveat to these findings is that, once respondents know that there is no incentive to participate, the interpretation of the treatment effects changes, as task performance no longer measures the willingness of participants to perform a compensated job, but rather their willingness to perform an act that is not rewarded. This is uncommon in organizations and firms, although it may be applicable to the performance of in-kind services for one’s organization (going beyond one’s job description). In that sense, the influence of gender and language on altruistic acts could have some applicability to corporate settings, but it would be unwise to generalize these findings to understand leader-follower relationships in firms. Furthermore, this interpretation in an altruistic context could help explain why weak language is more effective to achieve performance among male requesters, although it does not offer insights into why female requesters are more effective when using strong language.

Three limitations of this study must be highlighted. First, results were affected by changes in the experimental design, but these changes were a consequence of the lack of information regarding the opportunity costs of participants in MTurk experiments. An appropriate experimental design will require participants to be indifferent between performing or not, such that the treatment effects alone can explain the performance decision. Therefore, rewards must be large enough to incentivize performance, yet not large enough to guarantee that all participants will perform regardless of their individual preferences. As online experiments become more frequently used in the literature, this opens opportunities for further research.

A second limitation relates to the sample size, as two issues arose that led to having too small power to identify the expected effects. First, sample size was reduced when the experiment was redesigned three times in order to address the low variance in the dependent variable. Second, the literature does not have previously identified estimates of gender and language effects in performance that would have allowed to conduct a power analysis prior to establishing the size of the samples. In this regard, this study contributes initial expected effects to the literature, which can be used as a benchmark to estimate the effective sample sizes required in follow up experiments to achieve significant results.

Finally, a third important limitation of this study is the changes in the interpretation of the treatment effects caused by the explicit modification of the reward structure in the third experimental design. These changes limit the ability to interpret these results in the context of managerial leadership and performance; hence, the existence of a monetary reward is needed to maintain the broad applicability of the results to the business environment.

In terms of future research, there are opportunities to continue the inquiry into the influence of language and communication style on performance, a topic that has not yet been fully explored in the economic literature, theoretically or empirically. Additionally, there is potential to replicate this research design in workplaces and organizations or in an experiment that links performance to effort, to evaluate these findings in real business contexts.

References

Adams, R. B., & Funk, P. (2012). Beyond the glass ceiling: Does gender matter? *Management Science, 58*(2), 219-235.

Bartholomae, S., Kiss, D. E., Jurgenson, J. B., O'Neill, B., Worthy, S. L., & Kim, J. (2019). Framing the human capital investment decision: Examining gender bias in student loan borrowing. *Journal of Family and Economic Issues*, *40*(1), 132-145.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351-368.

Bertrand, M., Goldin, C., & Katz, L. F. (2010). Dynamics of the gender gap for young professionals in the financial and corporate sectors. *American Economic Journal: Applied Economics*, *2*(3), 228-255.

Bertrand, M., & Hallock, K. F. (2001). The gender gap in top corporate jobs. *ILR Review*, *55*(1), 3-21.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, *94*(4), 991-1013.

Bruckmüller, S., Ryan, M. K., Rink, F., & Haslam, S. A. (2014). Beyond the glass ceiling: The glass cliff and its lessons for organizational policy. *Social Issues and Policy Review*, *8*(1), 202-232.

Bruckmüller, S., Hegarty, P., & Abele, A. E. (2012). Framing gender differences: Linguistic normativity affects perceptions of power and gender stereotypes. *European Journal of Social Psychology*, *42*(2), 210-218.

Bruckmüller, S., & Abele, A. E. (2010). Comparison focus in intergroup comparisons: Who we compare to whom influences who we see as powerful and agentic. *Personality and Social Psychology Bulletin, 36*(10), 1424-1435.

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

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*(3), 573-598.

Ellingsen, T., Johannesson, M., Mollerstrom, J., & Munkhammar, S. (2013). Gender differences in social framing effects. *Economic Letters*, *118*(3), 470-472.

Emami, A. (2017). Gender risk preference in entrepreneurial opportunity: Evidence from Iran. *International Journal of Entrepreneurship and Small Business*, *30*(2), 147-169.

Groeneveld, S., Bakker, V., & Schmidt, E. (2020). Breaking the glass ceiling, but facing a glass cliff? The role of organizational decline in women's representation in leadership positions in Dutch civil service organizations. *Public Administration*, *98*(2), 441-464.

Grossman, P. J., Eckel, C., Komai, M., & Zhan, W. (2019). It pays to be a man: Rewards for leaders in a coordination game. *Journal of Economic Behavior & Organization*, *161*(C), 197-215.

Grossman, P. J., Komai, M., & Jensen, J. E. (2015). Leadership and gender in groups: An experiment. *Canadian Journal of Economics*, *48*(1), 368-388.

Guvenen, F., Kaplan, G., & Song, J. (2014). *The glass ceiling and the paper floor: Gender differences among top earners, 1981-2012* (No. 20560). National Bureau of Economic Research.

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., &Bigham, J. P. (2018). *A data-driven analysis of workers' earnings on Amazon Mechanical Turk, proceedings of the CHI Conference on Human Factors in Computing Systems Paper No. 449, Montréal, 2018*. New York, NY: Association for Computing Machinery.

Hasseldine, J., & Hite, P. A. (2003). Framing, gender and tax compliance. *Journal of Economic Psychology*, *24*(4), 517-533.

Huff, C., & Tingley, D. (2015). "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research and Politics*, *2*(3), 1-12.

Huang, Y., & Wang, L. (2010). Sex differences in framing effects across task domain. *Personality and Individual Differences*, *48*(5), 649-653.

Jian-jun, Z., Ji-ping, Y., Dan-hui, Z., & Hong-yun, L. (2011, September). The effects of frame, gender, and task difficulty on individual crisis decision-making. *2011 International Conference on Management Science & Engineering 18th Annual Conference Proceedings* (pp. 1245-1252). IEEE.

Joshi, A., Son, J., & Roh, H. (2015). When can women close the gap? A meta-analytic test of sex differences in performance and rewards. *Academy of Management Journal*, *58*(5), 1516-1545.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy*, *98*(6), 1325-1348.

Keltner, D., Gruenfeld, D. H., & Anderson, C. (2003). Power, approach, and inhibition. *Psychological Review, 110*(2), 265-284.

Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes*, *76*(2), 149-188.

Levin, I. P., Jasper, J. D., & Gaeth, G. J. (1996). Measuring the effects of framing country-of-origin information: A process tracing approach. *Advances in Consumer Research, 23,* 385-389.

Miller, D. T., Taylor, B., & Buck, M. L. (1991). Gender gaps: Who needs to be explained? *Journal of Personality and Social Psychology, 61*(1), 5-12.

Marcus, B. (2011, December 9). Do you sabotage yourself by using weak language? *Forbes*. Retrieved from https://www.forbes.com/sites/bonniemarcus/2011/12/09/do-you-sabotage-yourself-by-using-weak-language/#11cf8f2a1987.

Pande, R., & Ford, D. (2014). *Gender quotas and female leadership: A review*. World Development Report on Gender. Cambridge, MA: Harvard University.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker.* Cambridge, UK: Cambridge University Press.

Peng, J., Li, H., Miao, D., Feng, X., & Xiao, W. (2013) Five different types of framing effects in medical situation: a preliminary exploration. *Iran Red Crescent Med J.*, *15*(2), 161-165.

Ruggles S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., & Sobek, M. (2020). IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN. https://doi.org/10.18128/D010.V10.0.

Sacavem, A., Martinez, L. F., da Cunha, J. V., Abreu, A. M., & Johnson, S. K. (2017). Charismatic leadership: A study on delivery styles, mood, and performance. *Journal of Leadership Studies*, *11*(3), 21-38.

Simon, B., & Oakes, P. (2006). Beyond dependence: An identity approach to social power and domination. *Human Relations*, *59*(1), 105-139.

Sisense (2018). *What baby names tell us about ethnic and gender trends*. Retrieved from https://cdn.sisense.com/wp-content/uploads/What-Baby-Names-Tell-Us-About-Ethnic-and-Gender-Trends.pdf.

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. New York, NY: Penguin Group.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453-458.

Van Engen, M. L., & Willemsen, T. M. (2004). Sex and leadership styles: A meta-analysis of research published in the 1990s. *Psychological reports*, *94*(1), 3-18.

Weissman, J. (2011a, December 7). Replace meaningless words with meaningful ones. *Harvard Business Review*. Retrieved from https://hbr.org/2011/12/replace-meaningless-words-with.

Weissman, J. (2011b, September 14). Never ask “does that make sense?” *Harvard Business Review*. Retrieved from https://hbr.org/2011/09/never-ask-does-that-make-sense.

Appendix A: Survey

Implied Informed Consent

You are invited to participate in a brief research study being conducted as a requirement for the Master of Science degree at St. Cloud State University.

Background Information and Purpose

The purpose of this study is to collect demographic information from participants to better understand human behavior.

Procedures

If you decide to participate, you will be asked to complete a 10-question demographic survey, which is completely anonymous so no one will be able to identify a specific individual's responses.

Risks

There are no foreseeable risks associated with participation in this study.

Benefits

If you choose to participate, you will be compensated $0.10 through the Amazon Mechanical Turk platform. Additionally, it is my hope that the information gained in this study will help me advance current knowledge in the social and behavioral sciences.

Confidentiality

All data collected in this study will remain anonymous and the results will only be reported in aggregated form. Your information will be confidential and no answers that could identify you individually will be used.

Research Results

If you are interested in learning about the results of this study, feel free to contact the researcher at ajplachejo@stcloudstate.edu.

Contact Information

If you have additional questions about the study or your participation in it, please contact the researcher at ajplachejo@stcloudstate.edu.

Voluntary Participation/Withdrawal

Participating in this study is completely voluntary. If you decide to fill out the survey and there are any questions that you are not comfortable answering, you do not need to answer them. If you decide to participate, you are free to withdraw at any time without penalty.

Acceptance to Participate

Your completion of the survey indicates that you are at least 18 years of age and you consent to participate in the study.

Demographic Survey

1. What is your race? Mark (X) one or more boxes.

[ ] White

[ ] Hispanic

[ ] Black or African American

[ ] American Indian or Alaska Native

[ ] Asian

[ ] Middle Eastern of North African

[ ] Native Hawaiian or Pacific Islander

[ ] Other

2. What is your sex?

[ ] Male

[ ] Female

3. What is your age?

4. What is the highest degree or level of school you have completed? Mark ONE box. If currently enrolled, mark the previous grade or highest degree received.

[ ] Less than High School -- NO DIPLOMA

[ ] Regular high school diploma

[ ] GED or alternative credential COLLEGE OR SOME COLLEGE

[ ] Some college credit, but less than 1 year of college credit

[ ] 1 or more years of college credit, no degree

[ ] Associate's degree (for example: AA, AS)

[ ] Bachelor's degree (for example: BA, BS)

[ ] Master's degree (for example: MA, MS, MEng, MEd, MSW, MBA)

[ ] Professional degree beyond a bachelor's degree (for example: MD, DDS, DVM, LLB, JD)

[ ] Doctorate degree (for example: PhD, EdD)

5. LAST WEEK, did you work for pay at a job (or business)?

[ ] Yes

[ ] No - Did not work (or retired)

6. What was your total income during the past 12 months or income declared in your last tax fillings? (In US$ Dollars)

7. What is your marital status?

[ ] Now married or cohabitation

[ ] Widowed

[ ] Divorced

[ ] Separated

[ ] Never married

8. Generally speaking, do you usually think of yourself as a Republican, a Democrat, or something else?

[ ] Republican

[ ] Democrat

[ ] Other [Explain]

9. In which state do you currently reside?

10. In which setting do you currently reside

[ ] Small urban area (less than 100,000 people)

[ ] Large urban area (100,000 people or more)

[ ] Suburban area

[ ] Rural area

Experiment 1. With monetary compensation.

Treatment: Strong language

Thank you for completing the demographic survey. This study is important to advance knowledge in the social sciences. I am convinced that your participation will make a difference. I want to offer you the opportunity to complete an additional task for a bonus payment of $0.10.

My experience as a researcher makes me confident that your participation will impact my findings. If you choose not to do it, you will still receive compensation for the demographic survey. [Nathan Johnson/Michelle Johnson].

Will you complete the additional task?

[ ] Yes

[ ] No

Treatment: Weak language

Thank you for completing the demographic survey. I think this study may possibly advance knowledge in the social sciences. I think your participation will make a difference (let us hope so). I want to offer you the opportunity to complete just one additional task for a bonus payment of $0.10. As a researcher, I feel that your participation may somewhat impact my findings (keeping my fingers crossed!). If you choose not to do it, you will still receive compensation for the demographic survey. I hope this makes sense. [Nathan Johnson/Michelle Johnson].

Will you complete the additional task?

[ ] Yes

[ ] No

Experiment 2. Without monetary compensation.

Treatment: Strong language

Thank you for completing the demographic survey. This study is important to advance knowledge in the social sciences. I am convinced that your participation will make a difference. I want to offer you the opportunity to complete an additional task. My experience as a researcher makes me confident that your participation will impact my findings. If you choose not to do it, you will still receive compensation for the demographic survey. [Nathan Johnson/Michelle Johnson].

Will you complete the additional task?

[ ] Yes

[ ] No

Treatment: Weak language

Thank you for completing the demographic survey. I think this study may possibly advance knowledge in the social sciences. I think your participation will make a difference (let us hope so). I want to offer you the opportunity to complete just one additional task. As a researcher, I feel that your participation may somewhat impact my findings (keeping my fingers crossed!). If you choose not to do it, you will still receive compensation for the demographic survey. I hope this makes sense. [Nathan Johnson/Michelle Johnson].

Will you complete the additional task?

[ ] Yes

[ ] No

Experiment 3. Without monetary compensation (explicit).

Treatment: Strong language

Thank you for completing the demographic survey. This study is important to advance knowledge in the social sciences. I am convinced that your participation will make a difference. I want to offer you the opportunity to complete an additional task. NO ADDITIONAL PAYMENT WILL BE PROVIDED. My experience as a researcher makes me confident that your participation will impact my findings. If you choose not to do it, you will still receive compensation for the demographic survey. [Nathan Johnson/Michelle Johnson].

Will you complete the additional task?

[ ] Yes

[ ] No

Treatment: Weak language

Thank you for completing the demographic survey. I think this study may possibly advance knowledge in the social sciences. I think your participation will make a difference (let us hope so). I want to offer you the opportunity to complete just one additional task. NO ADDITIONAL PAYMENT WILL BE PROVIDED. As a researcher, I feel that your participation may somewhat impact my findings (keeping my fingers crossed!). If you choose not to do it, you will still receive compensation for the demographic survey. I hope this makes sense. [Nathan Johnson/Michelle Johnson].

Will you complete the additional task?

[ ] Yes

[ ] No

Post-Survey Task: Why did you choose to complete or not to complete the additional task?