GENDER AND THE EVALUATION OF JOB APPLICANTS
IN NATURAL SETTINGS

A Thesis
Presented to the Faculty of the Graduate School
of Cornell University
in Partial Fulfillment of the Requirements for the Degree of
Masters of Arts

by
Esther Quintero
January 2007
ABSTRACT

Researchers have pointed out that occupational sex segregation will not be fully understood until detailed data are collected on natural (Biernat & Fuegen 2001) and non-scripted interactional settings (Ridgeway & Correll 2004) which are central to gender and employment (Ridgeway 1997). This paper makes a direct contribution to this goal by applying experimentally established theory to a natural environment that involves the live interaction of applicants and evaluators who are recruiting them for jobs.

Gender status theories argue that gender systematically shapes the way men and women are perceived in evaluative and task oriented contexts when gender differentiates actors in the setting (Correll & Ridgeway 2003; Foschi 2000). Specifically, Status Characteristics Theory (SCT) predicts that, in such settings, men will have an advantage over women because individuals hold higher performance expectations for men than for women (Berger et al. 1977; Foschi 1989.) In other words, because there are broadly shared cultural beliefs implying that men are better at the things that count, specific men will also appear more skilled than specific and equally competent women. If these mechanisms are at play in hiring settings, the implication is that employers will be more likely to hire the male applicant even when the female applicant is equally qualified.

Although hiring contexts are almost never accessible to researchers, this project identifies and takes advantage of a unique setting that (1) permits direct observation and data collection on real hiring decisions made in the course of direct interaction and (2) meets the scope conditions of SCT.
The context of this study is the Spanish exam system that is used to recruit candidates to fill important government jobs. Women are currently underrepresented in these positions filling only about 30% of them. Exams to become a government employee in Spain are public and involve the face to face interaction of evaluators and job applicants. Applicants go through a series of qualifying testing rounds; those who succeed at all stages are automatically hired. This setting is exceptional in that: (1) is accessible for direct observation and data collection, (2) the event of interest (i.e. exam) repeats sufficiently so as to evaluate theory-driven claims statistically, and (3) exams are fairly structured, which deems the lack of strict controls less problematic.

I examine quantitative pass/fail exam data and information gathered from direct observation of exam sessions. I use SCT and draw from (and extend) Ridgeway’s ideas about gender and social interaction to make the following predictions. First, following SCT, I predict that (1) male (female) applicants will be advantaged (i.e. pass at higher rates) in exams involving skills typically perceived as neutral (feminine). Second, I extend Ridgeway’s discussion of gender and interaction, by arguing that (2) the degree of applicant-evaluator interaction will shape the magnitude of prediction one. In other words, larger differences in passing rates between male and female applicants will be observed at exams involving a greater degree of applicant-evaluator interaction than at those exams characterized by minimal interaction.

Both hypotheses were confirmed empirically thereby suggesting that (a) SCT appropriately explains the outcome of interest, and (b) the degree of interaction shapes the size of SCT’s predictions. The results of my work suggest that the mechanisms discovered in controlled environments are at
play in actual hiring settings. Second, my results also suggest that it may be useful to conceptualize interaction as a continuous measure shaping the size of SCT's predictions.
BIOGRAPHICAL SKETCH

Esther Quintero received a B.A. degree in History from the University of Seville (Spain) in 2001, and is currently a Ph.D. candidate in the Department of Sociology at Cornell University. She has several graduate awards from Cornell University, most notably the Luigi Einaudi Dissertation Fellowship (2005) and The Sage Fellowship (2006) as well as a Graduate Fellowship from the University of Seville (2001).

Esther Quintero has served as a teaching assistant at Cornell University and has collaborated with the Indian Institute of Management (Bangalore, India).
ACKNOWLEDGMENTS

This research was made possible with support from Cornell University Institute of European Studies (Luigi Einaudi Fellowship 2005-2006). I wish to thank Shelley Correll and Stephen Benard for their useful suggestions.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIOGRAPHICAL SKETCH</td>
<td>iii</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>CHAPTER ONE: INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>CHAPTER TWO: PRIOR RESEARCH</td>
<td>6</td>
</tr>
<tr>
<td>CHAPTER THREE: THEORY</td>
<td>9</td>
</tr>
<tr>
<td>CHAPTER FOUR: BACKGROUND</td>
<td>13</td>
</tr>
<tr>
<td>CHAPTER FIVE: CONTEXT AND HYPOTHESES</td>
<td>21</td>
</tr>
<tr>
<td>CHAPTER SIX: RESULTS AND DISCUSSION</td>
<td>31</td>
</tr>
<tr>
<td>CHAPTER SEVEN: SUMMARY AND CONCLUSION</td>
<td>37</td>
</tr>
<tr>
<td>WORKS CITED</td>
<td>41</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table Number</th>
<th>Table Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>ISSP Survey Results on Gender Attitudes</td>
<td>16</td>
</tr>
<tr>
<td>Table 2</td>
<td>Percentage of Women in Group A</td>
<td>18</td>
</tr>
<tr>
<td>Table 3</td>
<td>Percentage of Women by Pay Level in Group A</td>
<td>19</td>
</tr>
<tr>
<td>Table 4</td>
<td>Description of SCA Exams</td>
<td>27</td>
</tr>
<tr>
<td>Table 5</td>
<td>Exam Classification and Empirical Predictions</td>
<td>29</td>
</tr>
<tr>
<td>Table 6</td>
<td>Logistic Regression Results</td>
<td>34</td>
</tr>
<tr>
<td>Table 7</td>
<td>Predicted Probabilities of Passing</td>
<td>35</td>
</tr>
</tbody>
</table>
CHAPTER ONE
INTRODUCTION

Occupational sex segregation persists despite major changes in the composition of the labor force and in the legal environment (e.g. Bielby & Baron 1986; Jacobs 1989; Reskin 1993). While survey data are useful to substantiate macro level patterns of job segregation, experiments have identified some of the mechanisms that explain how gender often shapes the perception of competence in favor of men (Berger et al. 1977). Nonetheless, scholars have argued that the full dynamics of gender discrimination in the workplace will not be fully understood until detailed data are collected on natural (Biernat & Fuegen 2001) and less scripted interactional settings (Ridgeway & Correll 2004). Interactional settings are crucial to understand gender-based employment decisions because individuals automatically and unconsciously sex-categorize others to whom they must relate (Brewer and Lui 1989; Stangor et al. 1992). Interactionally driven sex-categorization triggers the activation and use of gender status beliefs encouraging actors to view men as more status worthy than women. As a result, men are typically granted a series of subtle privileges in a wide range of situations, including hiring.

Even though most employment decisions take place in the context of live interactions between employers and workers, existing studies fail to look at its impact. Indeed, nearly all studies of gender segregation begin with data on people who have already been hired, and very little empirical evidence has been provided on the workings of actual hiring processes. These research gaps exist, at least partially, because it is difficult to gain access to data and
settings that permit alternative and more in-depth approaches. This study identifies and takes advantage of a natural context where job applicants interact with employers who are evaluating them for jobs. Furthermore, I examine a recruitment process consisting of various rounds of testing that are much the same but, importantly, they differ in the degree of interaction between evaluators and candidates. As mentioned above, Ridgeway (1997) has argued that mixed-sex interaction will prompt sex-categorization which will in turn activate the use of gender status beliefs to guide attitudes and behavior. This study advances Ridgeway’s contribution by proposing to conceptualize interaction as a continuous measure. I argue that the degree of interaction will impact the extent to which status beliefs will be used to guide thought and action. In other words, evaluative settings where applicants and employers interact more will disadvantage female applicants more than settings where interaction is minimal. Assessing this claim would require comparing decisions made in contexts that only differ in degree of interaction. The setting I describe in the following paragraphs provides an exceptional opportunity to explore this proposition.

The context of this study is the Spanish exam system that is used to recruit candidates to fill important government jobs. Exams to become a government employee are public in Spain, and involve the face-to-face interaction of evaluators and job applicants. I focus on a specific competition that takes place annually and involves four qualifying rounds of testing (i.e. exams). Exams are highly similar but differ in the degree of applicant-evaluator interaction. I examine applicants’ pass/fail rates at each testing round for a period of three years (i.e. three competitions). My main prediction is that female applicants will experience greater disadvantages (i.e. will fail at higher
rates) in exams involving a higher level of interaction than in exams where interaction is minimal.

Given the interactive context, the salience of gender, and the importance of the task, I use 'status characteristics' theory (SCT) to explain why men are preferred or evaluated more positively than women. SCT has been developed and established mostly through controlled laboratory experiments. Although this study uses a non-experimental design, it is worth noting that the context I examine is sufficiently controlled in an unintended way. First, applicants do not generally differ in characteristics such as level of education, age, ethnicity, or nationality, which would complicate the interpretation of results. Variation among applicants is mostly limited to the characteristic of interest, namely their gender and that of their evaluators. Second, testing rounds are very similar (i.e. location, evaluating committees, study guide etc.) with the key exception that they do vary in the degree of applicant-evaluator interaction. Third, because testing rounds involve a large number of applicants, the same event (i.e. exam) repeats sufficiently so as to allow the collection of multiple instances of comparable cases suited to statistical analysis. Finally, evaluators do not have access to applicants’ CVs or any kind of additional information that might affect their assessment of performance at the actual exams. In fact, evaluators know very little about applicants (e.g. name, date of birth) precisely to facilitate objective evaluations. In sum, although the setting I selected has limitations, it is not very far distant from the kinds of conditions one would want to recreate in a controlled environment, with the important exception of real world consequences.
Social psychology has identified some mechanisms by which generalized cultural beliefs about gender disadvantage women whenever gender is salient and related to the task at hand (for reviews see Correll & Ridgeway 2003; Foschi 2000). I use experimentally established theory to understand hiring outcomes in a natural context that involves the direct interaction of job applicants and employers who are selecting them for jobs. Ridgeway (1997) and others have argued that our understanding of gender inequalities in employment will be enhanced by looking at the interaction-based processes that produce and perpetuate the observed inequalities (i.e. gender segregation). Ridgeway’s main argument is that interaction pushes actors to sex-categorize and thus use hegemonic gender beliefs to guide what they think and do. Following a non-experimental strategy (see Cohen 1982), this study contributes to expanding Ridgeway’s (1997) argument by comparing the same cohort of applicants across comparable testing rounds that differ in the degree of face-to-face interaction between evaluators and job applicants. I predict that exams involving greater levels of interaction will activate gender beliefs more strongly. As a result, gender differences in passing rates will be larger in more interactive exams than in exams where interaction is minimal.

In the next sections I review the empirical literature on sex segregation and evaluation biases against women. Next, I lay out the theoretical framework that guides this work. In the background section I first situate gender attitudes and beliefs in Spain in comparative perspective, and then describe the more specific context of the study and rehearse my empirical predictions. Third, in the results section I evaluate these predictions by examining pass/fail rates by gender using exam data on a specific civil service
competition. Finally, I discuss and summarize the main findings of the study and describe future work.
CHAPTER TWO
PRIOR RESEARCH

Sex segregation refers to the unequal distribution of men and women in industries, occupations, or jobs. Occupational segregation is a major source of labor market rigidity and economic inefficiency (Anker 1997). In addition, the segregation of men and women into different occupations produces a broad range of workplace inequalities. First, occupational segregation is a major cause of the gender gap in wages, benefits, and retirement income (Perman & Stevens 1989; Reskin & Hartman 1986). Second, female-typed jobs offer fewer promotion (Steinberg et al 1990) and on-the-job training opportunities (Bielby & Baron 1986; Farkas et al 1991). Feminized jobs are also associated with greater vulnerability to repeated unemployment (Reskin & Hartmann 1986). Finally, sex segregation not only reflects hegemonic gender beliefs but also it contributes to perpetuate them (Ridgeway 1997).

Although the proportion of men and women in the labor force is approaching parity, survey data suggest that the level of segregation is still high (for a review see Anker 1997; Reskin 1993). Even though a decrease of sex segregation was documented in the 1970s and 1980s, case studies of occupations in which women registered the greatest representational gains indicate that occupational desegregation did not ensure job-level integration (Reskin & Roos 1990). Segregation indexes computed across broad occupational categories miss substantial within category segregation and understate the extent of job-level segregation (Reskin 1988). In short, the workplace continues to be highly differentiated by sex; about forty percent of men or women would have to change major occupational categories to
achieve equal representation of men and women in all jobs (Reskin & Roos 1990).

One necessary condition for gender segregation is the existence of gender-based selection decisions in organizations. Thus, gender segregation, at least partially, is the result of aggregate individual-level gender-based judgments. Studies using posthire data (i.e. data on individuals already hired) cannot adequately identify prehire sorting mechanisms. In contrast, social psychology has made important contributions to explaining why employers might prefer males over females to fill the best jobs. Although cognitive approaches do not directly analyze job segregation, they examine processes that have clear implications for it. A review of this work will be presented below.

Laboratory experiments have demonstrated that equally competent performance by men and women is perceived as more indicative of skill and ability in men than in women (Deaux & Emswiller 1974). The use of different standards to evaluate men and women’s competence (henceforth ‘double standards’) has been confirmed in a variety of laboratory settings (for a review see Foschi 2000). For instance, in an experiment Foschi and colleagues (1994) recreated features of a hiring decision that involved the examination of files of fictitious applicants for professional jobs. Subjects had to make a recommendation about hiring the male or the female applicant. Although female subjects did not display the use of double standards when evaluating applicants’ resumes, the results from male subjects indicated that the male applicant with the slightly better record was chosen more often than the female applicant when she was in the same position (Foschi, Lai, & Siegerson 1994). Other studies suggest that both men and women rate the quality of men’s
work higher than that of women when they are aware of the sex of the person to be evaluated, but not when the same person’s gender is unknown (O’Leary & Wallston 1982).

In an audit study, Steinpreis and associates examined whether faculty would be influenced by the gender of the name on a CV in determining hireability and tenurability. Fictitious CVs were submitted to real academics. Both male and female faculty were significantly more likely to hire a potential male colleague than an equally qualified potential female colleague. In addition, both male and female faculty were more likely to positively evaluate the research, teaching, and service contributions of male applicants than of female applicants with identical records (Steinpreis, Anders, & Ritzke 1999). Taken together these studies show that gender status processes mediate how men and women are perceived. If at play in actual hiring settings, the implication is that real employers draw on generalized beliefs about women’s lower status when assessing female applicants, perceive them as less competent as a result, and ultimately end up hiring a disproportionately high number of males.

In closing, almost all research on gender segregation focuses on (1) data about people who have already been hired\(^1\), or (2) mechanisms that, if/when at play in real hiring settings, would explain employers’ preference for male hires even when male and female applicants are equally qualified.

---

\(^1\) For an exception see Fernández & Sosa 2005.
CHAPTER THREE
THEORY

This study uses status characteristics theory (SCT) to explain evaluation biases in a natural setting where applicants interact with employers who are evaluating them for jobs. The theoretical claim to be evaluated here is that individuals face disadvantages in evaluation contexts (e.g. job interviews) when (1) their personal attributes (e.g. gender) have attached to them negative connotations broadly shared in the culture, and (2) such attributes are salient or differentiate actors in the setting (e.g. mixed-sex groups).

Attributes such as race or gender for which there are consensually shared cultural beliefs are “status characteristics.” More formally, status characteristics are categorical distinctions among people; different instantiations of such distinctions (e.g. for gender, male/female) have attached to them hegemonic beliefs associating greater status and competence to one category of the distinction (e.g. male) than others (Berger et al. 1977). Gender status beliefs are a component gender stereotypes and reflect a cultural system representing what we think “most people” accept as true about men and women. Although stereotypes contain status beliefs, they also encompass notions unrelated to status such as the kinds of traits, attributes, or behaviors that can be or should be expected of men and women (Deaux & Kite 1987). By contrast, gender status beliefs are beliefs that men are more socially valued and diffusely more competent at the things that matter most (Wagner & Berger 1997). Because status beliefs function as cultural schemas (Ridgeway 1997), even individuals who do not personally endorse the content of the beliefs are likely to be aware of their existence and thus have their judgment
and behavior affected by them (Foschi 1996; Lovaglia et al. 1998; Steele 1997).

Gender is a diffuse status characteristic; that is, gender beliefs include expectations that men are generally more competent than women at most things, as well as specific assumptions that men are particularly better than women some tasks (e.g. mechanical tasks) while women are better than men at others (e.g. nurturing tasks) (Conway, Pizzamiglio, & Mount 1996; Wagner and Berger 1997; Williams & Best 1990). Even though beliefs about gender have evolved, empirical studies continue to find that men are thought to be generally more capable and competent than women (Williams & Best 1990).

No status characteristic disadvantages actors in all settings. A status characteristic becomes salient when it distinguishes actors in a setting (e.g. mixed-sex groups) or when the characteristic is perceived to be related to the task. In other words, the impact of gender beliefs is highly dependent on the structure of the context and ranges from imperceptible to substantial (Ridgeway & Smith-Lovin 1987). For example, men are thought to be more competent than women, except when the skills to be evaluated are perceived as feminine (Conway, Pizzamiglio, & Mount 1996; Wagner & Berger 1997; Williams & Best 1990). Meta-analyses of gender differences in specific skills such math and verbal ability have established that differences between men and women’s actual performance are negligible or non-existent (see Hyde et al. 1990 for mathematical ability, and Hyde & Linn 1988 for verbal ability). However there seems to be a consensus that women possess greater verbal skills than men (Hyde & Linn 1998) and that men are more skilled than women at mathematics. Correll (2001) found that females make higher self assessments of their verbal ability thereby proving that males do not globally
assess their task competence higher than females, regardless of the gender association of the task. Instead, cultural beliefs (which need not reflect de facto gender differences) associated with a particular task bias individuals perceptions of their abilities and the abilities of others (Correll 2001).

Salient status characteristics have been shown to impact attitudes and behavior in collective and task oriented settings. Later advances of the theory have demonstrated that status characteristics shape behavior in a broader range of social contexts than originally specified by the theory’s scope conditions. In fact, status characteristics matter in “social relational contexts” (Correll & Ridgeway 2004:511) or contexts where actors are compelled to anticipate (thus form “performance expectations”) their own behavior relative to others and/or the behavior of others (Lovaglia et al. 1998; Steele 1997; Foschi, Lai & Sigerson 1994; Correll 2004). When gender is salient in social relational and task oriented settings, SCT predicts that men will have an advantage over women insofar as they will be expected to perform better (Berger et al. 1977; Foschi 1989.) Higher performance expectations for men lead to some main theoretical predictions; relative to women, men: (1) will be given more opportunities to participate or make a contribution, (2) will have their mistakes judged by more lenient standards, and (3) will have their performances evaluated more positively, and (4) will command more influence over other group members.

Although a setting need not involve direct interaction for SCT’s predictions to operate, Ridgeway (1997:231) has argued that relating to “a concrete other” is sufficient to trigger gender status processes. Ridgeway (1997) understands social interaction as a complex phenomenon that requires to be simplified before it can be coordinated. Simplification begins to occur
when individuals develop a minimal definition of who “self” and “other” are in a given context; preliminary definitions are reached by contrasting self and other on dimensions where similarities and differences are perceived to exist. Empirical evidence demonstrates that sex serves as a primary categorization system in Western society (Fiske 1992) and that individuals automatically and unconsciously sex categorize any specific other to whom they relate (Brewer & Lui 1989; Stangor et al. 1992). Subsequent categorizations such as occupational roles become nested in gender (Brewer & Lui 1989), taking on slightly different meanings as a result. Most importantly, sex categorization prompts the use of gender stereotypes (including status beliefs) to guide attitudes and behavior (Blair & Banaji 1996).

Status beliefs cued by sex categorization can range from vague cognitive backgrounds to powerful determinants of actors’ expectations and behavior (Ridgeway 1997). I will argue that contexts involving greater levels of direct interaction will encourage actors to use gender status beliefs to a greater extent. Thus, while SCT’s predictions will work regardless of the level of interaction, I argue that the magnitude of these predictions might be greater when actors interact more naturally than when they do so in more scripted ways.
CHAPTER FOUR
BACKGROUND

This work draws from theory based on the fact that widely held beliefs about gender exist and are used constantly to organize social relations. Current hegemonic gender beliefs accord men greater status worthiness than women and individuals draw from this constantly in their assessment of social situations and behavior. This way, shared beliefs about gender translate into tangible hurdles and disadvantages for women in areas as crucial as education and employment, which in turn contributes to perpetuate and create the inequalities we observe. Although the process described is not a closed one (i.e. existing gender inequalities foment the persistence and renewal of hegemonic beliefs favoring the social perception of men), it does originate in deeply established and shared cultural beliefs about men and women. Generalized beliefs about gender are likely to vary in content across time and cultures. Importantly though, these beliefs will generally confer men and the things men do a greater status relative to women and the things women do. Nonetheless, it is relevant to examine how gender attitudes and beliefs in Spain may or may not differ from gender attitudes in the USA and other European countries.

In the following paragraphs, I will examine and discuss results suggesting that gender beliefs are much the same in the Spain and the rest of Europe/USA. Next, I will describe the specific context of this study, namely the characteristics of civil service recruitment systems in Spain. Finally, I will elaborate on the specific competition I examined and present the study’s hypotheses.
I analyzed data from the International Social Survey Programme (ISSP), which is a collaboration between different nations (a total of 41) conducting surveys about topics of ample interest for social science research. The results of the surveys provide a cross-national and cross-cultural perspective to individual national studies. I examined the 2002 module on gender attitudes to situate Spain in the broader picture with regards to gender beliefs. What are gender beliefs like in Spain relative to the USA and other European countries? I examined a set of attitudinal survey items that will provide a general picture of what gender beliefs are like in Spain relative to the USA and rest of Europe\(^2\).

A total of 2,471 respondents from Spain and 19,309 from Europe and the US combined participated in a survey about gender attitudes in 2002. Respondents were asked to rate a series of statements on 1 to 5 point scales where 1= “strongly agree” and 5= “strongly disagree”. Below I present some descriptive results concerning respondents’ attitudes toward men and women’s perceived roles and preferences.

About 32% (N=785) of Spaniards believe that working mothers cannot have a warm relationship with their children. About 21% (N=3,964) of respondents shared this view in other European countries and the USA combined. Similarly, about 52% (N=1,253) of Spanish respondents agreed or strongly agreed that children suffer if their mother works outside the home. About 43% (N=7,984) of respondents in Europe and the US answered likewise. In the same vein, 54% (N=1,331) of respondents in Spain agree or

\(^2\) Europe includes EU15 member states (except Greece, Belgium, and Luxemburg, which were not surveyed for gender attitudes in 2002), Norway, and Switzerland.
strongly agree that family life suffers if women work outside the home. About 41% (N=7,711) of respondents in Europe and the US answered the same.

Regarding the social perception of women’s preferences and aspirations, 42% (N=991) of Spanish respondents affirmed that what women “really want” is to stay home and take care of their children. About 34% (N=6,073) of US and European respondents shared the same views. Roughly 24% (N=601) of Spaniards agree that men’s job is outside the home and women’s job is in the household. In the USA and Europe combined, 19% (N=2,977) of respondents think likewise.

Table 1 summarizes some of the above results. Percentages represent the proportion of respondents who “strongly agreed” or “agreed” with the statements on the left hand column (i.e. “Children suffer if mother works”). In addition, Table 1 displays the mean values for both Spanish and USA/European respondents. All differences found were statistically significant. Higher values indicate less agreement with the statements on the left (i.e. 1= “strongly agree” and 5 = “strongly disagree”). As can be seen in Table 1, Spanish respondents seem somewhat more traditional in their gender beliefs than respondents in the rest of Europe and the USA.
Table 1 ISSP Survey Results on Gender Attitudes in Spain and the United States and Europe, 2002

<table>
<thead>
<tr>
<th>Strongly Agree/Agree That...</th>
<th>Spain</th>
<th>US/Europe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children suffer if mother works</td>
<td>52%(^1)</td>
<td>43%</td>
</tr>
<tr>
<td>N</td>
<td>1,253</td>
<td>7,984</td>
</tr>
<tr>
<td>Mean(^2)</td>
<td>2.84</td>
<td>3***</td>
</tr>
<tr>
<td>Family life suffers if women work</td>
<td>54%</td>
<td>41%</td>
</tr>
<tr>
<td>N</td>
<td>1,331</td>
<td>7,711</td>
</tr>
<tr>
<td>Mean</td>
<td>2.75</td>
<td>3.05***</td>
</tr>
<tr>
<td>What women &quot;really want&quot; is home/children</td>
<td>42%</td>
<td>34%</td>
</tr>
<tr>
<td>N</td>
<td>984</td>
<td>6,073</td>
</tr>
<tr>
<td>Mean</td>
<td>3.07</td>
<td>3.2***</td>
</tr>
<tr>
<td>Women’s place is the household</td>
<td>24%</td>
<td>19%</td>
</tr>
<tr>
<td>N</td>
<td>601</td>
<td>2,977</td>
</tr>
<tr>
<td>Mean</td>
<td>3.61</td>
<td>3.72***</td>
</tr>
</tbody>
</table>

\(^1\) Percentage of respondents who answered “strongly agree” or “agree”.

\(^2\) Where 1= strongly agree and 5= strongly disagree.

Note: *** significant at 0.5%

In sum, the results above suggest that gender beliefs are much the same in Europe and the US. If anything, gender attitudes in Spain appear slightly less progressive than in the US and the other European countries. To the extent that evaluators in Spain are aware that these beliefs exist, they will subconsciously draw from them to orient their attitudes and behavior even if they do not personally endorse the content of these beliefs.
The first challenge for researchers interested in the study of hiring practices is to gain access to actual data. Virtually all contexts where job applicants are evaluated for jobs are restricted to outsiders. This study takes advantage of a unique exception, and identifies a hiring setting where access is permitted making it feasible to observe and collect data on actual evaluations and hiring decisions. In Spain, exams to become a government employee are public and consist of oral exams that involve the face to face interaction of job applicants and judges who are recruiting them for important government positions.

A fixed number of government vacancies are announced annually in Spain for specific jobs. Access to these positions is first determined by applicants’ educational attainment. For example, individuals holding a 5-year college degree$^3$ may opt to compete for the top jobs or positions within Group A, the top group that is. The actual competition and selection procedures involve passing a series of qualifying exams that usually involve face to face evaluations of prospective job applicants by employers or civil service judges. After a brief probationary period, the highest scoring examinees automatically become permanent government employees. The focus here is on top-level jobs within the civil service; jobs that are regarded as socially prestigious occupations and where women are clearly underrepresented. Although some government jobs are perceived to be predominantly female (i.e. administrative jobs), the best positions in the public sector are largely filled by men. In Spain, specific civil service jobs fall into one of 5 broad categories: Groups A, B, C, D, and E. Group A jobs are the best paying and most prestigious, while Groups D

---

$^3$ The equivalent of a B.A. degree in the United States.
and E jobs require less skill and pay less well. As Table 2 shows, in 2003 women filled about 34% of Group A jobs.

Table 2 Percentage of Women in Group A Civil Service in Spain by Year

<table>
<thead>
<tr>
<th>Year</th>
<th>% Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>29.5</td>
</tr>
<tr>
<td>1997</td>
<td>29.7</td>
</tr>
<tr>
<td>1998</td>
<td>30.4</td>
</tr>
<tr>
<td>1999</td>
<td>30.9</td>
</tr>
<tr>
<td>2000</td>
<td>31.9</td>
</tr>
<tr>
<td>2001</td>
<td>32.4</td>
</tr>
<tr>
<td>2002</td>
<td>33.4</td>
</tr>
<tr>
<td>2003</td>
<td>34.1</td>
</tr>
</tbody>
</table>


Further inequalities can be detected if Group A aggregate figures are broken down by pay level. As Table 2 shows, the higher the salary the smaller the proportion of women.
Table 3 Percentage of Women by Pay Level in Group A Civil Service in Spain

<table>
<thead>
<tr>
<th>Level</th>
<th>2003</th>
<th>2002</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>29 (High)</td>
<td>2034</td>
<td>23.3</td>
</tr>
<tr>
<td>28</td>
<td>5744</td>
<td>30.3</td>
</tr>
<tr>
<td>27</td>
<td>2732</td>
<td>35.9</td>
</tr>
<tr>
<td>26 (Low)</td>
<td>4564</td>
<td>37.8</td>
</tr>
</tbody>
</table>


In Spain public employment means above anything else work stability. According to a report by ANECA\textsuperscript{4}, about 45% of university students in Spain report they intend pursue a career in government upon graduation. While it is natural for both men and women to pursue the best paying and higher status jobs, the Spanish labor market has a number of characteristics that make the absence of women in top public employment positions the more puzzling. First, Spain has one of the highest female unemployment rates of the EU-15\textsuperscript{5} - i.e. 12.2%, 5 points above the unemployment rate for males. Second, low job security disproportionately affects women in Spain; of workers with part-time contracts, 78% are women. Similarly, 60% of workers with indefinite work

\textsuperscript{4} Agencia Nacional de Evaluación de la Calidad y Acreditación: agency created in 2002 to design and implement quality controls of higher education systems in Spain.

\textsuperscript{5} Includes member countries in the European Union prior to the accession of ten candidate countries on May 2004.

The EU15 comprised the following 15 countries: Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden, United Kingdom.
contracts are men. These figures indicate that Spanish female workers often have a more precarious employment situation than that of male workers. Thus, government jobs represent one of women’s best employment alternatives inasmuch as they guarantee work stability (i.e. public jobs are jobs for life) and permit the conciliation of work and family. Finally, university graduates may compete for Group A jobs which offer an entry level salary of about 36,000 € per year, which is above the average income of workers with the same level of education working in the private sector (i.e. 33,000 € per year). In sum, these are good jobs in terms of pay, work conditions, and other benefits. Women account for 60% to 65% of the initial applicant pool for many Group A competitions, which further confirms that women do indeed want these jobs. If so, why is it that female applicants are failing exams that would lead to their recruitment?

---

CHAPTER 5
CONTEXT AND HYPOTHESES

This study focuses on a specific group, namely State Civil Administrators (SCA). SCA are officials with broad administrative knowledge and ample responsibility in areas such as budget, human resources, and contracts. SCA are in charge of designing and drafting top-level government proposals, as well handling public policy. They hold management-level positions in different government offices and sub-offices, and European Union agencies. The highest ranking SCA work closely with politicians (i.e. in political cabinets) and many become prestigious and influential politicians themselves (Crespo 2004:62).

The SCA recruitment process was selected based on several considerations. First, data from exploratory interviews suggested that SCA recruitment is perceived to be unbiased. In-depth interviews were conducted with 12 applicants (58% women) in May 2005 and multiple informal interviews were conducted with candidates during September-December 2005. Interviewees often referenced other competitions such as those to enter the diplomatic corps as examples of exams where evaluators explicitly show a strong bias against female candidates (e.g. judges ask female applicants (but not male applicants) personal questions about how they will conciliate family and career if they were recruited etc). Gender status theories do not presume employers are explicitly sexist or prejudiced; rather, these theories argue that, under certain conditions, gender beliefs implicitly affect the perception of competence in a way that usually (but not always) disadvantages women. If evidence for the theory’s predictions is found here, it would confirm the
theory’s adequacy to explaining the particulars of this setting. Second, although this is not a controlled setting in the strictest sense, it is a natural environment that closely resembles some of the characteristics one would want recreate in the laboratory. For example, examinees do not generally differ in their ethnicity, nationality, level of education or age. Since gender is the characteristic of interest here, variation in any of the above would complicate the interpretation of results. Likewise, all exams take place under similar conditions (physical location, format of the exam etc), which allows for repeated observation of the same event with gender (of applicant and judges) being the one aspect that varies. Third, the selection of applicants is not done by an individual evaluator but rather by committees of five judges. About 20 judges participate in SCA recruitment every year; in addition, judges vary from one year to the next. If decisions were made by a single evaluator, it would be hard to rule out the possibility that the outcomes observed are simply the reflection of one person’s preferences. Finally, this study examines processes at various rounds of testing; the range of ability among applicants is muted after round one since the lowest performing applicants have presumably been eliminated from the competition.

In the following paragraphs I will offer a detailed description of the SCA recruitment process. Interview and observational data were crucial to gain an in-depth understanding of the SCA hiring process and make fine distinctions among exams. An examination of the official regulations and procedures of SCA recruitment would have proven insufficient to establish actual differences among testing rounds since, on paper, exams look much more alike than they actually are. Learning about and being able to establish real differences between exams was crucial to formulate a set of the study’s hypotheses. I will
first provide the official exam description and then supplement it with information gathered via direct observation. Finally, I will explain the discrepancies between official and actual procedures using information gathered from personal interviews with exam judges.

SCA applicants must go through a total of 4 qualifying exams that take place throughout the year in Madrid. Examining boards are usually composed of 5 members whose votes are independent and have equal value regardless of rank. All exam sessions are public although observers are asked to leave the exam premises when judges deliberate and assign a score which is made public shortly after. All four testing rounds involve the face to face interaction of evaluators and applicants. However, as pointed out earlier in the paper, exams differ in the level of applicant-judge interaction. Level or degree is not understood here as frequency but rather more or less interaction is accorded as a function of qualitative aspects that define such interactions – i.e. scripted v. non-scripted. I will elaborate on this point later.

Exams 1, 2, and 4 are largely written exams although the exercises are not read nor evaluated by individual judges. Applicants themselves read their exams out loud to an evaluating committee (henceforth “reading session”) on a specific day and time assigned to them at random. Exam 3 is different from all others because it is purely oral; applicants are given one hour to verbally rehearse four questions drawn at random from a study guide composed of 180 questions.

In exam 1 applicants are given 4 hours to write a general knowledge essay that will later be read in front of the evaluating committee on a randomly assigned date. Although essay questions are broad and may be approached in different ways, applicants are specifically required to relate their answers to
the contents of the official study guide provided to them. Exam 2 is a foreign language test where candidates are evaluated on their translation and listening comprehension skills. In exam 4 applicants are given 4 hours to resolve several practical cases and are required to justify their answers. Applicants may consult their books and materials when writing exam 4. Exams 1, 2, and 4 are similar in format. First, all three involve a written part and, then a public reading session. Second, exams 1, 2, and 4 have a similar duration, between 20 and 30 minutes. In exams 2 and 4, evaluators have 15 minutes to engage in a dialogue with applicants and ask them exam-related questions. I will refer to this part as Q&A portion. Exam 1 does not include a Q&A portion.

Exam 3 is purely oral; applicants are summoned to exam 3 at a specific time/date based on a first random draw. In exam 3 applicants are given 1 hour to answer 4 questions drawn at random from questions in the study guide. Applicants have 20 minutes to write out an outline that is used to aid their later 1 hour uninterrupted performance. No study materials may be consulted in any part of exam 3. Evaluators have 15 minutes to ask exam-related questions at the end.

In sum, all four exams involve the face to face interaction of applicants and evaluators. All but exam 1 include a dialogue or Q&A portion at the end. Although this is what official procedures call for, the actual exam sessions either digress from the official version and/or are characterized by features only direct observation would permit identify. In the next paragraphs I will describe this in greater detail and will explain why differences exist.

I conducted direct observation of exam sessions in May 2005 for exam 1, and from September 2005 to December 2005 for exams 2, 3, and 4. I observed and collected data on about 8% of applicants in exam 1, 40% in
exam 2, and 85% in exam 3 and 70% in exam 4. Although nothing new was learned about exam 1, observational data provided important insights about all other 3 exams. First, about 60% of applicants were not asked any questions at the end of exam 4 in the 2005 SCA competition. Second, when applicants were asked questions the Q&A rarely lasted more than 5 minutes on average (SD=2.7 minutes) and seldom involved the participation of more than one judge. In contrast, the Q&A portion of exam 3 was systematically used, involved all or most judges, and lasted an average of 13 minutes.

The Q&A in exam 2 was slightly different from that of exams 3 and 4. All applicants were asked questions, however, the nature of questions was different from the nature of questions in exams 3 and 4. As mentioned earlier, exam 2 is a foreign language test and questions were intended to provide opportunities for applicants to express themselves in a language other than Spanish. Questions in exam 2 were drawn from a list of questions evaluators had previously come up with; thus, questions repeated frequently and were somewhat scripted insofar as their sole purpose was to encourage examinees to talk. Questions in exam 2 were generally posed by one of the evaluators, a language expert, brought specifically for assisting in exam 2.

In January 2006 I interviewed two judges, a man and a woman, who participated in the 2005 SCA recruitment process. In-depth interviews lasted about 90 minutes and consisted of a series of open-ended questions about the recruitment process. Interviews included questions such as evaluators’ perception of exams (i.e. which exams seem more/less difficult to evaluate and why?), how decisions are made when judges deliberate, and about the inconsistencies between official exam procedures and those that were
followed in actual exams. The judges I interviewed provided similar explanations to the latter question.

Evaluators tacitly agree not to ask questions at the end of exam 4 because of several reasons. First, questions are intended to measure reflexes, and spontaneity. Applicants typically consult what they wrote at exam 4 with their personal trainers and prepare in advance answers to possible questions they may be asked at the actual oral exam. This explanation was confirmed by the applicants I interviewed in May 2005. Therefore, questions asked at exam 4 do not necessarily measure spontaneity or what applicants knew/wrote at exam 4, but rather the extent to which applicants prepared for potential questions that could come up in the reading session of exam 4. In addition, applicants are summoned to exam 4 reading session on different days assigned at random. Therefore, applicants summoned earlier will have less time to prepare than those who do the public reading later. The judges I interviewed regard this as an unfair disadvantage and thus most evaluating committees informally agree to not ask questions at exam 4. This information also suggests that the Q&A portion is generally viewed as an opportunity for judges to gauge applicants’ natural responses and reaction to questions. When the ability to measure this is undermined because questions may have been prepared beforehand, questions are no longer important.

In sum, observational and interview data helped provide a very detailed description of the exam process, one that would not have been gained by looking at the written procedures alone. Observational data in particular were crucial to establish differences in the degree of applicant-judge interaction at each of the exams. As I explained above, all exams involve face to face interactions. One would think that exam 3 is different from all others because
candidates must rehearse their arguments verbally rather than just read a piece they have previously written like in exams 1, 2, and 4. Direct observation, however, made it clear that exam 3 is more similar to the others than one would intuitively think based on the official description. Due to the highly memory-oriented nature of exam 3 and the strict time constraints, applicants’ rehearsal of answers ends up being highly scripted, fast-paced, and matter-of-fact. Thus, applicants’ speech patterns are in no way more natural than in the reading sessions of exams 1, 2, and 4. However, the most important differences between exams are (a) the presence/absence of questions at the end and (b) the nature of the questions asked. Table 4 summarizes some key characteristics of exams.

Table 4 Description of SCA Exams

<table>
<thead>
<tr>
<th>Start Date</th>
<th>Process Duration (Weeks)</th>
<th>Exam Duration (Mins.)</th>
<th>Type</th>
<th>Q&amp;A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exam 1</td>
<td>June</td>
<td>8</td>
<td>30</td>
<td>Written Read</td>
</tr>
<tr>
<td>Exam 2</td>
<td>Sept</td>
<td>4</td>
<td>20</td>
<td>Written Read</td>
</tr>
<tr>
<td>Exam 3</td>
<td>Oct</td>
<td>6</td>
<td>95</td>
<td>Oral</td>
</tr>
<tr>
<td>Exam 4</td>
<td>Dec</td>
<td>2</td>
<td>30</td>
<td>Written Read</td>
</tr>
</tbody>
</table>

In light of the findings and description above exams were classified according to the level of applicant-judge interaction (from less to more): exam 1, exam 4, exam 2, and exam 3. All exam 1 sessions and most exam 4
sessions do not include a Q&A portion. Second, although both exam 2 and 3 have a Q&A part at the end, the questions asked in exam 2 are more scripted and are intended to encourage applicants to speak. The kind of applicant-judge dialogue in exam 2 is therefore not completely natural (i.e. judges, for instance, refrain from interrupting or participating in the conversation). Conversely, the Q&A portion of exam 3 constitutes a truly natural discussion: most judges ask questions, the sessions last longer, the questions respond to the specific performance of each given applicant and so forth. As a result, a more natural discussion ensues in the Q&A part of exam 3. Second, exams were classified according to the gender typing of the abilities evaluated. Exams 1, 3, and 4 involve skills that are not stereotypically perceived as masculine (e.g. computer/mathematical expertise) nor feminine (e.g. nurturing skills). Exam 2 does involve abilities that are stereotypically perceived as female, language skills. Table 5 summarizes the above classification as well as the empirical predictions I will rehearse in the next paragraphs.
Recall that SCT predicts that when gender is salient in task oriented settings, men will be evaluated more positively than women so long as the skills being assessed are perceived to be masculine or neutral.

Hypothesis 1: the group advantaged at each exam will vary as a function of the gender typing of the task to be evaluated.

Hypothesis 1a: male applicants will score higher or pass at higher rates than female applicants when the skills evaluated at a given exam are perceived as masculine or neutral, that is at exams 1, 3, and 4.

Hypothesis 1b: female applicants will score higher or pass at higher rates than male applicants when the skills evaluated at a given exam are perceived as feminine, that is at exam 2.

Second, I have argued throughout the paper that the degree of applicant-judge interaction at exams will impact the magnitude of H1 predictions. If so:
Hypothesis 2: the magnitude of gender differences in scores or passing rates will be affected by the degree of interaction that characterizes a given exam.

Hypothesis 2a: small differences will be observed between male and female’s passing rates in exams characterized by minimal applicant-evaluator interaction (i.e. exams 1 and 4).

Hypothesis 2b: larger differences will be detected between male and female’s passing rates in exams involving a greater degree of applicant-evaluator interaction (i.e. exams 2 and 3).
CHAPTER SIX
RESULTS AND DISCUSSION

I gathered and examined data for 1,476 SCA applicants (514 men and 955 women) who participated in SCA competitions between 2003 and 2005. Since 2003 the results of major civil service competitions are posted in PDF documents at official government websites. The information that is made public via the Internet includes applicants’ names, their personal identification number\(^7\), and exam results at each testing round. In addition, these records contain the date and order in which candidates take exams. Thus the variables of interest can be grouped as follows. First, measures about applicants’ characteristics (e.g. name, gender, and SSN). Second, applicants’ scores at all four exams (candidates who fail do not receive a numeric score). Finally, variables relating to the relative order in which applicants go through testing rounds (order is always assigned based on a first random draw).

All exams involve the physical presence of evaluators and applicants but exams 1 and 4 are significantly more scripted than 2 and 3 insofar as they lack a Q&A part, which is the exam portion where applicants and judges engage in a true dialogue. The non-Q&A part of all exams is fairly rigid; even that of exam 3.

These are multilevel data with three levels: candidate, exam, and year. I use logistic regression to evaluate the hypothesis that the degree of interaction mediates the effect of gender pass and fail outcomes. More specifically, I argue that the exams involving a greater level of applicant-judge interaction will disadvantage female applicants more than less interactive

\(^7\) The Spanish equivalent of the Social Security Number, a unique identifier issued by the government to each Spanish citizen.
exams. Recall that SCT does not predict that women will always be evaluated less positively relative to men. Rather, SCT argues that men will be perceived as diffusely more competent than equally qualified women so long as the task evaluated is not one where women are stereotypically thought to be better at (e.g. verbal skills). There are reasons to believe that in Spain, women are typically thought to be better at verbal skills in general and foreign languages in particular. For example, about 75% of university students majoring in English, French, Spanish, and Italian philology are women. Similarly, more than 80% of college students majoring in translation and interpretation are women. This suggests that generalized beliefs exist in Spain regarding women’s superior ability at foreign languages. It is then important to show that female applicants are not disadvantaged at all exams. Thus, I expect that the main effect of sex (female=1) on the dependent variable pass will not be significant. I run the following logistic regression model:

\[ y = a x_1 + e \]

where \( y \) is the dependent variable pass and \( x_1 \) is applicant’s sex (1=female). In this and all subsequent models the data were clustered by applicant ID to control for the fact some candidates reenter the competition after having failed in previous years and, as such, observations are not independent. As predicted, the logistic regression coefficient for the main effect of sex is not significant (results not shown) thereby proving that being female does not invariably have a negative effect on the likelihood of passing an exam.

By contrast, SCT would predict that male applicants will pass a greater rates than female applicants across all exams except exam 2 (i.e. foreign language):

\[ y = a + b x_1 + c x_2 + d x_1 x_2 + e \]
where $y$ is the dependent variable pass, $x_1$ is applicant’s sex (1=female), $x_2$ is exam2. If SCT’s predictions apply here, sex will have a negative main effect on pass but the interaction effect of sex and pass will be positive. In this model, sex has a negative main effect on pass but, when interacted with exam2 the effect is positive (results not shown). This suggest that female applicants have a higher probability of passing exam 2 but not others. The predicted probabilities of passing exam 2 are 72% for male applicants and 83% for female applicants. Similarly, the predicted probabilities of passing other exams are 55% for male applicants and 48% for female applicants.

While SCT predictions are confirmed here, I have argued that exams are characterized by different levels of interaction between applicants and judges. If exams were ordered according to the degree interaction between applicants and evaluators from low to high we would have: exam 1, exam 4, exam 2, exam 3. As shown earlier, SCT predicts that women will score lower at exams 1, 3, and 4. If degree of interaction matters, it should be possible to explore whether the negative impact of sex on pass varies across these three exams (i.e. 1, 3, and 4). I predict that the interaction effect of sex and exam will be negative for exams 1, 3, and 4, but that the effect will be considerably greater in exam 3 than in 1 and 4. To evaluate this the following logistic regression model was rehearsed:

$$y = a + b x_1 + c x_2 + d x_3 + e x_4 + f x_1 x_2 + g x_1 x_3 + h x_1 x_4 + e$$

where $x_1$ denotes applicants’ sex and $x_2, x_3, x_4$ denote the type of exam (1, 3, and 4 respectively).
Table 6 Logistic Regression Results of Sex, Exam, and Sex*Exam on Pass (Observations Clustered by ID)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>(2.25)*</td>
<td>(2.32)*</td>
</tr>
<tr>
<td>Exam 1</td>
<td>-0.99</td>
<td>-1.13</td>
</tr>
<tr>
<td></td>
<td>(3.73)**</td>
<td>(3.52)**</td>
</tr>
<tr>
<td>Exam 3</td>
<td>-0.42</td>
<td>-0.13</td>
</tr>
<tr>
<td></td>
<td>(1.21)</td>
<td>(0.30)</td>
</tr>
<tr>
<td>Exam 4</td>
<td>0.51</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>(1.09)</td>
<td>(0.58)</td>
</tr>
<tr>
<td>Exam1*Female</td>
<td>-0.73</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>(2.17)*</td>
<td>(2.17)*</td>
</tr>
<tr>
<td>Exam3*Female</td>
<td>-1.63</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>(3.86)**</td>
<td>(3.93)**</td>
</tr>
<tr>
<td>Exam4*Female</td>
<td>-0.97</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>(1.66)</td>
<td>(1.58)</td>
</tr>
<tr>
<td>Year</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Year*Exams</td>
<td></td>
<td>Yes</td>
</tr>
</tbody>
</table>

Note: N = 1372. Clusters = 424. Absolute values of Z statistics in parentheses. * Significant at 5% ** Significant at 1%

Because I argue that the negative effect of female is not invariant across exams 1, 3, and 4, the model also includes the interaction of sex and the exam dummies. As Table 6 (column 1) shows, there is a positive main effect of sex on pass because exam 2 is the baseline comparison (i.e. women are evaluated more positively in exam 2). While all interaction terms logistic regression coefficients have a negative impact on pass as expected, the size of the effect varies considerably across exams as predicted.

--

8 Exam 1 = 0.67 - 0.73 = - 0.06, Exam 4 = 0.67 - 0.96 = - 0.29, Exam 3 = 0.67 - 1.63 = - 0.96
As Table 7 shows, male applicants are somewhat more likely to pass exams 1 and 4 relative to female applicants. Although statistically significant, differences between male and female applicants passing rates at exams 1 and 4 are small. Recall that exam 3 is the only exam involving a truly natural interaction between examinees and judges. As Table 9 shows, gender differences in passing rates in exam 3 are substantial. In exam 3 male applicants are 1.5 times more likely to pass than female applicants. Finally, female applicants enjoy a moderate advantage over male applicants in exam 2 which is more interactive than exams 1 and 4 but involves the assessment of stereotypically female abilities (e.g. verbal skills).

In addition, the model in Table 6 (Column 2) controls for year’s fixed effects and the interaction effect of years and exams. None of these variables coefficients were statistically significant thereby suggesting that the pattern is stable across time.

Table 7 Predicted Probabilities of Passing for Men and Women Based on the Logistic Regression Model in Table 6 (Column 1)

<table>
<thead>
<tr>
<th></th>
<th>% Pass</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Men</td>
<td>Women</td>
</tr>
<tr>
<td>Exam 1</td>
<td>0.45</td>
<td>0.44</td>
</tr>
<tr>
<td>Exam 2</td>
<td>0.72</td>
<td>0.84</td>
</tr>
<tr>
<td>Exam 3</td>
<td>0.71</td>
<td>0.48</td>
</tr>
<tr>
<td>Exam 4</td>
<td>0.79</td>
<td>0.75</td>
</tr>
</tbody>
</table>
In sum, the analysis above suggest that (1) as SCT predicts being female does not always impact pass outcomes negatively; rather, the advantaged or disadvantaged group will vary as a function of the gender typing of the abilities evaluated, and (2) classifying exams by their degree of applicant-judge interaction seems useful to explain variation in the magnitude of SCT’s predictions across exams.
CHAPTER SEVEN
SUMMARY AND CONCLUSIONS

This paper is part of a larger research agenda whose end goal is to understand the workings of the hiring process. Specifically, this research seeks to identify the interactionally-based mechanisms that may be driving employers to recruit more men than women to fill the most desirable jobs. Empirical evidence demonstrates that sex segregation is still high and several important explanations have been proposed to explain it. However, almost all empirical research on sex segregation is based on data of people who have already been hired. Therefore, it is hard to separate out sorting mechanisms operating at the point of hiring. Laboratory experiments have made important contributions about what these sorting mechanisms might look like. This research applies experimentally established theory to a natural setting where actual applicants are being evaluated for jobs. The results of this work suggest that the mechanisms discovered in controlled environments could also be at play in actual hiring processes, at least in the specific one I examine.

SCT argues that since men are diffusely perceived as more competent and skilled than women, concrete men will also appear to do things better than equally qualified women when gender is salient and the skills to be evaluated are perceived as masculine or neutral. In this paper I have described and examined a setting that permits evaluating these claims in a natural environment. The hiring process I examined involves exams where both neutral and feminine skills are assessed. I found that female applicants do better at exams involving verbal skills, which are stereotypically viewed as
female, while men do better all other exams, which involve abilities stereotypically perceived as neutral.

The results presented here also suggest that it may be useful to conceptualize interaction as a continuous measure. Researchers have argued that the sex-categorization that takes place in interaction prompts the use of gender stereotypes to guide attitudes and behavior. I have argued here that the degree of interaction will impact the extent to which these stereotypes will be used. Specifically I have proposed that settings involving a greater degree of interaction will encourage actors to use gender status beliefs more than settings where interaction is minimal. SCT does not make specific claims about how variance in the degree of interaction might affect its predictions. In fact, SCT’s predictions have been shown to hold in a variety of non-collective settings. By considering how interactional degree might impact outcomes or predictions I do not wish to challenge the theory, but rather, exploring whether the magnitude of SCT’s predictions might be partly explained by the level of interaction present in a given situation. The hiring process I examine involves exams characterized by different degrees of interaction between evaluators and applicants. Gender differences in passing rates follow the direction predicted by SCT, namely men do better in exams that evaluate neutral skills and women do better in exams where feminine skills are assessed. In addition, I found that the magnitude of these differences in passing rates varies greatly from one exam to another as predicted. Specifically, differences in passing rates between men and women are larger in exams that involve a greater degree of interaction between evaluators and applicants.

While SCT specifies very precisely the mechanisms that lead to such outcomes, the data I presented here does not yet allow us to evaluate specific
SCT hypotheses. In order to examine whether SCT's mechanisms apply to concrete natural settings, more detailed data on the interactions of applicants and judges would be necessary. For example, SCT argues that men are generally given more opportunities to contribute to task-oriented discussions precisely because these contributions, insofar as they come from men, will be more valuable than, for example, the potential contributions of women. This specific prediction could be evaluated in the hiring setting I examine by looking at measures such as speech time allowed to male and female candidates during the Q&A portion of exams. If judges believe the responses of male applicants will be more relevant or interesting, judges will be inclined to let them talk for longer by limiting their interruptions. Similarly, if the contributions of female applicants are perceived as less valuable, judges will have less tolerance for long speeches and will therefore interrupt females more often. Future work will address this and other specific SCT predictions by examining the content of interactions in exam 3.

In sum, the results presented here have limitations that will be addressed in future work. A second set of limitations comes from the fact that I examine and natural and thus more complex environment where strict controls are not possible. However, the setting I picked is exceptional in that: (1) is accessible for direct observation and data collection, (2) the event of interest (i.e. exam) repeats sufficiently so as to evaluate theory-driven claims statistically, and (3) exams are fairly structured, which deems the lack of strict controls less problematic. This work does not attempt to test causal relationships but rather evaluate whether established causal links can be proven to be at work in this real hiring setting. Although work remains to be done, results so far indicate that SCT mechanisms are seemingly at play. A
criticism of experiments has been that it is sometimes difficult to argue that in the laboratory we observe behavior as it would happen in natural situations or that the findings of experiments may not generalize outside the laboratory.

A second long term goal of this research is to add external validity to theories and mechanisms discovered, and repeatedly verified in controlled settings. No matter how simple, natural settings are more messy and complex than controlled ones. It would be difficult to approach the study of natural settings without prior theoretical guidance or an idea of the causal relationships that may be at play. Thus, the theoretical and empirical contributions of social psychology have been crucial in my approach. To ascertain whether or not experimental findings generalize outside controlled environments is a question that may only be answered by finding and examining a carefully selected real-world setting. This is not a challenge for experimentalists but rather for those who believe that theory developed in controlled environments is essential to guide research in more chaotic environments.

Finally, while this analysis does not explicitly seek to make a theoretical contribution, it does suggest new directions for research. In particular, I have proposed that the level of interaction in a given setting may impact the magnitude of SCT's predictions. While my results suggest this may be the case, only a controlled study design will be able to establish a causal relationship unequivocally.
WORKS CITED


