GENDER DIFFERENCES IN THE LABOR MARKET: FOUR ESSAYS ON SUPPLY-SIDE DETERMINANTS AND CONSTRAINTS

kumulative Dissertation

Universität Hamburg

Fakultät für Wirtschafts- und Sozialwissenschaften

Dissertation

Zur Erlangung der Würde der Doktorin der Wirtschafts- und Sozialwissenschaften

(gemäß der Promotionsordnung vom 17. Januar 2017)

vorgelegt von

Melanie Schröder

aus Velbert

April, 2018

Vorsitzender: Prof. Thomas Siedler

Erstgutachterin: Prof. Miriam Beblo

Zweitgutachter: Prof. Hermann Ribhegge (Emeritus)

Datum der Disputation: 25.10.2018

„When economic outcomes are different for women than for men, not all of these may be due to differences in the constraints they face, the skills they possess, or the discrimination they may encounter. It is conceivable that, in similar circumstances, the economic behavior of women may differ from that of men".

(Eswaran, 2014:3)

Contents

Abbreviations

e.g.	for example
i.e.	id est
ICD	International Classification of Diseases
ISCO	International Standard Classification of Occupations
OECD	Organisation for Economic Co-Operation and Development

List of Figures

List of Tables

1 Introduction

Only recently, the Organisation for Economic Co-operation and Development (OECD) report on gender equality proclaimed “a stark call to action” to make “gender equality a reality” in OECD countries (OECD, 2017b:4). This gender equality does not only represent a human right, it is also “a keystone of a prosperous, modern economy that provides sustainable inclusive growth” (OECD, 2017b:3). Although labor market outcomes represent only one dimension of gender inequality, this facet is considered to be a very important one in terms of men’s and women’s economic well-being (Eswaran, 2014). Women have made significant progress toward gender equality in the labor market in recent decades, but labor market outcomes are still different between men and women (OECD, 2017b). Germany, the country where the empirical data for this dissertation was collected, is by no means an exception. In this dissertation I shed light on supply-side determinants and constraints of gender differences in the labor market.

1.1 In which labor market outcomes do men and women differ?

To begin with, men and women differ in their employment levels. Men are still more likely to be in paid employment than women in every OECD country (OECD, 2017b). In 2015, OECD average labor force participation rates were 63.1% for working-age women and 79.8% for working-age men. Referring to labor force participation rates in full-time equivalents, the gap is even larger: The employment rates in full-time equivalents are 51% for working-age women and 74% for working-age men in the OECD, with comparable rates for Germany (OECD, 2017a). Consequently, it is not surprising that part-time work is more prevalent in the female working population (around 26%) compared to men (9%) in the OECD (OECD, 2017a). While the share of men in part-time work is comparable to the OECD average, women’s representation in part-time work (37%) lies above the OECD average.

Moreover, the labor market is characterized by vertical and horizontal segregation by gender. Horizontal segregation refers to the concentration of men and women in different occupations. While women are underrepresented in the industry sector, they are overrepresented in the service sector (Eurostat, 2017) . Additionally, women are less likely to work in top managerial positions (vertical segregation): In Europe, about 35% of managers were female in 2014 while the share of women in managerial positions is only 22% in Germany (Eurostat, 2017).

The resulting gender gap in pay is substantial, sticky and frequently addressed by politics and society. On average, the raw gender pay gap in the European Union amounted to 16.3% in 2015, with Germany lying well above the average at 22.0% (Eurostat, 2018b).

Another fact, which has received only little attention in economics literature is that, in many European countries, women and men differ in their absences from the labor market due to sickness (Barmby et al., 2002; Mastekaasa and Melsom, 2014; Scheil-Adlung and Sandner, 2010; Spasova et al., 2016). In the European Union, women's probability of being absent from work for health reasons exceeds men's probability by 20%. In Germany, the probability of absence for women is larger than men's, too, with a gap of 8% (Eurostat, 2018a).

1.2 What are the driving forces of gender differences in labor market outcomes?

According to Azmat and Petrongolo (2014) there are three driving forces of gender differences in labor market outcomes in a broader sense: productivity, preferences and discrimination. While productivity and discrimination were in the main scope of research in the late 1990s, there is only little evidence on the role of preferences, as Altonji and Blank's (1999) work on gender equality shows. There has been a stark increase in work on gender differences in preferences since the beginning of the millennium (Bertrand, 2011; Niederle, 2016).

Employer discrimination is the only driving force on the demand side. According to the most frequently cited theory of discrimination, employers discriminate against women either because they have a taste for discrimination against women, and hiring women creates a disutility to them, irrespective of women's and men's productivity (Becker, 1957), or because they have imperfect information about women's abilities, use sex as an easily observable characteristic that is assumed to be correlated with productivity and thereby discriminate statistically (Arrow, 1972; Phelps, 1972). Discrimination is mostly measured by the portion of gender differences in pay that cannot be attributed to differences in observable characteristics, such as human capital, and thus remains unexplained. Indeed, a substantial part of the gender gap in hourly wages remains unexplained in many OECD countries (OECD, 2017b). Since this measure of discrimination is subject to criticism, newer studies try to disentangle the isolated effect of employer discrimination by using laboratory and field experiments.[1]

[1] See Blau and Kahn (2017) for a critique of the interpretation of the unexplained part and a comprehensive overview of studies on gender discrimination in the labor market.

A gender difference in productivity that represents one driving force on the supply side, is mainly based on differences in human capital accumulation and family constraints (Azmat and Petrongolo, 2014). The starting point for most arguments is a biologically based comparative advantage of women in child-bearing and child-rearing compared to men (Becker, 1991). According to this "intrinsic difference" (Becker, 1991:32), women are constrained to spend less time in market activities and more time in non-market activities compared to men. To this end, it is rational that women who plan to have children invest less in human capital compared to men. Moreover, it is not only rational to a mother: Anticipating gender discrimination in the labor market, it is rational for every woman to invest less in human capital. This reasoning accounts not only for potential gender differences in years of schooling but also for differences in on-the-job training that affect human capital and thus both productivity[2] (Blau et al., 2014) and differences in occupational choices. Given their career interruptions and fewer working hours, women are assumed to be less likely to invest in occupations that are highly rewarded in the market, such as law, medicine and engineering (Altonji and Blank, 1999).[3] However, the major and enormous progress in gender equality in OECD countries is registered in this area: The gender gap in educational attainment has been reversed, and, meanwhile, girls now obtain, on average, more schooling than boys (OECD, 2017b).

Another driving factor of gender differences in labor supply decisions – gender differences in preferences – has gained the interest of labor economists relatively recently. With the influence of behavioral economics, which detected heterogenous and non-standard preferences, representing a deviation from the standard economic model[4], reasoning about differences in preferences has entered labor economists' research agenda (Dohmen, 2014).[5] Then, with the beginning of the millennium, there has been a stark increase in the number of research papers investigating gender differences in preferences (Bertrand, 2011; Niederle, 2016). A recent paper (Shurchkov and Eckel, forthcoming) reviews the dimensions of gender differences in preferences alongside behavioral traits that are supposed to be relevant for labor market outcomes, such as risk taking, competitiveness, the propensity to negotiate and social attitudes.

[2] However, it is criticized that the relationship between earnings and labor market experiences is solely explained by an increase in human capital and not by tenure itself (Blau et al.,2014). Nevertheless, the implications for gender differences remain.

[3] However, demand-side effects on gender differences in educational choices should not be neglected (Shurchkov and Eckel, forthcoming).

[4] See DellaVigna (2009) for an excellent overview of empirical evidence on deviations of economic models.

[5] In his comment on Dohmen's article, Winter-Ebmer (2014) acknowledges that the idea of heterogeneous effects and preferences already existed before the entry of behavioral economics into the discipline of labor economics. His comment provides a more pessimistic view of the importance of behavioral economics in labor economics compared to Dohmen (2014).

As they conclude, there is indeed evidence on gender differences in these dimensions: women are, on average, less risk and competition loving, show less desire to negotiate and exhibit stronger social preferences (Shurchkov and Eckel, forthcoming).

Regarding the sources of these gender differences – inherent biological or socially constructed differences – the evidence is not clear. However, most of the surveyed studies "point to at least some role of [societal differences]" (Shurchkov and Eckel, forthcoming:3) and thus confirm the feminist economists' argument of gender differences as being socially constructed and not given by birth.[6] One framework that incorporates the idea of socially constructed gender norms leading to gender differences in preferences into an economic model has been developed by Akerlof and Kranton (2000) and is commonly referred to as "identity economics". Although several other concepts of the role of identity in decision making exist – see, for example, Bénabou and Tirole (2011) and an overview by Davis (2006) for alternative concepts – this framework seems to be the most influential one, since it gives rise to a range of empirical and theoretical papers. Moreover, this concept has been extended to explain the association between social identity and performance in an economic model (Dee, 2014). Accordingly, gender norms not only shape preferences; they also have an effect on performance and productivity – an argument that is common for psychologists but relatively new for economists.

At least, institutions matter. Formal norms, such as laws and regulations, shape the decisions of men and women. Attempts to promote gender equality can be found in all areas of public policy (OECD, 2017b): To increase women's participation in intensive and extensive margin, public childcare education and care are considered to be crucial for gender equality in the labor market. Moreover, in some countries, parental leave programs incentivize the more equal participation of both mothers and fathers in caregiving activities. Policies addressing gender segregation already operate in the pre-market phase: Several countries have programs to encourage girls and young women to enter into STEM (science, technology, engineering, and mathematics) disciplines. Although a matter under discussion (see Schmitt, 2015 for an overview), affirmative action policies, such as quotas for women in managerial positions or public leadership to reduce vertical segregation, are a reality in a majority of OECD countries. At least, two-thirds of all OECD countries have specific policies that aim to close the gender wage gap, such as antidiscrimination laws and wage transparency projects. However, there are also regulations that are suspected to reinforce gender differences in the labor market. A system of joint income taxation of married couples that is still present in Germany levies higher marginal

[6] See Eswaran (2014) for an excellent overview of the debate about nature and nurture arguments as she discusses several approaches.

tax rates for secondary earners, who are mostly women, and smaller marginal tax rates for the primary earners, who are mostly men, compared to a system of individual taxation. Thus, such a system reinforces traditional family arrangements, with women specializing in household work and creates disincentives for women's participation in the labor market (OECD, 2012b).

While the majority of outcomes – such as gender differences in participation (extensive or intensive), occupational choice and in pay – can be explained by the determinants and constraints mentioned above, the case is somewhat different for gender differences in work absenteeism. From an economist's point of view, work absenteeism is a matter of labor supply decisions, and thus an outcome of an individual's decision between work and leisure that is voluntary and can be influenced by incentives (Barmby et al., 2002). Contrary to that, epidemiological and sociological research interpret work absenteeism as a simple response to a medical condition – thus involuntary (Mastekaasa and Melsom, 2014). Distinguishing between these two concepts and thus the involuntary or voluntary character of a sick note seems to be impossible. Although one could refer to the duration of a sick note to distinguish, this argumentation is not straightforward, as Beblo and Ortlieb (2012) point out. Short-term absences could also be involuntary, and an individual can decide to return to work after several weeks of absenteeism. Therefore, there seems to be agreement on the fact that work absenteeism reflects both the health status of an individual – thus a constraint – and her labor market related health behavior (Kröger, 2017). Empirically, it has been shown that a gender gap in absenteeism is related to gender differences in participation rates (Angelov et al., 2013), women's higher care burden (e.g. Angelov et al., 2013; Beblo and Ortlieb, 2012) and occupational gender segregation (e.g. Mastekaasa and Melsom, 2014).

1.3 How does this thesis contribute to the field of research?

In this dissertation I shed light on the supply-side determinants and constraints of gender differences in the labor market and provide empirical evidence. In the research agenda of empirical investigation of gender differences in labor market decisions, clearly distinguishing each of the driving forces of labor supply decisions remains a key challenge, since not only outcomes but also explanations are interrelated (Azmat and Petrongolo, 2014). This dissertation approaches this challenge. To achieve this, most papers in this dissertation refer to the experimental approach. The experimental setting is characterized by tight control of the environment, and it enables researchers to investigate the "slippery world of preferences" (Ichino, 2014:41), isolated from other factors (Azmat and Petrongolo, 2014; Charness and

Kuhn, 2011; Croson and Gneezy, 2009). Additionally, one chapter provides empirical evidence by using administrative data: To disentangle effects, Chapter 5 presents the use of an econometric method adopted from the literature on the gender pay gap but innovative in this strand of the literature.

One strategy to isolate the determinants of interest is to create an experimental setting in a way that eliminates confounding factors. This strategy has been adopted in two chapters.

In Chapter 2, which presents joint work with Norma Burow, we observe labor supply decisions in intensive margin and the allocation of housework within heterosexual couples. In the world outside the lab, it has been shown that the outcomes are gendered: Men are more attached to the labor market, while housework remains women's work (OECD, 2012a). The reasoning points to differences in constraints that men and women face. Additionally, investigating intra-couple time-allocation decisions in the real world often implies problems of endogeneity of wages (Laczó, 2011; Triebe, 2013). We question whether the same patterns will be found in real-world data in a setting where these constraints and problems of endogeneity are absent. To achieve this, we created an experimental setting where these gender differences in constraints and factors that bias labor supply decisions, such as gendered care responsibilities and endogeneity of wages, are absent: To create this gender-neutral setting, that we called the "gender-neutral lab", we assigned roles of a primary earner with a higher piece-rate compensation and a secondary earner with a lower piece-rate compensation randomly across real heterosexual couples (married and cohabiting). Therefore, our sample entails both traditional couples with a male breadwinner and non-traditional couples with a female breadwinner, alike. We observe work-effort as a proxy for labor supply, and the allocation of an uncompensated task (that reduces time for the compensated task) serves as a proxy for housework. In such a gender-neutral lab, we cannot confirm real-world gender gaps in labor supply, nor could we confirm the general conviction that housework is women's work. The allocation of unpaid work in our experiment indeed follows an economic rationale, with opportunity costs determining couples' decisions. Moreover, women and men do not differ in terms of labor supply, on average, but when we take their marriage status into account. Married men provide more labor supply than cohabiting men, and married women less than cohabiting women, which matches real-world findings (Barg and Beblo, 2012). We conclude that this points to the stability of specialization within married couples, which seems to overcome the gender-neutral lab.

In chapter 3, which presents joint work with Miriam Beblo, Denis Beninger and Norma Burow, we observe differences in pay between men and women in an experimental setting. Real-world and experimental data (e.g., Blau and Kahn, 2017; Heinz et al., 2016; Schwieren, 2012) have already showed that women earn less than men, which can be explained by supply- and demand-side factors. The aim of our work is to isolate supply-side determinants of men's and women's earnings, such as preferences for payment schemes and productivity. To achieve this, we fully ruled out effects from the demand side by the proper design of an online experiment. Women and men were asked to perform a compensated effort task based on a payment scheme (contract) that they had chosen before. Women and men were offered the same contracts to choose from, one of which was then drawn randomly, and they were asked to perform the same task. The design ensures that we can fully rule out favoritism by the employer. In this experimental setting, we observe a surprisingly high gender pay gap in view of the fact that this gap is solely supply driven. A decomposition of this experimental gender pay gap reveals that most of the gap can be attributed to the fact that women and men differ in their preference for characteristics of a payment scheme and thus select into different contracts. Interestingly, we find that women and men differ not only in preferences for risk, which has already been shown by the literature, but also in preferences for accepting losses, thus loss aversion, representing a psychological trait that has not yet been in the scope of the research agenda of the gender pay gap. On the productivity side, we find that, while controlling for potential productivity, a substantial part of the gap can be attributed to the fact that women and men differ in performance and thus productivity conditional on a given payment scheme. In particular, we find women to underperform men although they face the same incentives and ability is controlled for. We therefore refer to the psychological explanation of a stereotype threat that harms women's productivity which has as yet received only little attention by economists.

Another strategy that is able not only to eliminate confounding factors but also to derive causal evidence is the introduction of an experimental variation and thus treatment. An experimental variation to identify causal effects is used in two chapters.

In chapter 2, which presents joint work with Norma Burow, we further introduced an experimental variation in the gender-neutral lab to investigate the effect of different taxation systems. In particular, we investigate how men and women react to a change in taxation schemes and compare joint taxation – which is present in Germany although it is criticized for harming gender equality – with a system of individual taxation. Commonly, researchers use natural experiments on changes in tax schemes to document the impact of individual compared to joint taxation (e.g. Selin, 2014). Since Germany still adheres to this system, solely

microsimulation studies exist to predict the impact of a switch to individual taxation (e.g. Steiner and Wrohlich, 2008). The findings in the gender-neutral lab provide the first experimental evidence on this topic. We varied the taxation systems across stages. In one stage, couples faced individual taxation, and in another stage, they faced joint taxation. We find only little evidence on an effect of individual taxation on work-effort, but strong results regarding the allocation of the uncompensated task: Individual taxation encourages a more egalitarian allocation of the unpaid task that proxies housework, compared to a system of joint taxation.

In chapter 4, which presents joint work with Miriam Beblo, Denis Beninger and Norma Burow, we refer to an experimental manipulation to investigate the impact of norms that are linked to a social identity on individual decision making. To achieve this, we refer to a priming manipulation that stimulates and exogenously varies a social – in our case, gender – identity. This priming stimulus makes the gender identity more salient and evokes the behavior to move closer to an action that is prescribed for this gender identity (Benjamin et al., 2016). Thus, it is assumed to cause a behavioral change that captures a social identity effect on behavior. In our experiment we observe men's and women's selection into payment schemes that entail risk and competition, which are assumed to be influenced by gender norms (for an overview see Shurchkov and Eckel, forthcoming). To capture whether and how gender identity affects behavior, we primed half of the participants in an online experiment by asking them to indicate their gender at the beginning of the experiment. To reinforce the priming effect, a screen wiper with a pictogram of a man or a woman appeared constantly on the screen for the whole remainder of the experiment. We find that men and women differ in their preferences for these payment schemes, and we find a weak association between stated gender norms (on an appropriate behavior regarding risk and competition for men and women that we collected two weeks before the experiment) and selection into these payment schemes. However, we do not find any statistically significant priming effect. To find explanations for these null findings of a priming stimulus, which are not uncommon, according to a systematic literature overview, we discuss the effectiveness of social identity priming techniques from theoretical and experimental perspectives by using insight from economics and social psychology. Thus, this paper provides not only (a lack of) experimental evidence on the effect of gender identity norms on behavior but also a critical reflection of null findings that priming research in economics should consider in designing and interpreting experiments in the future.

Chapter 5 provides evidence on the association between gender segregation in the labor market and a gender gap in absenteeism by using administrative data from the largest statutory health insurance in Germany. I adopt the decomposition method, which is commonly used to

investigate the gender pay gap, and present a new approach in the research agenda of gender differences in absenteeism. This enables me to provide evidence on the effects of both gender differences in exposure and gender differences in vulnerability, which presents two sets of reasoning to explain the relationship between segregation in the labor market and gender differences in work absenteeism. The effect of gender differences in exposure refers to the fact that women and men are employed in different occupations that presumably entail different health risks. Although some studies have identified this effect by interpreting the change of the gender variable when they control for employees' occupations in estimating a measure of work absenteeism (Mastekaasa and Dale-Olsen, 2000; Melsom and Mastekaasa, 2017), my work contributes by providing detailed evidence and distinguishing between, for example, vertical and horizontal segregation. However, the crucial point of my contribution is that I provide first direct empirical evidence on the argument of gender differences in vulnerability, which suggests that there are gender differences in reactions to (certain characteristics of) an occupation. The results of the decomposition show that gender segregation works in favor of women, since the gender gap in absenteeism increased in a world without gender segregation. Within the same occupation, women are on average more vulnerable than men but, and as the gender-difference-in-vulnerability argument suggests, the-within occupational differences vary across characteristics of an occupation such as the complexity of the task in the occupation and the occupational area.

1.4 Abstracts of each chapter

To conclude, I give a brief description follows by the abstracts of each of the four papers that make up this dissertation, given by the abstracts.

Couple's Labor Supply, Taxes, and the Division of Housework in a Gender-Neutral Lab

We use a lab-in-the-field experiment to investigate intra-couple labor supply decisions and the division of housework under individual and joint income taxation systems. In order to eliminate problems of endogenous intra-couple time-use decisions, we exogenously varied not only the taxation system but also the intra-couple roles of primary and secondary earners. With work-effort used as a proxy for labor supply, 62 established couples, both cohabiting and married (124 participants), performed real-effort tasks under a piece-rate payment system within a given time. Prior to this paid task, couples had to decide on the allocation of an unpaid task serving as our proxy for housework. In our gender-neutral lab, we find tax effects only on men's labor supply but not on women's, and no gender differences in the allocation of housework. Instead, the allocation of housework follows a purely economic rationale, with the majority of secondary earners taking responsibility. This is confirmed by a shift to a more egalitarian allocation when individual taxation applied. However, one result replicates real-world findings with married male participants providing more labor supply than cohabiting men and married women less than cohabiting women. This result hinges on the stability of specialization in married couples, which seems to overcome the gender-neutral lab.

Self-Selection and Conditional Performance: The Gender Pay Gap in a Choice Experiment

We investigate the gender pay gap in a choice experiment, run online all over Germany, where, by design, we rule out employer-side discrimination, and are able to isolate the labor supply-side determinants of earnings. Almost 900 participants performed an effort task based on their preferred compensation scheme (linear piece-rate vs. piece rate with performance bonus/competition premium, at varying difficulty levels and with/without a risk premium). We observe a gender gap of 23% in the compensation of female and male participants. An Oaxaca-Blinder decomposition reveals that one-quarter of this gap can be explained by selection into contract types (selection effect), and about one-half can be attributed to the participants' performance conditional on these contracts (contract effect). The selection and contract effects increase in the presence of a bonus payment. The contract effect is largest when the payment

scheme includes a risk premium and it is robust to productivity differences between participants. We argue that the observed pay difference is driven by women being more loss averse than men and possibly underperforming in a stereotype threat situation.

Social Identity Priming in Economics Experiments: No Clear Evidence

Using a priming stimulus to vary exogenously the salience of a social identity and its impact on individual decision making is a recent trend in experimental economics. However, results are mixed, and significant priming effects are regularly lacking. This paper reviews existing priming experiments in economics and compares them with a large-scale field experiment we conducted in Germany. In particular, we discuss the frequent absence of priming effects based on recent theoretical insights in economics and social psychology. Consequently, this paper can be seen as a critical reflection of experimental (null-)findings when social identity priming techniques are used.

Identifying Gender Differences in Exposure and Vulnerability – A Decomposition Analysis of the Gender Absenteeism Gap in Germany

This paper provides evidence on the relationship between occupational segregation by gender and the gender gap in absenteeism in Germany. There are two explanations: One argues that women and men are employed in different jobs that differ in health risks (difference in exposure) and the other that women and men differ in their vulnerability to the characteristics of an occupation (difference in vulnerability). Performing a detailed decomposition of the gender gap in absenteeism, which is new in this strand of the literature, confirms previous findings on gender segregation working in favor of women. Additionally, it provides the first direct evidence on the gender-difference-in-vulnerability argument, revealing that women seem to be, on average, more vulnerable than men within an occupation, while the direction and the magnitude of effects are heterogeneous across occupational areas and job levels.

2 Couple's Labor Supply, Taxes and the Division of Housework in a Gender-Neutral Lab[7]

2.1 Introduction

General wisdom suggests that women and men behave differently in the labor market. Although gender gaps in the labor market have been narrowing over the course of the past century, they are still substantial in employment levels and in pay (OECD, 2012a). When it comes to married women, these gaps are even more pronounced (Blau and Kahn, 2007). Most explanations point to traditional family arrangements, which are characterized by rather unequal divisions of family chores that, consequently, are found to be the main drivers of gender differences in the labor market (Ichino, 2014). Accordingly, women and men face different constraints (Blau and Kahn, 2007), no matter whether this is explained by comparative advantages at home or gender norms supporting the work division puzzle (Cochard et al., 2015). Indeed, data show that in all OECD countries women do more unpaid work than men (OECD, 2012a).

At the same time, an institution like joint income taxation is suspected to reinforce these gender differences in the labor market, as it affects intra-couple time allocation. A joint income taxation system does not use the individual spouse's labor income as its basic tax unit but rather the split total labor income of the couple. Therefore, compared to an individual taxation system, it levies higher marginal tax rates for secondary earners within couples, which is why it creates larger disincentives to work. Since it is most commonly women who earn less than their spouses, such a system reinforces traditional family arrangements, with women specializing in household work and being absent from the workforce (OECD, 2012b).

Since estimating labor supply decisions is challenged by the endogeneity of wages and self-selection into the labor market, correction methods are commonly used in studies that employ survey data (Laczó, 2011; Triebe, 2013). Contrary to these, we investigate labor supply decisions of couples in a controlled laboratory experiment that rules out these problems by design. Consequently, we ask how couples' labor supply and the allocation of housework are determined (under different taxation systems), when endogeneity of wages and self-selection are truly eliminated.

We conduct a framed field experiment (Harrison and List, 2004), in which 62 established hetero-sexual couples perform under piece-rate payment on real effort tasks (i.e., solving

[7] This paper was developed by Melanie Schröder as the main author together with Norma Burow and has been published as DIW Discussion Paper No. 1593 (Schröder and Schmitt, 2016).

mazes) within a given time and with work-effort (i.e., number of solved mazes) serving as our proxy for labor supply. The concept of labor supply is usually based on the measure of hours of work, but we observe work-effort instead, because it "describes many short-run labor supply decisions" (Dickinson, 1999: 640) and is a good proxy for today's real-world labor contracts (Meghir and Phillips, 2009). There were two types of mazes, differing in complexity level (hard, easy), with corresponding wages (high, low), thus randomly and exogenously determining who is the primary earner (i.e., hard mazes with a higher piece-rate wage) and the secondary earner (i.e., easy mazes with a lower piece-rate wage) within the couple. In addition, we exogenously assign individual and joint taxation, with each couple facing individual taxation in one stage and joint taxation in the other. To investigate the allocation of housework, each couple had to decide upon who of the two takes over an unpaid but compulsory task that noticeably reduces time for the paid task and has to be completed prior to the compensated task. It is exactly this implementation of these exogenous variations, possible only in an experiment, that creates a "gender neutral" setting.

Interestingly, in such a "gender-neutral lab," we cannot confirm real world gender gaps in labor supply, nor could we confirm the general conviction that housework is women's work. The allocation of unpaid work in our experiment indeed follows an economic rationale, with opportunity costs determining couples' decisions. This is why our experimentally applied individual taxation system encourages a more egalitarian allocation of the unpaid work that proxies the "unloved" housework outside the lab. However, we indeed find some interesting results when it comes to the institution of marriage. In fact, we herewith confirm findings from survey data demonstrating that married male participants provide more labor supply than cohabiting men, but married women provide less labor than cohabiting women (Barg and Beblo, 2012; Blau and Kahn, 2007). We argue that sorting into specialization (Barg and Beblo, 2012), with a strong emphasis on traditional gender norms, is the main driver of this result.

This paper is organized as follows: The literature review in Section 2.2 is followed by a presentation of theoretical background and some hypotheses in Section 2.3 and the description of the experiment in Section 2.4. In Section 2.5, we present our results, followed by a discussion and conclusion in Section 2.6.

2.2 Review of the literature

This is the first experimental economics paper focusing on the impact of the different income taxation systems on couples' labor supply and the division of housework.[8] However, there is a rewarding stand of empirical literature using survey data that refers to labor supply and labor supply elasticity of individuals living as a (married) couple. The main descriptive findings for many Western countries are: (1) there is a gap in labor force participation, in that men have higher rates compared to women (OECD, 2012a), (2) married men participate more in the labor market than cohabiting men (Barg and Beblo, 2012); (3) married women participate less in the labor market than do cohabiting women (Barg and Beblo, 2012), (4) labor supply elasticity is larger for women compared to men, and (5) this latter gap increases when it comes to married couples (Bargain et al., 2014). As Ichino (2014) points out, these gender differences are largely determined by the unequal division of household chores. Or, according to Blau and Kahn (2007), women face constraints that men do not. What they mean is that, typically, men tend only to substitute market work with leisure, while women face an additional market work substitute: housework.

Consequently, the gendered division of housework is a well-established empirical fact: Women do the bulk of household work (OECD, 2012a) and there are several explanations of why. In a world with gender-based pay gaps, the first economic explanation at hand is that the difference in opportunity costs leads to a gendered allocation of housework. Interestingly, Brines (1994) and Haberkern (2007) show that housework remains women's work regardless of the intra-couple income differences, even if women earn more than their husbands. Referring to the concept of "doing gender," West and Zimmerman (1987) state that women display their femininity by doing housework. Gender identity considerations, as conceived by Akerlof and Kranton (2000), translate into societal expectations like (1) husbands should earn more than their wives, and (2) wives are unwilling to earn more than their husbands – both relating to the norm of a male breadwinner. Indeed, Bertrand et al. (2015) confirm that these expectations have severe effects on the intra-couple division of housework. They show that, unlike what could be expected, primary-earner women take over most of the domestic work – more than their husbands are expected to do with respect to their comparative disadvantage. Contrary to these findings, Auspurg et al. (2017) and Cochard et al. (2015) do not show any systematic gender differences in the division of housework within couples. However, the couple's reference is

[8] To our knowledge, there is only one experimental paper, by Cochard et al. (2015), that studies real couples' work division in an artificial field setting. Moreover, they exogenously assigned intra-couple roles by creating an "advantaged" and "disadvantaged" partner by varying the earnings from their private accounts such that investing in the household public good is not efficient for the advantaged player.

important, as demonstrated in the lab by Görges (2015). She finds that women are significantly more likely to perform an unpaid task when they play with their beloved compared to playing with a stranger and explains this with social gender norms. Using the UK Time Use Survey, Stratton (2012) discovers that it is not only opportunity costs but also preferences for "evil" housework tasks that help explain the division of housework tasks within households.

As regards the influence of income taxation systems on couples' labor supply, the focus of the literature is on the disincentives to work in general. In particular, a joint income taxation system is usually implemented in order to realize horizontal tax equity no matter the intra-couple income distribution. It incorporates individual income capabilities by relieving taxes for the spouse who is in the workforce while the other is at home, working less and/or earning less. However, it supports a breadwinner model, since in such a system taxes are applied on the split total earnings of married couples, which under a progressive tax function results in lower marginal tax rates for the primary earner and higher rates for the secondary earner.[9] LaLumia (2008), Crossley and Jeon (2007) and Selin (2014) use natural experiments in the United States (change from individual to joint taxation), Canada and Sweden (joint to individual taxation), respectively, showing that the system of joint taxation is associated with a lower labor force participation of married women, since it is mostly women who are in the role of secondary earners and, therefore, face disincentives to (increase) labor market work. For Germany, which still adheres to a system of joint taxation, microsimulation studies predict an increase in married women's labor supply if individual taxation is introduced (Bach et al., 2011; Beninger et al., 2007; Steiner and Wrohlich, 2004). On the contrary, husbands would reduce their hours worked, as well as their participation rate, but their labor supply effects in total would be much smaller than the effects for married women. Moreover, Decoster and Haan (2014) demonstrate, with the help of a structural model, that an individual taxation system would additionally increase households' disposable income. Kabatek et al. (2014) additionally integrate the housework domain in their simulation study by using the French Time Use Survey and show that a shift from joint to individual taxation could contribute to equalizing the within-couple housework allocation.

We contribute to the existing literature in that we investigate couples' labor supply and the allocation of housework under different taxation systems under what we call a "gender-neutral" setting. First, we exogenously assign intra-couple roles and balance them over gender, thus

[9] Joint income taxation systems are also referred to as "family based" taxation (OECD, 2015), "income splitting" systems (Steiner and Wrohlich, 2004), or "joint filing" as opposed to "separate filing" (Pollak, 2011). An overview of countries adhering to such systems is provided in OECD (2015).

ruling out problems of self-selection and endogeneity. Second, survey data often encounter the problem of a small sample size of men working part-time or male secondary earners when investigating labor supply choices. In our setting, half of the sample consists of male secondary and female primary earners. Third, each couple, whether married or not, experiences both individual taxation and joint taxation. This presents a big advantage, since joint taxation normally is the privilege of married couples, and self-selection into specialization by marriage is shown in the literature (Barg and Beblo, 2012). Fourth, by using work effort we observe changes from a short-term perspective since work effort instead of hours worked "describes many short-run labor supply decisions" (Dickinson, 1999: 640). This makes it a good proxy for today's real-world labor contracts, offering employers the opportunity to substitute on-the-job-leisure for work effort (Meghir and Phillips, 2009). Even though Dickinson (1999) calls for caution when generalizing work effort to more traditional measures of labor supply, Doerrenberg and Duncan (2014) conclude that work effort is indeed a good proxy for labor supply. Additionally, Meghir and Phillips (2009) stress that hours of work is just one dimension of work effort for many individuals, especially for workers with a high level of autonomy in their work. Moreover, these changes in short-term are clearly supply driven, without potential influence from the employer side.

2.3 Theoretical background and hypotheses

In order to meet the challenge of endogeneity of wages and selection into the labor market, we exogenously assign intra-couple roles with the help of tasks and corresponding income capabilities. Each couple consists of one primary earner (PE), whose job is to solve harder tasks with a higher gross wage rate, and a secondary earner (SE), whose job is to solve easier tasks for a lower gross wage rate. That assignment remains unchanged throughout the duration of their participation in the experiment. To observe the reaction to a change in tax system, every couple, whether married or not, faces both tax systems during the experiment.

To investigate labor supply, we use work effort as a proxy and refer to the Intensity Model of Dickinson (1999). The model assumes utility to be a function of consumption (c), productive hours of work (h_w) and hours of on-the-job-leisure (h_l) by $U_c > 0$, $U_{h_l} > 0$, $U_{h_w} < 0$. Hours of work (h) can then be denoted with $h = h_l + h_w$. Since hours of work are fixed in our experiment, the subject's choice variable instead is h_w, work effort: This is the intensity that participants choose in working on the paid task within a given time to maximize utility, recognizing that they also could engage in their market work substitute, which is on-the-job

leisure. Since both men and women have the same market work substitute and no other duties, as in a real-world setting – such as domestic and caring tasks – they face the same constraints. Therefore, we do not expect behavioral differences between men and women to occur.

Table 1 presents the taxation systems designed for our experiment: individual and joint income taxation. We hold the gross wage rates, denoted as w_g^{SE}(i.e., gross wage rate for secondary earner) and w_g^{PE} (i.e., gross wage rate for primary earner), as well as the progressive tax function constant across both taxation systems. In order to create marginal tax rates that differ between the systems, we simply allocate basic allowances, which determine initial tax-free income ranges, differently. Under individual taxation, each partner gains from a basic allowance E, while under joint taxation both partners' basic allowances are assigned to only the primary earner and the secondary earner is taxed beginning with the first euro. Basically, our tax scheme, in both cases, consists of individual taxation, but due to assigning basic allowances differently we simulate the typical differences in the marginal tax rates between both systems.[10] This mirrors the well-known real-world constellation under joint income taxation with a progressive tax function, where both spousal incomes are summed and divided (equally)[11] as the basic unit for assessing income taxes.

Table 1 Net Individual Incomes Conditionally on Taxation System

	Individual Taxation	Joint Taxation
Secondary Earner (SE)	$I_I^{SE} = nw_g^{SE} - \tau(nw_g^{SE} - E)$	$I_J^{SE} = nw_g^{SE} - \tau(nw_g^{SE})$
Primary Earner (PE)	$I_I^{PE} = nw_g^{PE} - \tau(nw_g^{PE} - E)$	$I_J^{PE} = nw_g^{PE} - \tau(nw_g^{PE} - 2E)$

Note: E - basic allowance, τ - progressive tax function, w - wage rate, J - joint taxation, I - individual taxation, PE - primary earner, SE - secondary earner.

With regard to labor supply choices in our experiment, we expect a positive substitution effect. This means an increase in work effort when the own net wage increases due to a lower marginal tax rate, i.e., from individual taxation to joint taxation for primary earners and from joint taxation to individual for secondary earners, and vice versa. However, labor supply choices also emerge endogenously from intra-couple bargaining, that is, choices are also made jointly in a

[10] Implementing a joint taxation system experimentally via assigning the tax allowance to one partner within a couple is based on the income tax class combination in Germany, III and V. Here, one spouse is grouped in tax class III receiving basic and lump-sum allowances, while the other spouse, receiving no allowances, is grouped in tax class V (Stöwhase , 2011). Following Stöwhase (2011), this tax class combination can be seen as an early realization of a "splitting advantage."

11 Or by a certain factor conditionally on the number of children, as in France (Steiner and Wrohlich , 2008).

couple. Therefore, we could also expect an income effect to occur with a change in the couple's total income, which differs between the taxation systems.[12] On account of the fact that both partners' incomes change simultaneously (but independently) by design – i.e., an increase or decrease in own net wage and a decrease or increase in the partner's wage occur at the same time – we are not able to distinguish between the two effects. Since both effects point in the same direction because of the fact that an increase of own net wage goes along with a potential decrease of the partner's net wage, this issue is negligible. In sum, we expect an increase in work effort with decreasing marginal tax rates, as shown by researchers using both survey (see Meghir and Phillips, 2009 for an overview) and experimental data (see Alm, 2010 for an overview). We assume that these effects do not differ over gender since both genders face the same market work substitute in our experiment and could only consume on-the-job-leisure.

Besides the choice of work effort, participating couples also faced the decision of substituting market work with a non-market alternative – an unpaid but compulsory task, our proxy for housework. As in a real-world setting, housework – although unpaid and undesired – must be done, thus reducing precious time for earning money or consuming leisure. Where productivity differences might exist in reality, we ensure that productivity differences could not occur and announced that the so-far-unknown task is "not difficult at all with no special previous knowledge required." However, since the unpaid task in the experiment was indivisible, couples had to agree jointly upon the allocation of housework to only one of the two. Following cooperative models of intra-family decision making (see Donni and Chiappiori, 2011 and Grossbard, 2011 for excellent overviews), which all point to similar predictions regarding a rational allocation of housework, we predict the majority of couples would choose the secondary earner to take it over, irrespective of gender. In the absence of productivity advantages, only a comparative disadvantage in market production or a bargaining disadvantage in negotiations may be held responsible for that decision.

As the unpaid task in our experiment is designed to be gender neutral, in the sense that it is unknown to participants and requires no previous knowledge, couples should not expect productivity differences in advance. To this end, the intra-couple allocation should be totally unrelated to gender. However, it might also be reasonable to expect a gendered allocation of the unpaid task when we follow the "doing gender" concept (West and Zimmerman, 1987) or

[12] From a couple's perspective, individual taxation yields slightly lower total income compared to joint taxation with a gap of 3 % in the average couple's income. Due to the small magnitude of the gap, a change in tax systems can be considered as almost income neutral with respect to the total income. Consequently, we can ignore a potential income effect.

the "Identity Economics" approach of Akerlof and Kranton (2000), if couples bring their social gender norms from outside into the lab (Kimbrough and Vostroknutov, 2016).

Last, but not least, as income taxes determine labor supply choices, the division of housework must also be affected. Comparing the two taxation systems, we predict that individual taxation encourages a more egalitarian division of housework, as already shown by Kabatek et al. (2014). Primary earners are expected to take over this task more frequently under individual taxation compared to the situation of joint taxation, since higher marginal tax rates reduce primary earners' net piece-rate wage and, hence, decrease opportunity costs of market work.

2.4 Description of the experiment

We conduct a "framed field experiment" with non-standard subjects participating in a lab experiment with field context in that we framed the information set concerning the taxation of income as stemming from participants' natural environment (Harrison and List, 2004). We invited both cohabiting and married, heterosexual couples who had been living together for at least one year in the area of Frankfurt (Oder), Germany, to participate in our experiment.[13] Contrary to Güth et al. (2004), who invited standard subjects (students), but in line with others conducting real couple experiments (Bateman and Munro, 2009, 2005; Palma et al., 2009),[14] we used couples. By doing so, we realized a high level of control in the lab with a "subject pool from the market of interest" (Beblo and Beninger, 2017:786) Moreover, as Fochmann and Weimann (2013) show, it was required that participants had personal experience with income taxes; we would have had probably too few cases among a student sample. All experimental sessions were carried out at the European University Viadrina in Frankfurt (Oder), Germany, in the evenings and on weekends throughout the summer and autumn of 2012.[15]

2.4.1 Experimental procedure

After arriving in the classroom, participants were seated in pairs with partition screens that separated couples from one another in order to prevent interaction and provide privacy. Subjects were informed that they were taking part in a scientific study that consisted of two stages in which they could accumulate income by solving tasks (with an additional show-up fee of 2.50 €), but only one of the two stages (random selection) would be relevant for payoff. Immediately

[13] We recruited subjects by distributing postcards, publishing a call for participation in the local press (newspaper and radio) and visiting parents evenings at local kindergartens and schools.

[14] For an overview of couple experiments in economics, see Beblo (2015) and Munro (2015)

[15] For participants with children, professional childcare was provided during the experiment in cooperation with local kindergarten teachers.

before each stage, subjects were informed about the type of task they had to perform. All instructions were handed out and read aloud.[16]

In both stages, each partner's compensated task was to solve mazes[17] by using paper and pencil within a period of 15 minutes. One person in each couple was assigned to be the secondary earner, with easy-level mazes and a lower piece-rate wage (0.50 €), while the other was assigned to be the primary earner, with hard-level mazes and a higher piece-rate wage (1.50 €). Choosing mazes that differ in complexity level is advantageous in many ways. With regard to productivity-oriented wage setting, it is plausible from a participant's point of view that for a task of a higher complexity level, a complexity-premium is applied that leads to a higher overall remuneration. Additionally, there are no effects coming from the preference for a specific task when the same type of task is assigned for both partners.

Income taxes were collected during both stages. The couple's income was taxed individually in one stage and jointly in the other. In both situations, the same progressive tax function applied. Under individual taxation, both partners gain from a basic allowance of 4.50 €, while under joint taxation the basic allowance (E) for both partners of 9 € was assigned to only the primary earner. The tax description sheets, which were handed out at the beginning of each stage, included an effort-income table and a short explanation of the tax system. Following Fochmann and Weimann (2013), who emphasize that complex tax environments may cause biases, we kept our experiment's tax schemes and instructions as simple as possible. First, we made use of a tax function that is piecewise linear and progressive with increasing marginal tax rates (20%, 40%, 60%, 80%, 90%, 95%).[18] Second, we represented net wages instead of a tax rate (Sillamaa, 1999). Third, we illustrated the tax burden per unit graphically with the help of a pie chart (Fochmann and Weimann, 2013). There was also a clearly written description of the tax system, summarized as, for individual taxation "Both partners' income is taxed to the same degree," and, for joint taxation, "Both partners' income is taxed to a different degree. The tax burden of the one with the lower wage rate is higher, and the tax burden of the one with the higher wage is lower." To make sure that subjects were acquainted with both tax sheets and both income opportunities, they had to answer control questions concerning their own and their partner's potential income.

[16] The translated instructions are provided in the Appendix.

[17] We used a collection of mazes from the web: http://www.onebillionmazes.com. Unfortunately, this website provides different content today. "Easy" refers to mazes of a low difficulty level, "hard" to mazes of a slightly higher difficulty level.

[18] Piecewise linear tax systems are very common in reality, although in Germany non-linearity is additionally implemented. Apps et al. (2014) present an analysis of the main characteristics of a piecewise linear tax system. See also Apps and Rees (2009) for a general overview of household taxation systems.

Additionally, one partner had to fulfill an unpaid but compulsory task, which was easy but reduced time for the paid task from 15 to 12 minutes. To avoid effects from a preference for a specific task, the kind of task was unknown to the participants. In one stage, subjects connected dots (with paper and pencil) that should yield a picture. In the other stage, subjects were asked to decode numerical series into words by substituting the numbers with letters, by using an encryption table that assigned a number to each letter of the alphabet, similar to Erkal et al. (2011). The couple had to decide who of the two undertakes this task prior to the following paid work part. To allow for on-the-job leisure (Dickinson, 1999), we arranged a selection of magazines, a daily newspaper, sweets and drinks on each of the couples' desks.

After having performed in two stages, each single participant was asked to fill out a post-experimental questionnaire that contained questions about their individual socio-demographic and couple-related characteristics. To avoid communication and interaction while they were filling out the questionnaire, we seated the partners apart from each other at this stage. Directly after the experiment, each couple received their payments, plus the show-up fee, in another room. Since outcomes of only one stage were relevant for payoff, and to secure random selection, one partner of each couple had to draw a ball from an urn with red and yellow balls that represented the potential income of each stage to determine the final household total income.

2.4.2 Experimental design

In our within-between-subjects design, we assigned two types of mazes, with corresponding piece-rate wages to create two different roles within each couple for the duration of the experiment. The higher piece-rate wage defines the primary earner (PE), and the lower, the secondary earner (SE). As depicted in Figure 1, in experimental groups 1.1 and 2.1, the couples consisted of a male primary and a female secondary earner. In the other groups (1.2 and 2.2), we reversed these intra-couple income roles, creating couples with a female primary and a male secondary earner.

To control for learning and boredom effects that could occur when solving the same task in both stages and that would confound tax effects, we inverted the order of the tax conditions in the second treatment as compared to the first treatment.

Figure 1: Experimental Design

	Treatment 1		Treatment 2	
	Group 1.1 male PE & female SE	Group 1.2 female PE & male SE	Group 2.1 male PE & female SE	Group 2.2 female PE & male SE
Stage 1	individual taxation		joint taxation	
Stage 2	joint taxation		individual taxation	
	post-experimental questionnaire, urn-decision			

Note: PE – primary earner, SE – secondary earner

2.5 Results

We conducted 24 sessions with 124 participants (62 couples). Each session took about one hour, and the average payment was 27.24 € per couple. A sample description can be found in the Appendix. One couple cheated by exchanging their assigned mazes, and we dropped these observations. Furthermore, we restrict the sample to participants with tax experience.

Before performing the compensated task, each couple had to decide who of the two will undertake the unpaid but compulsory task. This task has to be done prior to the compensated task and reduces time for the paid task to 12 instead of 15 minutes for the one who undertakes it. As the couples' time use decisions include both paid and unpaid work, we divide the results section into two parts: Section 2.5.1 presents the intra-couple unpaid task allocation, while Section 2.5.2 considers the work-effort choices of each individual after controlling for housework responsibilities – that is, the take-over of the unpaid task.

2.5.1 Allocation of the unpaid task

Figure 2 presents the raw data results for the allocation of the uncompensated task. In line with our hypothesis, 75% of the couples choose the secondary earner to take over this task. Comparing the shares of primary earners who take over this task under individual and joint taxation reveals a significant difference (binomial test, two-sided; $p<.001$). While only 14% of the primary earners take over this task in the case of joint taxation, this share more than doubles with individual taxation. As a result, the intra-couple allocation of the unpaid task is more equal under individual taxation compared to joint taxation. The men's share is, surprisingly slightly

larger than that of the women, but it is not significantly larger than 50 % (binomial test, two-sided; p=.441), with, therefore, no gender gap shown to exist.[19]

Figure 2: Allocation of the Unpaid Task

Note: Displayed are the proportions of those who undertake the unpaid task and corresponding binomial tests. PE = primary earner, SE = secondary earner.

These results are confirmed by using multivariate estimation analysis. Table 2 presents estimation results for the allocation of the unpaid task by using a pooled linear probability model with cluster robust standard errors at the individual level, since subjects made decisions in two subsequent stages.[20] Figure 3 shows corresponding marginal effects with the full model (column 3). We observe allocation decisions conditional on the assigned role (dummy for primary earner: pe, reference secondary earner) and participant's gender (dummy for men: male, reference women). In order to investigate tax effects, we use a dummy for the tax system (joint, reference individual), which takes the value of 1 if joint taxation is applied and 0 if individual taxation is applied. Stage controls (interaction of stage and pe) are considered in all estimations, and individual controls are added in column 2 (including participant's highest educational attainment, labor market status, personal gross income, age, whether the person was born in East Germany and his/her satisfaction with the assigned role). Couple controls (interaction of marital status and gender) are then added in column 3.

Results clearly indicate that couples follow the economic rationale. First, being assigned to the primary earner role significantly decreases the probability of the participant taking over the

[19] These effects do not change, when we restrict the sample to married couples.
[20] Running logit regressions leads to similar results and tables are available upon authors request.

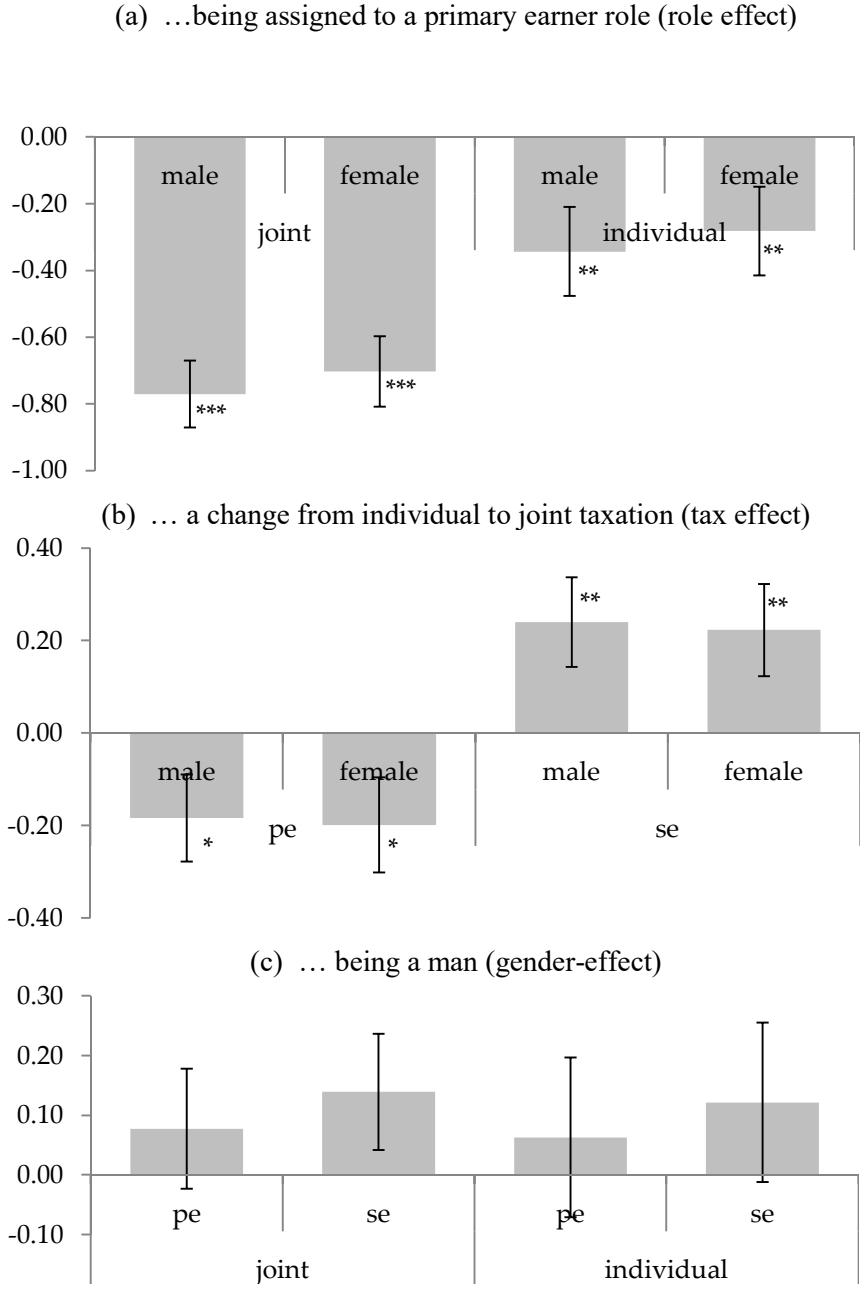
unpaid task for men and women in both tax conditions (Figure 3a). Interestingly, the effects seem to differ over both tax conditions, with the primary earner being significantly less likely to take over this task under joint taxation. Second, as Figure 3b depicts, joint taxation indeed causes a reallocation of the unpaid task between the intra-couple roles. The probability of taking over the task is 20% lower for the group of primary earners when joint taxation applies. And finally, in accordance with our hypothesis and the raw data findings from above, gender appears not to play a role in the assignment of the unpaid work in our experiment since gender differences are not statistically different from zero (Figure 3c).

Table 2: Regression Results for Taking Over the Unpaid Task

DV= unpaid task	(1)	(2)	(3)
joint	0.223**	0.223**	0.223**
	(0.098)	(0.099)	(0.100)
male	0.108	0.103	-0.052
	(0.129)	(0.132)	(0.185)
joint x male	0.010	0.018	0.018
	(0.134)	(0.138)	(0.139)
pe	-0.272*	-0.310**	-0.302**
	(0.139)	(0.140)	(0.139)
joint x pe	-0.465***	-0.421***	-0.421***
	(0.137)	(0.143)	(0.143)
male x pe	-0.084	-0.035	-0.063
	(0.184)	(0.185)	(0.187)
joint x male x pe	0.049	-0.003	-0.003
	(0.190)	(0.196)	(0.197)
constant	0.599***	0.760***	0.877***
	(0.102)	(0.203)	(0.218)
stage controls	yes	yes	yes
individual controls	no	yes	yes
couple controls	no	no	yes
observations	226	216	216
R-squared	0.321	0.329	0.340
adj R-squared	0.293	0.279	0.284

Note: Displayed are the coefficients of pooled linear probability models with cluster robust standard errors on the individual decision to take over the unpaid task ((0-1) choice counted on the individual level, jointly agreed upon within the couple). * p<.100, ** p<.050, *** p<.010. Full table in Appendix, Table 4.

Figure 3: Marginal Effects of ...



Note: Displayed are average marginal effects on the decision to take over the unpaid task ((0-1) choice made by the individual, jointly agreed upon within the couple) and error bars that represent the 95 % confidence interval. Basis is the full model of column 3 in Table 2. * p<.10, ** p<.05, *** p<.01.

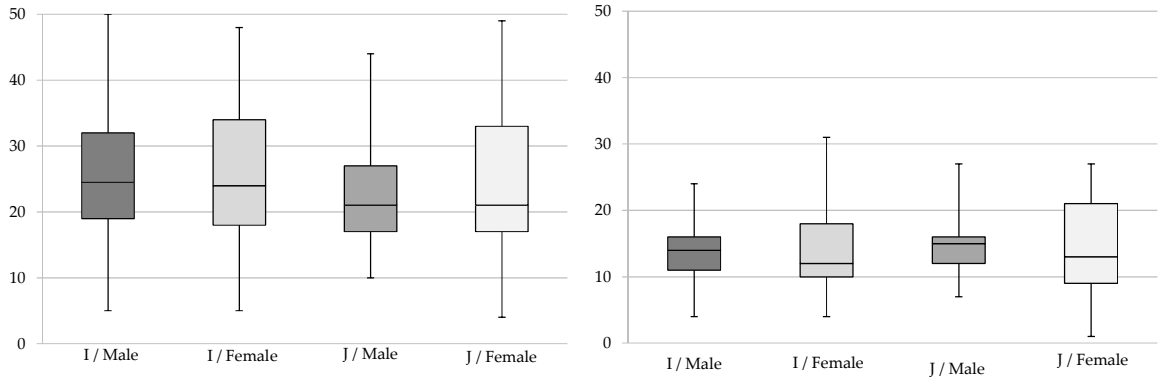
2.5.2 Work effort

To analyze labor supply decisions in our experiment, we consider paid work, that is the individual work-effort as measured by the number of solved mazes. Figure 4 gives an overview of participants' work effort by intra-couple role, taxation system and gender. At first glance, it seems odd that secondary earners are more productive than primary earners. This finding simply reflects the difference in the complexity levels of the tasks assigned to primary earners and to

secondary earners. This is exactly the reason why we do not compare secondary and primary earners' work-effort.[21] To begin with, Mann-Whitney tests do not detect any significant gender differences or tax effects in work-effort decisions. Nor do men and women differ in their work effort, nor does a change in the taxation system disclose any behavioral effect.

Figure 4: Work-Effort by Role, Tax and Gender

(a) Secondary Earner (b) Primary Earner



Note: Displayed are the box plots of work effort by intra-couple role, taxation system and gender.

Table 3 then provides multivariate estimation results from pooled linear regression models with cluster robust standard errors. The dependent variable is work effort, i.e., the number of solved mazes. We estimate labor supply decisions conditionally on the assigned role (dummy for primary earner: pe, reference secondary earner) and participant's gender (dummy for men: male, reference women). In order to investigate tax effects, we again use a dummy for the tax system, along with a dummy that captures stage effects. Furthermore, we control the allocation of housework (interaction of taking over the unpaid task, male and pe) in all estimations. Column 2 then adds personal characteristics, such as participant's highest educational attainment, labor market status, personal gross income, age, whether the person was born in East Germany and his/her satisfaction with the assigned role. In column 3, we additionally consider marital status in a dummy (married, reference cohabiting).[22]

[21] One might claim that the difference in complexity levels challenges the implementation of the intra-couple role assignment. However, we overcompensated the complexity level in the following way: A hard-level maze equals 1.72 easy-level mazes, whereas a hard-level maze is remunerated 3 times higher than an easy-level maze. We thereby ensured that secondary earners' income is always lower than primary earners' income. Table 6 in the Appendix demonstrates that the intra-couple role implementation holds true for the majority of the sample. T-tests indicate that the incomes of primary earners are significantly higher than those of secondary earners ($p<.001$).

[22] A full table can be found in the Appendix.

As Table 3 displays, the negative sign of pe mirrors the higher complexity level of the task, which is why primary earners' work effort is significantly lower than that of secondary earners (see the first paragraph of this subsection for a discussion of this point). Besides this, none of the other coefficients seems to play a role until individual characteristics are included in column 2, which leads to the interaction of joint taxation and male becoming significant, thus suggesting tax effects and/or gender differences. Adding marital status to the model in column 3 reduces the male coefficient to a large extent, with the coefficient becoming significant. This indicates that marriage seems to drive results, thus playing a role in our experiment.

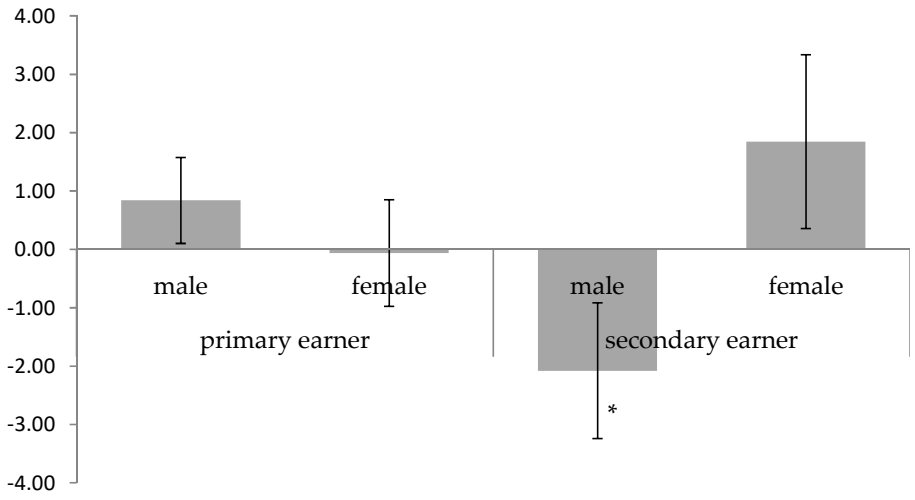
Table 3: Regression Results on Work-Effort

DV= work effort	(1)	(2)	(3)
joint	1.643	1.856	1.849
	(1.541)	(1.485)	(1.488)
pe	-14.74***	-12.13***	-11.64***
	(3.729)	(2.875)	(2.908)
joint x pe	-0.515	-1.599	-1.913
	(1.828)	(1.714)	(1.717)
male	-0.025	-0.430	-6.056*
	(4.662)	(2.997)	(3.189)
joint x male	-2.710	-4.141**	-3.926**
	(2.102)	(1.995)	(1.968)
pe x male	1.708	0.716	-0.377
	(4.971)	(3.374)	(3.512)
joint x pe x male	2.295	4.725**	4.830**
	(2.417)	(2.294)	(2.258)
married			-4.516*
			(2.318)
married x male			8.180***
			(2.277)
Constant	29.04***	44.44***	48.32***
	(3.393)	(4.300)	(4.260)
stage controls	yes	yes	yes
housework controls	yes	yes	yes
individual controls	no	yes	yes
Observations	226	212	212
R-squared	0.347	0.671	0.700
adj R-squared	0.307	0.635	0.663

Note: Displayed are coefficients of pooled linear regression models with cluster robust standard errors at the individual level on work effort. Full table in Appendix Table 5. * p<.100, ** p<.050, *** p<.010.

Following our hypotheses, joint taxation is expected to decrease secondary earners' work effort but to increase primary earners' work effort. Columns 2 and 3 reveal an effect that confirms our hypothesis, at least for men. To illustrate this, Figure 5 plots the marginal tax-effects on work effort conditionally on the intra-couple role and subject's gender by using the estimation results in column 3, which account for individual characteristics and the couples' marital status. This figure nicely highlights the average tax effect when joint taxation replaces individual taxation, accounting for individual heterogeneity. Men in the role of a secondary earner decrease their work-effort significantly when taxed jointly. In joint taxation, they solve two mazes fewer, on average, than under an individual taxation regime. In contrast, tax effects for women are not statistically significantly different from zero. F-tests confirm that men and women respond differently to a change in tax rates for the secondary earner group but not for the primary earner group.

Figure 5:Marginal Effects of Joint Taxation on Work-Effort



Note: Displayed are the average marginal effects and error bars of the 95 % confidence intervals of the joint taxation effect on participants' work effort by intra-couple role and gender. Basis is the full model of column 3 in Table 3. p-values are from F-tests.

Including an interaction of male and marriage into our analysis increases the strength of the male coefficient, which illustrates gender differences. Indeed, studies using observational data already show that married and cohabiting women as well as men, differ in their labor supply in general. According to Barg and Beblo (2012), married women's working hours are smaller than

cohabiting women's working hours and married men's working hours, are larger than cohabiting men's working hours, at least in Germany.[23]

And indeed, the overall labor supply patterns of married and unmarried couples in our experiment are identical to survey data findings. The coefficients of the marriage interaction in column 3 of Table 3 depict the male marriage surplus in labor supply. Married men solve an average of 4 mazes more than the cohabiting men, while married women solve, on average, 8 mazes fewer than cohabiting women. Thus, married men provide substantially and significantly more work effort than cohabiting men, while married women provide less work effort than cohabiting women, even though factors that might affect underlying ability (e.g. age, education, labor market experience) are controlled for.

2.6 Discussion and conclusion

In a laboratory where constraints are absent, we cannot find many of the gender differences that are present outside the lab in the real-world labor market. Without real-world constraints, women and men do not behave all that differently.

First, our findings on the allocation of the housework task are in line with Auspurg et al. (2017) and Cochard et al. (2015) but contrary to real-world findings, since we do not confirm that housework has a female label. We acknowledge it to the gender-neutral framing of the task as being "not difficult at all [with] no special previous knowledge [being] required." However, one might claim that exposure to the unknown nature of the task counteracts the expected gender differences in the intra-couple allocation of the unpaid task. In fact, if women's risk and uncertainty aversion had influenced our results, the gender ratios in the allocation of this task would differ over the stages. Nevertheless, knowledge is de facto less imperfect in the second stage, as the couples experienced the truly easy nature of the unpaid task in the first stage and should have updated their beliefs accordingly. Since we, in the end, do not find allocation differences over the stages, a potential uncertainty aversion has indeed not contradicted our results. Another reason could be that the participants simply prefer a relatively equal allocation of the unpaid task in the situation of that artificial setting and therefore altered their household decision making (Munro et al., 2008) from the real world to inside the lab. This is supported by answers from the post-experimental questionnaire, where the male participants report taking on

[23] Inspired by these studies and Bargain et al. (2014), one could assume that the reaction to a change in tax rates is stronger for married couples and thus the gender gap in elasticities becomes more visible. This is exactly what we try to capture when extending the tax interaction by a dummy for being married in another estimation and comparing the tax-effects between married and cohabiting subjects. At a first glance, F-tests indicate that the tax-effect of male secondary earners seems to be driven by married men but the difference between married and unmarried men is not statistically significant.

less than their fair share of housework in their daily life, while the women report doing more than their fair share (Wilcoxon-Signed Rank test; p>.001).[24] Thus, we assume that the subjects are indeed encouraged to perform a fair share of "housework" in our experiment because its realization is less costly compared to their daily life.

Second, although we do not find gender differences in labor supply generally, men and women differ in their reaction to a change in tax rates. Secondary-earner men react to a change while secondary-earner women are not affected. Interestingly, this gender difference in labor supply elasticity does not match the well-established fact of a larger labor supply elasticity of women compared to men outside the lab (Bargain et al., 2014). Revealing quite the opposite leads us to follow Keane's (2011) argumentation of men's labor supply being "more elastic than conventional wisdom suggests"(Keane, 2011: 1071).

Finally, we find strong gender differences when it comes to the institution of marriage, which matches real-world findings. As already shown by survey data, married men provide more work effort than cohabiting men, while married women provide less work effort than cohabiting women. Sorting into specialization and the impact of specialization-enhancing institutions for married couples in Germany are found to explain this observation. First, as Barg and Beblo (2012) argue, sorting into specialization happens if couples get married who have already planned to divide housework and labor market work in traditional ways. In short, the ones who anticipate specialization marry. Consequently, we interpret our finding as exactly a display of the differences in specialization preferences between married and cohabiting men and women. Second, since for each couple the same rules of the game apply in our experiment – i.e., both married and unmarried couples' income is charged to both joint and individual taxation – the experimental institutions cannot be the driver of the observed behavioral differences between them. However, there could be an indirect influence from the taxation system given outside the lab: Joint taxation is, in the real world, applied only to married couples, whereas individual taxation is applied to cohabiting couples. Following Lewis (2002), who argues that institutions could not only offer incentives but also shape attitudes and nourish the social norm of a male breadwinner, as well as Akerlof and Kranton (2000), who advocate that social norms translate into norm-conforming behavior, we could conclude that the different behavior of married couples in the lab is a result of experiencing such specialization-enhancing institutions outside the lab. Kimbrough and Vostroknutov (2016) have showed that participants might import their

[24] We adopted this PAIRFAM (Panel Analysis of Intimate Relationships and Family Dynamics) question in our post-experimental questionnaire: "Looking at both housework and paid work, how fair is the division of labor between you and your partner?".

social norms from outside into the lab, an effect that is more pronounced, the higher an individual's "strength" of adherence to social norms, i.e., the norm sensitivity outside the lab. This is exactly what we see: A stronger specialization for married couples despite the fact that the rules are the same for married and cohabiting couples.

Furthermore, we add a qualitative observation made during the experiment that supports the assumption of the existence of the social norm of the male breadwinner in the sample: During the experiment, we had the impression that the subjects were puzzled when we assigned potentially atypical roles (female primary and male secondary earner). The couples frequently asked whether anything went wrong or whether the assignment was truly random when they recognized their wages (and thus their roles) from reading the instructions. Some women asked if they could trade the task, and some men asked if they could help their female partner. In fact, one couple cheated by exchanging the mazes when the experimenter was out of sight (we dropped this observation). Remarkably, this kind of behavior did not appear in the experimental group with presumably typical roles, i.e., couples consisting of male primary earners and female secondary earners.

While the effect of joint taxation on subjects' work effort is not very strong in our experiment, the effect on the allocation of housework is. Individual taxation significantly increases the probability of the primary earners taking on this task. Thus, in view of the world of gender gaps that result in the majority of couples consisting of male primary and female secondary earners, the abolition of joint taxation in favor of individual taxation could be a fruitful way to relax women's constraints. It could contribute to reducing the unequal division of family chores that otherwise "is likely to be the primary determinant of most if not all the gender differences in the labor market" (Ichino, 2014: 41).

2.7 Appendix

2.7.1 Results

Table 4: Regression Results for Taking Over the Unpaid Task (Full Table)

DV= unpaid task	(1)	(2)	(3)
joint	0.223**	0.223**	0.223**
	(0.098)	(0.099)	(0.100)
male	0.108	0.103	-0.052
	(0.129)	(0.132)	(0.185)
joint x male	0.010	0.018	0.018
	(0.134)	(0.138)	(0.139)
pe	-0.272*	-0.310**	-0.302**
	(0.139)	(0.140)	(0.139)
joint x pe	-0.465***	-0.421***	-0.421***
	(0.137)	(0.143)	(0.143)
male x pe	-0.084	-0.035	-0.063
	(0.184)	(0.185)	(0.187)
joint x male x pe	0.049	-0.003	-0.003
	(0.190)	(0.196)	(0.197)
stage	-0.013	-0.009	-0.009
	(0.067)	(0.069)	(0.070)
stage x pe	0.050	0.040	0.040
	(0.095)	(0.098)	(0.098)
age		-0.003	-0.004
		(0.003)	(0.003)
education		0.029	0.025
		(0.081)	(0.082)
employed		-0.155	-0.160
		(0.117)	(0.119)
children		0.018	-0.036
		(0.102)	(0.116)
income		0.021	0.026
		(0.041)	(0.041)
east		-0.077	-0.098
		(0.123)	(0.124)
married			-0.039
			(0.160)
married x male			0.231
			(0.169)
constant	0.599***	0.760***	0.877***
	(0.102)	(0.203)	(0.218)
observations	226	216	216
R-squared	0.321	0.329	0.340
adj R-squared	0.293	0.279	0.284

Note: Displayed are coefficients of pooled linear probability models with cluster robust standard errors on the individual decision to take over the unpaid task ((0-1) choice counted on the individual level, jointly agreed upon within the couple). * p<.100, ** p<.050, *** p<.010.

Table 5: Regression Results for Work-Effort (Full Table)

DV= work-effort	(1)	(2)	(3)
joint	1.643	1.856	1.849
	(1.541)	(1.485)	(1.488)
pe	-14.74***	-12.13***	-11.64***
	(3.729)	(2.875)	(2.908)
joint x pe	-0.515	-1.599	-1.913
	(1.828)	(1.714)	(1.717)
male	-0.0250	-0.430	-6.056*
	(4.662)	(2.997)	(3.189)
joint x male	-2.710	-4.141**	-3.926**
	(2.102)	(1.995)	(1.968)
pe x male	1.708	0.716	-0.377
	(4.971)	(3.374)	(3.512)
joint x pe x male	2.295	4.725**	4.830**
	(2.417)	(2.294)	(2.258)
stage	0.793	0.297	0.332
	(0.875)	(0.917)	(0.922)
stage x pe	-2.959***	-2.600**	-2.672**
	(1.039)	(1.067)	(1.070)
unpaid	-7.525*	-8.402**	-8.376**
	(4.054)	(3.295)	(3.252)
unpaid x pe	8.888*	5.945	4.326
	(5.140)	(3.599)	(3.642)
unpaid x male	1.470	4.284	3.480
	(5.547)	(3.910)	(3.885)
unpaid x male x pe	-6.753	-5.029	-2.622
	(6.526)	(4.394)	(4.435)
age		-0.326***	-0.323***
		(0.045)	(0.047)
education		1.368	1.439
		(1.084)	(1.033)
employed		-1.430	-0.821
		(1.515)	(1.380)
children		-1.050	-1.018
		(1.572)	(1.637)
income		0.376	0.411
		(0.530)	(0.485)
east		-4.453***	-5.772***
		(1.387)	(1.469)
change		4.953**	4.895**
		(2.423)	(2.448)
change x pe		-7.724***	-7.939***
		(2.799)	(2.790)
married			-4.516*
			(2.318)
married x male			8.180***
			(2.277)
Constant	29.04***	44.44***	48.32***
	(3.393)	(4.300)	(4.260)
Observations	226	212	212
R-squared	0.347	0.671	0.700
adj R-squared	0.307	0.635	0.663

Note: Displayed are coefficients of pooled linear regression models with cluster robust standard errors at the individual level. * p<.100, ** p<.050, *** p<.010.

2.7.2 Instructions

General Instructions (page 1, for all participants)

Welcome and thank you for your participation! You are an important part of our study, which we are conducting with 250 people from the area of Frankfurt (Oder). You are participating as a couple because we are interested in how you jointly, together, make decisions. The study consists of two rounds and a questionnaire. In both rounds you will make decisions and solve tasks. Your decisions and the performance on these tasks determine your income. At the end you will receive the income of one round, which will be chosen randomly. The questionnaire is important for our analysis. Therefore, we ask you to complete it carefully. After filling out the questionnaire, you will receive a voucher from the supervisor that entitles you to collect your payoff in the next room.

It is essential that you read the instructions carefully. In case of any doubts or concerns, please address your questions to the supervisor. Please indicate your concern by raising your hand. We will come to your seat in order not to disturb the other participants. Your anonymity is assured during all times. As participant, you will receive a code number that is written in the upper right-hand corner of each page.

Stage Instructions (page 2, identical in both stages, SE)

Task:

Your task is to solve mazes printed on paper within 15 minutes. The aim of the paper-and-pencil game is to draw a route through the maze from the start to finish without being hindered by dead ends. The maze is solved after having drawn a continuous line from the starting point (S) to the finishing point (F). The inner and outer frames of the maze should not be touched or crossed by the pencil line. There are easier and harder mazes. You will have to solve the easy mazes. Your partner's task will be to solve the harder ones. Every maze that is solved correctly yields earnings as follows: Your salary for each easy maze is 0.50 €.

Taxation:

As in real life, your income is taxed. Your net income (salary minus taxes) on each solved maze depends on the total number of solved mazes and the tax rate. The more mazes you solve, the higher your income, and also the higher the tax burden. In other words, the tax burden is progressive. A table will show you how the tax affects

- your net wage per solved maze,

- the tax burden per solved maze, and
- your accumulated income.

This way, you always know what wage to expect when you decide to solve another maze. Shortly, we will show you how to read the table's information. In case of any doubts or concerns, please address your questions to the supervisor. Please enjoy the drinks, cookies and the magazines as pastime!

Stage Instructions (page 2, identical in both stages, PE)

Task:

Your task is to solve mazes printed on paper within 15 minutes. The aim of the paper-and-pencil game is to draw a route through the maze from the start to finish without being hindered by dead ends. The maze is solved after having drawn a continuous line from the starting point (S) to the finishing point (F). The inner and outer frames of the maze should not be touched or crossed by the pencil line. There are easier and harder mazes. You will have to solve the harder mazes. Your partner's task will be to solve the easy ones. Every maze that is solved correctly yields earnings as follows: Your salary for each hard maze is 1.50 €.

Taxation:

As in real life your income is taxed. Your net income (salary minus taxes) on each solved maze depends on the total number of solved mazes and the tax rate. The more mazes you solve, the higher your income, and also the higher the tax burden. In other words, the tax burden is progressive. A table will show you how the tax affects

- your net wage per solved maze,
- the tax burden per solved maze, and
- your accumulated income.

This way, you always know what wage to expect when you decide to solve another maze. Shortly, we will show you how to read the table's information. In case of any doubts or concerns please address your questions to the supervisor. Please enjoy the drinks, cookies and the magazines as pastime!

Tax Description and Effort-Income Table (page 3, SE, individual taxation)

Note:

The more mazes you solve and thus the more income you generate, the higher is your tax burden. That is why your net piece rate declines the more mazes you solve.

The income of both partners is taxed equally.

Example:

The person who solves the easier mazes and thus receives a lower wage earns a net piece rate of 0.50 € when solving the 5th maze. For the 15th maze, the person receives 0.40 €. With 15 solved mazes, the person receives 6.90 € in total.

The person who solves the more difficult mazes and therefore receives a higher wage earns a net piece rate of 1.20 € for the 5th maze. For the 15th maze, the person receives 0.30 €. With 15 solved mazes, the person receives 13.50 € in total.

Number of mazes	1- 3	4-6	7-9	10-12	13-15	16-18	19-21	22-24	25-27	28-30	31-33	34-36	37-39
Net wage	0.50 €	0.50 €	0.50 €	0.40 €	0.40 €	0.40 €	0.30 €	0.30 €	0.30 €	0.20 €	0.20 €	0.20 €	0.10 €
Tax burden	0.00%	0.00%	0.00%	20.00%	20.00%	20.00%	40.00%	40.00%	40.00%	60.00%	60.00%	60.00%	80.00%
													
Net income	1.50 €	3.00 €	4.50 €	5.70 €	6.90 €	8.10 €	9.00 €	9.90 €	10.80 €	11.40 €	12.00 €	12.60 €	12.90 €

Tax Description and Effort-Income Table (page 3, PE, individual taxation)

Note:

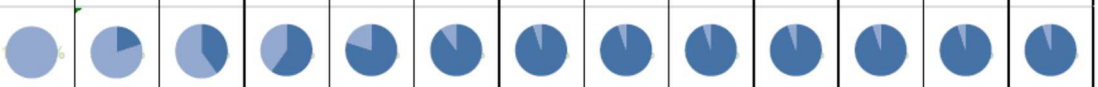
The more mazes you solve and thus the more income you generate, the higher is your tax burden. That is why your net piece rate declines the more mazes you solve.

The income of both partners is taxed equally.

Example:

The person who solves the easier mazes and thus receives a lower wage earns a net piece rate of 0.50 € when solving the 5th maze. For the 15th maze, the person receives 0.40 €. With 15 solved mazes the person receives 6.90 € in total.

The person who solves the more difficult mazes and therefore receives a higher wage, earns a net piece rate of 1.20 € for the 5th maze. For the 15th maze, the person receives 0.30 €. With 15 solved mazes, the person receives 13.50 € in total.

Number of mazes	1- 3	4-6	7-9	10-12	13-15	16-18	19-21	22-24	25-27	28-30	31-33	34-36	37-39
Net wage	1.50 €	1.20 €	0.90 €	0.60 €	0.30 €	0.15 €	0.08 €	0.08 €	0.08 €	0.08 €	0.08 €	0.08 €	0.08 €
Tax burden	0.00%	20.00%	40.00%	60.00%	80.00%	90.00%	95.00%	95.00%	95.00%	95.00%	95.00%	95.00%	95.00%
													
Net Income	4.50 €	8.10 €	10.80 €	12.60 €	13.50 €	13.95 €	14.18 €	14.40 €	14.63 €	14.85 €	15.08 €	15.30 €	15.53 €

Tax Description and Effort-Income Table (page 3, SE, joint taxation)

Note:

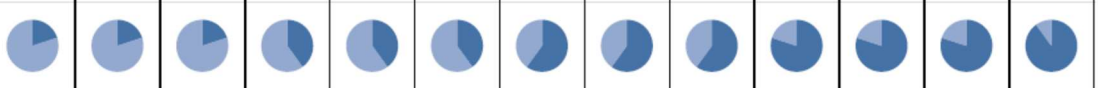
The more mazes you solve and, thus, the more income you generate, the higher is your tax burden. That is why your net piece rate declines the more mazes you solve.

The income of each partner is taxed differently.

Example:

The income of the person with the easier task is taxed higher. The income of the person with the difficult task is taxed at a lower level The person who solves the easier mazes and thus receives a lower wage earns a net piece rate of 0.40 € when solving the 5th maze. For the 15th maze, the person receives 0.30 €. With 15 solved mazes, the person receives 5.40 € in total.

The person who solves the more difficult mazes and therefore receives a higher wage, earns a net piece rate of 1.50 € for the 5th maze. For the 15th maze, the person receives 0.60 €. With 15 solved mazes, the person receives 17.10 € in total.

Number of mazes	1- 3	4-6	7-9	10-12	13-15	16-18	19-21	22-24	25-27	28-30	31-33	34-36	37-39
Net wage	0.40 €	0.40 €	0.40 €	0.30 €	0.30 €	0.30 €	0.20 €	0.20 €	0.20 €	0.10 €	0.10 €	0.10 €	0.05 €
Tax burden	20.00%	20.00%	20.00%	40.00%	40.00%	40.00%	60.00%	60.00%	60.00%	80.00%	80.00%	80.00%	90.00%
													
Net income	1.20 €	2.40 €	3.60 €	4.50 €	5.40 €	6.30 €	6.90 €	7.50 €	8.10 €	8.40 €	8.70 €	9.00 €	9.15 €

Tax Description and Effort-Income Table (page 3, PE, joint taxation)

Note:

The more mazes you solve and thus the more income you generate, the higher is your tax burden. That is why your net piece rate declines the more mazes you solve.

The income of each partner is taxed differently.

Example:

The income of the person with the easier task is taxed higher. The income of the person with the difficult task is taxed at a lower level. The person who solves the easier mazes and thus receives a lower wage earns a net piece rate of 0.40 € when solving the 5th maze. For the 15th maze, the person receives 0.30 €. With 15 solved mazes, the person receives 5.40 € in total.

The person who solves the more difficult mazes and therefore receives a higher wage earns a net piece rate of 1.50 € for the 5th maze. For the 15th maze, the person receives 0.60 €. With 15 solved mazes, the person receives 17.10 € in total.

Number of mazes	1- 3	4-6	7-9	10-12	13-15	16-18	19-21	22-24	25-27	28-30	31-33	34-36	37-39
Net wage	1.50 €	1.50 €	1.20 €	0.90 €	0.60 €	0.30 €	0.15 €	0.08 €	0.08 €	0.08 €	0.08 €	0.08 €	0.08 €
Tax burden	0.00%	0.00%	20.00%	40.00%	60.00%	80.00%	90.00%	90.00%	90.00%	90.00%	90.00%	90.00%	90.00%
													
Net income	4.50 €	9.00 €	12.60 €	15.30 €	17.10 €	18.00 €	18.45 €	18.68 €	18.90 €	19.13 €	19.35 €	19.58 €	19.80 €

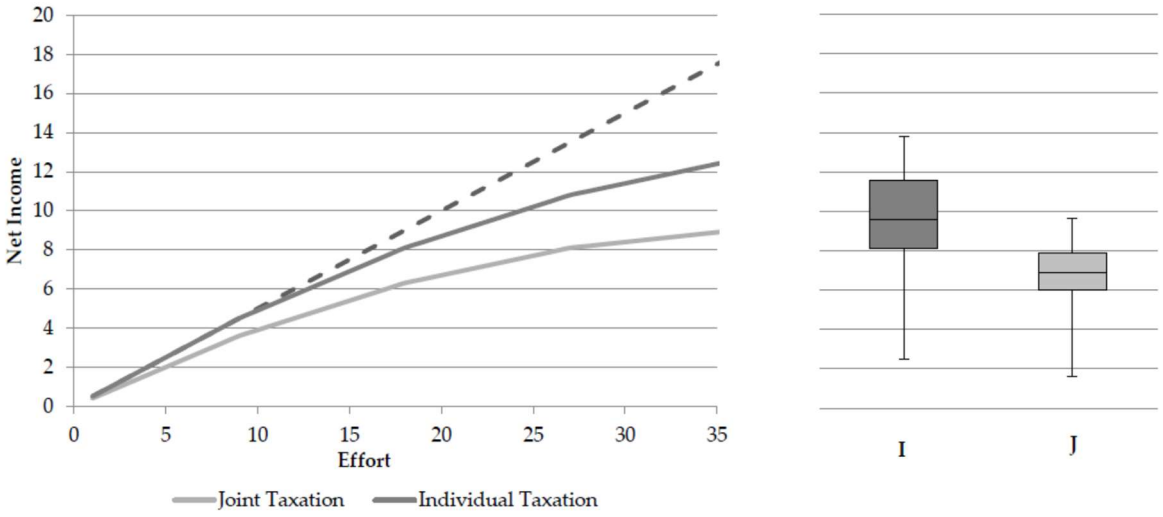
Decision on the Uncompensated Task (page 4, identical in both stages, for all participants)

Before you can start, you are asked to decide jointly if you by yourself or your partner by him/herself will solve an unpaid task. This task is not difficult at all, and no special previous knowledge is required. Whoever you decide on will have to solve the task by him/herself alone. The solving takes 3 minutes. It thereby shortens the available total time for solving the mazes by 3 minutes. After finishing this task, this person can also start to solve the paid task and thus generate income. Please check the box if you are the person solving this mandatory task.

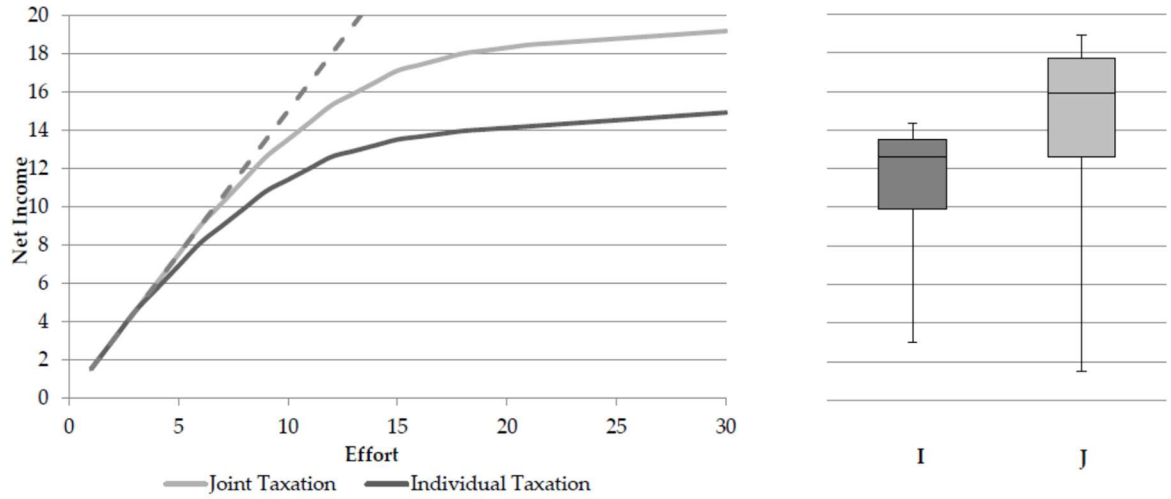
2.7.3 Tax schemes

Figure 6: Tax Schedule and Actual Net Income Distribution

(a) Secondary Earner



(b) Primary Earner



2.7.4 Sample descriptives

Table 6: Sample Descriptives

	Women	Men	p-value (gender differences)
Age	41.97	44.83	.310
	(15.52)	(15.84)	
Share married	0.73	0.73	1
Living together (years)	17.51	17.95	.880
	(15.46)	(15.60)	
Share of couples with at least one child	0.73	0.73	1
Highest educational attainment	2.14	2.28	.380
	(0.91)	(0.89)	
Full-time employment	0.40	0.50	.031
Part-time employment	0.18	0.05	.023
Personal gross income	3.08	3.46	.087
	(1.25)	(1.21)	
Tax experience	0.94	0.95	.700
N	62	62	

Note: Displayed are means and standard deviations. "Highest educational attainment" denotes mean of the highest educational attainment (0=none, 1=vocational education without A-level, 2=vocational education with A-level, and university degree=3). 'Full-time employment' and 'part-time employment' denote the share of participants with this employment status. Others are either in pension, unemployed, in maternity leave, in education programs or working in marginal employment. Personal gross income is classified in categories as follows: 1=0€-500€, 2=501€-1000€, 3=1001€-2000€, 4=2001€-3000€, 5=<3001 €. "Tax experience" denotes the share of people having experience with income taxes, being defined as experienced if they had ever made a tax declaration or know which tax class they have.

3 Self-Selection and Conditional Performance: The Gender Pay Gap in a Choice Experiment[25]

3.1 Introduction

In 2015, the raw gender pay gap in the European Union amounted to 16.3%, with Germany lying well above the average at 22.0% (Eurostat, 2018b). In Blau and Kahn's (2017) literature survey on the determinants of the gender pay gap, explanations range from the commonly observed factors – such as differential labor market participation (extensive and intensive margin), education or job experience – to the more recently studied in experimental economics – such as psychological attributes and preferences. The gender pay gap investigated in experimental labor markets (Heinz et al., 2016; Schwieren, 2012) mirrors the real-world observation that women earn less than men, and shows that both supply and demand factors contribute to the gap. In our experiment, we are able to isolate the supply-side factors. From a labor market perspective, we can fully rule out favoritism by the employer and are able to discriminate solely among the supply-side determinants. We thereby focus on the impacts of individual selection into payment schemes and returns to these payment schemes on the gender pay gap. Payment schemes are characterized by the extent of competition and risk involved, as these attributes have been shown to be highly relevant for labor market outcomes (Shurchkov and Eckel, forthcoming).

We use a large-scale online experiment run all over Germany. In total, 883 participants performed a real effort task after having decided on their preferred compensation contract. The choice task consisted of a series of choice sets, with a compulsory choice between two contract types in each of the choice sets: one always providing a linear piece-rate payment for the task (solving mazes) and the other one varying either over (i) a bonus payment based on a performance threshold, (ii) a payment based on competition with all participants or (iii) a payment based on competition with participants of the same gender. Additional variations were introduced by (iv) the difficulty level of the task and (v) a higher payment spread (risk premium) that increases potential gains but also losses with regard to the alternative linear piece-rate compensation.

We find that the total gender pay gap resulting from the participants' preferred type of contract and their later performance in solving mazes amounts to 23 %. Since compensation is based on a randomly drawn choice set at the end of the experiment and is fully independent of the

[25] This paper was developed together with Miriam Beblo, Denis Beninger and Norma Burow.

participants' gender (all participants anonymously perform the same task under the same instruction set), the gap cannot be attributed to systematic favoritism of the male participants. An Oaxaca-Blinder decomposition reveals that some part of the pay gap may be attributed to the participants' ability in the task, measured by individual performance in test runs at the beginning of the experiment, prior to the choice task. Still, most of the gap can be attributed to the types of contracts female and male participants choose (*selection effect*) as well as to the returns on these contracts (*contract effect*) – i.e., the participants' conditional performance in the maze task once their random choice set has been drawn and they have been informed about the applying compensation scheme.

Our experiment contributes to the existing literature in three ways. First, studies on the gender pay gap along payment schemes often concede discrimination as part of the explanation (Grund, 2015; Manning and Saidi, 2010), but are not able to control properly for the impact of discrimination. We try to rule out employer-side discrimination by design and isolate the effects of supply-side determinants on gendered earnings. Our approach can thus be seen as a complement to audit studies (i.e., Goldin and Rouse, 2000) that aim to disclose discriminatory practices on the demand side while ruling out supply-side effects (see Azmat and Petrongolo, 2014 for an overview). Second, with having information on the participants' performance in test runs prior to the choice and the performance task, we have a good measure of their potential productivity. Third, as we perform an online experiment, we are able to combine the features of a controlled setting with a relatively large and heterogeneous sample, thus addressing internal as well as external validity concerns. In addition, our setting can be considered as being close to a real-world situation, as non-student subjects participated on their personal computers and in their personal surroundings.

The remainder of the paper is set out as follows: Section 3.2 reviews the relevant literature. Sections 3.3 and 3.4 describe the design of the experiment and our sample. Section 3.5 reports the experimental compensation gap and investigates it from various angles. This is followed by robustness checks in Section 3.6 and conclusions in Section 3.7.

3.2 Related literature

The link between the competitiveness of payment schemes and gendered earnings has been addressed in a number of non-experimental and experimental studies. Starting with the survey evidence, Manning and Saidi (2010) investigate the impact of different payment schemes on the gender pay gap by using the British Workplace Employment Relations Survey (WERS). They find women less likely to work under performance pay contracts, but the effect of

performance pay on earnings is modest. Based on two cohorts of the U.S. National Longitudinal Surveys of Youth (NLSY 79 and NLSY97), McGee et al. (2015) also show that women in the United States are less likely to be employed in jobs with the most competitive forms of performance pay (commissions and bonuses), but this explains only a small fraction of the observed gender pay gap. Grund (2015) detects a relatively high gender pay gap within the relatively homogenous group of professionals and managers of the German chemical sector (2008–2012), with the gap being more pronounced for bonus payments than for fixed salaries. Although these studies suggest that the size of the gender pay gap may also depend on the contract type, little is known about whether this is driven by supply- or demand-side factors.

Focusing on supply-side explanations, there seems to be ample evidence that women shy away from competitive environments.[26] In the influential experimental study by Niederle and Vesterlund (2007) participants have to choose compensation schemes before performing on a task. They find that women prefer a non-competitive compensation scheme (piece-rate pay) for this task over a competitive compensation scheme (tournament) while the majority of men choose the competitive payment scheme.[27] Datta Gupta et al. (2013: 832), hence, conclude that "women compet[e] too little."

Differences in preferences for competition seem to translate into different career choices. Kleinjans (2009) shows that women's greater distaste for competition decreases their educational achievement, which explains part of the gender segregation in occupational fields. Buser et al. (2014) use data from classroom experiments on preferences for competition as an explanation for educational choices. According to their results, about 20% of the observed gender differences in educational choices can be attributed to gender differences in competitiveness.

The empirical evidence on the impact of a competition-based payment scheme on an individual's performance is mixed. Niederle and Vesterlund (2007) find a positive impact of a tournament scheme on both men's and women's performance but no gender difference at all. Gneezy et al. (2003) detect gender differences with an increase in performance for men when pay is competitive compared to a piece-rate pay. In a meta-analysis on 18 experimental studies, Bandiera et al. (2016) test whether women and men differ in their behavioral responses to performance pay. They find that performance pay induces both groups similarly to improve performance. Besides this experimental evidence, some authors refer to administrative data in

[26] See Nelson (2015) for a general critique on such statements about gender differences.
[27] See Shurchkov and Eckel (forthcoming) for an overview of experiments on gender differences in risk and competition.

studying the impact of competition on performance. Among these, Jurajda and Münich (2011) examine multiple university entry exams taken by the same individuals in the Czech Republic. They find that female students perform less well compared to similar male students when the admission rate of the institution is relatively low. When admission rates are higher and the situation can be perceived as being less competitive, the gender gap in the success rate vanishes. Ors et al. (2013) use a series of academic examinations that differ in the level of competition. While women outperform men in the baccalauréat exam, the reverse holds true for the entrance exam for the same cohort for admission to the Master of Science in management, which the authors characterize as being more competitive. Iriberri and Rey-Biel (2016) investigate performance in a math contest in Spain. Although the female participants yield higher math grades at school, men outperform women in the contest. Moreover, the gender gap increases with increasing competitiveness.

In our experiment, we build on this evidence and deal with some of its shortcomings. We rule out employer-side discrimination and shift the focus to isolating the effects of supply-side determinants on gendered earnings. In particular, we investigate the contracts chosen by women and men and the performance induced by these contracts.

3.3 Experimental design

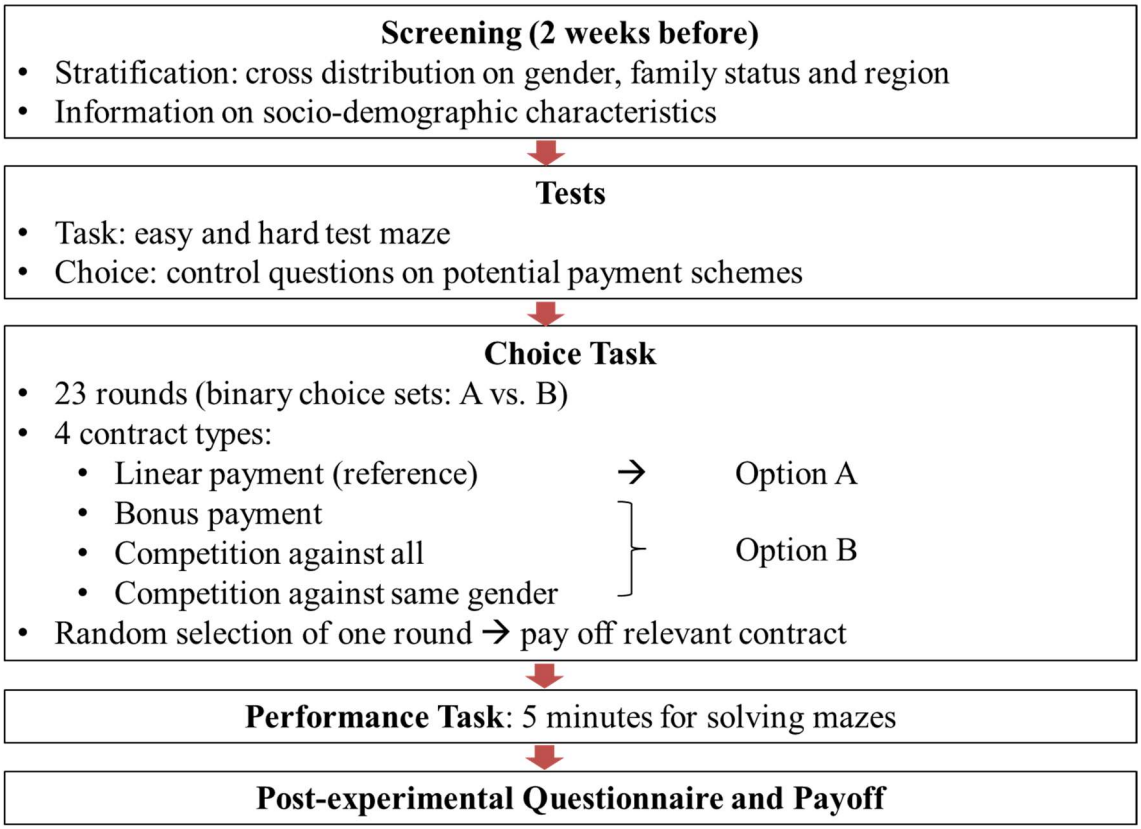
We used a non-standard subject pool recruited through random sampling from the original data panel of a market research institute. The full panel includes basic and confidential information on socio-demographic and economic characteristics of about 250,000 clients.

We chose an online solution, so that participants take part in their familiar environments, which nonetheless ensure anonymity. In addition, the large sample size of our experiment addresses one of the major criticisms against lab experiments: that the results are limited to the sample population which is often very small (Reips, 2000). It may be argued that online experiments suffer from lack of control; however, Horton et al. (2011) show that results of laboratory experiments on framing, social preferences and labor supply can be replicated online.

Figure 7 pictures the experimental procedure. In a screening step, a market research institute selected a gross sample of 1,444 people by a uniform stratification procedure based on gender, family status and region for our purpose. Those selected filled out a questionnaire on their socio-demographic characteristics as well as their risk attitudes and were invited to participate in “a scientific study.” Our full net sample includes 883 people, almost equally distributed across gender (men vs. women), family status (single vs. in a relationship) and region (East vs. West).

In the test part, participants were introduced to the task by solving one easy and one difficult trial maze. They were subsequently informed about their individual time performance on the screen to make sure that they knew their absolute productivity. Afterward, they answered three control questions on hypothetical contracts to ensure that they understood how the payment worked in the experiment. Before the choice task, half of the participants were asked about their gender. To strengthen the gender identity priming, a screen wiper with a pictogram of a man or a woman appeared constantly on the screen for the remainder of the experiment. Although the priming turned out to have no systematic effect on the participants' behavior (Schröder et al., 2017), we include it as a control variable in our estimations.

Figure 7: Experimental Procedure



Note: Although in some of the 23 choice rounds option B does not imply bonus payment, competition against all or competition against same gender, this applies to our set of meaningful rounds that we specify in Section 4 below.

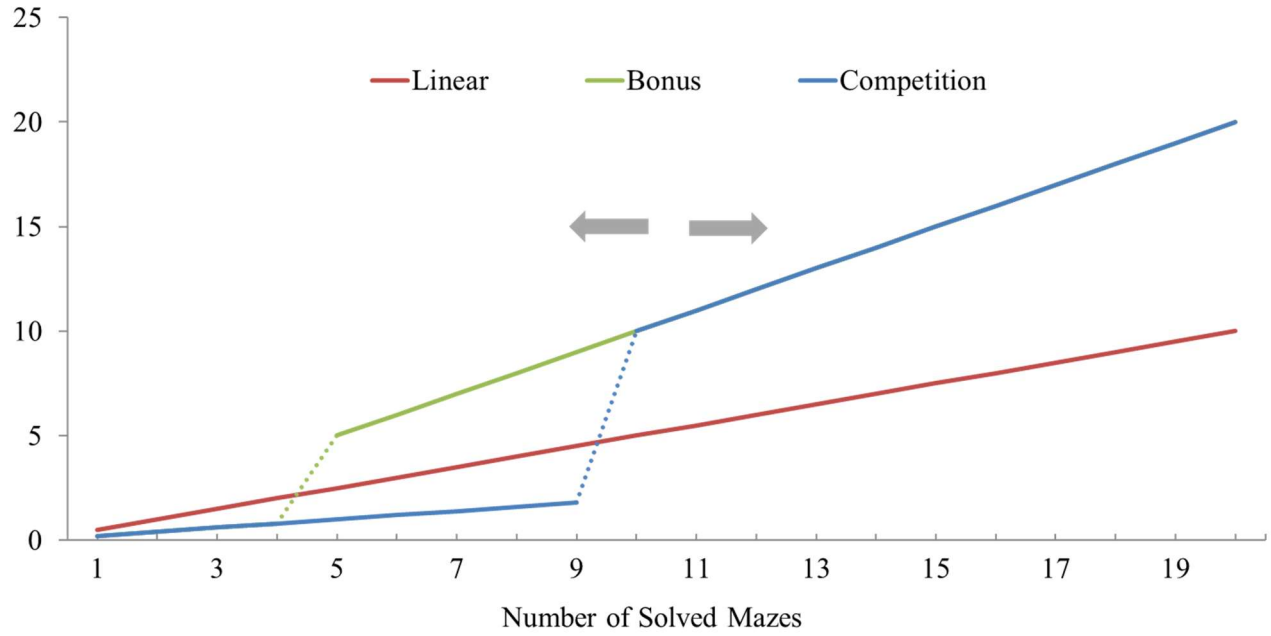
In the choice task, participants were asked to choose one of two options (A vs. B) in 23 successively presented choice sets.[28] The choices reveal their preferences regarding different

[28] A selection of choice sets can be found in the Appendix.

compensation contracts. There are four contract types (linear, bonus, competition against all participants, competition against same gender), visualized in Figure 8.

In a linear payment contract, subjects receive a piece-rate pay for each maze solved. In the bonus contract a participant receives a high piece-rate payment if she meets the threshold of five mazes or more but a low piece-rate payment if she fails to reach that threshold. This "threshold feature" may be compared to bonuses paid in the real world when a certain performance target, whether fixed or self-determined, is reached (Dalton et al., 2015).[29]

Figure 8: Contract Types



Note: Displayed are expected payoffs (low wage spread) depending on the number of solved mazes and the compensation contract. A person who solves, for example, seven mazes and chooses the linear payment contract receives €0.50 per maze. In the bonus contract, this person will receive €1 per solved maze since her performance exceeds the threshold of five mazes and €0.20 if she does not meet the threshold. In the competition contract, she receives €1 per maze if she places herself among the best 30%. Since the 30th percentile of the whole sample was ten mazes in our experiment, she receives only €0.20 per maze. The arrows indicate that the performance threshold in competition is not fixed; it depends on the performance of the competitor's group, which is either all participants or participants of the same gender.

The contract-type competition distinguishes between competition against all participants and competition against a same-gender group. A participant receives a high piece-rate payment if she places herself among the best 30% of either all or the same-gender group but a low piece-

[29] According to the WorlddatWork (2016a, 2016b) over 85% of American privately held companies and 86% of American nonprofit or government organizations report that they use an annual incentive plan that includes bonuses (that are defined at the beginning of the business cycle). Moreover, paying individual bonuses is very common in the salesforce practice (Kräkel and Schöttner, 2016) .

rate payment if she fails. Unlike the bonus payment, the number of people who will be rewarded in this contract type is limited, in that only the top 30% will receive a bonus. Thus, when choosing this contract type, a participant enters competition over scarce resources, with individual performance being evaluated in relation to the other participants' performances. This corresponds to a tournament or "winner-takes-all" work environment. In every contract type, we additionally vary the difficulty level of the mazes (easy/hard) and the wage spread that impels a risk premium.

When all choices had been made, one out of the 23 choice sets was drawn randomly. Each participant was informed about the selected choice set and the respective contract type by a short notice appearing on the computer screen. The participants then had five minutes to solve as many mazes as possible (performance task). We chose mazes as the real-effort task since previous studies (for example, Datta Gupta et al., 2013; Gneezy et al., 2003) had found them to be gender neutral in performance. In particular, a maze task is simple to communicate and easy to understand and requires little knowledge or experience. It involves little randomness and allows measurement of performance either by the time needed to solve one maze or by the number of mazes solved correctly in a given period of time (see Gill and Prowse 2012). The payoffs were paid by bank transfer from the market research institute to the participants, to help guarantee the anonymity of the participants with regard to the authors.[30]

3.4 Our sample

For the purpose of this paper, we focus on choice sets with a meaningful and comparable reference option A i.e. linear piece rate, and where the difficulty level is stable over A and B. This selection leaves 14 choice sets and a sample of 522 participants. The participants' average age is 36. 55% of the women and 33% of the men live with children, the women less often have professional education (61% vs. 68% for men) and are less often in employment (70% vs. 84% for men). See more details in Table 10 in the Appendix.

[30] In addition to their performance-related compensation every participant received a "show-up fee" of €2.

Table 7: Structure of Choice Sets

Choice Set	A	B	
	Difficulty		Risk Premium
1	€0.5	€1	0
2	€0.5	€2	0
	Bonus		
3+4	€0.5	€0.2 €1	0
5+6	€0.5	€0.05 €2	1
	Competition (all)		
7+8	€0.5	€0.2 €1	0
9+10	€0.5	€0.05 €2	1
	Competition (same)		
11+12	€0.5	€0.2 €1	0
13+14	€0.5	€0.05 €2	1

Table 7 provides a detailed overview of the choice sets. Across all, reference option A offers a piece rate of €0.50 per solved maze. Option B is characterized by either a bonus payment scheme (bonus = 1), competition against all participants (competition (all) = 1) or competition against participants of the same gender (competition (same) = 1). These contract types go along with different risk premia. The low risk setting with a relatively small wage spread serves as a reference (risk premium = 0). The risk premium is characterized by a larger wage spread (risk premium = 1). The choice sets, 3+4, 5+6, 7+8, 9+10, 11+12 and 13+14, vary by difficulty level, which is stable over both options and thus not a matter of choice. To gather information on the participants' preferences regarding the difficulty level of the task, two piece-rate scenarios with differing difficulty are offered in choice sets 1 and 2. Here, option A entails solving easy mazes and earning €0.5 per solved maze and option B yields a higher piece rate of €1 or €2 for solving harder mazes (difficulty = 1).

3.5 Results

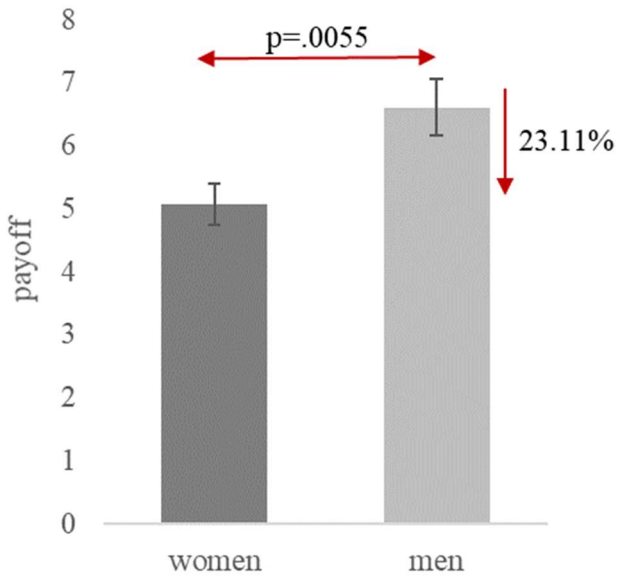
This section guides describes the gender differences in payoffs and performance observed in the experimental data, followed by a regression-based decomposition of the gender gap in experimental payoffs.

3.5.1 The experimental pay gap

The average gender pay gap resulting from the participants' preferred payment contracts and their conditional performance amounts to a statistically significant 23% (t-Test, $p<.01$), where women earn less than men (see Figure 9).

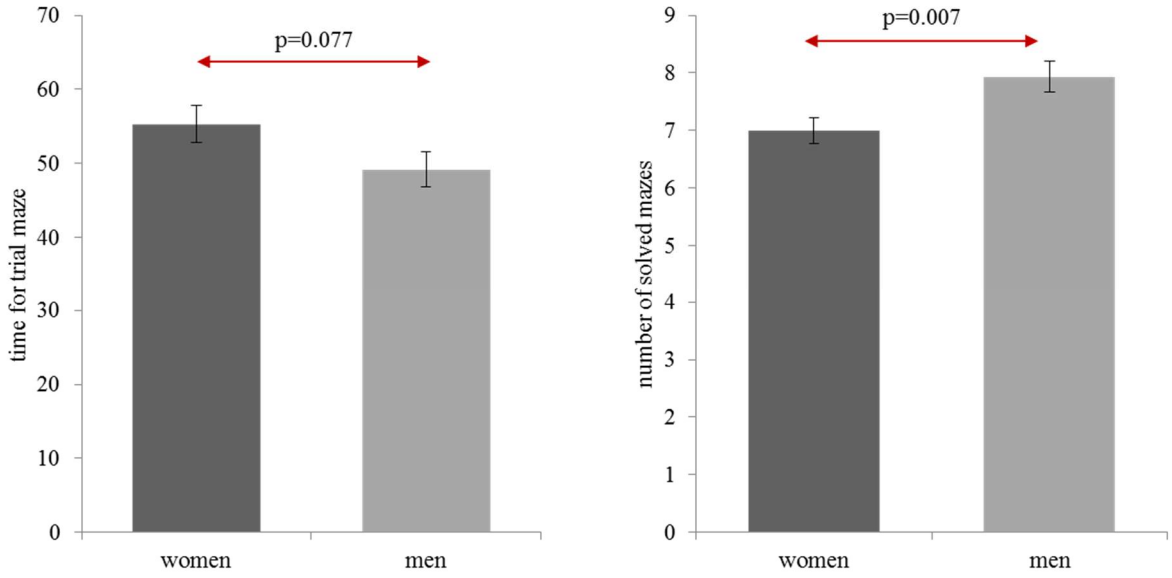
It may be claimed that part of this gap is due to performance differences in the maze task. Indeed, as shown on the left-hand side in Figure 10, women need, on average, significantly more time to solve a trial maze. On the right-hand side, we see the performance gap in the number of mazes solved in the real effort task. Similar to the gender gap in the trial maze, men solve more mazes in the real effort task. Interestingly, the performance gap conditional on the contract type in the real effort task is more pronounced, since this gap is significant at the 1% level, while the gender gap in the trial maze performance is significant only at the 10% level. Thus, the size of the gender difference in performance seems to be related to the payment scheme.

Figure 9: Average Payoff by Gender



Note: Displayed are the means in payoff for men and women. P-value is from t-test.

Figure 10: Average Performance by Gender

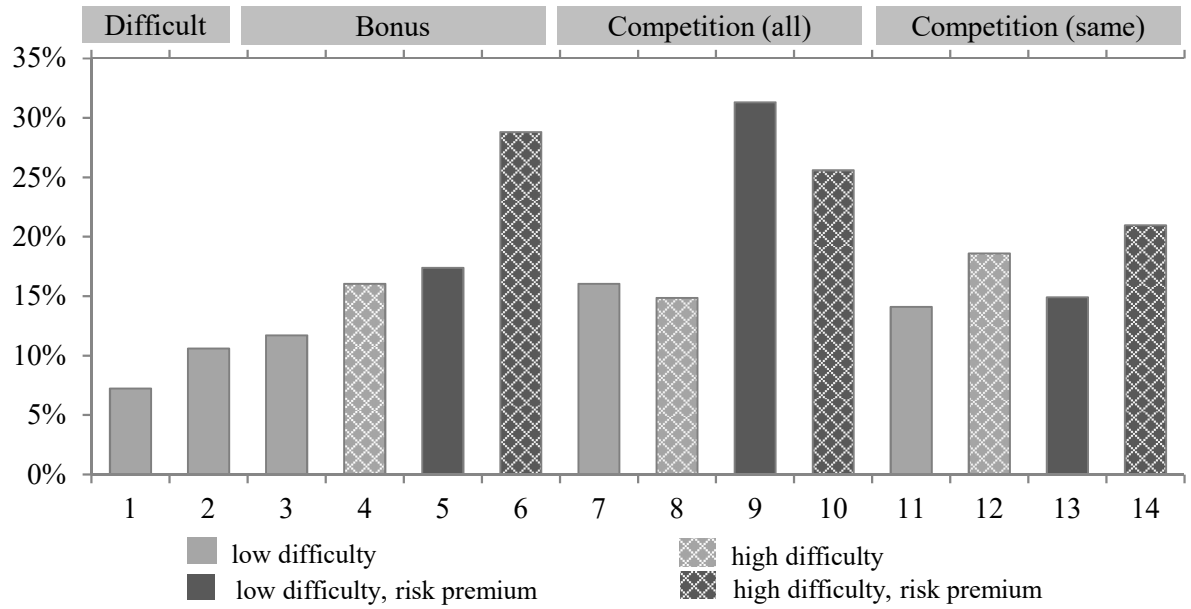


Note: Displayed are the average time for solving an easy and a hard trial maze and the average numbers of solved easy and difficult mazes in the real effort task. P-values are from t-tests.

3.5.2 Counterfactual pay gaps

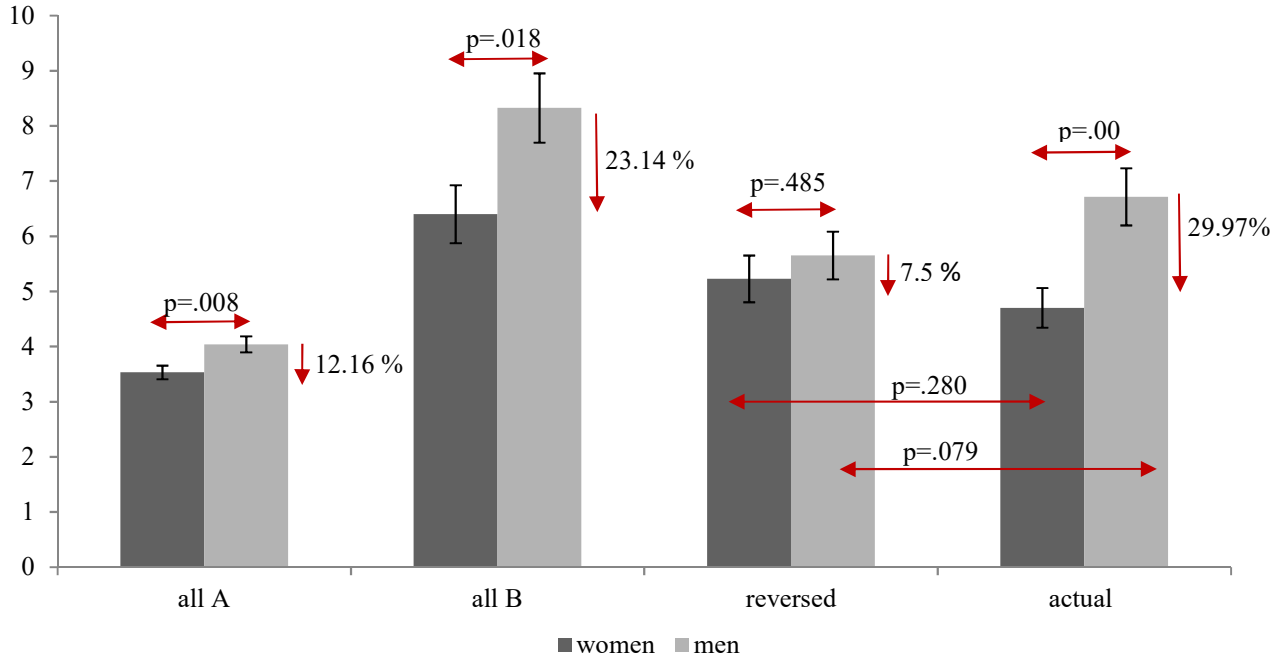
To investigate the relative importance of the contract types selected and the performance on the contracts, we create several hypothetical scenarios and their implied payment gaps. We first calculate a hypothetical gender pay gap for each choice set, given the participants' actual choices and their conditional performance in the experiment (Figure 11). The resulting gender pay gap is 17%, on average, and ranges from 6% to 29%. If we aggregate choice sets by contract attributes, the gap is largest for competition against all participants (19%; choice sets 7,8,9 and 10) and a risk premium is involved (21%; choice sets 5, 6, 9, 10, 13 and 14). The gap is lowest for choice sets where competition against the same gender is an option (14%; choice sets 11, 12, 13 and 14) or when the risk premium is absent (15%; choice sets 3, 4, 7, 8, 10 and 11). In addition, the gap is larger, on average, when the difficulty level is high (21%; choice sets 4, 6, 8, 10, 12 and 14) compared to low (15%; choice sets 3, 5, 7, 9, 11 and 13).

Figure 11: Gender Pay Gap in Each Round



We next assume that all participants choose the same option and compare actual and hypothetical payoffs (the first pair of pillars in Figure 12). One scenario assumes that all participants choose option A *(all A)* in each choice set, which is always the linear payment scheme. In another scenario, we assume that all participants choose B *(all B)* in each choice set. Comparing these hypothetical payoffs for men and women who are all assigned to option A or option B indicates that choosing option B is superior to option A, on average (second pair of pillars in Figure 12). Both men and women would have had the opportunity to increase their payoffs by always choosing option B. However, men take this chance more often; men's probability of choosing option B is higher compared to women's in most of the choice sets. Additionally, the gender pay gaps in both scenarios are statistically significantly different from zero. Since differences in selection into contracts are not relevant in these two hypothetical scenarios because we assigned identical contracts to men and women, the gender pay gaps can be explained only by men's better conditional performance.

Figure 12: Actual and Hypothetical Payoffs



Note: Displayed are the payoffs in three hypothetical scenarios and the actual payoff. Since we have to hold the difficulty level constant over both choice options, we have to drop two choice rounds (round 1+2). P-values are from t-tests.

In a third scenario, we simulate a situation where the participants are assigned to the alternative, not-chosen option to assess the effect of selection into contracts. We therefore assign participants who choose option A to option B and vice versa. The last four bars in Figure 12 compare the actual payoff (actual) with this counterfactual payoff (reversed). The gender pay gap drops from almost 30%[31] to 7.50% and is no longer statistically significant. From the fact that the actual gender pay gap is significantly different from zero and this hypothetical one is not, we conclude that the participants' choices and, hence, differences in preferences, are driving part of the gap.

The three hypothetical scenarios presented suggest that the observed gender gap in payments results from the decisions that women and men make when they can choose between contract types as well as from their performance in the task, once the attributes of a contract are given. In order to disentangle these two effects – the selection into a contract and the performance conditional to this contract – we undertake a decomposition analysis of the gender pay gap in the next section.

[31] This number differs from the one in Figure 9 since it does not include choice sets 1 and 2.

3.5.3 Oaxaca-Blinder decomposition

As we are interested in the differences in mean payoffs between men and women, the Oaxaca-Blinder decomposition (Blinder, 1973; Oaxaca, 1973) seems most suitable. It serves as a standard tool in the gender pay gap literature, making our results most comparable to previous findings.[32] Following the Oaxaca-Blinder approach, the gender gap is typically decomposed into two parts, as shown in Equation 1. The first component, called the explained part or the *endowment effect*, accounts for differences in observable characteristics. The second component is characterized as the unexplained part or the *remuneration effect*, as it reflects differences in remunerations to these characteristics between men and women.[33] For our purpose, we extend the basic Oaxaca-Blinder technique and focus on the behavioral components of gendered earnings *(choice)* while controlling for socio-economic characteristics *(char)* separately, as reflected in Equation 2.

$$w^M - w^F = (\bar{X}^M - \bar{X}^F)\beta^M + (\beta^M - \beta^F)\bar{X}^M \tag{1}$$

$$\text{raw gender pay gap} = \text{endowment effect} + \text{remuneration effect}$$

$$\Leftrightarrow (\bar{X}^M - \bar{X}^F|_{char})\beta^M + (\bar{X}^M - \bar{X}^F|_{choice})\beta^M + (\beta^M - \beta^F|_{char})\bar{X}^M + (\beta^M - \beta^F|_{choice})\bar{X}^M \tag{2}$$

$$\Leftrightarrow \text{endowment of obs. char.} + \text{selection effect} + \text{remuneration to obs. char.} + \text{contract effect}$$

Our experimental data enables us to decompose the difference in mean payoffs – that is, the raw gender pay gap – into four parts. The *effect of observable characteristics* and the *remuneration of observable characteristics* capture participants' socioeconomic characteristics. These variables include information on participants' ages, their family status (single vs. in a relationship), whether they have children in the household, the region where they live (East vs. West Germany), their employment status (full- and part-time employed vs. other), their level of education (low, mid, high and currently in education) and a measure of their general risk

[32] An overview of common decomposition methods as well as a discussion of the Oaxaca-Blinder technique can be found in Fortin et al. (2011).

[33] Since this interpretation has been subject to criticism (see Shurchkov and Eckel, forthcoming) and our experiment rules out the potential influence of discrimination by design, we do not adhere to the common interpretation of the unexplained part as being a measure of discrimination.

aversion (1–10 scale). To control for the subjects' ability we add a variable that measures the time participants needed to solve the trial mazes. [34]

The *selection* and the *contract effect* refer to the choices participants make during the experiment. The selection effect reflects differences in the payment contracts chosen by men and women – that is, differences in their selection into contract types. The contract effect reflects gender differences in individual performances conditional on the chosen contract. Since we control for potential differences in ability in the estimation, the contract effect can be considered as a performance reaction conditional on the contract type. Decomposing this way, we are able to distinguish factors that are purely supply-side-driven (selection and contract effects) from endowment and remuneration effects that may still potentially be biased by pre-market discrimination.

Figure 13: How Selection and Contract Effect Contribute to the Gender Pay Gap

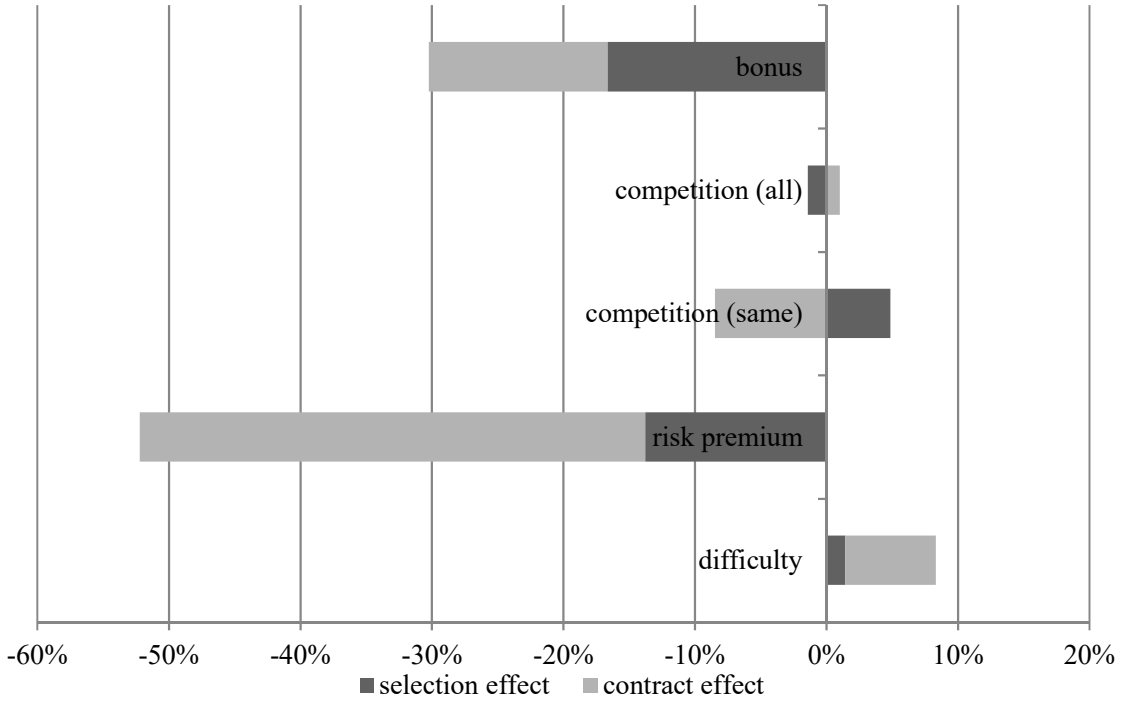


Figure 13 displays the selection and contract effects based on the decomposition analysis using all variables (the full table can be found in the Appendix).

The cumulative selection effect amounts to 26% of the gender pay gap observed in the experiment. This means that if women choose the same contracts as men, the gap will be about one-quarter smaller. The cumulative contract effect is negative, too, revealing that men's conditional performance surpasses women's conditional performance in the aggregate. This

[34] Further control variables include features of the design, such as the priming treatment.

conditional performance effect is responsible for about half of the experimental gender pay gap. Nonetheless, the effects vary across contract types, as described in the following.

The topmost bar indicates that the selection and contract effects of the bonus option are driving the gender pay gap. A gender difference in preferences for the bonus option – i.e., men choose this option more frequently – explains 17% of the gap. Furthermore, if women choose the bonus option as often as men, the gap will decrease. The negative contract effect discloses that if a bonus contract applies, men's conditional performance will surpass women's.

With regard to the competition contract, the effects differ between competition against all participants and competition against participants of the same gender. While women are slightly less prone than men to choose competition against all participants, they engage in competition relatively more often compared to men when the competitor group is female. While the effect of selection into a competition contract against all participants does not contribute to the gap, differences in selection into same-gender competition decrease the gap by about 5%. The respective contract effects differ, too. The contract effect of competition against all participants is positive (though small), meaning that women's performance in competition against all tends to reduce the pay gap, if anything. In contrast to this, women's higher preference for competition with the same gender does not seem to pay off, as it contributes to their lower pay.

The contract types also vary by difficulty level and risk premium. As the bottommost bar in Figure 13 indicates, the gender difference in preferences for the difficult mazes decreases the pay gap. Women are more likely to choose harder mazes than men. Thus, if men choose the difficult mazes as often as women do, the gender pay gap will be even larger. Since the contract effect is positive, too, implying that the women's choices of hard mazes are related to higher performance and pay, their preference for difficulty seems to pay off relatively well.

The effect of the risk premium exceeds all other effects in this figure. First, the difference in preferences for a risk premium explains 14% of the gap, with men having a stronger preference for this contract option. The contract effect is negative, too. If women perform on the risk premium as well as men do, the gender pay gap will be 38% smaller.

Possible explanations for the differences in preferences for bonus- and competition-involving contracts may be drawn from expected utility theory, in combination with the assumption that women more likely fear the risk of failing or that women believe in performance differences between men and women.[35] However, the large difference in the selection into the risk premium

[35] Burow et al. (2017) address the behavioral puzzle of women's preference for competition when competitors are women only, by using the same experimental data.

option does not seem compatible with this explanation. Since choice sets without and choice sets with a risk premium both carry the same rules and thus the same probability to fail (p), the risk premium yields a higher expected income, holding everything else constant. A person who chooses option B (i.e., bonus, competition with same gender or competition with all) in a risk premium-free setting should be expected do so in a risk premium setting, unless the person is extremely risk averse.[36] Interestingly, there is a tendency among all participants to restrain from a contract type if it comes with a risk premium (this applies particularly to the bonus contracts and competition against same gender).[37] Since in the risk premium setting not only gains but also losses are larger, this observation may be explained by loss aversion (Tversky and Kahneman, 1992). A loss-averse individual is more likely to reject a higher expected gain when it comes at the cost of a higher potential loss. Our observation, that particularly the female participants react to the risk premium and thereby expose more loss aversion is consistent with previous studies (Gächter et al., 2010; Rau, 2014; Schmidt and Traub, 2002). The selection effect of a risk premium amounts to 14%, meaning that 14% of the gender gap in pay can be explained by women being more loss averse than men. It is unlikely that this effect is induced by a gender difference in general risk aversion for two reasons: First, we control for general risk aversion by using the participants' stated risk preferences in our estimations. Second, the selection effect of the risk premium is driven by high earners, for whom the risk of losing is higher than it is for low earners. When we divide the sample into quartiles along the experimental payoff distribution, it becomes evident that selection into risk premium happens in the third (28%) and fourth quartiles (57%) only.

The contract or conditional performance effect, which is negative and relatively large in the aggregate, especially in the risk premium, is also worth investigating in detail. Remember that gender differences in ability should not be the reason, since we control for individual ability by using the participants' performance in the trial mazes. Why do the men still outperform the women most in this particular environment? The literature suggests a "competitive pressure" (Iriberri and Rey-Biel, 2016) or stereotype threat (Gneezy et al., 2003; Niederle and Vesterlund, 2007). Individuals experience a stereotype threat in a situation when they are at risk of confirming a negative stereotype about their social group (Steele, 1997). The motivation to disconfirm the negative stereotype about the group then leads to underperformance. As Spencer

[36] $p * 0.05 + (1 - p) * 2 \geq p * 0.2 + (1 - p) * 1$ is true for $p \geq .9$. This means that a risk-neutral participant will avoid option B when it comes with a risk premium but will opt for it under low risk, only if she expects a probability to fail that is equal to or larger than 90%.

[37] The likelihood of choosing the bonus contract when a risk premium applies decreases by 2 percentage points and the likelihood of choosing competition with same gender by 0.5 percentage points, on average.

et al. (2016) argue, a stereotype threat does not necessarily impair performance in all tasks or every situation. Only when tasks are challenging and an "extra pressure to succeed" (Spencer et al., 2016: 420) is brought into play is a stereotype threat said to harm the performance of the relevant group. A stereotype threat has been frequently observed to harm women's performance in competitive environments and thereby drive a gender gap. In our experiment, we expect the same mechanism to work in the risk premium environment as has been found for competitive environments so far. The risk premium can be considered an extra pressure to succeed given the risk of falling back from €2 to €0.05 per maze in case of failing (i.e., not achieving five mazes or placing among the best 30%). We therefore advocate that the stereotype threat effect additionally harms women's performance relative to men's and is thus responsible for at least 38% of the observed gender pay gap in our experiment (see contract effect of risk premium). Only in cases where loss aversion matters less, as with the attribute difficulty, for example, does the contract effect prove positive.

3.6 Robustness checks

As pointed out earlier, the interpretation of our findings relies on two assumptions: (i) Demand-side discrimination does not exist, and (ii) potential productivity differences are controlled for in the estimation model and do not affect the adjusted gender pay gap. In this section, we will take a closer look at the validity of both assumptions.

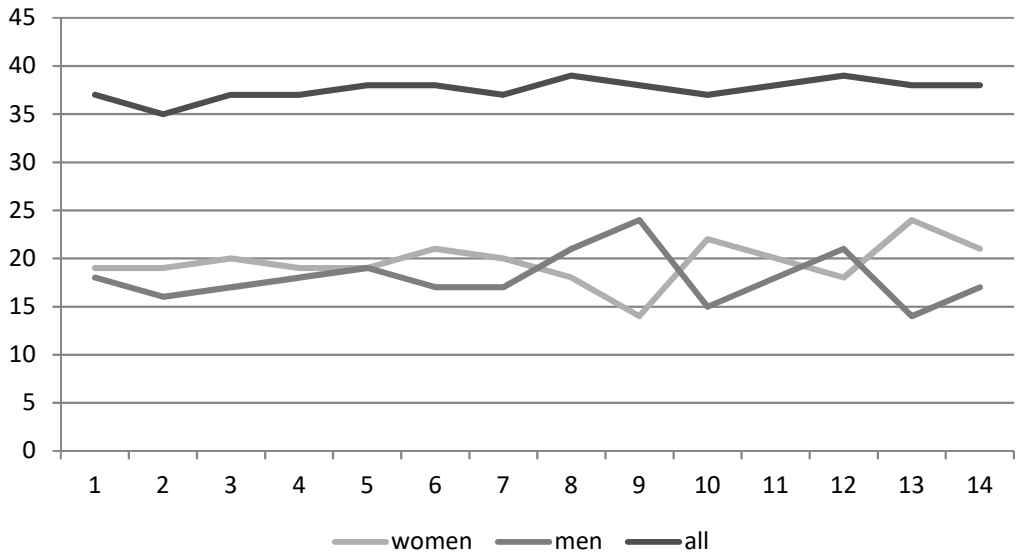
3.6.1 Randomness

To rule out any bias stemming from employer-side discrimination we have to make sure, that the assignment of choice sets to participants has been truly random in our experiment, not only by design but also in practice. In a real job market, an employer could have a preference for one specific contract type and therefore be more likely to pick a choice set that involves this contract type. In our experimental job situation, we let the computer assign the rounds and thus choice sets to the participants randomly. Theoretically, we expect all choice sets to be equally often assigned. Figure 14 illustrates the actual assignment of choice sets and, hence, contract types in our experiment. The topmost line in Figure 14 confirms that the number of individuals assigned to a specific choice set is relatively stable over all choice sets and ranges between 37 and 39. That is, for the full sample, the assignment is truly random (Pearson's chi squared test: p=.851).

In a real labor market, however, an employer may also assign contracts differently to male and female employees. To fully rule out this potentially discriminatory behavior in our experiment, we compare the assignments of men and women separately, represented by the two gray lines

in Figure 14. We use Binomial tests to compare the women's assignment share to a particular choice set with the men's. In accordance with the visual impression from the gray lines in Figure 8, the tests do not reject an equal representation of genders, i.e., random assignment, except in choice sets 9 and 13. To gauge a potential bias arising from this slightly asymmetric representation, we re-run our estimation and decomposition analysis, leaving out these two choice sets, and compare the results with the results given in Figure 13 (and now summarized as "baseline" in the last column of Table 8).

Figure 14: Assignment of Rounds



As the numbers in Table 8 reveal, the selection and contract effects remain totally robust in sign (contributing to a pay gap) and partially in magnitude. If any, we tend to underestimate the selection effect and overestimate the contract effect when using all choice sets. In each of the three alternative specifications, the selection effect tends to be larger and the contract effect to be smaller compared to the baseline estimation. Hence, the results of our baseline decomposition analysis may be interpreted as a lower-bound estimate for selection into a payment scheme and an upper-bound estimate for conditional performance.

Table 8: Decomposition Results for Specific Rounds

	9 out	**13 out**	**9 & 13 out**	**baseline**
GPG	0.210	0.219	0.195	0.231
Selection effect	-0.259	-0.335	-0.217	-0.255
Contract effect	-0.375	-0.584	-0.412	-0.525
Observations (f/m)	485 (259/226)	484 (249/235)	447 (235/212)	522 (273/249)

3.6.2 Productivity differences

The interpretation of the contract effect as a conditional performance effect relies crucially on our choice of proxy for attempting to capture the participants' potential productivity. Since we control for this potential productivity in the estimation, by using the participants' mean performance in the trial mazes, any existing gender differences should not bias our results. To fully rule out any potential bias from productivity differences between men and women in our baseline analysis, we re-run our estimation and decomposition analysis with more homogenous samples of female and male participants in terms of the time they needed for the trial mazes. Our experiment provides us with two measures of potential productivity: the time needed for an easy maze and the time needed for a hard maze. To create a relatively homogeneous subsample in terms of participants' potential productivity, we start from the first and third quartiles of the time men and women needed to solve one maze and successively widen the sample in 5-percentile steps. We stop enhancing the sample as soon as the gender differences start becoming statistically significant ($p<.05$). Then we re-run our estimation and decomposition analysis using these homogenous samples in terms of potential productivity, and compare the results with the baseline results given in Figure 13 (and summarized in the last column of Table 9).

Table 9: Selection and Contract Effect in Subsamples

	hard (p25-p75)	**easy** (p25-p75)	**hard** (p5-p95)	**easy** (p10-p90)	**baseline**
GPG	0.235	0.369	0.237	0.266	0.231
Selection effect	-0.021	-0.552	-0.222	-0.377	-0.255
Contract effect	-0.992	-0.608	-0.740	-0.518	-0.525
Observations (m/f)	263 (118/145)	253 (105/148)	462 (218/244)	420 (193/227)	522 (249/273)

Table 9 displays the gender pay gap, the selection and the contract effect of four different subsamples. The first two are limited to the first and the third quartiles of the trial maze performance distribution, and the next two are defined by the successive enhancement procedure. As Table 9 shows, the magnitude of the gender pay gap, as well as the selection and the contract effect, differs across the subsamples. The selection effect is larger in the homogeneous easy-maze productivity samples and smaller in the hard-maze ones. The contract effect, on the contrary, always tends to be larger in the homogeneous productivity samples, no matter which difficulty level applies. It is responsible for 100% of the gender pay gap in the

hard-maze homogenous sample. That is, differences in pay are fully explained by conditional performance. We therefore conclude that the gender pay gap in our sample is indeed driven by a performance reaction on the given contract that exists irrespective of the participants' potential productivity. In this respect, the baseline analysis seems to provide a lower-bound estimate.

3.7 Conclusions

Our experiment yields a gender pay gap of more than 20%, which is close to the actual one observed in the German labor market. At first glance, this comes as a surprise, as we ruled out employer discrimination by design. Female and male participants are offered the same payment schemes to choose from, one of which is then drawn randomly, and they are asked to perform the same task. As it turns out, women prefer different payment contracts than men, and, when exposed to a given contract, women's performance is weaker, on average. Though the latter applies to any contract offered, the gap in earnings is largest at any deviation from the basic piece rate, – that is when either a threshold, competition with all participants or a risk premium (or a combination of these) is involved. Interestingly, the performances of female and male participants differ slightly in the test mazes, but the gap becomes economically and statistically significant only when the participants are paid according to their selected contract types. The fact that unconditional performance differs less between the genders than conditional performance is a first indication that incentive-driven efforts depend on gender. We further observe that the selection and contract effects are largest when a payment scheme includes a risk premium, and we conclude that women are more loss averse than men and more likely to underperform in what we would call a stereotype threat situation. Overall, the stereotype threat effect might be responsible for at least 38% of the observed gender pay gap in our experiment.

Our findings are particularly suggestive, as the participants made their experimental choices on their own computers. This feature of an artefactual field experiment combines the advantages of a non-artificial environment with those of a controlled setting. Together with the anonymity that comes with an online experiment, this underlines the credibility of the preferences revealed by the participants.

However, before drawing policy conclusions from our findings, the caveats shall be discussed as well. Although we made sure to avoid any explicit gender discrimination on the experimental employer's side, the way we designed the experimental design may still not be free of implicit discriminatory practices, for example, with regard to the experimental environment and the implementation of risk and relative performances. The decomposition analysis revealed that

this setting produced large gender differences in conditional performance from rather small starting differences in potential productivity, which translated into a sizeable experimental gender pay gap in the end. As a consequence, a re-design, or a de-biasing of the rules of the game – or the "playing field" as Bohnet (2016) puts it – may provide an ever more gender-neutral decision environment. One issue would be the existence and display of risk involved in the decision problem (Bohnet, 2016). In our experiment, the risk premium option in the choice sets counts as responsible for about 50% (14% selection and 38% contract effect) of the gender pay gap. Going without it may eliminate the impact of (women's higher) loss aversion and, as a result, reduce the gender pay gap by one half. Providing the participants with explicit feedback on their absolute and relative performance in the experiment is another option to prevent men and women from making a potentially biased self-assessment and to reduce gender differences in selection into competitive payment schemes (Bohnet, 2016). According to our experiment, the gap in compensation would be 26% smaller if women chose the same contracts as men. A final caveat of our design may be seen in the type of effort task we chose. This choice was guided by the desire for gender neutrality in performance, and several studies seemed to suggest equal ability of women and men in the maze task. However, notwithstanding actual ability, the believed ability may still be gendered, and some gendering has in fact been documented for the maze task (Günther et al., 2010). Does this pose a threat to our results? We do not think so. Even if the participants of our experiment share the belief of gender-specific abilities in solving mazes, this may just be the reason why women and men choose different payment schemes, and it would still confirm our interpretation of a stereotype threat effect. Hence, while the experiment was designed to mimic the reality of most of today's job situations, the political starting points in order to further gender equality seem to have been reconfirmed.

3.8 Appendix

Table 10: Sample Descriptives

Variable	Women	Men	p-value (gender differences)	Cohen's d
Characteristics				
Single*	.487	.510	.603	
Living in East-Germany*	.469	.514	.302	
Age (years)	36.31	37.31	.123	
Children in household*	.549	.333	<.001	.445
Education: basic*	.029	.020	.500	
Education: lower professional*	.403	.446	.323	
Education: higher professional*	.216	.233	.646	
Education: university / polytechnics*	.289	.237	.176	
In education*	.044	.044	1	
Labor force participation*	.696	.843	<.001	-.353
Choices				
Competition (same)*	.317	.316	.986	
Competition (all)*	.252	.314	.041	-.179
Bonus*	.641	.718	.009	-.228
Risk premium*	.404	.500	.051	-.171
Difficult*	.692	.687	.871	
Controls				
Time for easy trial maze (sec.)	48.88	46.94	.493	
Time for hard trial maze (sec.)	60.73	51.37	.044	.177
Risk aversion (1-10 scale)	4.51	5.26	<.001	-.329
Number of Obs.	273	249		

Note: * indicates a dummy variable.

Figure 15: Choice Set Bonus

Choose one of the two options.
If this choice set is drawn at the end of this study,
the selected option will be the basis of your payment.

Option A ◯	Option B ◯
level of difficulty: easy	level of difficulty: easy
50 cents for each solved maze	1€ for each solved maze (at least 5 mazes) 20 cents for each solved maze (less than 5 mazes)
Explanation: You receive **50 cents** for each solved maze with an **easy level of difficulty**.	Explanation: You receive **1€** for each solved maze with an **easy difficulty level** if you solve at least 5 mazes. **If you don't** you will get **20 cents** for each solved maze.

Figure 16: Choice Set Competition

Choose one of the two options.
If this choice set is drawn at the end of this study,
the selected option will be the basis of your payment.

Option A ◯	Option B ◯
level of difficulty: easy	level of difficulty: easy
50 cents for each solved maze	1€ for each solved maze (among the best 30%) 20 cents for each solved maze (if you'r not among the best 30%)
Explanation: You receive **50 cents** for each solved maze with an **easy level of difficulty**.	Explanation: You receive **1€** for each solved maze with an **easy difficulty level** if you rank among the best 30% of all participants. **If you don't** you will get **20 cents** for each solved maze.

Table 11: Full Decomposition

	(1)		(2)		(3)		(4)		(5)	
	Explained	**Unexplained**	**Explained**	**Unexplained**	**Explained**	**Unexplained**	**Explained**	**Unexplained**	**Explained**	**Unexplained**
competition (same)*	0.0111	-0.135	0.0190	-0.166	0.0190	-0.170	0.0315	-0.103	0.0742	-0.129
	(0.0471)	(0.203)	(0.0472)	(0.199)	(0.0474)	(0.200)	(0.0516)	(0.197)	(0.0787)	(0.224)
competition (all)*	0.0712	0.0345	0.0697	0.0325	0.0697	0.0302	0.0642	0.0543	-0.0216	0.0157
	(0.0753)	(0.141)	(0.0730)	(0.137)	(0.0732)	(0.138)	(0.0710)	(0.135)	(0.0730)	(0.155)
bonus*	-0.235	-0.121	-0.224	-0.127	-0.224	-0.132	-0.230	-0.169	-0.254	-0.207
	(0.226)	(0.231)	(0.216)	(0.226)	(0.216)	(0.228)	(0.221)	(0.224)	(0.244)	(0.230)
risk premium*	-0.221	-0.615**	-0.213	-0.594**	-0.213	-0.591*	-0.213	-0.682**	-0.210	-0.586*
	(0.183)	(0.311)	(0.176)	(0.302)	(0.177)	(0.303)	(0.176)	(0.303)	(0.174)	(0.299)
difficulty*	0.00801	0.199	0.00589	0.170	0.00589	0.169	0.0133	0.103	0.0218	0.105
	(0.0234)	(0.160)	(0.0195)	(0.156)	(0.0195)	(0.156)	(0.0336)	(0.153)	(0.0523)	(0.160)
time for trial maze			-0.235	1.021*	-0.235	1.013*	-0.193	1.065*	-0.209	1.047*
			(0.143)	(0.606)	(0.143)	(0.609)	(0.123)	(0.621)	(0.130)	(0.625)
risk aversion					0.000681	0.101	-0.0152	-0.401	-0.0125	-0.346
					(0.115)	(0.853)	(0.114)	(0.842)	(0.114)	(0.837)
single*							-0.0330	-0.673	-0.0335	-0.653
							(0.0663)	(0.501)	(0.0674)	(0.496)
child in household*							0.165	-0.533	0.226	-0.701
							(0.197)	(0.583)	(0.198)	(0.580)
age							0.128	-0.663	0.138	-0.105
							(0.0967)	(2.259)	(0.102)	(2.253)
employed*							-0.335*	-1.280	-0.371**	-1.527*
							(0.175)	(0.839)	(0.179)	(0.833)
education low*							0.134	1.769**	0.133	1.767**
							(0.155)	(0.873)	(0.154)	(0.866)
education mid*							0.0478	0.919*	0.0474	0.956*
							(0.109)	(0.497)	(0.108)	(0.494)
education high*							-0.124	0.958	-0.121	0.982
							(0.132)	(0.648)	(0.130)	(0.643)
education current*							3.57e-05	0.210	-1.94e-05	0.183
							(0.00296)	(0.140)	(0.00167)	(0.136)
east*							0.00140	-0.335	0.00555	-0.215
							(0.0303)	(0.395)	(0.0304)	(0.391)

													(0.0182)		(0.973)	
	(1)			**(2)**			**(3)**			**(4)**			**(5)**			
	Differ-ential	**Ex-plained**	**Unex-plained**	**Differ-ential**	**Ex-plained**	**Unex-plained**	**Differ-ential**	**Ex-plained**	**Unex-plained**	**Differ-ential**	**Ex-plained**	**Unex-plained**	**Differ-ential**	**Ex-plained**	**Unex-plained**
round controls	yes			yes			yes			yes			yes			
design controls	no			no			no			yes			yes			
underlying preferences	no			no			no			no			yes			
Total		**-0.327**	**-1.198*** **		**-0.533**	**-0.992** **		**-0.532**	**-0.993** **		**-0.528**	**-0.997***		**-0.126**	**-1.399** **
		(0.405)	**(0.451)**		**(0.429)**	**(0.444)**		**(0.447)**	**(0.461)**		**(0.540)**	**(0.529)**		**(0.563)**	**(0.546)**
Prediction 1	5.075***			5.075***			5.075***			5.075***			5.075***		
	(0.329)			(0.329)			(0.329)			(0.332)			(0.334)		
Prediction_2	6.600***			6.600***			6.600***			6.600***			6.600***		
	(0.451)			(0.451)			(0.452)			(0.456)			(0.458)		
Difference	-1.525***			-1.525***			-1.525***			-1.525***			1.525		
	(0.558)			(0.558)			(0.559)			(0.564)			(0.567)		
Constant			-1.323			-1.994*			-2.076			-1.676			-2.197
			(0.934)			(1.080)			(1.396)			(3.540)			(3.698)
Observations	522	522	522	522	522	522	522	522	522	522	522	522	522	522	522
N_1	273	273	273	273	273	273	273	273	273	273	273	273	273	273	273
N_2	249	249	249	249	249	249	249	249	249	249	249	249	249	249	249

Note: * indicates a dummy variable. * p<.100, ** p<.050, *** p<.010

4 Social Identity Priming in Economics Experiments: No Clear Evidence

Individual behavior is influenced by social norms and identities. As much as this seems to be common knowledge in social sciences, only a few papers have yet been able to provide causal evidence regarding the relation between norms, identity and behavior (Chang et al., 2014).

A framework that helps to describe the impact of social identities on economic behavior has been introduced by Akerlof and Kranton (2000) and is commonly referred to as "identity economics." Their seminal work gave rise not only to papers that illustrate the association between social identities and economic decision making but also to studies that derive causal evidence. To achieve this, this type of studies increasingly uses priming as a technique to stimulate and exogenously vary a social identity (gender, ethnicity, etc.). While priming is a relatively new trend in experimental economics, it serves as a standard tool in social psychology (see, for example, Bargh et al., 1996; Davies et al., 2002; Steele and Aronson, 1995). A priming stimulus unconsciously activates mental associations in subjects' minds and makes a social identity (such as gender or race) more salient (Benjamin et al., 2016). As such, it evokes the behavior to move closer to the prescribed action that is associated with the particular social identity (Benjamin et al., 2016).

However, Kahneman (2012) started questioning the robustness of priming results due to multiple replication failures in psychology. Similarly, in economics, experimental results are mixed when a priming stimulus is used. The aim of this paper, therefore, is to reflect on priming effects in economics with theoretical insights and recent research from social psychology on moderators of a priming effect. Thus, we discuss the effectiveness of social identity priming techniques from theoretical and experimental perspectives in a two-step procedure. First, we review the economic literature, with an emphasis on the theoretical background and existing experimental studies. We focus on a comparison of the priming stimuli that are used, the social identity that is primed, and the results, depending on whether priming shows any impact. In the second step, we illustrate the priming issue in more detail with own results from our experiment on selection into compensation schemes.

First, and in line with others, we show that the participating men and women differ in their selection into compensation schemes. Second, our subgroup analyses reveal a weak association between gender norms and selection. However, we fail to produce effects from activating these gender norms by using a priming stimulus. Since we do not interpret the absence of priming

effects as definite evidence against the effect of gender norms, we address some factors that could be responsible for a failed activation in the discussion.

This paper is organized as follows: The next section reviews the experimental priming literature, and Section 4.2 presents the design, hypotheses and main results of our own experiment. Section 4.3 discusses some aspects that could explain the absence of priming effects in other experiments as well as in ours, to address important factors for future priming experiments in economics.

4.1 Social identity priming in economics

In this section, a brief theoretical classification is followed by a review of experimental papers in economics that make use of a priming stimulus.

4.1.1 Theoretical Outline

Akerlof and Kranton (2000) introduce the concept "identity," which had originally been developed in social psychology, into an economic decision model. They argue that people derive utility from complying with social norms and disutility from deviating from socially desired behavior. Their model expands the standard utility function with a non-pecuniary term representing the social identity. The additional term reflects people's assignment in pre-supposed social categories. Individuals' choices thus depend on self-identity and preferences, beyond purely economic incentives. Akerlof and Kranton (2000) apply their model to many economic issues, such as labor market discrimination, household division of labor, the economics of education and contract theory.

To investigate the effect of social identity on individual decision making, a priming stimulus can be used. This technique is imported from socio-psychological studies on the effect of mental representations or concepts on judgments and behavior (review by Doyen et al., 2014). The priming stimulus calls social situations or relationships into mind and subtly influences behavior even when individuals do not link it to the current action (Molden, 2014). Behavior will evolve without the awareness of such an external determinant (Bargh, 2006). Thus, a priming stimulus unconsciously activates associations in subjects' minds and makes a social identity more salient.

Benjamin et al. (2010) provide an economic framework based on the identity economics approach by Akerlof and Kranton (2000) to explain how priming leads to a behavioral response. Accordingly, an individual belongs to a social category (such as gender or race), and there are actions that are preferred in the absence of identity considerations and actions that are

prescribed for this social category. Moreover, there is a weight placed on the social category's influence on an individual's decision making. A priming stimulus that activates a social category consequently increases the strength of the social category weight in the decision process and evokes the behavior to move closer to the prescribed action that is associated with the particular social identity. As a result, priming promises to be a useful tool to study how social identity affects preferences and thus behavior (Benjamin et al., 2010b).

4.1.2 Experimental findings

To review economic studies that use a priming stimulus, a systematic search was performed by using the RePec database on IDEAS. We ran a simple search using "identity" and "priming" as keywords. Based on these studies, additional publications were searched on Google Scholar with backward and forward referencing.

We restrict our review to papers exploring gender, racial or religious priming, since these social identities are most frequently used to address questions on the relationship between social norms and economic behavior. Moreover, we focus economic decision-making studies and do not consider priming experiments that focus on performance such as Carr and Steele (2010) and Dee (2014). In contrast to Cohn and Maréchal's (2016) literature survey, which is limited to economic experiments with incentivized decisions, we also consider studies that include stated preferences, as Dohmen et al. (2011) and Lönnqvist et al. (2015) demonstrated the behavioral validity of stated preferences in the area of risk taking. Since religious priming effects have been explored in an extensive meta-analysis (Willard et al., 2016), only exemplary studies are mentioned in this dimension.

4.1.2.1 Risk preferences

Meier-Pesti and Penz (2008) used a priming stimulus to investigate the effects of both gender and biological sex on stated risk preferences. They presented pictures of either a young woman looking after a baby, serving as a feminine prime, or a young man in business clothing and with appliances (phone and filofax), serving as a masculine prime. In the control group, a picture of six young people was presented. After having seen the pictures, participants were asked to figure out what these people might think and feel and what they would do in the evening. To check whether the priming worked successfully (manipulation check), participants had to fill the blanks of five incomplete sentences that included presumably masculine (e.g., willing to take risks, rational, ambitious), feminine (e.g., sensitive to others needs, compassionate, tender) and gender-neutral attributes (e.g., reliable, impolite, sociable). They find that men show a lower identification with masculine attributes if they have been exposed to the feminine prime.

This corresponds to the gender priming effects on risk taking in the male sample they observe: men in the masculine prime group behave more risk loving compared to the control group and men in the feminine prime group. However, gender priming does not affect the female sample – through neither identification nor risk preferences. The authors argue that this could be an unintended effect of the feminine prime on the student sample. Since students are preparing themselves for the workforce, a picture of a woman holding a baby "seems to oppose women's plans to enter the workforce" (Meier-Pesti and Penz, 2008: 191) and may thus increase awareness of this issue, which, in turn, contradicts expected priming effects.

Benjamin et al. (2010) observed revealed instead of stated risk preferences in two experiments using a sample of students of white and Asian descent in the United States. In one experiment, they employed a background questionnaire as an ethnic prime. The questionnaire included questions about languages spoken at home and how many generations of their family had lived in the United States. The control group was questioned about the school meal plan and cable television subscription. However, Benjamin et al. (2010) cannot find a priming effect. They consider the absence of a priming effect in their experiment as being "perhaps not surprising" (Benjamin et al., 2010: 5) since findings on risk aversion among Asians are ambiguous.

In the second experiment, racial and gender primes were used to elicit differences in risk preferences. In the racial prime condition, primed subjects were asked for their race and their opinion about living with roommates of the same race. Subjects who were gender primed had to state their gender and their opinion about living on a mixed or a single-sex dormitory floor. The control group was asked only whether they lived on or off campus. In the priming condition, they find blacks to be more risk averse than in the control condition, which is in line with their hypothesis, while there is no effect on whites. Making gender salient had an effect on neither females' nor males' risk aversion.

Boschini et al. (2014) primed their participants, a large random Swedish adult sample, who were interviewed by telephone, by asking them to indicate their gender. They observed binary choices between a risky and a safe option, with one of them having been randomly selected for payoff. They do not find gender differences in risk taking but they find treatment effects: Men behave – contrary to what has been expected – less risk averse in the priming treatment. These results lead the researchers to announce a test in "differences in results [between their and others' findings] in future work" (Boschini et al., 2014: 35).

Weaver et al. (2013) used a prime to investigate the effect of masculinity on men's willingness to take financial risks. In a fictitious product test, students tried hand lotions (gender threat

condition) or a power drill (gender affirmation condition). They were subsequently asked to close their eyes for 10 seconds and think about the quality of the product. After a manipulation check that disclosed men in the gender threat condition felt less masculine compared to men in the gender affirmative condition, they had to play a gambling game. All participants were videotaped during the experiment to increase the public nature of the decision. They find that men who experienced a threat to their masculinity by using the hand lotion are more willing to take risks compared to men whose manhood had "not been questioned" (Weaver et al., 2013:186). According to the authors' discussion, the findings are not in line with the idea of priming a masculine stereotype by experiencing the power drill, because opposite results would be expected. Instead, their effects are driven by anxiety caused by "threatening manhood" (Weaver et al., 2013:189).

D'Acunto (2015) examined the effect of gender identity on risk taking by using mTurk – an online labor market platform. In the within-subjects designed experiment, the subjects were asked to make different lottery choices before and after the experimental treatment and a manipulation check. The experimental manipulation included reading a short text that discussed principles of a healthy lifestyle for the control group or feminine and masculine behavior for the female- and male-primed participants, respectively. One of the incentivized tasks was to write a short essay of 5 to 10 sentences on how they were feeling in detail. The average number of times subjects wrote about a stereotype associated with male or female individuals serves as a manipulation check. Indeed, male-primed subjects report more male stereotypes compared to subjects in other conditions, and female-primed subjects report more female stereotypes compared to other subjects. A difference-in-difference approach applied to the risk preferences before and after the manipulation reveals that the men in both priming conditions increase their risk willingness. On the contrary, priming does not affect the women's risk preferences.

Benjamin et al. (2016) investigated the effect of religious identity on risk aversion by using a sentence-unscrambling task. Subjects were asked to drop one word out of a five-word group and rearrange the words in a meaningful way. In the priming condition, sentences included religious content. The participants were then asked to rethink the groups they belong to and list five of the most important characteristics that define who they are. This manipulation check indicated that the priming instrument increased the salience of the religious identity. Moreover, priming increases risk taking for Catholics, while Protestants are not affected. This is only partially in line with their hypotheses because Catholicism actually promotes gambling, and priming indeed increased risk taking in their setting, while Protestantism is associated with anti-gambling norms, which has not been confirmed by the data (Benjamin et al., 2016).

4.1.2.2 Preferences for competition

Boschini et al. (2014) primed subjects by asking them for their gender and then investigate gender differences in subjects' preferences for competition in math and language tasks. Before performing on the task, the participants had to choose between an individual piece-rate payment scheme and a tournament payment, which involved competing against a random counterpart. Winning the tournament doubled the individual's piece rate, but losers received nothing. While the authors do not find gender differences in the language task or the baseline treatment, preferences for competition in the math task exist in the priming treatment, with women being less competitive.[38] A between-treatment analysis suggests no priming effects at all.

Cadsby et al. (2013) used a relatively homogeneous sample of MBA students at an elite Canadian business school. In the gender/family priming treatment, the participants had to indicate their gender or whether they had children. In the professional priming treatment, they were asked about their Graduate Management Admission Test (GMAT) score or their salary expectations after completion of their degree. Before performing on a real effort task, they had to choose between piece rate and a tournament that yielded a quadrupled piece rate in case of winning and nothing in case of losing. While there are no gender differences in the professional priming treatment, men are found to be more competitive than women in the gender/family priming treatment. Moreover, women are more competitive in the professional priming group compared to women in the gender priming group, while they detect no differences for men. They conclude that an identity conflict exists for women but not for men.

4.1.2.3 Altruism

Asking subjects to indicate their gender is also common in investigations of gender differences in altruism. Boschini et al. (2012) used this priming technique just before students played a dictator game. Besides the gender priming the experimental manipulation included a gender-mixed environment (women and men were in the lab) and a single-gender environment (only men or women in the lab). They show that priming decreases men's generosity in the mixed environment, while there are no statistically significant priming effects for women.

Within a random sample of the Swedish adult population, Boschini et al. (2015; see Section 4.1.2.1) find the same gender-priming technique to result in gender differences in altruism when using the dictator game in the priming condition, while gender differences are not present in the control group. Again, this effect seems to be driven by treatment effects in subgroups: Men

[38] Besides the priming treatment, there were two other treatments where a male or female counterpart was assigned. Gender differences in preferences for competition were present only when the counterpart was female and only for the math task.

who are primed and placed in a mixed-gender context (counterpart is female) give significantly less than men who are primed and have a male interview partner. Women with a male counterpart give significantly more when they were primed.

Benjamin et al. (2016) examine the effect of a religious identity on altruism by using the same sentence-unscrambling task described in Section 4.1.2.1. They find that priming a religious identity causes neither Protestants nor Catholics to give significantly more.[39]

4.1.2.4 Time preferences

Besides risk preferences, Benjamin et al. (2010) also investigate the role of ethnic and gender identity (Section 4.1.2.1) in time preferences. Participants had to choose between a certain amount of money they will receive immediately after the experiment and a larger delayed amount. Priming the ethnic identity with background questionnaires (see Section 4.1.2.1) led Asians to make less impatient choices, which was in line with their hypothesis. However, this priming procedure does not affect the time preferences of black, white, male or female subjects.

Weaver et al. (2013) primed men by asking them to recall ten (gender threat condition) or two (gender affirmation condition) past actions that demonstrated that they are "real men." While it is easy to recall only two actions, recalling ten actions is relatively hard, and, thus, men were assumed to be less secure in their self-perceived masculinity in this gender threat condition. Manipulation checks indicated that men feel indeed significantly less masculine in the threat condition compared to the affirmation condition. Weaver et al. (2013) additionally varied the publicity of the choice. In one treatment, men were told that they had to justify their answers on videotape, whereas in the other treatment their choices were private. By observing the subjects' choices between a smaller but immediate reward and a later reward, the authors evaluate men's patience (depending on the manipulation) and find them to be less patient when they are threatened in their manhood – but only when their choice was kept private. If their choice was public, threatened men are less impatient than men in the gender-affirmative condition.

Religious identity priming by using a sentence-unscrambling task as presented by Benjamin et al. (2016; see Section 4.1.2.1) has no effect on time preferences.

[39] However, they were not able to replicate the findings of Shariff and Norenzayan (2007) of higher generosity in the priming treatment, although they used the same priming instrument. Similarly, Gomes and McCullough (2015), who replicated the design by Shariff and Norenzayan (2007), do not find priming effects.

4.1.2.5 Cooperation

Preferences for cooperation are often observed in classic public goods experiments. The religious priming instrument that uses a sentence-unscrambling task by Benjamin et al. (2016; see Section 4.1.2.1) causes behavioral effects. While priming religious identity increases the contribution to a public good for Protestants, Catholics contribute significantly less when primed.

Horton et al. (2011) test the effect of a religious prime on cooperation in an online laboratory. The primed participants read a bible passage on the importance of charity while the control group read a text about three species of fish. Results from the prisoner's dilemma game indicate that priming does not increase cooperation per se. Only in the subgroup of believers who report to have experience that convinced them of god's existence (post-experimental questionnaire) cooperation increased.

Boschini et al. (2014) elicit preferences for cooperation in a prisoner's dilemma game. As described in Section 4.1.2.1, primed participants were asked to indicate their gender before their decisions, while unprimed were asked afterward. Moreover, the researchers varied the counterparts' gender in two other experimental conditions. They neither find gender nor treatment effects. Their findings are in line with the results of previous work on gender differences in cooperation preferences, supporting their hypotheses.

Chen et al. (2014) primed Asian-American and Caucasian students with a questionnaire that included questions about languages spoken at home, the family's home country and how many generations of the family had lived in the United States which is similar to Benjamin et al. (2010), before playing a prisoner's dilemma game. To reinforce the primes during the experiment, participants were exposed to photos of architecture from China and Europe. They find that ethnic priming does not affect cooperation in the prisoner's dilemma game in the aggregate, which is not in line with their hypotheses.

4.1.3 Presence and absence of priming effects

The above-described experiments differ in various dimensions. As the overview in Table 12 shows, priming can take many forms. Interventions ranges from a simple question on the participant's gender, age or ethnic background, to work search puzzles or reading texts, to looking at pictures, to using hand lotion versus experiencing a drill. Moreover, all experiments use subtle instead of blatant priming. The advantage of a subtle priming method is that subjects are not aware of the connection between the social category and the related behavior, whereas

a blatant prime increases their awareness (Shih et al., 2002) and may additionally produce contrast effects (for an overview see Loersch and Payne, 2014).

Most experiments that are listed in Table 12 were conducted in the laboratory; some were conducted online, and one via telephone. Since treatment effects mostly occur in the lab, it seems that priming works relatively well in this setting, which is marked by a high degree of internal validity. However, Horton et al. (2011) were able to replicate priming effects from a lab experiment with students in an online experiment, and Boschini et al. (2015) show gender priming to have an effect in a setting where the adult participants were interviewed on the phone.

Table 12 shows that priming effects are absent in many studies. Moreover, some of these studies also lack in terms of explanations for or discussions on these null findings or contradictory results. For example, Meier-Pesti and Penz (2008), who do not find gender-priming effects for women, explain this in terms of supposed career objectives among the female student sample. Their female prime, making the participants view a picture of a young woman looking after a baby, could have challenged their career objectives, increased their awareness and contradicted the priming effect. This could be an issue in other experiments that have tried to manipulate female students with a gender prime.

Generally, null findings could be explained by the unsuccessful activation of mental representations in subjects' minds. To explore this, Meier-Pesti and Penz (2008), Weaver et al. (2014), D'Acunto (2015) and Benjamin et al. (2016) implemented manipulation checks.

Table 12: Overview of Experimental Studies

Authors	Objective	Social Identity	Priming Technique	Target Group/ Setting	Treatment Effects
Benjamin et al. (2010)	Risk preference	gender, race (blacks, Asian, white)	background questionnaire (indicate race or gender and roommates race or gender)	students in the lab	Asians -, Whites -, Blacks ✓, Men -, Women -
	Time preferences		background questionnaire (questions on languages spoken at home, family's home country and how many generations of the family had lived in the United States)		Asians ✓, Blacks ✓, Men -, Women -
Benjamin et al. (2016)	Risk preference	religion	sentence-unscrambling task	students in the lab	Catholics ✓, Protestants -
	Time preferences				Catholics -, Protestants -
	Cooperation				Catholics ✓, Protestants ✓
	Altruism				Catholics -, Protestants -
Boschini et al. (2012)	Altruism	gender	asking for gender	students in the lab	Men ✓, Women -
Boschini et al. (2014)	Risk preference	gender	asking for gender	random adult population via telephone	Men ↯, Women -
	Competition				Men -, Women -
	Cooperation				Men -, Women -
Boschini et al. (2015)	Altruism				Men -, Women ✓
Cadsby et al. (2015)	Competition	gender/family prime vs. professional prime	background questionnaire (gender- and family-related concerns vs. career planning issues)	students in the lab	Gender prime: Men ✓, Women – Professional prime: Men -, Women ✓
Chen et al. (2014)	Cooperation	ethnic prime	background questionnaire (questions on languages spoken at home, family's home country and how many generations of the family had lived in the United States)	students in the lab	Asian-Americans -, Caucasians -
D'Acunto (2015)	Risk preference	gender	reading a text		Men ✓, Women -
Horton et al. (2011)	Cooperation	religion	reading a text	adult sample online	subgroup of believers ✓
Meier-Pesti and Penz (2008)	Risk preference	gender	picture	students in the lab	Men ✓, Women -
Weaver et al. (2013)	Risk preference	manhood	product test	students in the lab	Men ↯
	Time preferences	threat in manhood	recall past actions that demonstrate being a real man		Men ✓

4.2 Gender Priming and Selection into Incentive Schemes: No-Evidence from the Field

In the following, we give a brief description[40] of our experiment. Primed and non-primed participants were asked to choose between incentive schemes for a maze task that they had to solve afterward. Primed participants had to indicate their gender on an additional screen at the start of the experiment. In these aspects, our priming intervention is close to the experiments by Boschini et al. (2014) and Cadsby et al. (2013). To reinforce the priming effect, a screen wiper with a pictogram of a man or a woman appeared constantly on the screen for the whole remainder of the experiment which is similar to the technique used by Chen et al. (2014).

However, our procedure is original in other ways. First, we chose an online solution, which guaranteed the anonymity of the participants but created a familiar environment. Although online experiments present a risk of lacking control, it has been shown that the findings of classical lab experiments and even priming effects can be replicated online (Horton et al., 2011). Second, our sample is a relatively large non-student sample. We exploited a uniformly stratified random sample by gender, geographical location (East vs. West) and marital status (single vs. in a couple) of 883 German individuals. Third, we refer to the participants' gender norms instead of assuming exogenous ones: Two weeks prior to the experiment, the participants had to answer questions on behavioral prescriptions for men and women. They had to indicate their agreement level with the following four questions on a scale of 1 to10: "Should a man/woman take a challenging job?", "Should a man/woman take a high-risk job?" The aim was to obtain an exogeneous measure of gender traditionalism when comparing the agreement levels, depending on the gender of the group addressed by these questions. Therefore, we were able to perform separate analyses for the subpopulation holding traditional views (i.e., women should be less risk loving and less competitive) versus those with more egalitarian views on gender roles. This is an important issue, as priming effects depend on the social groups and the norms associated with them (Benjamin et al., 2016; Horton et al., 2011).

The priming procedure was followed by a choice task that disclosed the subjects' preferences for payment schemes, which included bonus and competition schemes. In a first set of rounds, participants had to choose between piece-rate compensation (linear payment) and a bonus scheme with a high piece-rate payment if the participant reaches the threshold of five mazes but a low piece-rate payment if the participant fails to reach the threshold. In a second set of

[40] A full description of the experimental procedure can be found in the Appendix.

rounds, the linear payment came along with competition that yielded a high piece-rate payment if the participant is among the best 30 % and a low piece-rate payment otherwise. In the next set of four rounds, piece-rate payment came along with competition against a same-gender group respectively. Since the subjects were made aware of their competitors' gender and, thereby, conceivably their own gender, which might induce priming effects, we excluded these rounds from the analysis in this paper.[41] Within the sets of rounds, attributes varied by the difficulty level of the task (easy and hard mazes) and a risk premium.

4.2.1 Hypotheses

In our analyses, we consider both within- and between-treatment effects. Women are commonly expected to be more risk averse[42] and to avoid competition.[43] Moreover, it is assumed that this behavior is driven by gender identities that are made salient through a priming stimulus. Therefore, we assume the following hypotheses:

H1: Men are, on average, more likely than women to choose bonus (H1a) and the competition scheme, respectively (H1b).

H2: Traditional men are more likely to choose bonus (H2a) and competition (H2b) schemes than non-traditional men, while traditional women are less likely to choose bonus (H2c) and competition (H2d) schemes than non-traditional women.

H3: Priming positively affects the likelihood of men selecting into bonus (H3a) and competition schemes (H3b). Women's likelihood is impacted negatively by priming (H3c and H3d).

H4: The priming effect on selection into bonus (H4a) and competition (H4b) schemes is stronger in the subgroup of people holding traditional views.

4.2.2 Effects of priming: statistics and regressions

In this section, we present descriptive evidence as well as results of a multiple regression analysis on priming effects. Table 13 summarizes the choices of bonus and tournament payment schemes depending on gender (man or woman) and the participants' stated gender norms (holding traditional or non-traditional norms). We measure gender norms by comparing the agreement levels between the two sets of two questions (11-point scale) on behavioral job prescriptions for men and women – "Should a man take a challenging job?" and "Should a

[41] Burow et al. (2017) address the behavioral puzzle of women's preference for competition when competitors are women only, with the same experimental data.

[42] Although there is evidence on the statistically significant differences in means between men and women, two recent meta-analyses show that the magnitudes of the gender differences in risk taking are negligible (Filippin and Crosetto,2016 ; Nelson, 2015).

[43] See Niederle (2016) for an overview.

woman take a challenging job?" – as well as agreement levels between "Should a man take a high risk job?" and "Should a woman take a high-risk job?" Gender role traditionalism is then defined by a higher agreement level with men's behavioral prescriptions compared to women's behavioral prescriptions, on average. This means that traditionalists assess a challenging or a high-risk job as a masculine behavior. For non-traditionalists, the agreement level with men's and women's behavioral prescriptions is either equal, or the agreement level with women's behavioral prescription is even higher than the agreement level with men's behavioral prescription.

Table 13: Subgroup Differences in Selection into Bonus and Competition

		all	trad	non-trad	trad vs. non- trad
bonus	**women**	0.6783	0.6891	0.6758	p=-650
	(# obs.)	(1632)	(312)	(1320)	
	men	0.7413	0.7522	0.7370	p=.553
	(# obs.)	(1612)	(452)	(1160)	
	gender differences	p<.001	p=-055	p<.001	
competition	**women**	0.2672	0.2404	0.2735	p=.235
	(# obs.)	(1632)	(312)	(1320)	
	men	0.3089	0.3429	0.2957	p=.065
	(# obs.)	(1612)	(452)	(1160)	
	gender differences	p=.009	p=.002	p=.221	

Note: Displayed are the mean relative frequencies to choose bonus or tournament options. Each participant was observed four times in the bonus rounds and four times in the competition rounds. p-values are from t-tests on mean-differences. Mann-Whitney test yields similar results.

Comparing men's and women's selection into bonus and competition payment schemes shows that, in the most cases, the male participants are on average significantly more likely to choose these payment schemes. This supports hypotheses *H1a* and *H1b*. However, these gender differences do not occur in all subgroups. Men are more likely to choose bonus and competition in the sample of people holding traditional norms. Within the subgroup of people holding non-traditional norms, significant gender differences exist only for selection into bonus.

Nonetheless, choices of bonus and competition schemes occur almost similarly often between traditional and non-traditional subjects. *H2a, H2c* and *H2d* are therefore rejected. An exception is male participants in the competition rounds. As suggested by *H2b,* traditional men are more likely to engage in competition than non-traditional men.

In the following, we investigate priming effects in the whole sample and in subgroups. From Table 14, we conclude that priming effects are absent in the whole sample. Even in the subgroup

of subjects holding traditional gender norms, where we expected the priming effect to be stronger, we do not find any statistically significant priming effect. We therefore reject hypotheses *H3* to *H4*.

Table 14: Priming Effects in Subgroups

		all			non-traditionalists			traditionalists		
		prime	**control**	**prime vs. control**	**prime**	**control**	**prime vs. control**	**prime**	**control**	**prime vs. control**
bonus	**women** (# obs.)	0.6856 (808)	0.6712 (824)	p=.560	0.6913 (664)	0.6600 (656)	p=.226	0.6597 (144)	0.7143 (168)	p=.300
	men (# obs.)	0.7363 (804)	0.7429 (808)	p=.648	0.7292 (576)	0.7449 (584)	p=-544	0.7544 (228)	0.7500 (224)	p=.914
	gender differences	p=.025	p=.017		p=.143	p=.001		p=.048	p=.429	
competition	**women** (# obs.)	0.2785 (808)	0.2561 (824)	p=.307	0.2786 (664)	0.2683 (656)	p=.674	0.2778 (144)	0.2083 (168)	p=.153
	men (# obs.)	0.3085 (804)	0.3094 (808)	p=.967	0.2917 (576)	0.2997 (584)	p=.766	0.3509 (228)	0.3348 (224)	p=.720
	gender differences	p=.186	p=.017		p=.612	p=.221		p=.143	p=.006	

Note: Displayed are the mean relative frequencies to choose bonus or tournament/competition options Each participant was observed four times in the bonus rounds and four times in the competition rounds. p-values are from t-tests on mean differences. Mann-Whitney test yields similar results.

To incorporate the heterogeneity of the sample and control for possible correlations between variables, we perform OLS regressions on the probability of choosing the non-linear option.[44] Determinants we investigate are the treatment-indicator variable (*prime*), a gender indicator (*male*), an indicator for belonging to the group of people holding traditional attitudes (*tradis*), interactions depending on the subgroups, and an additional set of socio-demographic control variables.[45]

The regression analyses in Table 15 show that in our experiment, male subjects are, on average, more likely to select into bonus schemes and, when sociodemographic characteristics and other control variables are taken into account, into competition. Although these gender differences vary in magnitude across primed and non-primed and traditional and non-traditional participants, they do not vary systematically. As a consequence, the regression analysis

[44] Running logit regressions leads to similar results.
[45] A full table can be found in the Appendix.

confirms that priming does not lead to any statistically significant behavioral effect. Thus, the regression analyses in Table 15 are mostly in line with descriptive statistics: First, men, are on average, more likely than women to choose bonus or competition schemes. Second, there is no difference between traditional men and non-traditional men and between traditional women and non-traditional women. Third, we do not find any priming effects, either in the whole sample or in subgroups.[46]

Table 15: Regression Analysis on Selection into Bonus and Competition

	bonus		**competition**	
	(1)	(2)	(3)	(4)
prime	0.0312	0.0221	0.0103	0.0049
	(.0347)	(.0352)	(.0372)	(.0380)
male	0.0848**	0.196***	0.0314	0.163**
	(.0367)	(.0630)	(.0411)	(.0699)
prime x male	-0.0469	-0.0475	-0.0183	-0.0232
	(.0509)	(.0521)	(.0565)	(.0584)
tradis	0.0542	0.0410	-0,0600	-0.0517
	(.0552)	(.0556)	(.0527)	(0533)
prime x tradis	-0.0858	-0.0804	0.0591	0.0720
	(.0854)	(.0842)	(.0795)	(.0821)
male x tradis	-0.0491	-0.0419	0.0951	0.0680
	(.0736)	(.0755)	(.0795)	(.0815)
male x tradis x prime	0.106	0.109	-0.0351	-0,0452
	(.109)	(.110)	(.115)	(.0119)
Constant	0.660***	0.912***	0.268***	0.239***
	(.0257)	(.0902)	(.0268)	(.104)
Controls	No	Yes	No	Yes
Observations	3244	3044	3244	3044
R-squared	0.006	0.105	0.004	0.046

Note: * p<.100, ** p<.050, *** p<.010

In the next section we focus on explaining the absence of a priming effect and consider both theoretical explanations derived from the model by Benjamin et al. (2010, 2016) and insights

[46] Sensitivity analyses that use different definitions of traditionalists reveal a difference between traditional and non-traditional men in selection into competition in that traditional men select into competition more frequently. However, these sensitivity analyses did not lead to any priming effect.

from social psychology on mediators and moderators of priming effects, in explaining ours (and others') findings.

4.3 Discussion

The question whether priming works and when it occurs has been increasingly raised by social psychologists (Molden, 2014), at least since Kahneman's (2012) open letter to scholars in psychology. Still, this question does not seem to have become a big issue in economics yet – even though priming effects have been shown to be ambiguous here too.

In line with Benjamin et al. (2010), we do not interpret the absence of a priming effect "as definite evidence against the existence of social category effects in a particular domain" (Benjamin et al., 2016: 11). Beyond this background, we discuss some factors that are related to the absence of priming effects in our and related experiments.

First, priming effects can differ within heterogeneous groups. Priming will have a stronger effect when one of the two actions – the action that is preferred in the absence of identity considerations and the action that is prescribed for members of this social category – is more extreme compared to the other (Benjamin et al., 2016). This is in line with the religious priming effect that occurs in a group of people who believe in God while it is not present for the non-believers, as shown by Horton et al. (2011). We address this issue in our design by eliciting subjects' gender norms and running subgroup analyses for traditionalists and non-traditionalists. This potentially explains the null results among the latter subgroup. Nonetheless, priming leads to no behavioral effect in the subgroup of people holding traditional gender norms.

Second, null priming effects can also occur when the action in absence of identity considerations matches the action that is prescribed for this social category (Benjamin et al., 2016). If, for example, a man's decision to choose a risky option is optimal without identity considerations due to a higher expected gain, making salient the social norm that prescribes men to choose risky options should lead only to small or no behavioral effects. Thus, only if a conflict exists between the action in absence of a certain identity and the action prescribed by this identity, will priming lead to a behavioral effect. In line with this argumentation, Cadsby et al. (2013) do not find differences in preferences for competition between men who are primed with a gender/family stimulus and men who are primed with a professional prime. They argue that this discloses the absence of an identity conflict between the professional and the gender identity for men (Cadsby et al., 2013). For our experiment, we expect the conflict between the

action in absence of identity considerations and the action prescribed by the gender identity to be largest for women who are relatively good at solving mazes. Without identity considerations, a woman who is able to reach five mazes or to win the competition should go for the bonus or competition schemes. The action prescribed for her social identity, however, is to choose neither the bonus nor the competition scheme. The theoretical dilemma is solved by a subsample analysis of women who are able to reach the threshold of five mazes, as we do not observe a statistically significant priming effect for bonus and competition for this subgroup.

Third, the effectiveness of the salience manipulation itself can explain the absence of priming effects. Insights from social psychology suggest that the absence or presence of response-relevant objects in the current environment is relevant for a prime to be effective (Cesario and Jonas, 2014). This reasoning can serve as an explanation for differences in priming effects that occur when action is (supposed to be) public (Weaver et al., 2013) or public to a certain group, as in Boschini et al. (2014), where the gender of the interviewer plays a role. Since choices were private in our experiment, the lack of publicity of the choice could have biased our results or could be responsible for null results.

Fourth, Cesario and Jonas (2014) have further showed that the physical environment mediates priming effects. While a potential variation in the physical environment, which may reasonably be assumed for online experiments, has obviously not biased the results by Horton et al. (2011), it could, however, provide an explanation for the null findings in our study. We do not have any information about the situation participants were in while participating. One could, for example, suppose that activating a gender identity produces different effects depending on whether men or women are at home looking after their children or at work. In this respect, one could expect generally to find stronger gender priming effects in a relatively controlled setting, as in the lab, compared to online or field experiments. A recent meta-analysis on religious priming effects by Willard et al. (2016) only partially supports this assumption. The effects are strongest in the field, only medium in the lab and smallest in online experiments.

Fifth, Loersch and Payne (2014) report on a range of moderator variables identified in previous priming studies in social psychology. For example, high time pressure and multiple-task environments increase the impact of primes. We do not have any information about time pressure subjects might have experienced during our experiment or during the experiments mentioned in Section 2. In our experimental setting, people were not restricted to a certain time limit in choosing payment schemes. However, we could expect time pressure to be present when the duration of time subjects needed for the choice task was comparably short. Restricting

the sample to people who chose the payment schemes relatively quickly did not change the results.

Finally, primes that ask subjects to recall (relevant) past behavior or feelings and thus "self-generate the behavioral content from their memory" seem to be more effective than others (Loersch and Payne, 2014: 141). In this regard, the priming technique used by Weaver et al. (2013) serves as a good example: Men were asked to recall past behavior that demonstrated that they are "real men."

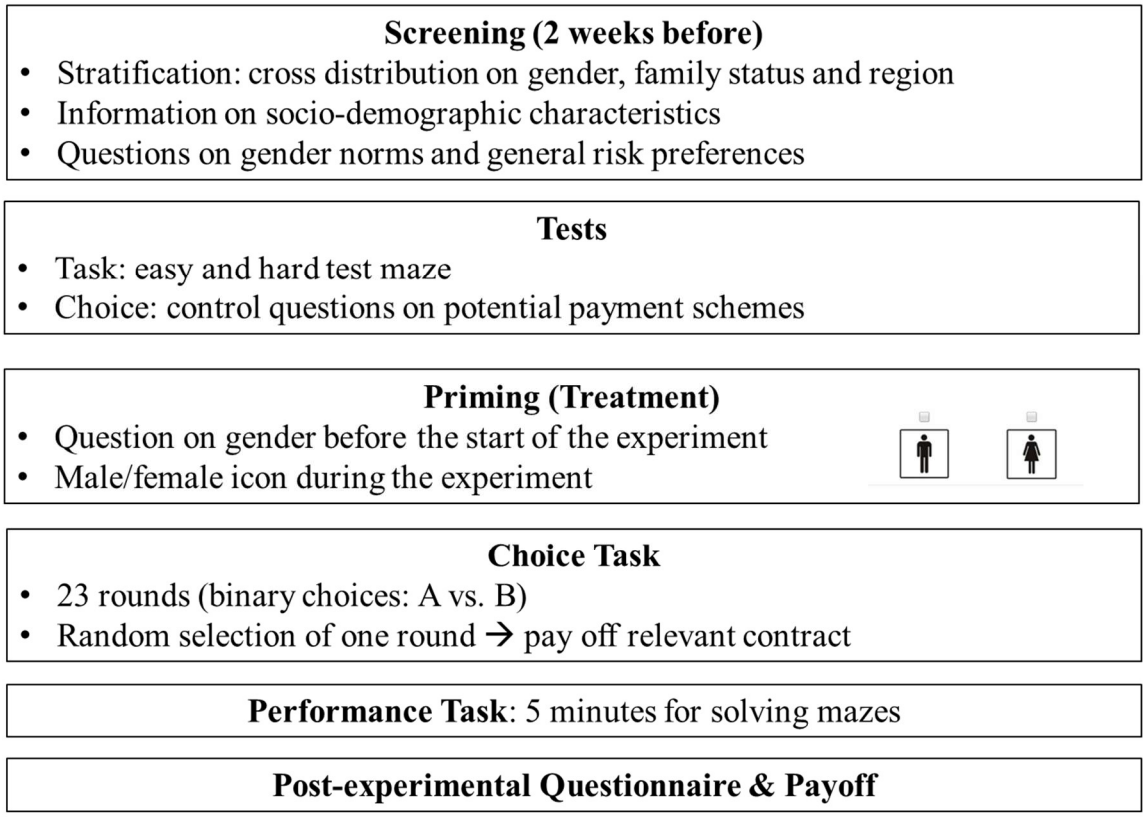
Some of the biases discussed here may also be held responsible for the absence of priming effects in our experiment. Future priming research in economics should consider these factors in designing and interpreting experiments to be able to conclude from the absence or presence of priming effects as sound evidence against or for social category effects.

4.4 Appendix

Experimental Procedure

Figure 17 pictures the experimental procedure. In a screening step, a market research institute selected a gross sample of 1,444 people by a uniform stratification procedure based on gender, family status and region for our purpose. Those participants filled out a questionnaire on their socio-demographic characteristics as well as their risk attitudes, and were invited to participate in "a scientific study". Our full net sample includes 883 people, almost equally distributed across gender (men vs. women), family status (single vs. in a relationship) and region (East vs. West).

Figure 17: Experimental Procedure



In the test part, participants were introduced to the task by solving one easy and one difficult trial maze. They were subsequently informed about their individual time performance on the screen, to make sure that they knew their absolute productivity. Afterward, they answered three control questions on hypothetical contracts, to ensure that they understood how the payment worked in the experiment.

Before the choice task, half of the participants were asked about their gender. To reinforce the gender identity priming, a screen wiper with a pictogram of a man or a woman appeared constantly on the screen for the remainder of the experiment.

In the choice task, participants were asked to choose one of two options (A vs B) in 23 successively presented rounds. There are four payment schemes: In the linear payment, subjects receive piece-rate pay for each maze solved. In the bonus contract, a participant receives a high piece-rate payment if she meets the threshold of five mazes or more but a low piece-rate payment if she fails to reach that threshold. The contract type competition distinguishes between competition against all participants and competition against a same-gender group. A participant receives a high piece-rate payment if she places among the best 30% of all or the same gender group but a low piece-rate payment if she fails. In every payment scheme, we additionally vary the difficulty level of the mazes (easy/hard) and the wage spread that provides a risk premium. The low risk setting with a relatively small wage spread (low piece-rate payment of €0.20 and high piece-rate payment of €1) serves as a reference. The risk premium is characterized by a larger wage spread (low piece-rate payment of €0.05 and high piece-rate payment of €2).

For the purpose of this paper, we focus on choice rounds with meaningful and comparable reference options. To reach this set of meaningful binary choices, we first dropped rounds where the piece-rate compensation is not the reference option. Second, we dropped rounds that included competition against a same-gender group since this could bias the gender priming effects. Third, only participants that passed a rationality check in the last round were considered.

When all choices had been made, one out of the 23 choice sets was drawn randomly. Each participant was informed about the selected choice set and the respective contract type by a short notice appearing on the computer screen. The participants then had five minutes to solve as many mazes as possible (performance task). We chose mazes as the real effort task since previous studies (for example Datta Gupta et al., 2013; Gneezy et al., 2003) had found them to be gender neutral in performance. In particular, a maze task is simple to communicate; and easy to understand and requires little knowledge or experience. It involves little randomness and allows measurement of performance either by the time needed to solve one maze or by the number of mazes solved correctly in a given period of time (see Gill and Prowse 2012). The payoffs were paid by bank transfer from the market research institute to the participants, to help guarantee the anonymity of the participants with regard to the authors.[47]

[47] In addition to their performance-related compensation every participant received a "show-up fee" of €2.

Figure 18: Sample Descriptives

Characteristics	**Women** (primed)	**Women**	**Men** (primed)	**Men**
Single*	.509	.495	.498	.500
	(.035)	(.035)	(.035)	(.035)
Living in East Germany*	.510	.481	.507	.495
	(.035)	(.035)	(.035)	(.035)
Age (years)	35.63	36.13	37.23	37.81
	(.538)	(.518)	(.531)	(.511)
Children in household*	.530	.558	.343	.332
	(.035)	(.035)	(.034)	(.033)
Income class	3.47	3.36	3.78	3.84
	(.106)	(.096)	(.112)	(.108)
Education: basic*	.015	.029	.035	.010
	(.009)	(.012)	(.013)	(.007)
Education: lower professional*	.465	.432	.408	.441
	(.035)	(.035)	(.035)	(.035)
Education: higher professional*	.168	.204	.229	.208
	(.026)	(.028)	(.030)	(.029)
Education: university /polytechnics*	.267	.277	.254	.277
	(.031)	(.031)	(.031)	(.032)
In education*	.054	.039	.060	.045
	(.016)	(.013)	(.017)	(.015)
Employed full Time*	.465	.461	.756	.777
	(.035)	(.035)	(.030)	(.029)
Employed part Time*	.272	.228	.060	.059
	(.031)	(.029)	(.017)	(.017)
Number of Obs.	202		201	

Note: * indicates a dummy variable.

Table 16: Full Table

	bonus	**competition**
prime*	0.0221	0.0049
	(.035)	(.038)
male*	0.196***	0.163**
	(.063)	(.070)
prime x male	-0.0475	-0.0232
	(.052)	(.058)
tradis*	0.0410	-0.0517
	(.056)	(.053)
prime x tradis	-0.0804	0.0720
	(.084)	(.082)
male x tradis	-0.0419	0.0680
	(.076)	(.082)
male x tradis x prime	0.109	-0.0452
	(.110)	(.012)
risk aversion*	0.0201***	0.0161**
	(.007)	(.007)
male x risk aversion	-0.0226**	-0.0128
	(.010)	(.011)
time for trial mazes	-0.0002	0.0005*
	(.000)	(.000)
male x time for trial mazes	0.0002	0.0005*
	(.000)	(.000)
spread*	0.0276*	0.0302*
	(.0156)	(.0168)
difficulty*	-0.200***	-0.139***
	(.018)	(.017)
spread x difficulty	-0.0710***	-0.0118
	(.021)	(.021)
single*	0.0087	0.0449
	(.027)	(.029)
east*	-0.0368	-0.005
	(.023)	(.026)
age	-0.0047***	0.0005
	(.002)	(.002)
children*	0.0211	-0.0027
	(.029)	(.030)
income	0.0063	0.0069
	(.011)	(.011)
employed fulltime*	-0.0382	-0.104**
	(.035)	(.041)
employed parttime*	-0.0057	-0.0298
	(.039)	(.044)
education low*	-0.1090*	-0.0289
	(.050)	(.061)
education mid*	-0.0172	-0.0402
	(.053)	(.066)
education high*	-0.0154	-0.0098
	(.052)	(.064)
education current*	-0.128*	-0.1080
	(.068)	(.081)
Constant	0.912***	0.239***
	(.090)	(.104)

Observations	3044	3044
R-squared	0.105	0.046

Note: * indicates a dummy variable. * p<.100, ** p<.050, *** p<.010.

5 Identifying Gender Differences in Exposure and Vulnerability: A Decomposition Analysis of the Gender Absenteeism Gap in Germany

5.1 Introduction

It has frequently been showed that women are, on average, more absent from work due to sickness than men in many European countries (Barmby et al., 2002; Mastekaasa and Melsom, 2014; Scheil-Adlung and Sandner, 2010; Spasova et al., 2016). Explanations for a gender gap in work absenteeism[48] are diverse and range from biological differences (Ichino and Moretti, 2009; Rockoff and Herrmann, 2009) to household context (e.g. Angelov et al., 2013; Beblo and Ortlieb, 2012) to working conditions (e.g. Beblo and Ortlieb, 2012; Casini et al., 2013; Sterud, 2014).

Moreover, previous studies indicate that occupational gender segregation in the labor market may be related to the gender gap in absenteeism. Two explanations for such a relationship are given in the literature. The first explanation argues that men and women are exposed to different health risks because they are employed in different occupations. There is evidence on this "difference in exposure" argument with mixed results: Some studies show that gender segregation favors women in terms of work absenteeism, in that the gap in absenteeism increases when the researchers control for employees' occupations (Mastekaasa and Dale-Olsen, 2000; Melsom and Mastekaasa, 2017). In contrast, Laaksonen et al. (2010) show a decrease of the gap once they control for occupations. The second explanation refers to gender differences within an occupation and argues that women and men differ in their reactions to certain characteristics of the occupation. However, there is no direct evidence on this "difference in vulnerability" argument regarding occupations or characteristics of the occupation.

Addressing these two explanations – gender differences in exposure and gender differences in vulnerability – this study contributes to the existing literature in three ways: First, the decomposition technique that is used is not only able to identify the contribution of gender differences in exposure to jobs (characteristics effect). Given the coefficients effect that measures gender differences in "reaction" to occupations in terms of sickness absence, it reveals a gender difference within an occupation and thus a gender difference in vulnerability regarding the occupation. Therefore, this study provides the first direct evidence of gender differences in

[48] Bekker et al. (2009) provide a comprehensive overview on determinants of gender differences in work absenteeism.

vulnerability effect regarding an occupation. Second, the large dataset offers a detailed occupational classification that includes information on both horizontal and vertical dimensions of a job. Therefore, I am able to distinguish between horizontal and vertical segregation throughout the analysis. Third, I use a current administrative dataset on certified sickness absences from the largest statutory health insurance in Germany (AOK), which results in a relatively large sample of approximately 7 million observations. It additionally includes information on diagnoses behind each sickness absence period which is not available in survey data.

This study finds that women are, on average, 6% more absent from work, when using the total number of sickness days in 2016. The decomposition shows that gender segregation works in favor of women since the gap would increase if women and men worked in the same jobs. This effect is driven by the horizontal dimension of the job, while vertical segregation explains a part of the gap. When holding the same occupation, women seem to be, on average, more vulnerable than men. However, as the difference-in-vulnerability argument suggests, the gender differences vary across characteristics of the occupation. Women seem to be more vulnerable in occupations with lower complexity levels while men seem to be more vulnerable in occupations that entail tasks with higher complexity levels and in the majority of occupational areas.

The remainder of the paper is organized as follows: Section 5.2 reviews related literature, and Section 5.3 describes methods and data. Section 5.4 reports results, followed by a discussion in Section 5.5. Section 5.6 concludes.

5.2 Occupational segregation and work absenteeism

There are several papers that address the relationship between an occupational gender segregation in the labor market and a gender gap in work absenteeism. The literature provides two lines of reasoning in explaining this relationship: the “difference in exposure” argument and the “difference in vulnerability” argument (Beblo and Ortlieb, 2012; Bekker et al., 2009; Mastekaasa and Melsom, 2014).

The first argument, of gender differences in exposure, relates to the fact that women and men are employed in different occupations, which presumably entail different health risks. Therefore, women and men are exposed to different health risks that result in differences in work absenteeism. However, whether women or men are worse off as regards health risks measured by work absenteeism is not clear, and the literature provides two opposing views

(Mastekaasa and Dale-Olsen, 2000; Mastekaasa and Melsom, 2014; Melsom and Mastekaasa, 2017): Given employer discrimination in terms of income, prestige and career opportunities, one can assume that the jobs where women find themselves can also be characterized by being less healthy. Contrary to this notion, there is also reason to assume that gender segregation works in favor of women in terms of work absenteeism. Assuming a traditional specialization within a couple, in that women are responsible for effort-intensive and energy-consuming housework while men experience less-energy-consuming leisure, Becker (1985) argues that women economize on energy and avoid effort-intensive and energy-consuming jobs. Since illness is one of the major threats to energy, women avoid less healthy jobs (Becker, 1985).

Empirical evidence is mixed, and the majority of papers point to segregation working in favor of women. To begin with, Mastekaasa and Dale-Olsen (2000) examine the impact of an individual's occupation or job on the probability of absence by using Norwegian administrative data. Although they assume women to work in less healthy occupations and jobs, their results detect an increase in gender differences when they control for employees' occupations and work-places. They conclude that gender differences in absenteeism are not due to women working in less healthy jobs than men. Moreover, they showed that the effect differs across diagnoses: While gender segregation seems to work in favor of women in terms of sickness absence due to musculoskeletal diseases, the opposite is true for psychological diseases (Mastekaasa and Dale-Olsen, 2000).

Using pooled data taken from EU Labor Force Surveys, Mastekaasa and Melsom (2014) support the overall finding for the majority of 17 European countries. They find an increase in gender differences in sickness absences when they control for 147 detailed occupational categories, and they conclude that women work in more healthy jobs than men in most of the countries, including Norway and Germany. However, they detect opposing effects for Denmark, Finland, Sweden and the United Kingdom. In a relatively recent study, that uses longitudinal administrative data for the entire population of Norwegian employees, Melsom and Mastekaasa (2017) show that the gender segregation in the labor market contributes to smaller gender differences, and, thus, it seems to work in favor of women. Contrary to this, Laaksonen et al. (2010) show that differences between occupations explain a substantial part of the observed gender differences in absenteeism by using a sample of municipal employees of Helsinki. Mastekaasa and Olsen (1998) find a decrease of the gap once they control for occupational categories for a sample of employees from government offices in Norway and an increase of the gap for a sample of employees in the Norwegian state railway.

The argument of gender difference in vulnerability suggests that there are gender differences in reactions to certain characteristics of a job. Since each occupation offers a mix of characteristics, within-occupational gender differences should vary across occupations (Mastekaasa and Melsom, 2014). There is only one paper that addresses the argument of gender difference in vulnerability regarding the occupation and provides empirical evidence. Mastekaasa and Melsom (2014) assume a female-dominated occupation to be better adapted to women and a male-dominated occupation to be better adapted to men, leading women to be more affected and thus more absent in male-dominated occupations and men to be more affected and thus more absent in female-dominated occupations. However, the results do not support their hypotheses. They show that gender differences are quite similar across gender-dominated and gender-balanced occupations.[49] Although not addressing the effect of vulnerability in their paper, Mastekaasa and Olsen (1998) provide evidence that gender differences in absenteeism within occupations varies across two samples of employees in different occupational areas in the labor market (government offices vs. Norwegian State Railway).

The present paper provides evidence on both, the effect of gender differences in exposure and the effect of gender differences in vulnerability regarding an occupation. For this purpose, I perform a decomposition analysis, which is common in the literature on the explanations of the gender gap in pay but unique in this strand of the research. This method reveals the effect of both gender segregation on the gender gap in absenteeism, which has already been shown, and gender differences within occupations, for which no direct evidence exists. Furthermore, I distinguish between a vertical and horizontal dimension of an occupation in my analysis.

5.3 Data and methodology

Although research seems to agree on a measure of work absenteeism as describing non-attendance of work due to self-certified or physician-certified sickness absence, there are several concepts of what work absenteeism expresses. From an economist's point of view, work absenteeism is not a simple response to a medical condition, as it is treated in epidemiological research, but an outcome of an individual's choice between work and leisure that is voluntary

[49] Though these studies do not address the "difference in vulnerability" argument, a range of studies investigate the impact of the gender composition of the occupation in gender differences in sickness absence. The results are mixed: Melsom and Mastekaasa (2017) and Kröger (2017) find a U-shaped relationship, in that a minority position in the occupation is related to higher sickness absence rates, thus resulting in a larger gender gap in absenteeism. Mastekaasa (2005) found no relation between sickness absence and gender composition for men and only weak effects for women.

and can be influenced by incentives (Barmby et al., 2002). Thus, the distinction lies in the voluntary or involuntary character of sickness absence. Some authors make a distinction by referring to the duration of the absence period: Short-term absences are considered to be more voluntary, and absences with longer durations are considered to be more involuntary. As Beblo and Ortlieb (2012) point out, this distinction is not straightforward since short-term absences could be also involuntary, for example, in the case of a severe headache, and long-term absences can also be voluntary, since an individual can choose to return to work even after several weeks of absenteeism (Beblo and Ortlieb, 2012). Consequently, Beblo & Ortlieb (2012) conceptualize work absenteeism as periods of no attendance at work due to illness-induced lack of capacity, irrespective of the duration (Beblo and Ortlieb, 2012). In the same vein, Kröger (2017) describes work absenteeism as reflecting both the health status of an individual and the person's labor market-related health behavior. In line with their arguments and as usual in this strand of the literature (Laaksonen et al., 2010; Mastekaasa, 2014; Mastekaasa and Melsom, 2014; Melsom and Mastekaasa, 2017), I do not distinguish between the voluntary or involuntary character of sickness absences.

5.3.1 Sample and variables

I use administrative data on sickness absences in 2016 – the most recent data available – from the largest German statutory health insurance (AOK). In Germany, roughly 90% of the population are members of the statutory health insurance, with the AOK covering one-third of this population.

The dataset on physician-certified sickness absences is restricted to the working population and does not include self-employed persons or retirees. For the purpose of this paper, I restrict the analysis to voluntarily or mandatorily insured employees between 20 and 60 years of age who were covered by the health insurance from 1January until 31December 2016. I exclude all persons in vocational training because deviations from general regulations regarding sick notes apply to the dual vocational training system in Germany. To reduce biological differences in the sample, I dropped all observations for whom pregnancy-related illnesses can be identified (ICD-10 O00–O99). These restrictions result in a sample of 6,856,746 employees.

Information about sickness absences is taken from physicians' sickness certifications. According to German regulations, employees are allowed to take self-certified absence spells with a maximum of three days' duration (§5 Entgeltfortzahlungsgesetz). This dataset does not include these short spells that do not require certification from a physician.

The dataset provides information about the employment relationship of workers, including employment status (part-time or full-time employed), the (temporary) nature of a contract and the five-digit occupational code of each individual's job, according to the 2010 occupational classification of the German Federal Statistical Office ("Klassifikation der Berufe", KldB 2010). An individual's occupation is measured at the three-digit level of this classification and represents the horizontal dimension of the job. There are at least 144 different occupational categories in the dataset. The fourth and fifth digits of this classification capture information about the vertical dimension of a job. The fourth digit gives information on whether the job includes supervisory and executive tasks. The fifth digit of the classification captures the complexity level of a job. There are four different categories, including unskilled or semiskilled activities, specialist activities, complex specialist activities, and highly complex activities (Paulus and Matthes, 2013).[50]

As can be deduced from Table 17, men are, on average, absent 15.98 days from work during 2016, and women's average total absence days amounted to 16.98 days. This results in a statistically significant gender absenteeism gap of a full day, or 6% for this sample. Although the direction of the gap is similar, the gap in the AOK sample is smaller compared to the gap in the whole population of the statutory health insurance, which amounted to 8% in 2016. This is driven by a smaller number of men's absence days in the AOK sample, while women's absence days are relatively equal (Busch, 2017). The difference can be attributed to characteristics of the AOK sample: Based on the origins of the AOK, this sample provides a relatively high share of employees in production, a sector that is dominated by men and characterized by a high number of absent days (Meyer et al., 2017).

As can be further deduced from Table 17, gender segregation occurs in this sample. The share of men in executive or supervisory positions (fourth digit of the classification) is nearly twice as large as women's, which is comparable to the German labor market (see tables in Eisenmenger and Schweinert-Albinius, 2014). Additionally, men are overrepresented in jobs that require specialist, complex or highly complex activities and underrepresented in jobs that include unskilled or semiskilled activities (dummies based on the fifth digit of the classification). To illustrate horizontal segregation in this sample, Table 17 shows the shares of men and women within occupational areas. For illustrative reasons, I present the effects on the aggregated level of occupational areas (sectors) on the one-digit level of the occupational classification although the three-digit code (captured by 144 dummy variables) is used in the

[50] A supervisory or executive activity applies only for occupations with specialist activities, complex specialist activities and highly complex specialist activities.

analysis. Table 17 shows that men are more frequently employed in the first five occupational areas, agriculture, production, construction, natural sciences and traffic, whereas the share of women is larger in commercial services, business organization, health care and humanities. The proportions of women and men are comparable to the proportions in the German labor market (see tables in Eisenmenger and Schweinert-Albinius, 2014).

Table 17: Descriptive Statistics by Gender

(Set of) Variables	Women	Men	p-value (gender difference)
Dependent variable			
Total annual absent days	16.98	15.98	<.0001
Vertical dimension of a job			
Executive or supervisor position*	.0355	.0604	<.0001
Complexity of the task			
Unskilled or semiskilled activities*	.2479	.2012	<.0001
Specialist activities*	.6150	.6400	<.0001
Complex specialist activities*	.0757	.0952	<.0001
Highly complex activities*	.0614	.0635	<.0001
Horizontal dimension of the job (one-digit level)			
Occupations in agriculture, forestry, farming, and gardening*	.0110	.0251	<.0001
Occupations in production of raw materials and goods, and manufacturing*	.1215	.3970	<.0001
Occupations in construction, architecture, surveying and technical building services*	.0057	.1240	<.0001
Occupations in natural sciences, geography and informatics*	.0124	.0310	<.0001
Occupations in traffic, logistics, safety and security*	.1418	.2328	<.0001
Occupations in commercial services, trading, sales, the hotel business and tourism*	.1874	.0669	<.0001
Occupations in business organization, accounting, law and administration*	.2104	.0655	<.0001
Occupations in health care, the social sector, teaching and education*	.2913	.0450	<.0001
Occupations in philology, literature, humanities, social sciences, economics, media, art, culture, and design*	.0184	.0126	<.0001
Military occupations*	.0001	.0001	.331
Other socio-economic characteristics			
Age in years	42.067	42.179	<.0001
Living in East Germany*	.1781	.1503	<.0001
Children *	.1132	.0459	<.0001
Employed full time*	.4972	.9059	<.0001
Temporary contract*	.1820	.1459	<.0001
Number of observations	**3,099,791**	**3,757,082**	

Note: Displayed values are means. * indicates a dummy variable. Values are from t-tests.

Regarding the employment level, the patterns in the dataset at hand are similar to the German labor market, too. Women are less often employed full time and more often employed temporarily. Additionally, information on the individual's age (in years) and the region of living (western states of Germany and Berlin vs. eastern states) is given. Women in this sample are slightly younger and more frequently live in East Germany. Since the dataset does not include information on children within the household, I proxy individuals' care responsibilities by using information on sickness absences due to children's illnesses since 2010. Thus, if a person was absent with a sick note reporting a children's illness between 2010 and 2016, the individual is considered to care for at least one child. According to that, women are responsible for the care of an ill child twice as often as men.

5.3.2 Decomposition method

In order to distinguish between the effects of gender differences in exposure and gender differences in vulnerability and the contribution of each of these effects in explaining the gender gap in work absenteeism, I refer to a decomposition technique.[51] Typically, the aim of a decomposition that is based on the technique by Blinder (1973) and Oaxaca (1973) is to divide the difference in mean outcomes of a variable between two groups – for example, men and women – into various explanatory factors. This technique is commonly used in the gender pay gap literature, where the mean difference in pay between men and women is divided into a part that can be explained by gender differences in the observable characteristics (characteristics effect or explained part), such as, for example, education, employment status, occupation and work experience. The other part reflects differences in remuneration to these characteristics (coefficients effect or unexplained part) between men and women (Fortin et al., 2011).

$$\begin{array}{lcll} \bar{Y}_W - \bar{Y}_M & = & [F(\overline{X_M \beta_W}) - F(\overline{X_W \beta_W})] & + [F(\overline{X_M \beta_W}) - F(\overline{X_M \beta_M})] \\ \textit{Raw gender gap} & = & \textit{characteristics effect} & \textit{coefficients effect} \end{array} \quad (1)$$

The mean difference in absenteeism Y between women (W) and men (M) is expressed by equation (1). The first component, the characteristics effect, reveals the part of the gap that can be explained by differences in men's and women's observable characteristics, such as their occupation and other socio-economic characteristics. The question that is answered is: How will the gender gap in absenteeism change if men and women are assumed to be similar in terms of the observed characteristics? The second component, called the coefficients effect, reflects

[51] An overview of common decomposition methods can be found in Fortin et al. (2011).

differences in the coefficients estimated for men and women and therefore depicts behavioral responses to the observables (Powers et al., 2011).

To examine the contribution of the gender-differences-in-exposure effect and the gender-differences-in-vulnerability effect in terms of the vertical and horizontal dimension of a job, I further decompose the gap into six parts (see equation (2)).

$$
\begin{aligned}
&\bar{Y}_W - \bar{Y}_M \\
&= \underbrace{[F(\overline{X_M^{complex}\beta_W^{complex}}) - F(\overline{X_W^{complex}\beta_W^{complex}})]}_{\textit{difference in exposure (job level)}} + \underbrace{[F(\overline{X_M^{complex}\beta_W^{complex}}) - F(\overline{X_M^{complex}\beta_M^{complex}})]}_{\textit{difference in vulnerability (job level)}} \\
&+ \underbrace{[F(\overline{X_M^{occ}\beta_W^{occ}}) - F(\overline{X_W^{occ}\beta_W^{occ}})]}_{\textit{difference in exposure (occ. group)}} + \underbrace{[F(\overline{X_M^{occ}\beta_W^{occ}}) - F(\overline{X_M^{occ}\beta_M^{occ}})]}_{\textit{difference in vulnerability (occ. group)}} \\
&+ \underbrace{[F(\overline{X_M^{char}\beta_W^{char}}) - F(\overline{X_W^{char}\beta_W^{char}})]}_{\textit{characteristics effect (others)}} + \underbrace{[F(\overline{X_M^{char}\beta_W^{char}}) - F(\overline{X_M^{char}\beta_M^{char}})]}_{\textit{coefficients effect (others)}}
\end{aligned}
\tag{2}
$$

The cumulative characteristics effect is broken down into the characteristics effect of each set of variables that describe the dimensions of interest. The left-hand side of the right side of the equation depicts the effect of gender differences in exposure in terms of the vertical and horizontal dimension of a job – the job level (fourth and fifth digits of the occupational classification) and the occupation (three-digit level of occupational classification), respectively. Moreover, the question answered is: What is the effect of vertical and horizontal segregation, or occupational segregation as a whole as the sum of both parts, on the gender absenteeism gap? Or, putting it differently: How would the gender gap in absenteeism change if an occupational segregation did not exist?

The right-hand side captures the effect of gender differences in vulnerability regarding the vertical and horizontal dimensions of a job. The question answered is: How do men and women differ in responses to these occupational dimensions? Are women or men more vulnerable to certain dimensions of the occupation? Again, adding these effects expresses the vulnerability regarding the occupation without distinguishing between the vertical and horizontal dimensions of the occupation.

Additionally, I use a set of control variables that are assumed to influence gender differences in absenteeism. As usual in this strand of the literature, I control for age, employment status (full time vs. part time) and a proxy for being responsible for care during children's illnesses, to capture effects of the presence of children in the household. Moreover, I add the temporary nature of work contracts, as is the region of living in Germany, since it has been shown that

these are related to work absenteeism (Meyer et al., 2017), and the occurrences differ between men and women in this sample. Although other studies control for an individual's education, I refrain from including education, since the vertical dimension of the job, which is not included in these studies, is correlated with the level of education.

5.4 Results

Since the dependent variable is a count variable, I base the decomposition of the gender absenteeism gap on a negative binomial model and adopt the detailed decomposition procedure for nonlinear models by Powers et al. (2011).[52] I analyze the gender gap in work absenteeism irrespective of the diagnosis first, which is followed by analyses of the gender gap in absenteeism for two main diagnostic categories, since research has shown that the relationship between occupational segregation and the gender gap in work absenteeism to differ across diseases (Mastekaasa and Dale-Olsen, 2000; Melsom and Mastekaasa, 2017).

5.4.1 Decomposition results irrespective of diagnosis

Table 18 presents the results for the decomposition of the gender absenteeism gap irrespective of the diagnosis. The numbers reveal the relative effects of each (set) of variable(s) on the gender absenteeism gap. Aggregated effects regarding the dimensions of interest are given in bold numbers.

As the bold numbers for the gender-difference-in-exposure effect in the first row in Table 18 show, the fact that men and women are employed in different occupations increases the gap by 77%. This indicates that the gender absenteeism gap would be 77% larger if occupational segregation did not exist in this sample. Thus, occupational segregation as a whole seems to work in favor of women, which is in line with the majority of studies. Moreover, the magnitude is comparable to the findings of Mastekaasa and Melsom (2014), who used EU Labor Force Survey data, including Germany among other countries: Taking comparable occupational categories (three-digit International Standard Classification of Occupations) into account leads to an increase of 89% of the gender pay gap in the probability of sick leave in Mastekaasa and Melsom's analysis.

[52] This procedure overcomes the problem of coefficients that are invariant to the choice of dummy variables' reference category, as mentioned by Oaxaca and Ransom (2011) by a normalization of dummy variables as proposed by Yun (2005). Moreover, it addresses the problem of sensitivity in decomposition results to the order of explanatories in a non-linear estimation by using a strategy of sequential replacement and randomization (Powers et al., 2011), proposed by Jann (2008).

Does this hold for vertical and horizontal segregation, too? As the aggregated effects of gender difference in exposure regarding these dimensions show, this is not true: Only horizontal segregation works in favor of women since the gap nearly doubled in a world without it. In contrast, vertical segregation explains 11% of the gap. Thus, in a world without vertical segregation, the gender gap would be 11% smaller. Investigating these effects in more detail reveals that the overrepresentation of women in unskilled and semiskilled tasks is responsible for the largest part of the effect. If women and men were equally employed in these occupations, the gender absenteeism gap would be smaller.

Table 18: The Effect of Gender Differences in Exposure and Vulnerability on the Gender Absenteeism Gap

(Set of) Explanatories	Exposure	Vulnerability
Occupation	**.77**	**.77**
Vertical dimension of an occupation	**-.11**	**1.05**
Executive or supervisor position	.01	-.00
Unskilled or semiskilled activities	-.16	.22
Specialist activities	.08	.99
Complex specialist activities	-.02	-.01
Highly complex activities	-.01	-.15
Horizontal dimension of an occupation	**.88**	**-.28**
Occupations in agriculture, forestry, farming, and gardening	-.01	.01
Occupations in production of raw materials and goods, and manufacturing	.41	-.16
Occupations in construction, architecture, surveying and technical building services	.14	.01
Occupations in natural sciences, geography and informatics	-.05	.01
Occupations in traffic, logistics, safety and security	.16	-.01
Occupations in commercial services, trading, sales, the hotel business and tourism	.23	-.31
Occupations in business organization, accounting, law and administration	.44	.08
Occupations in health care, the social sector, teaching and education	-.43	.12
Occupations in philology, literature, humanities, social sciences, economics, media, art, culture, and design	-.01	-.01
Military occupations	.00	-.00

Note: For illustrative reasons, I present the components of the occupations aggregated on the one-digit level (occupational areas) of the occupational classification although a more detailed classification of the occupations (three-digit code, 143 occupational codes) is used in the analysis. A full table showing using occupational areas, including all controls, can be found in the Appendix.

Regarding the occupational areas, striking results on the effect of gender differences in exposure can be found in occupations in health care, the social sector, teaching and education. In these

occupations, the overrepresentation of women does not seem to work in favor of women. The gender gap in absenteeism would decrease by nearly one-half (-43%) if gender segregation in these occupations did not exist.

The third column presents how gender differences in absenteeism within occupations contribute to the gender gap in absenteeism. Overall, when women and men hold the same occupation, women are more often absent than men, which drives the gap by 77%. However, the effects of vulnerability differ across the vertical and horizontal dimensions of an occupation, as assumed by the argument of gender differences in vulnerability described next.

While women seem to be more vulnerable, on average, when holding the same job level – since the gender effect in vulnerability drives the gap by 105% – men seem to be, on average, more vulnerable when holding the same occupational group, since this reduces the gap by 28%. Investigating these effects in detail discloses interesting results: While women are more often absent in lower occupational levels, such as occupations characterized by unskilled, semiskilled and specialist activities, the opposite is true for higher job levels. Actually, men are more often absent than women when occupations are characterized by tasks with higher complexity levels, since this effect of vulnerability decreases the gap by 16%. Interestingly, the gap is not affected by gender differences within occupations with executive and supervisory tasks. This highlights that men and women do not differ in terms of work absenteeism within supervisory and executive positions.

Regarding the occupational area, men seem to be more vulnerable, on average, but the effects differ across the occupational areas. Again, striking results are given by occupations in health care, the social sector, teaching and education, where women seem to be more vulnerable than men which drives the gap by 12%.

In summary, as the difference in exposure effect shows, occupational gender segregation in the labor market in this sample favors women in terms of work absenteeism and women seem to select into healthy occupations. In a world without gender segregation, the gap would be larger than the observed raw gap. As the overall difference in vulnerability effect shows, women seem to be more vulnerable, given the same occupation. The effects of gender differences in both exposure and vulnerability vary across the dimensions describing the occupation.

5.4.2 Decomposition results for two main disease categories

In the following, I take the diagnosed diseases behind each sickness absence period into account. Since Mastekaasa and Dale-Olsen (2000) and Melsom and Mastekaasa (2017) showed heterogeneous effects of occupational segregation on the gender gap in absenteeism across

psychological and physical diseases using Norwegian data, I assume to find heterogeneous effects across these diseases in my analysis, too. Table 19 reports decomposition results of the gender gap in absenteeism with two disease categories based on the International Classification of the Diseases and Related Health Problems (ICD-10): diseases of the musculoskeletal system (ICD-10: M00-M99) and mental and behavioral disorders (ICD-10: F00-F99). These diseases serve as an example for physical and psychological diseases and are comparable to the categories used in Melsom and Mastekaasa (2017) and Mastekaasa and Dale-Olsen (2000). Moreover, these two main disease categories were responsible for one-third of all absence days in the AOK in 2016 (Meyer et al., 2017).

While women are, on average, 70% more absent with mental and behavioral disorders, men are, on average, 14% more absent with diseases of the musculoskeletal system (see Table 20 in the Appendix for mean absent days in 2016 by gender and the four most frequent diagnoses that account for 47% of all absent days). While the first diagnoses drive the gender gap, the latter limits the gender gap in absenteeism. As the gender-difference-in-exposure effect for diseases of the musculoskeletal system reveals, this gender gap in absenteeism decreased by 130% in a world without occupational segregation. Following, as regards exemplary physical diagnoses, occupational segregation as a whole does not work in favor of men since this gap increased in a world without it, but of women, which is in line with the overall results presented in Table 18. Contrary to that, the gender-difference-in-exposure effect for mental and behavioral disorders shows that the gender gap in absenteeism decreased in a world without occupational segregation by 45%. Thus, as regards exemplary psychological diseases, occupational segregation does not work in favor of women, thus women seem to select into unhealthy occupations in this respect, which matches Mastekaasa and Dale-Olsen's (2000) findings for Norway.

Table 19: The Effect of Gender Differences in Exposure and Vulnerability on the Gender Absenteeism Gap for two disease categories

	Diseases of the musculo-skeletal system (M00-M99)		**Mental and behavioural disorders (F00-F99)**	
Gender gap in absent days	-.14		.70	
	Exposure	**Vulner-ability**	**Exposure**	**Vulner-ability**
Occupation	**-1,30**	**-0,05**	**-0,45**	**0,36**
Vertical dimension of an occupation	**0,06**	**-0,79**	**0,00**	**0,08**
Executive or supervisor position	-0,01	0,01	0,00	0,00
Unskilled or semiskilled activities	0,10	-0,04	-0,01	0,02
Specialist activities	-0,04	-0,84	0,01	0,07
Complex specialist activities	0,01	-0,02	0,00	0,00
Highly complex activities	0,01	0,11	0,00	-0,01
Horizontal dimension of an occupation	**-1,36**	**0,74**	**-0,45**	**0,28**
Occupations in agriculture, forestry, farming, and gardening	-0,01	-0,01	-0,01	0,00
Occupations in production of raw materials and goods, and manufacturing	-0,34	0,15	-0,06	-0,03
Occupations in construction, architecture, surveying and technical building services	-0,27	-0,02	-0,08	-0,01
Occupations in natural sciences, geography and informatics	0,07	-0,01	0,00	0,00
Occupations in traffic, logistics, safety and security	-0,13	-0,04	0,01	0,01
Occupations in commercial services, trading, sales, the hotel business and tourism	-0,11	0,45	-0,03	0,03
Occupations in business organization, accounting, law and administration	-0,45	-0,11	-0,04	0,08
Occupations in health care, the social sector, teaching and education	-0,11	0,29	-0,24	0,19
Occupations in philology, literature, humanities, social sciences, economics, media, art, culture, and design	-0,02	0,02	-0,01	0,00
Military occupations	0,00	0,00	0,00	0,00

Regarding the gender-difference-in-vulnerability effect, Table 19 indicates that women are, on average, more often absent from work due to diseases of the musculoskeletal system and mental diseases than men within the same occupation. However, when taking diseases of the musculoskeletal system into account, the effects differ across the vertical and horizontal dimensions of an occupation: Men are less absent when holding the same job level but seem to be more vulnerable when holding the same occupational group which is quite similar to what was found for the gender gap in absenteeism irrespective of the diagnosis. As regards work absenteeism with mental and behavioral disorders, women seem to be more vulnerable in nearly all job levels and all occupational areas which deviates from the overall picture irrespective of the diagnosis.

5.5 Discussion

The present paper provides empirical evidence on the effect of gender differences in exposure and gender differences in vulnerability on the gender gap in absenteeism. The gender gap in physician-certified days of absenteeism amounts to 6% in this sample. As already discussed, the gender gap in absenteeism in this AOK sample is relatively small compared to the gap in the whole population of members of the statutory health insurance thus presenting a lower bound for the gap. Moreover, the dataset includes only certified absence spells; it does not include short spells that did not require certification from a physician. Unfortunately, there is no German dataset available that enables a distinction between self-certified and physician-certified absence days. According to Norwegian data, where similar regulations apply, an employee does not need a sickness certification for short spells up to three days,[53] the gender gap in sickness absences is larger for physician-certified absence days compared to self-certified absence days of maximum three days (Mastekaasa and Olsen, 1998). Thus, the gender gap in days of work absenteeism with a physician's certification in the dataset at hand is presumably larger than a gender gap in total days of absenteeism irrespective of the certification.

The decomposition disentangles the effect of gender differences in exposure and the effect of gender differences in vulnerability. As the gender-difference-in-exposure effect reveals, occupational segregation, as a whole, works in favor of women. Thus, other papers, as well as this one, conclude that women seem to select into healthy occupations. However, this interpretation has to be treated with caution due to endogeneity issues. It is not clear whether the gender-difference-in-exposure effect is driven by relatively unhealthy women or men in this occupation or by the occupation to provide a relatively unhealthy environment. A more objective measure of health risks within an occupation could be a solution. Nevertheless, I use the same methodology used in the majority of papers on the gender pay gap and ignore this endogeneity issue. Moreover, one should bear in mind that the results are presumably selectivity biased, since health status and health behavior could be related to labor market participation decisions. Albeit, this bias pertains to all studies analyzing work absenteeism.

This paper provides the first evidence on the effect of gender differences in vulnerability regarding different occupations. It shows that within-occupational gender differences exist and, as assumed by the gender-differences-in-vulnerability argument, that men and women differ in

[53] Meanwhile the possibility of self-certified absences has been expanded in Norway and employees do not need certifications until the seventh day.

vulnerability regarding the characteristics of an occupation, such as the job level, the personnel responsibility (executive and supervisory tasks) and the occupational area. I show that women seem to be more vulnerable in occupations that can be characterized as having lower complexity levels, while men seem to be more vulnerable when the activities are more complex. Moreover, women seem to be more vulnerable in occupations in health care and business organizations, while men are more vulnerable in occupations in construction and commercial services.[54]

As the heterogeneity of the gender-difference-in-vulnerability effect across the characteristics of an occupation shows, "a refinement of [the] general view" (Bekker et al., 2009: 409) of women being more often absent than men seems to be necessary. From a feminist economists' view, this finding is particularly important since, acting as a stereotype, it could bias employees' hiring decisions and thus lead to male favoritism due to (statistical) gender discrimination. In this respect, the fact that men and women in this sample do not differ in terms of work absenteeism when holding supervisory and executive positions is of particular importance.

As the large constant indicates, a relatively large part of the gender gap in sickness absences cannot be explained by the variables in the dataset at hand. First, the dataset lacks information on children that live in the household and on marital status that is shown to drive the gender gap in absenteeism (Barmby et al., 2002). To control for care responsibilities that could differ across genders, I created a dummy using sickness absences due to children's illnesses from 2010 to 2016 as a proxy for being considered to care for at least one child. Therefore, only 5% of men and 11% of women are considered to have care responsibilities. However, the results are as expected: The fact that women more frequently are considered to have care responsibilities compared to men explains 6% of the gap (see Table 5 in the Appendix). Since I do not observe the existence of children in the household but the perception of care responsibilities in case of a child's illness, I assess the effect of care responsibilities in my analysis as providing a lower bound. Second, although I used the most detailed occupational classification available, systematic gender differences in working conditions within one occupational category, such as workplace characteristics, are likely to occur. However, it seems unlikely that the main results will be challenged, since Mastekaasa and Olsen (1998) find that workplaces have only a minor impact compared to occupations. Investigating characteristics behind occupational categories, such as workplaces or objective health measures that are related

[54] Interestingly, women's higher vulnerability occurs in an occupational area that is female dominated, and men's vulnerability is larger in occupations in construction, a male-dominated area. Thus, my findings challenge the gender-minority hypothesis.

to an occupation, in explaining the gender gap in absenteeism could be a fruitful path for future study.

Sources of these gender differences, such as women being less healthy or women being more inclined than men to stay at home when experiencing the same health problems (Mastekaasa and Melsom, 2014), that could also bias the results are beyond the scope of this work.

5.6 Conclusion

The use of a decomposition analysis to identify the effects of both gender differences in exposure and gender differences in vulnerability constitutes a unique method in this strand of the literature. The results indicate that, on average, gender segregation works in favor of women. Women seem to select into healthier occupations, since the gender gap in absenteeism would decrease if gender segregation did not exist. While gender segregation seems to work in favor of women in terms of work absenteeism due to musculoskeletal diseases, the opposite is true for psychological diseases. As the gender difference within an occupation shows, women seem to be more vulnerable, on average, but the effects vary strongly across characteristics of the occupation. Women are more vulnerable than men in lower job levels but less vulnerable in higher job levels. Moreover, within occupational gender differences strongly vary across occupational areas. Consequently, these heterogeneities call for a "a refinement of [the] general view" (Bekker et al., 2009: 409) of women being more often absent from work than men.

5.7 Appendix

Table 20: Mean Days of Absence in 2016 by Gender and Most Frequent Diagnoses (ICD-10)

ICD-10	Women	Men	Gender Absenteeism Gap
Mental and behavioral disorders	2.51	1.48	.70
Diseases of the respiratory system	2.89	2.51	.15
Diseases of the musculoskeletal system and connective tissue	4.37	5.11	-.14
Diseases of the digestive system	0.87	1.04	-.17
Injury, poisoning and certain other consequences of external causes	1.61	2.81	-.43

Table 21: Full Table of Decomposition Results

(Set of) Explanatories	Characteristics Effect	Coefficients Effect
Vertical dimension of an occupation	**.11**	**-1.05**
Executive or supervisor position*	-.01	.00
Complexity of the task	.12	-1.05
Unskilled or semiskilled activities	.16	-.22
Specialist activities	-.08	-.99
Complex specialist activities	.02	.01
Highly complex activities	.01	.15
Horizontal dimension of an occupation	**-.88**	**.28**
Occupations in agriculture, forestry, farming, and gardening	.01	-.01
Occupations in production of raw materials and goods, and manufacturing	-.41	.16
Occupations in construction, architecture, surveying and technical building services	-.14	-.01
Occupations in natural sciences, geography and informatics	.05	-.01
Occupations in traffic, logistics, safety and security	-.16	.01
Occupations in commercial services, trading, sales, the hotel business and tourism	-.23	.31
Occupations in business organization, accounting, law and administration	-.44	-.08
Occupations in health care, the social sector, teaching and education	.43	-.12
Occupations in philology, literature, humanities, social sciences, economics, media, art, culture, and design	.01	.01
Military occupations	.00	.00
Other socio-economic characteristics	**-1.06**	**-1.17**
Age in years	-.04	-.79
Living in East Germany	.00	.27
Children	.06	-.14
Employed full time	-1.06	-.62
Temporary contract	-.03	.11
Constant		4.77
Total	**-1.83**	**2.83**

6 Conclusion

In this dissertation I provide empirical evidence on supply-side determinants and constraints of gender differences in the labor market by using experimental and administrative data from Germany. The dimensions of investigation range from determinants and constraints of time-allocation decisions of heterosexual couples within an experimental setting, to supply-side driving forces of an experimental gender pay gap, to explanations for gender differences in days of absenteeism due to illness, based on administrative data. Beyond this, one paper tries to provide causal experimental evidence on the impact of gender identity on gender differences in selection into payment schemes.

In particular, in Chapter 2, we detect that real-world findings regarding couples' time-allocation decisions cannot be found in a "gender neutral" lab: Women and men do not differ in terms of labor supply, on average, and the allocation of housework follows an economic rationale with opportunity costs determining couples' decisions. In Chapter 3 we investigate supply-side determinants of gender differences in pay between men and women in an experimental setting. We show that women prefer different payment schemes than men; men's preferences for risk and competition are larger, on average, than women's; and men and women perform differently conditional on a given contract, in that men outperform women, on average. In Chapter 4 we refer to a priming technique to investigate the impact of gender norms on selection into payment schemes. Although we find stable gender differences in preferences, evidence for a relationship between stated gender norms and gender differences in selection is weak, and we fail to produce effects from activating these gender norms by using a priming stimulus. In a decomposition analysis undertaken in Chapter 5 I show that gender segregation, commonly associated with lower income, prestige and career opportunities, actually works in favor of women in terms of work absenteeism. Additionally, women are, on average, more often absent than men when holding the same occupation.

Although the topics and methods of investigations are diverse, I summarize that this dissertation reveals that women and men indeed behave differently, on average. However, two aspects regarding these gender differences have to be mentioned: First, I examined gender differences on average, which does not mean that women's behavior is distinct from men's behavior (Nelson, 2015). We cannot preclude the existence of similarities and overlaps between distributions of men's and women's outcomes, as Nelson (2015) finds this to be the case in documented gender differences in risk preferences. Thus, generalizing statements should be avoided. This is particularly important to note because these generalizing statements could lead

to stereotypical thinking about men and women, which in turn could serve as a basis for employer discrimination, and, as a consequence, constrain women's opportunities.

A somewhat related second point is, that the context matters. Summarizing experimental findings on gender differences for a range of outcomes, Eswaran (2014) draws the conclusion that "apart from economics, culture matters, politics matters, biology matters, psychology matters, history matters, law matters and religion matters" (Eswaran, 2014: 57). In this dissertation I show, in Chapter 2, that gender differences in labor supply are restricted to married couples *only*. The gender difference in performance of the task that drives the experimental gender pay gap in Chapter 3 is *conditional* to the contract that applied. Thus, it is not a universal difference in productivity for all situations or all contexts. In Chapter 2 I find that occupational segregation working in favor of women in terms of work absenteeism and that women are, on average, more absent when holding the same occupation, but the results vary strongly across occupations. Taken together, gender differences discussed in this dissertation are heterogeneous across institutions and contexts.

Since we know that context and institutions matters, it seems reasonable to follow Bohnet's (2016) call for interventions on the institutional level rather than on the individual level to enhance gender equality in terms of labor supply decisions. She argues that people often make poor decisions that are biased by, for example, stereotypes, and that a de-biasing of each individual's mind would be difficult and expensive. She refers to the idea of choice architecture discussed by Thaler and Sunstein (2009) and argues that a de-biasing of institutions would be more effective. Therefore, a future research agenda in examining gender differences in supply-side determinants is recommended, to identify when and where these gender differences in behavior occur and detect the roots of gender differences in institutions. Once the roots are identified, researchers should experiment what might close gender gaps based on insights from behavioral economics and test a re-design of the environment (Bohnet, 2016).

7 References

Akerlof, G.A., Kranton, R.E., 2000. Economics and Identity. The Quarterly Journal of Economics 115 (3), 715–753.

Alm, J., 2010. Testing Behavioral Public Economics Theories in the Laboratory. National Tax Journal 63 (4 Part 1), 635–658.

Altonji, J.G., Blank, R.M., 1999. Race and gender in the labor market. Handbook of labor economics 3, 3143–3259.

Angelov, N., Johansson, P., Lindahl, E., 2013. Gender Differences in Sickness Absence and the Gender Division of Family Responsibilities. IZA Discussion Papers 7379.

Apps, P., Rees, R., 2009. Public economics and the household. Cambridge University Press, Cambridge and UK and and New York.

Apps, P.F., van Long, N., Rees, R., 2014. Optimal Piecewise Linear Income Taxation. Journal of Public Economic Theory 16 (4), 523–545.

Arrow, K.J., 1972. Models of Job Discrimination, in: Pascal, A.H. (Ed.), Racial Discrimination in Economic Life. Mass., Lexington Books, DC Heath and Co, pp. 83–102.

Auspurg, K., Iacovou, M., Nicoletti, C., 2017. Housework share between partners: Experimental evidence on gender-specific preferences. Social science research 66, 118–139.

Azmat, G., Petrongolo, B., 2014. Gender and the labor market: What have we learned from field and lab experiments? Labour Economics 30, 32–40.

Bach, S., Geyer, J., Haan, P., Wrohlich, K., 2011. Reform of Income Splitting for Married Couples: Only Individual Taxation Significantly Increases Working Incentives. DIW Economic Bulletin 1 (5), 13–19.

Bandiera, O., Fischer, G., Prat, A., Ytsma, E., 2016. Do women respond less to performance pay? Building evidence from multiple experiments. CEPR Discussion Paper (11724).

Barg, K., Beblo, M., 2012. Does Sorting into Specialization Explain the Differences in Time Use between Married and Cohabiting Couples? An Empirical Application for Germany. Annals of Economics and Statistics (105-106), 127–152.

Bargain, O., Orsini, K., Peichl, A., 2014. Comparing Labor Supply Elasticities in Europe and the United States New Results. Journal of Human Resources 49 (3), 723–838.

Bargh, J.A., 2006. What have we been priming all these years? On the development, mechanisms, and ecology of nonconscious social behavior. European journal of social psychology 36 (2), 147–168.

Bargh, J.A., Chen, M., Burrows, L., 1996. Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. Journal of Personality and Social Psychology 71 (2), 230.

Barmby, T.A., Ercolani, M.G., Treble, J.G., 2002. Sickness absence: An international comparison. The Economic Journal 112 (480).

Bateman, I.J., Munro, A., 2005. An Experiment on Risky Choice Amongst Households. Economic Journal 115, C176-C189.

Bateman, I.J., Munro, A., 2009. Household Versus Individual Valuation: What's the Difference? Environmental and Resource Economics 43 (1), 119–135. 10.1007/s10640-009-9268-6.

Beblo, M., 2015. Experimente in der Familienökonomik, in: Held, M. (Ed.), Reformen und ihre politisch-ökonomischen Fallstricke, vol. 14. Metropolis-Verl., Marburg, pp. 141–164.

Beblo, M., Beninger, D., 2017. Do husbands and wives pool their incomes? A couple experiment. Review of Economics of the Household 15 (3), 779–805.

Beblo, M., Ortlieb, R., 2012. Absent from work? The impact of household and work conditions in Germany. Feminist economics 18 (1), 73–97.

Becker, G., 1991. A Treatise on the Family, 1991. Harvard University Press, Cambridge.

Becker, G.S., 1957. The economics of discrimination. The American Catholic Sociological Review 18, 276.

Becker, G.S., 1985. Human capital, effort, and the sexual division of labor. Journal of Labor Economics, S33-S58.

Bekker, M.H.J., Rutte, C.G., van Rijswijk, K., 2009. Sickness absence: A gender-focused review. Psychology, Health & Medicine 14 (4), 405–418.

Bénabou, R., Tirole, J., 2011. Identity, morals, and taboos: Beliefs as assets. The Quarterly Journal of Economics 126 (2), 805–855.

Beninger, D., Laisney, F., Beblo, M., 2007. Welfare analysis of a tax reform for Germany: A comparison of the unitary and collective models of household labour supply. Journal of Population Economics 20, 869–893.

Benjamin, D.J., Choi, J.J., Fisher, G., 2016. Religious identity and economic behavior. Review of Economics and Statistics 98 (4), 617–637.

Benjamin, D.J., Choi, J.J., Strickland, A.J., 2010b. Social Identity and Preferences. American Economic Review 100 (4), 1913–1928.

Bertrand, M., 2011. New Perspectives on Gender, in: Handbook of Labor Economics. Elsevier, pp. 1543–1590.

Bertrand, M., Kamenica, E., Pan, J., 2015. Gender Identity and Relative Income within Households. The Quarterly Journal of Economics 130 (2), 571–614.

Blau, F.D., Ferber, M.A., Winkler, A.E., 2014. The economics of women, men, and work.

Blau, F.D., Kahn, L.M., 2007. Changes in the Labor Supply Behavior of Married Women: 1980-2000. Journal of Labor Economics 25, 393–438.

Blau, F.D., Kahn, L.M., 2017. The gender wage gap: Extent, trends, and explanations. Journal of Economic Literature 55 (3), 789–865.

Blinder, A.S., 1973. Wage Discrimination: Reduced Form and Structural Estimates. The Journal of Human Resources 8 (4), 436.

Bohnet, I., 2016. What works: Gender equality by design. The Belknap Press of Harvard University Press, Cambridge, Massachusetts, London, England, 385 pp.

Boschini, A., Dreber, A., Essen, E. von, Muren, A., Ranehill, E., 2014. Gender and economic preferences in a large random sample. Stockholm University, Department of Economics.

Boschini, A., Dreber, A., Essen, E. von, Muren, A., Ranehill, E., 2015. Gender priming and altruism in a random sample. Stockholm University, Department of Economics.

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

Brines, J., 1994. Economic dependency, gender, and the division of labor at home. American Journal of Sociology 100 (3), 652–688.

Burow, N., Beblo, M., Beninger, D., Schröder, M., 2017. Why Do Women Favor Same-Gender Competition? Evidence from a Choice Experiment. DIW Discussion Papers 1662.

Busch, K., 2017. Die Arbeitsunfähigkeit in der Statistik der GKV, in: Badura, B., Ducki, A., Schröder, H., Klose, J., Meyer, M. (Eds.), Fehlzeiten-Report 2017: Krise und Gesundheit - Ursachen, Prävention, Bewältigung. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 485–495.

Buser, T., Niederle, M., Oosterbeek, H., 2014. Gender, Competitiveness, and Career Choices. The Quarterly Journal of Economics 129 (3), 1409–1447.

Cadsby, C.B., Servátka, M., Song, F., 2013. How competitive are female professionals? A tale of identity conflict. Journal of Economic Behavior & Organization 92, 284–303.

Carr, P.B., Steele, C.M., 2010. Stereotype threat affects financial decision making. Psychological science 21 (10), 1411–1416.

Casini, A., Godin, I., Clays, E., Kittel, F., 2013. Gender difference in sickness absence from work: A multiple mediation analysis of psychosocial factors. The European Journal of Public Health 23 (4), 635–642.

Cesario, J., Jonas, K.J., 2014. Replicability and models of priming: What a resource computation framework can tell us about expectations of replicability. Social Cognition 32 (Supplement), 124–136.

Chang, D., Chen, R., Krupka, E., 2014. Social norms and identity driven choice. University of Michigan.

Charness, G., Kuhn, P., 2011. Lab labor: What can labor economists learn from the lab?, in: Handbook of labor economics, vol. 4. Elsevier, pp. 229–330.

Chen, Y., Li, S.X., Liu, T.X., Shih, M., 2014. Which hat to wear? Impact of natural identities on coordination and cooperation. Games and Economic Behavior 84, 58–86.

Cochard, F., Couprie, H., Hopfensitz, A., 2015. What if women earned more than their spouses? An experimental investigation of work-division in couples. Experimental Economics, 1–22.

Croson, R., Gneezy, U., 2009. Gender Differences in Preferences. Journal of Economic Literature 47 (2), 448–474. 10.1257/jel.47.2.448.

Crossley, T.F., Jeon, S.-H., 2007. Joint Taxation and the Labour Supply of Married Women: Evidence from the Canadian Tax Reform of 1988. Fiscal Studies 28 (3), 343–365.

D'Acunto, F., 2015. Identity, Overconfidence and Investment Decisions. University of Maryland.

Dalton, P.S., Gonzalez, V.H., Noussair, C.N., 2015. Paying with Self-Chosen Goals: Incentives and Gender Differences. SSRN Electronic Journal.

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

Davies, P.G., Spencer, S.J., Quinn, D.M., Gerhardstein, R., 2002. Consuming Images: How Television Commercials that Elicit Stereotype Threat Can Restrain Women Academically and Professionally. Personality and Social Psychology Bulletin 28 (12), 1615–1628.

Davis, J.B., 2006. Social identity strategies in recent economics. Journal of Economic Methodology 13 (3), 371–390.

Decoster, A., Haan, P., 2014. Welfare Effects of a Shift of Joint to Individual Taxation in the German Personal Income Tax. FinanzArchiv: Public Finance Analysis 70 (4), 599–624.

Dee, T.S., 2014. Stereotype Threat and the Student-Athlete. Economic Inquiry 52 (1), 173–182.

DellaVigna, S., 2009. Psychology and economics: Evidence from the field. Journal of Economic Literature 47 (2), 315–372.

Dickinson, D.L., 1999. An Experimental Examination of Labor Supply and Work Intensities. Journal of Labor Economics 17 (4), 638–670.

Doerrenberg, P., Duncan, D., 2014. Experimental evidence on the relationship between tax evasion opportunities and labor supply. European Economic Review 68, 48–70.

Dohmen, T., 2014. Behavioral labor economics: Advances and future directions. Labour Economics 30, 71–85.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., Wagner, G.G., 2011. Individual risk attitudes: Measurement, determinants, and behavioral consequences. Journal of the European Economic Association 9 (3), 522–550.

Donni, O., Chiappiori, P.-A., 2011. Nonunitary Models of Household Behavior: A Survey of the Literature, in: Molina, J.A. (Ed.), Household Economic Behaviors. Springer, New York, pp. 1–40.

Doyen, S., Klein, O., Simons, D.J., Cleeremans, A., 2014. On the other side of the mirror: Priming in cognitive and social psychology. Social Cognition 32 (Supplement), 12–32.

Eisenmenger, M., Schweinert-Albinius, S., 2014. Men twice as often in executive positions as women: Federal Statistical Office Press release 189.

Erkal, N., Gangadharan, L., Nikiforakis, N., 2011. Relative Earnings and Giving in a Real-Effort Experiment. American Economic Review 101 (7), 3330–3348.

Eswaran, M., 2014. Why gender matters in economics. Princeton University Press.

Eurostat, 2017. Gender statistics. http://ec.europa.eu/eurostat/statistics-explained/index.php/Gender_statistics#Further_Eurostat_information.

Eurostat, 2018a. Database. http://ec.europa.eu/eurostat/data/database.

Eurostat, 2018b. Gender pay gap in unadjusted form: % of average gross hourly earnings of men. http://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=sdg_05_20&plugin=1.

Filippin, A., Crosetto, P., 2016. A reconsideration of gender differences in risk attitudes. Management Science 62 (11), 3138–3160.

Fochmann, M., Weimann, J., 2013. The effects of tax salience and tax experience on individual work efforts in a framed field experiment. FinanzArchiv: Public Finance Analysis 69 (4), 511–542.

Fortin, N., Lemieux, T., Firpo, S., 2011. Decomposition methods in economics. Handbook of labor economics 4, 1–102.

Gächter, S., Johnson, E.J., Herrmann, A., 2010. Individual-level loss aversion in riskless and risky choices. CeDEx discussion paper series.

Gill, D., Prowse, V., 2012. A Structural Analysis of Disappointment Aversion in a Real Effort Competition. American Economic Review 102 (1), 469–503.

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

Goldin, C., Rouse, C., 2000. Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. American Economic Review 90 (4), 715–741.

Gomes, C.M., McCullough, M.E., 2015. The effects of implicit religious primes on dictator game allocations: A preregistered replication experiment. Journal of experimental psychology: General 144 (6), e94.

Görges, L., 2015. The Power of Love: A Subtle Driving Force for Unegalitarian Labor Division? Review of Economics of the Household 13 (1), 163–192.

Grossbard, S., 2011. Independent Individual Decision-Makers in Household Models and the New Home Economics, in: Molina, J.A. (Ed.), Household Economic Behaviors. Springer, New York, pp. 41–56.

Grund, C., 2015. Gender pay gaps among highly educated professionals -- Compensation components do matter. Labour Economics 34, 118–126.

Günther, C., Ekinci, N.A., Schwieren, C., Strobel, M., 2010. Women can't jump?--An experiment on competitive attitudes and stereotype threat. Journal of Economic Behavior & Organization 75 (3), 395–401.

Güth, W., Ivanova-Stenzel, R., Tjottax, S., 2004. Please, marry me! An experimental study of risking a joint venture. Metroeconomics 55, 1–21.

Haberkern, K., 2007. Zeitverwendung und Arbeitsteilung in Paarhaushalten. Zeitschrift für Familienforschung 19 (2), 159–185.

Harrison, G.W., List, J., 2004. Field Experiments. Journal of Economic Literature 42 (4), 1009–1055.

Heinz, M., Normann, H.-T., Rau, H.A., 2016. How competitiveness may cause a gender wage gap: Experimental evidence. European Economic Review.

Horton, J.J., Rand, D.G., Zeckhauser, R.J., 2011. The online laboratory: Conducting experiments in a real labor market. Experimental Economics 14 (3), 399–425.

Ichino, A., 2014. Comments on Gender and the labor market: What Have We Learned from Field and Lab Experiments. Labour Economics 30, 41–42.

Ichino, A., Moretti, E., 2009. Biological gender differences, absenteeism, and the earnings gap. American Economic Journal: Applied Economics 1 (1), 183–218.

Iriberri, N., Rey-Biel, P., 2016. Competitive Pressure Widens the Gender Gap in Performance: Evidence from a Two-Stage Competition in Mathematics. CEPR Discussion Papers 11493.

Jann, B., 2008. The Blinder-Oaxaca decomposition for linear regression models. The Stata Journal 8 (4), 453–479.

Jurajda, Š., Münich, D., 2011. Gender Gap in Performance under Competitive Pressure: Admissions to Czech Universities. American Economic Review 101 (3), 514–518.

Kabatek, J., van Soest, A., Stancanelli, E., 2014. Income taxation, labour supply and housework: A discrete choice model for French couples. Labour Economics 27, 30–43.

Kahneman, D., 2012. A proposal to deal with questions about priming effects. Nature 490.

Keane, M.P., 2011. Labor supply and taxes: A survey. Journal of Economic Literature 49 (4), 961–1075.

Kimbrough, E.O., Vostroknutov, A., 2016. Norms Make Preferences Social. Journal of the European Economic Association 14 (3), 608–638.

Kleinjans, K.J., 2009. Do gender differences in preferences for competition matter for occupational expectations? Journal of Economic Psychology 30 (5), 701–710.

Kräkel, M., Schöttner, A., 2016. Optimal sales force compensation. Journal of Economic Behavior & Organization 126, 179–195.

Kröger, H., 2017. The stratifying role of job level for sickness absence and the moderating role of gender and occupational gender composition. Social Science & Medicine.

Laaksonen, M., Mastekaasa, A., Martikainen, P., Rahkonen, O., Piha, K., Lahelma, E., 2010. Gender differences in sickness absence-the contribution of occupation and workplace. Scandinavian journal of work, environment & health, 394–403.

Laczó, S., 2011. Living Arrangements and Labour Supply: Evidence from the United Kingdom. http://www.eea-esem.com/files/papers/eeaesem/2011/1426/LaczoCouplesEvidence.pdf.

LaLumia, S., 2008. The effects of joint taxation of married couples on labor supply and non-wage income. Journal of Public Economics 92 (7), 1698–1719.

Lewis, J., 2002. Gender and welfare state change. European Societies 4 (4), 331–357.

Loersch, C., Payne, B.K., 2014. Situated inferences and the what, who, and where of priming. Social Cognition 32 (Supplement), 137–151.

Lönnqvist, J.-E., Verkasalo, M., Walkowitz, G., Wichardt, P.C., 2015. Measuring individual risk attitudes in the lab: Task or ask? An empirical comparison. Journal of Economic Behavior & Organization 119, 254–266.

Manning, A., Saidi, F., 2010. Understanding the gender pay gap: What's competition got to do with it? Industrial & Labor Relations Review 63 (4), 681–698.

Mastekaasa, A., 2005. Sickness absence in female- and male-dominated occupations and workplaces. Social Science & Medicine 60 (10), 2261–2272. 10.1016/j.socscimed.2004.10.003.

Mastekaasa, A., 2014. The gender gap in sickness absence: Long-term trends in eight European countries. The European Journal of Public Health 24 (4), 656–662.

Mastekaasa, A., Dale-Olsen, H., 2000. Do women or men have the less healthy jobs? An analysis of gender differences in sickness absence. European Sociological Review 16 (3), 267–286.

Mastekaasa, A., Melsom, A.M., 2014. Occupational Segregation and Gender Differences in Sickness Absence: Evidence from 17 European Countries. European Sociological Review 30 (5), 582–594.

Mastekaasa, A., Olsen, K.M., 1998. Gender, absenteeism, and job characteristics: A fixed effects approach. Work and Occupations 25 (2), 195–228.

McGee, A., McGee, P., Pan, J., 2015. Performance pay, competitiveness, and the gender wage gap: Evidence from the United States. Economics Letters 128, 35–38. 10.1016/j.econlet.2015.01.004.

Meghir, C., Phillips, D., 2009. Labour supply and taxes, in: Mirrlees, J.A. (Ed.), Dimensions of tax design. Oxford Univ. Press, Oxford, pp. 202–274.

Meier-Pesti, K., Penz, E., 2008. Sex or gender? Expanding the sex-based view by introducing masculinity and femininity as predictors of financial risk taking. Journal of Economic Psychology 29 (2), 180–196.

Melsom, A.M., Mastekaasa, A., 2017. Gender, occupational gender segregation and sickness absence. Acta Sociologica 22 (1), 000169931769158.

Meyer, M., Wehner, K., Cichon, P., 2017. Krankheitsbedingte Fehlzeiten in der deutschen Wirtschaft im Jahr 2016, in: Badura, B., Ducki, A., Schröder, H., Klose, J., Meyer, M. (Eds.), Fehlzeiten-Report 2017: Krise und Gesundheit - Ursachen, Prävention, Bewältigung. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 281–484.

Molden, D.C., 2014. Understanding priming effects in social psychology: An overview and integration. Social Cognition 32 (Supplement), 243–249.

Munro, A., 2015. Intra-household Experiments: A survey and some methodological observations. National Graduate Institute for Policy Studies.

Munro, A., McNally, T., Popov, D., 2008. Taking it in Turn: An Experimental Test of Theories of the Household. University Library of Munich, München.

Nelson, J.A., 2015. ARE WOMEN REALLY MORE RISK-AVERSE THAN MEN?: A RE-ANALYSIS OF THE LITERATURE USING EXPANDED METHODS. Journal of Economic Surveys 29 (3), 566–585.

Niederle, M., 2016. Gender, in: Kagel, J.H., Roth, A.E. (Eds.), The Handbook of Experimental Economics, Volume 2. The Handbook of Experimental Economics. Princeton University Press, pp. 481–553.

Niederle, M., Vesterlund, L., 2007. Do Women Shy Away From Competition? Do Men Compete Too Much? The Quarterly Journal of Economics 122 (3), 1067–1101.

Oaxaca, R., 1973. Male-female wage differentials in urban labor markets. International economic review, 693–709.

Oaxaca, R.L., Ransom, M.R., 1999. Identification in detailed wage decompositions. The Review of Economics and Statistics 81 (1), 154–157.

OECD, 2012a. Closing the gender gap: Act now. OECD, Paris.

OECD, 2012b. OECD Economic Surveys: Germany. OECD, Paris.

OECD, 2015. Public policies for families and children: PF1.4 Neutrality of tax-benefit systems. http://www.oecd.org/els/soc/PF1_4_Neutrality_of_tax_benefit_systems.pdf.

OECD, 2017a. Gender Data Portal. Organisation for Economic Co-operation and Development. http://www.oecd.org/gender/. Accessed 4 December 2017.

OECD, 2017b. The Pursuit of Gender Equality - An Uphill Battle. OECD Publishing, Paris.

Ors, E., Palomino, F., Peyrache, E., 2013. Performance gender gap: Does competition matter? Journal of Labor Economics 31 (3), 443–499.

Palma, A.D., Picard, N., Ziegelmeyer, A., 2009. Individual and couple decision behavior under risk: Evidence on the dynamics of power balance hal-00418899. http://ideas.repec.org/p/hal/wpaper/hal-00418899.html.

Paulus, W., Matthes, B., 2013. The German Classification of Occupations 2010: Structure, Coding and Conversion Table. FDZ-Methodenreport. Institut für Arbeitsmarkt-und Berufsforschung (IAB), Nürnberg [Institute for Employment Research, Nuremberg, Germany].

Phelps, E.S., 1972. The statistical theory of racism and sexism. The American Economic Review 62 (4), 659–661.

Pollak, R.A., 2011. Family Bargaining and Taxes: A Prolegomenon to the Analysis of Joint Taxation. CESifo Economic Studies 57 (2), 216–244.

Powers, D.A., Yoshioka, H., Yun, M.-S., others, 2011. mvdcmp: Multivariate decomposition for nonlinear response models. Stata Journal 11 (4), 556–576.

Rau, H.A., 2014. The disposition effect and loss aversion: Do gender differences matter? Economics Letters 123 (1), 33–36.

Reips, U.-D., 2000. The Web experiment method: Advantages, disadvantages, and solutions. Psychological experiments on the Internet, 89–117.

Rockoff, J.E., Herrmann, M.A., 2009. Biological gender differences, absenteeism, and the earnings gap: Comment. Working Paper. Columbia University.

Scheil-Adlung, X., Sandner, L., 2010. The case for paid sick leave. World health report.

Schmidt, U., Traub, S., 2002. An experimental test of loss aversion. Journal of Risk and uncertainty 25 (3), 233–249.

Schmitt, N., 2015. Towards a gender quota. DIW Economic Bulletin 5 (40), 527–536.

Schröder, M., Beblo, M., Beninger, D., Burow, N., 2017. Social Identity Priming in Economics: No clear Evidence, mimeo, Universität Hamburg.

Schröder, M., Schmitt, N., 2016. Couple's Labor Supply, Taxes, and the Division of Housework in a Gender-Neutral-Lab. DIW Discussion Papers 1593. DIW.

Schwieren, C., 2012. The gender wage gap in experimental labor markets. Economics Letters 117 (3), 592–595.

Selin, H., 2014. The Rise in Female Employment and the Role of Tax Incentives: An Empirical Analysis of the Swedish Individual Tax Reform of 1971. International Tax and Public Finance 21 (5), 894–922.

Shariff, A.F., Norenzayan, A., 2007. God is watching you: Priming God concepts increases prosocial behavior in an anonymous economic game. Psychological science 18 (9), 803–809.

Shih, M., Ambady, N., Richeson, J.A., Fujita, K., Gray, H.M., 2002. Stereotype performance boosts: The impact of self-relevance and the manner of stereotype activation. Journal of Personality and Social Psychology 83 (3), 638.

Shurchkov, O., Eckel, C.C., forthcoming. Gender Differences in Behavioral Traits and Labor Market Outcomes, in: Averett, Argys et al. (Hg.) 2018 – The Oxford Handbook of Women.

Sillamaa, M.A., 1999. How work effort responds to wage taxation: An experimental test of a zero top marginal tax rate. Journal of Public Economics 73 (1), 125–134.

Spasova, S., Bouget, D., Vanhercke, B., 2016. Sick pay and sickness benefit schemes in the European Union: Background report for the Social Protection Committee's In-depth Review on sickness benefits (17 October 2016). European Social Policy Network, Brussels.

Spencer, S.J., Logel, C., Davies, P.G., 2016. Stereotype Threat. Annual Review of Psychology 67 (1), 415–437.

Steele, C.M., 1997. A threat in the air: How stereotypes shape intellectual identity and performance. American psychologist 52 (6), 613.

Steele, C.M., Aronson, J., 1995. Stereotype threat and the intellectual test performance of African Americans. Journal of Personality and Social Psychology 69 (5), 797.

Steiner, V., Wrohlich, K., 2004. Household Taxation, Income Splitting and Labor Supply Incentives - A Microsimulation Study for Germany. CESifo Economic Studies 50 (3), 541–568.

Steiner, V., Wrohlich, K., 2008. Introducing Family Tax Splitting in Germany: How Would It Affect the Income Distribution, Work Incentives, and Household Welfare? FinanzArchiv: Public Finance Analysis 64, 115–142.

Sterud, T., 2014. Work-related gender differences in physician-certified sick leave: A prospective study of the general working population in Norway. Scandinavian journal of work, environment & health, 361–369.

Stöwhase, S., 2011. Non-minimization of source taxes on labor income: Empirical evidence from Germany. Review of Economics of the Household 9 (2), 293–306.

Stratton, L.S., 2012. The role of preferences and opportunity costs in determining the time allocated to housework. The American Economic Review 102 (3), 606–611.

Thaler, R.H., Sunstein, C.R., 2009. Nudge: Improving decisions about health, wealth, and happiness. Penguin, New York, NY, 312 pp.

Triebe, D., 2013. Wo(men) at Work? The Impact of Cohabiting and Married Partners' Earnings on Women's Work Hours. DIW Berlin, Berlin.

Tversky, A., Kahneman, D., 1992. Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and uncertainty 5 (4), 297–323.

Weaver, J.R., Vandello, J.A., Bosson, J.K., 2013. Intrepid, imprudent, or impetuous? The effects of gender threats on men's financial decisions. Psychology of Men & Masculinity 14 (2), 184.

West, C., Zimmerman, D.H., 1987. Doing Gender. Gender & Society 1 (2), 125–151.

Willard, A.K., Shariff, A.F., Norenzayan, A., 2016. Religious priming as a research tool for studying religion: Evidentiary value, current issues, and future directions. Current Opinion in Psychology 12, 71–75.

Winter-Ebmer, R., 2014. What is (not) behavioural in labour economics? Labour Economics 30, 86–87.

WorlddatWork, 2016a. Incentive Pay Practices: Nonprofit/Government Organizations. https://www.worldatwork.org/adimLink?id=79940.

WorlddatWork, 2016b. Incentive Pay Practices: Privately Held Companies. https://www.worldatwork.org/adimLink?id=79934.

Yun, M.-S., 2005. A simple solution to the identification problem in detailed wage decompositions. Economic Inquiry 43 (4), 766–772.

8 Appendix

8.1 Executive Summary

Although women made significant progress towards gender equality in the labor market in recent decades, disparities in the labor market between men and women still exist: Women and men differ in employment levels and in pay, they work in different occupations and positions. Moreover, they differ in absence from the labor market due to sickness. This dissertation sheds light on the supply side determinants and constraints of some of these gender differences in the labor market and provides empirical evidence using experimental and administrative data from Germany.

In Chapter 2, which presents joint work with Norma Burow, we use an experiment to investigate intra-couple labor supply decisions and the division of housework under individual and joint income taxation systems. In order to rule out problems of endogeneity, we created a gender-neutral lab, where we exogenously assigned the intra-couple roles of primary and secondary earners. With incentivized work-effort used as a proxy for labor supply, cohabiting and married couples performed real-effort tasks within a given time. Additionally, couples had to decide on the allocation of an unpaid task serving as our proxy for housework. In our gender-neutral lab, we find no gender differences in the allocation of housework as in the real world. Instead, the allocation of housework follows a purely economic rationale with the majority of secondary earners taking responsibility and a shift to a more egalitarian allocation when individual taxation applied. Moreover, we show a difference between married and cohabiting men and women in labor supply which hinges on the stability of specialization in married couples outside the lab.

In Chapter 3, which presents joint work with Miriam Beblo, Norma Burow and Denis Beninger, we investigate the gender pay gap in an online choice experiment. By design, we rule out employer-side discrimination and thus isolate the labor supply-side determinants of earnings. In the experiment, participants performed a real effort task after having decided on their preferred compensation contract that vary in the extent of competition and risk involved. We observe a gender pay gap of 23% in experimental pay and reveal that one-quarter of this gap can be explained by selection into contract types (selection effect), and about one-half can be attributed to the participants' performance conditional on these contracts (contract effect). Moreover, we show that the observed pay difference is driven by women being more loss averse and possibly underperforming in a stereotype threat situation.

In Chapter 4, which presents joint work with Norma Burow, Miriam Beblo and Denis Beninger, we review existing priming experiments in economics and compare them with an online

experiment we conducted. Using a priming stimulus to vary exogenously the salience of a social identity and its impact on individual decision making is a recent trend in experimental economics. However, a systematic literature review detects that results are mixed, and significant priming effects are regularly lacking. In particular, we discuss the absence of priming effects in others and our experiment based on recent theoretical insights in economics and social psychology. Consequently, we provide a critical reflection of experimental (null-)findings when social identity priming techniques are used.

In chapter 5, I provide evidence on the relationship between occupational segregation and the gender gap in work absenteeism in Germany and address two explanations for a relationship that are prominent in the literature: Gender differences in exposure and gender differences in vulnerability. Performing a detailed decomposition of the gap, which is new in this strand of the literature, I confirm previous findings on the differences-in-exposure argument. I show that gender segregation works in favor of women in terms of work absenteeism and thus women seem to select into healthier occupations, on average. Additionally, I provide the first direct evidence on the gender-difference-in-vulnerability argument, revealing that women seem to be, on average, more vulnerable than men within an occupation, while the direction and the magnitude of effects are heterogeneous across occupational areas and job levels.

Taken together, this dissertation reveals that women and men indeed behave differently, on average, but effects are heterogeneous across groups, contexts and institutions. Thus, one path for future research should identify when and where these gender differences in behavior occur and detect the roots of gender differences in institutions.

8.2 Zusammenfassung

Obwohl erhebliche Fortschritte auf dem Weg zur Gleichstellung der Geschlechter auf dem Arbeitsmarkt gemacht wurden, existieren immer noch Disparitäten: Frauen und Männer unterscheiden sich im Beschäftigungsniveau und Lohn, sie arbeiten in verschiedenen Berufen und Positionen. Darüber hinaus unterscheiden sie sich auch in krankheitsbedingten Fehlzeiten. Diese Dissertation beleuchtet die angebotsseitigen Determinanten und Restriktionen und liefert empirische Evidenz anhand von experimentellen und administrativen Daten aus Deutschland.

In Kapitel 2, das eine Zusammenarbeit mit Norma Burow darstellt, beobachten wir Paare im Experimentallabor, die Entscheidungen über das individuelle Arbeitsangebot und die Aufteilung der Hausarbeit treffen, unter individueller Besteuerung des Einkommens und unter gemeinsamer Besteuerung (Ehegattensplitting). Um Endogenitätsprobleme zu umgehen, haben wir ein sogenanntes geschlechtsneutrales Labor geschaffen, in dem die Rollen des Erst- und Zweitverdienenden innerhalb der Paare zufällig zugewiesen wurden. Im Experiment lösten dann verheiratete und unverheiratete Paare innerhalb einer bestimmten Zeit Aufgaben, die vergütet wurden, und die uns als Maß für das Arbeitsangebot dienen. Zusätzlich mussten Sie entscheiden, wer eine unbezahlte Aufgabe erfüllt, die stellvertretend für Hausarbeit fungiert. In unserem geschlechtsneutralen Labor finden wir keine geschlechtsspezifischen Unterschiede bei der Übernahme von Hausarbeit wie in der Welt außerhalb des Labors. Stattdessen folgt die Entscheidung einer rein ökonomischen Logik, wobei die Mehrheit der Zweitverdiener diese Verantwortung übernimmt, und die Aufteilung bei individueller Besteuerung egalitärer ausfällt. Darüber hinaus zeigen wir einen Unterschied zwischen verheirateten und unverheirateten Männern und Frauen im Arbeitsangebot, was wir als Hinweis auf die Stabilität der Spezialisierung von verheirateten Paaren außerhalb des Labors werten.

In Kapitel 3, das eine gemeinsame Arbeit mit Miriam Beblo, Norma Burow und Denis Beninger darstellt, untersuchen wir das geschlechtsspezifische Lohngefälle in einem Online-Experiment. Dabei schließen wir durch das Design des Experiments Diskriminierung auf der Arbeitgeberseite aus, sodass wir die Determinanten auf der Angebotsseite isolieren können. Im Experiment lösten die Teilnehmenden Aufgaben, nachdem sie sich zwischen Vergütungsverträgen, die im Ausmaß von Wettbewerb und Risiko variierten, für einen entschieden hatten. Wir beobachten ein geschlechtsspezifisches Lohngefälle von 23% im Experiment und zeigen, dass ein Viertel dieser Lücke durch die unterschiedliche Auswahl der Vertragsarten (Selektionseffekt) zwischen Männern und Frauen erklärt werden kann und etwa die Hälfte davon auf die unterschiedliche vertragsabhängige Leistung der Teilnehmenden

zurückzuführen ist (Vertragseffekt). Darüber hinaus zeigen wir, dass die beobachtete Lohndifferenz durch unterschiedliche Verlustaversion (loss aversion) zwischen Frauen und Männern und durch eine stereotype Bedrohungslage (stereotype threat) für Frauen getrieben wird.

In Kapitel 4, das eine gemeinsame Arbeit mit Norma Burow, Miriam Beblo und Denis Beninger darstellt, betrachten wir bestehende Priming-Experimente in der Ökonomik und vergleichen sie mit einem Online-Experiment, das wir durchgeführt haben. Die Verwendung eines Priming-Stimulus zur exogenen Veränderung der Salienz einer sozialen Identität und die Auswirkung auf die individuelle Entscheidungsfindung, stellt einen neuen Trend in der experimentellen Ökonomik dar. Ein systematischer Literaturüberblick zeigt jedoch, dass die Effekte heterogen sind und regelmäßig keine Effekte zu identifizieren sind. Wir diskutieren diese Abwesenheit von Priming-Effekten – in anderen und unserem Experiment – basierend auf neueren theoretischen Erkenntnissen aus der experimentellen Ökonomik und der Sozialpsychologie. Folglich liefern wir eine kritische Reflektion der (Null-)Ergebnisse von Priming-Experimenten.

In Kapitel 5 widme ich mich dem Zusammenhang zwischen der Geschlechtersegregation auf dem Arbeitsmarkt und den geschlechtsspezifischen Unterschieden bei krankheitsbedingten Fehlzeiten in Deutschland und adressiere zwei in der Literatur prominente Erklärungen für einen Zusammenhang: Geschlechtsspezifische Unterschiede, die auf unterschiedliche Berufe und damit verbunden unterschiedlichen Gesundheitsrisiken für Frauen und Männer zurückgeführt werden (gender difference in exposure) und das Argument eines geschlechtsspezifischen Unterschieds in der Anfälligkeit (gender differences in vulnerabilty). Ich führe eine detaillierte Zerlegung dieser Lücke auf Basis administrativer Daten durch, die eine neue Methode in diesem Teil der Literatur darstellt, und bestätige frühere Ergebnisse zum „gender difference in exposure“ Argument. Ich zeige, dass Frauen im Durchschnitt „gesündere“ Berufe wählen, sodass sich die geschlechtsspezifische Segregation zugunsten von Frauen in Bezug auf die krankheitsbedingten Fehlzeiten auswirkt. Darüber hinaus gebe ich erste Evidenz für das Argument der geschlechtsspezifischen Unterschiede in der Anfälligkeit (gender differences in vulerability) und zeige, dass Frauen in einem Beruf im Durchschnitt „anfälliger“ sind als Männer und häufiger fehlen. Allerdings variiert die Richtung und die Größe des Effekts sehr stark über die verschiedenen Berufe und Positionen.

Zusammenfassend zeigt diese Dissertation, dass es durchaus Evidenz für unterschiedliches Verhalten von Frauen und Männer gibt – allerdings „nur“ im Durchschnitt und mit großer Heterogenität über Gruppen, Kontexte und Institutionen. Dies zeigt, dass sich zukünftige

Forschung auch möglichen Wurzeln dieser geschlechtsspezifischen Unterschiede in Institutionen widmen könnte.