Perceptions of (mis)behaviour by gender:
Evidence from a global religious gathering and non-religious settings*

Sandra Maximiano† Susana Peralta ‡ João Pereira dos Santos§

António Silva¶

March 2026

Abstract

We provide the first causal field evidence on gender discrimination in moral judgments, a domain where informal social sanctions shape incentives and behavior. In two preregistered survey experiments, we asked 1,800 Catholic pilgrims during World Youth Day in Lisbon and 967 citygoers in Lisbon to morally evaluate norm violating vignettes. We randomly varied whether the norm violator was described with a female or male identity and considered violations related to honesty, sexual behavior, and parental care. In both settings, respondents judge female norm violators, on average, less harshly. We also elicit meta norms in the citygoers sample and find a large belief wedge: most respondents think others will judge women more harshly, even though our causal estimates point in the opposite direction. These results advance research on stereotypes, discrimination, and social norms by documenting (i) causal gender gaps in informal sanctioning, (ii) their robustness across distinct norm environments, and (iii) systematic misperceptions about the direction of discrimination that can distort behavior and policy demand.

JEL Codes: J16, Z12, C93

*The authors are indebted to Sascha O. Becker, Pedro Bordalo, Adeline Delavande, Miguel A. Fonseca, Diogo Geraldes, Victoire Girard, Nagore Iriberri, Camille Landais, Muriel Niederle, John de Quidt, and Anna Raute, as well as participants in the Lisbon Micro Group, 2025 Florence Meeting on Behavioural and Experimental Economics, the 4th Gender and Economics workshop (Université du Luxembourg), and the Global GLO-JOPE Conference for comments and suggestions. We also thank an army of research assistants for their motivation and outstanding work. This work was funded by Fundação para a Ciência e a Tecnologia (UIDB/00124/2020, UIDP/00124/2020, UID/00124, Nova School of Business and Economics and Social Sciences DataLab - PINFRA/22209/2016) and 2022.04821.PTDC. Pre-registration RCT ID AEARCTR-0011865. Errors are our own.

†ISEG – University of Lisbon. E-mail: smaximiano@iseg.ulisboa.pt.

‡Nova School of Business and Economics. E-mail: peralta@novasbe.pt

§ISEG – University of Lisbon; IZA; GLO. E-mail: joao.santos@iseg.ulisboa.pt.

¶ISEG – University of Lisbon. E-mail: apsilva@iseg.ulisboa.pt.

1 Introduction

Society imputes different social and moral responsibilities to men and women. These expectations shape how individuals are perceived and evaluated in domains such as morality, risk-taking, negotiation, and performance. These different ways of being and acting, which ultimately shape moral, social, risk, negotiation, and competition preferences (Niederle and Vesterlund, 2011; Bordalo et al., 2019). Moral judgment is based on cognitive and emotional processes that interact with social norms and gender stereotypes (Buckholtz et al., 2015). Glick and Fiske (1996)'s ambivalent sexism theory underscores how stereotypes can affect judgments in scenarios like parental care or sexual behavior, mediated by both hostile and benevolent sexism.

Gender stereotypes can be descriptive, referring to beliefs about how men and women are, or prescriptive, referring to beliefs about how men and women should behave. Prescriptive stereotypes are particularly relevant for moral judgment because they imply social sanctions for norm violations. These sanctions do not need to be uniformly negative. For example, perceiving women as more caring may generate leniency in some moral evaluations. Prescriptive stereotyping has been extensively studied in psychology and increasingly in economics, especially in labour market and leadership contexts where deviations from gendered expectations are penalised (Heilman et al., 2004, 2024; Rudman and Glick, 2001). These penalties may include lower wage recommendations (Heilman et al., 2004), or restricted access to social networks (Casciaro and Lobo, 2005).

Despite a large literature on gender stereotypes, there is limited empirical evidence on how prescriptive gender norms shape moral judgment and, to our knowledge, this is the first study to investigate this experimentally. Existing work has focused primarily on performance evaluations, wages, or leadership outcomes (Niederle and Vesterlund, 2007; Bertrand and Duflo, 2017; Kleven et al., 2019, 2024; Bohren et al., 2025). Related research shows that gender stereotypes affect the way people approach, interpret, and remember information about themselves and others (Ellemers, 2018; Coffman et al., 2021; Exley and Nielsen, 2024).

Much less is known about whether men and women are judged differently when they violate moral norms related to family responsibilities, sexual behaviour, or honesty, and whether such judgments depend on the gender of the evaluator.

We study gender discrimination in moral judgment using a field survey experiment conducted in two distinct settings in Lisbon. In both settings, respondents are asked to evaluate hypothetical norm-violating scenarios related to lying, sexual behaviour, and parental care, where the gender of the individual involved in the behaviour is randomly varied. The first sample consists of participants recruited during the 2023 World Youth Day, while the second comprises citygoers surveyed in public spaces throughout Lisbon. This design allows us to examine whether gendered patterns in moral judgment are consistent across contexts that differ markedly in demographics, motivations, and normative environments.

One of our data collection settings coincides with a large-scale religious event, which allows us to observe moral judgments in a context where norms may be particularly salient (Gundersen and Beck, 2020; Becker et al., 2021, 2025). Importantly, the inclusion of a second, non-religious sample enables us to assess the extent to which observed gender differences generalise beyond this setting.

We find that women are, on average, less judged than men. We find the same pattern in both religious and non-religious settings. In the religious setting, women are judged less harshly on scenarios of parental care (i.e. rarely picking up their children from school) and adultery (i.e. cheating on their partner). In the non-religious setting this only applies to the parental care setting. In both settings, we find no statistical significant discrimination on scenarios involving dishonesty (i.e. lying to your boss).

We explore the heterogeneity of effects across gender of respondent, nationality, educational level, and age. In the religious setting, we find that while men do not discriminate in the scenarios according to the gender of the person, women tend to be more lenient about other women's behavior. In contrast, in the non-religious setting, we find that women do not discriminate and that men are more lenient towards women. We do not find any significant

heterogeneity based on nationality, educational level, or age in either setting.

We also explore whether the gender of the survey enumerator has an impact on the results and do not find an effect, suggesting that experimenter demand effects may not be very relevant in this setting (Zizzo, 2010; De Quidt et al., 2018, 2019).

Finally, in the non-religious sample we investigate the meta-norms of gender discrimination. We ask respondents how they think other people will judge if the gender in the scenario is inverted. We find that across all scenarios the large majority of people think other people will judge women more harshly. Only 4 and 12 percent of the respondents think women will be judged better than men for cheating on their partner and not picking up their children, when in reality we find the opposite to be true.

Our paper is related to the nascent experimental literature on prescriptive stereotyping (Bordalo et al., 2016) and misperceptions about others (Bursztyn and Yang, 2022). Bursztyn et al. (2020) show that correction of men's beliefs about their male peers' greater acceptance of women working outside the home increases their willingness to support their wives seeking outside the home jobs. Conversely, women who learn about the true acceptance amongst males are more likely to take higher-paying outside-the-home jobs. This mismatch between perceived and actual norms is also found in the support for equality-promoting policies being more under-estimated by men in less equal countries (Bursztyn et al., 2023). Dhar et al. (2022) analyses a randomized classroom experiment involving discussions of gender equality, and finds a significant and long-lasting decrease in support for regressive attitudes.

Contrary to these two papers, our treatment does not aim at curbing the stereotype with information provision; rather, we focus on the impact of the gender of the person who judges females or males who fail to comply with prescriptive stereotypes usually associated with women. This approach resembles that of Asmat et al. (2024), who assess the extent to which females are harmed when females are judged using data from 133 years of piano competitions; however, their analysis is about performance and is unrelated to stereotyping. Moreover, they do not vary exogenously the gender of the judged individual, as we do in our

field experiment. Relatedly, Lee and Guven (2013) found that women are often perceived as more ethical in corporate decision-making.

The policy relevance of this topic lies in its implications for fairness and equality. Moral judgments are not made in a vacuum; they often inform decisions in governance, law, education, and healthcare. For example, biases in moral judgment could affect jury decisions, hiring practices, or parental custody rulings. As these moral judgments are likely to affect individual preferences, they can also have an impact on education and professional choices, for example, leading men and women to choose different fields of study and careers. Evidence of systematic gender discrimination in moral evaluations – particularly surprising findings like ours – calls for careful, nuanced approaches when developing fairer institutional frameworks.

This paper is organized as follows. In Section 2 we discuss the institutional setting and describe the data. In Section 3 we explain the experiment while in Section 4 we present the results. Section 5 concludes.

2 Institutional Setting and Data

We implemented a one-sheet, two-page survey, in A4 format, collecting sociodemographic information (age, gender, nationality, education) and questions on religiosity. The latter includes items on practices such as church attendance and prayer, allowing us to construct the centrality of religiosity scale (Huber and Huber, 2012). The treatment described in Section 3 was embedded in the same paper-based survey. We restricted participation to individuals aged 18 or above who were able to understand Portuguese and/or English.[1]

A key feature of our design is that we implemented the same experimental protocol across two settings in Lisbon: one linked to a large religious event (World Youth Day) and another drawn from citygoers in public spaces. This allows us to assess whether gendered patterns in moral judgment generalise across contexts that differ in demographics and in the likely salience of moral and social norms.

[1] All participants signed an informed consent agreement stating that they could stop at any point in time.

2.1 World Youth Day

World Youth Day (WYD) is a global event organised by the Catholic Church that gathers young people from around the world, established by Pope John Paul II in 1985. The 2023 WYD was held in Lisbon. For this experiment, we leveraged on a large installation of 150 private confession cabins in a public park in Lisbon, the so-called *Parque do Perdão* (Reconciliation Park), where we approached pilgrims.[2] This location was particularly suited for recruitment because it featured substantial waiting times and a calm environment, which increased willingness to participate. Research assistants were trained to approach the participants respectfully and to avoid disrupting religious activities.

Subjects were incentivised with 22 lotteries (two 500-euro prizes and twenty 50-euro prizes) to increase attentiveness; prizes were later randomly assigned and announced on the X-Lab website.[3]

We present descriptive statistics in Table 1. The WYD sample includes slightly more female respondents (58.2%) and is young (mean age 27), highly educated (66% with a bachelor degree), and internationally diverse. Responses were collected across the four days the *Parque do Perdão* operated.

Table 1: Summary statistics (World Youth Day sample)

	Mean	Std. dev.	Min	Max
female respondent	0.582	0.493	0	1
female assistant	0.473	0.499	0	1
age	27.286	10.548	18	79
in Lisbon for WYD	0.890	0.313	0	1
religiosity index	4.541	0.527	1.6	5
foreigner	0.713	0.452	0	1
graduate	0.658	0.474	0	1

Our sample comprises respondents from 80 nationalities, as shown in Table A1 in the Appendix. The five most common nationalities are Portuguese (29%), Spanish (12%), Brazilian (8%), French (8%), and American (6%).

[2] Confessions were available in the official languages of WYD Lisbon 2023 and in additional languages. More information about the park is available at `https://www.lisboa2023.org/en/forgiveness-park`.

[3] More information at `http://xlabonline.com/field-experiment-at-world-youth-event-2023/`.

2.2 Lisbon citygoers

We implemented the same experimental protocol with citygoers recruited in central areas of Lisbon, including shopping centres, train and bus stations, and tourist landmarks (*Oriente, Cais do Sodré, Baixa, Belém, Estrela*). Data collection was carried out between March and May 2025.

We present descriptive statistics in Table 2 and this sample includes respondents from 51 nationalities (see Table B1 in the Appendix). Relative to the WYD sample, citygoers are older on average and include a larger share of Portuguese nationals, providing a useful contrast in demographics and likely motivations for being in the study setting.

Table 2: Summary statistics

	Mean	Std. dev.	Min	Max
female respondent	0.506	0.500	0	1
female assistant	0.545	0.498	0	1
age	34.914	15.648	18	85
religiosity index	2.722	1.294	1	5
foreigner	0.337	0.473	0	1
graduate	0.468	0.499	0	1

3 Treatment

We measure the gendered dimension of moral judgment in prescriptive norm violations using three hypothetical scenarios, randomly varying the gender of the individual committing the norm-violating act. Respondents then evaluate the scenarios.

Our pre-specified outcome is the sum of the responses to the three scenario questions. We call this variable the *index*. Respondents were asked to: "Morally rate the following fictional situations on a scale from 1 (morally objectionable) to 6 (not morally objectionable):"

1. Alice/Henry missed work because she/he went to a party the day before, and lied to the employer saying she/he was sick

2. Diana/Adrian always works late and can rarely pick up the children from school

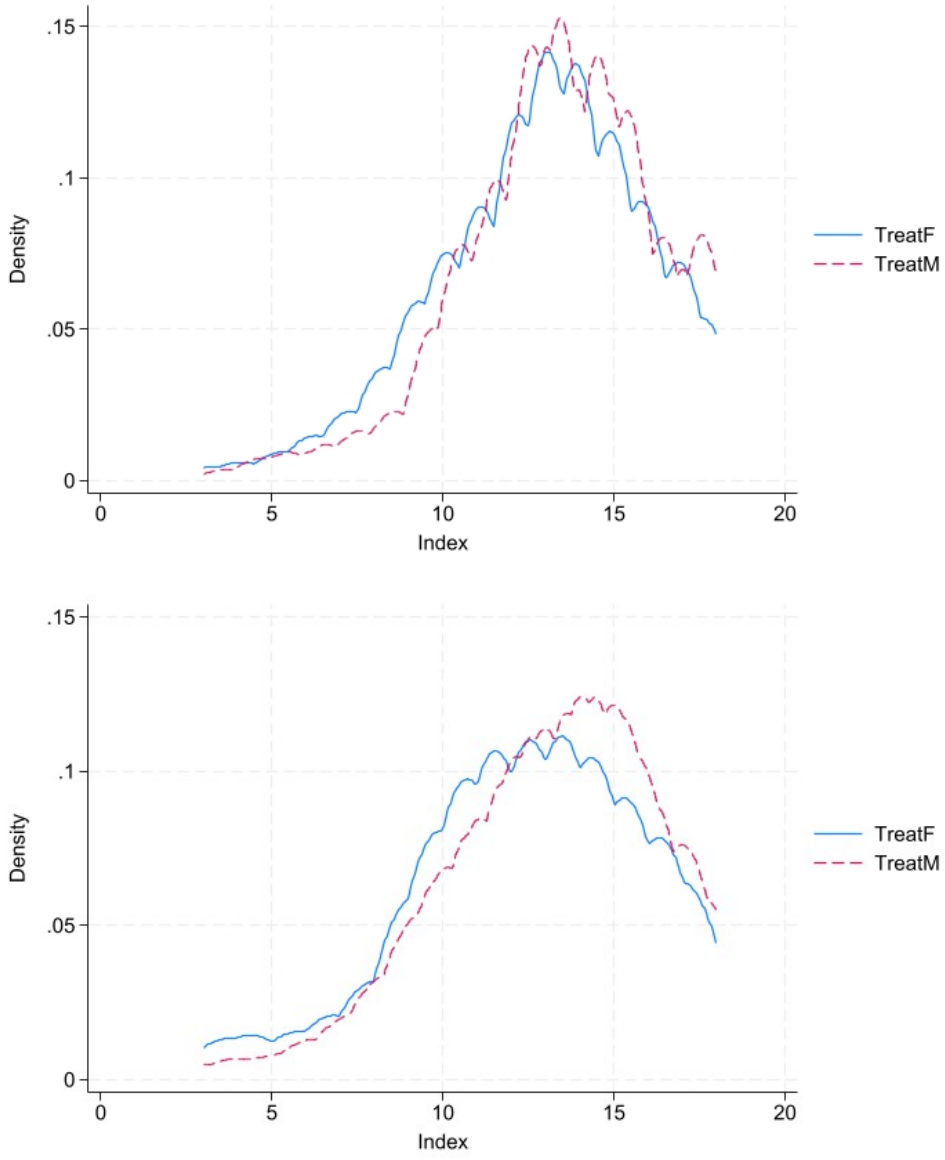
3. Caroline/Carl, who is married, went out at night and had sex with a stranger

Each respondent saw either three female or three male names. This is a between-subject design: participants are only exposed to one gender condition and are not shown alternative versions of the scenarios (Charness et al., 2012). The treatment condition uses female characters; the control uses male characters.

Enumerators drew surveys from a common alternating pile, ensuring that, on average, half of the respondents received the female-character version. Balance on observables is shown in Table A2 (WYD) and Table B2 (citygoers) in the Appendix.

The non-parametric distribution of the index is displayed in Figure 1, which suggests first-order stochastic dominance of the control distribution over the treatment distribution.

Figure 1: Distribution of responses by treatment. Top: Religious setting. Bottom: Non-religious setting



(

4 Experimental results

We present three sets of results. First, we estimate the main treatment effect: whether norm-violating behaviours are judged differently depending on the gender of the person in the

scenario. Second, we examine whether the treatment effect varies with respondent characteristics. Third, we test whether the gender of the enumerator affects outcomes as a diagnostic for experimenter demand effects (Zizzo, 2010; De Quidt et al., 2018, 2019).

Our main specification is:

$$Index_i = \alpha_t + \beta\, TreatF_i + X_i'\,\delta + \epsilon_i, \tag{1}$$

where $Index_i$ is the sum of the three scenario ratings by subject i. $TreatF$ takes value one if the respondent saw females names and zero if the respondent saw male names. In the WYD sample, α_t are day fixed effects. The pre-registered controls X_i include respondent gender, age, indicators for being a foreigner, having a bachelor degree, and (for WYD) being in Lisbon specifically for WYD; we also include religiosity using the centrality of religiosity scale (Huber and Huber, 2012). For the citygoers sample we use the analogous covariates available in that survey.

4.1 World Youth Day

Main treatment effects. Table 3 shows that respondents judge female norm violators less harshly than male norm violators. The point estimate on the index corresponds to roughly a 4% reduction relative to the sample mean (13.19). When decomposing the index, the significant differences appears in the parental care and adultery scenarios, while the lying scenario shows no detectable difference.

Table 3: World Youth Day: women were judged less harshly than men

	Index		Lied to boss		Did not pick kids		Adultery	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TreatF	-0.492***	-0.479***	-0.101	-0.093	-0.243***	-0.241***	-0.149**	-0.146**
	(0.141)	(0.138)	(0.070)	(0.069)	(0.077)	(0.075)	(0.062)	(0.061)
female respondent		-0.062		-0.185***		0.119		0.004
		(0.140)		(0.070)		(0.077)		(0.063)
age		0.043***		0.013***		0.029***		0.000
		(0.007)		(0.004)		(0.004)		(0.003)
in Lisbon for WYD		-0.227		-0.018		-0.149		-0.060
		(0.233)		(0.112)		(0.126)		(0.099)
religiosity index		0.582***		0.260***		0.099		0.222***
		(0.137)		(0.066)		(0.074)		(0.062)
foreigner		-0.123		-0.123		0.117		-0.116*
		(0.164)		(0.080)		(0.089)		(0.070)
graduate		-0.125		-0.076		0.007		-0.057
		(0.152)		(0.077)		(0.083)		(0.069)
Day Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
N	1,800	1,800	1,800	1,800	1,800	1,800	1,800	1,800
Mean Y	13.19	13.19	4.43	4.43	3.36	3.36	5.40	5.40

Notes: Robust standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Point estimates are similar with and without controls, indicating a successful randomisation. Religiosity is positively associated with harsher judgment on average. We do not find systematic differences in overall judgment by education, and age is associated with harsher judgment.

Heterogeneity by respondent characteristics. We report heterogeneity by respondent characteristics (gender of respondent, nationality, educational level and age.) in Table 4. We find that while men do not discriminate in the scenarios according to the gender of the person, women tend to be more lenient about other women's behavior. The interaction estimates indicate that, in this setting, the overall treatment effect is driven primarily by female respondents being more lenient toward female norm violators, particularly in the parental care scenario. We find no evidence that men systematically differentiate by the gender of the person in the scenario. We also do not find any significant heterogeneity based on nationality, educational level or age.

Enumerator gender. We test for experimenter demand effects by interacting the treatment with enumerator gender. As shown in Table A6 in the Appendix, coefficients are small

Table 4: World Youth Day: Heterogeneity

	(1) Index	(2) Index	(3) Index	(4) Index
TreatF	-0.146	-0.458*	-0.450**	-0.497***
	(0.210)	(0.239)	(0.228)	(0.188)
TreatF × female respondent	-0.572**			
	(0.280)			
TreatF × foreigner		-0.030		
		(0.292)		
TreatF × graduate			-0.045	
			(0.287)	
TreatF × above median age				0.019
				(0.277)
female respondent	0.219	-0.061	-0.061	-0.047
	(0.194)	(0.140)	(0.140)	(0.140)
age	0.043***	0.043***	0.043***	0.035***
	(0.007)	(0.007)	(0.007)	(0.009)
in Lisbon for WYD	-0.236	-0.228	-0.225	-0.227
	(0.231)	(0.232)	(0.233)	(0.233)
religiosity index	0.593***	0.582***	0.582***	0.570***
	(0.137)	(0.137)	(0.137)	(0.137)
foreigner	-0.115	-0.107	-0.123	-0.148
	(0.163)	(0.210)	(0.164)	(0.164)
graduate	-0.123	-0.125	-0.103	-0.212
	(0.152)	(0.152)	(0.204)	(0.166)
above median age				0.277
				(0.242)
Day Fixed Effects	Yes	Yes	Yes	Yes
N	1,800	1,800	1,800	1,800
Mean Y	13.19	13.19	13.19	13.19

and statistically indistinguishable from zero, suggesting that enumerator gender does not meaningfully shift responses in this setting.

4.2 Lisbon citygoers

Main treatment effects. We now turn to the citygoers sample. We find, again, that female norm violators are judged less harshly than male norm violators, with Table 5 showing a clear treatment effect on the index . Decomposing the index, we find no detectable gender difference in the lying scenario, while we observe significant differences in the parental care scenario and in the sexual adultery scenario. This pattern mirrors the WYD results for parental care and strengthens the evidence that gendered moral judgment is not specific to a religious setting.

Table 5: Lisbon citygoers: women were judged less harshly than men

	Index		Lied to boss		Did not pick kids		Adultery	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TreatF	-0.639***	-0.588***	-0.015	-0.007	-0.254**	-0.251**	-0.370***	-0.330***
	(0.215)	(0.211)	(0.101)	(0.100)	(0.112)	(0.111)	(0.101)	(0.098)
female respondent		0.258		-0.028		0.174		0.112
		(0.212)		(0.101)		(0.112)		(0.099)
age		0.021***		0.011***		0.018***		-0.008**
		(0.007)		(0.003)		(0.004)		(0.003)
religiosity index		0.333***		0.106***		0.033		0.194***
		(0.084)		(0.041)		(0.046)		(0.040)
foreigner		-0.780***		-0.300**		-0.127		-0.353***
		(0.247)		(0.118)		(0.123)		(0.115)
graduate		-0.028		0.069		-0.031		-0.066
		(0.220)		(0.106)		(0.116)		(0.103)
N	967	967	967	967	967	967	967	967
Mean Y	12.83	12.83	4.26	4.26	3.47	3.47	5.10	5.10

Heterogeneity by respondent characteristics. We zoom in on heterogeneous effects by respondent characteristics in Table 6. In contrast to the religious setting, here we find that women do not discriminate and that men are more lenient towards women. Again, we do not find any significant heterogeneity based on nationality, educational level or age.

Enumerator gender. Table B5 in the Appendix provides an analogous test for enumerator gender in the citygoers setting. As in WYD, interaction estimates are not statistically distinguishable from zero, showing no support for enumerator-driven demand effects.

4.2.1 Meta-norms about gender discrimination (citygoers)

In the citygoers survey, we also measure perceived social norms about gender discrimination. After respondents evaluated the scenarios, we asked how they believe "most people in your country" would judge the same situation if the gender of the person in the vignette were switched. Responses fall into three categories: "the same", "better", or "worse".

Across all three scenarios, only a small minority expect women to be judged less harshly, which is precisely what our findings show. For example, in the adultery scenario, as seen in Table 9, 91.8% of the respondents believe that women would be judged more harshly in the scenario than men. In contrast, only 8.2% expect women to be judged *better* than men. A similar pattern holds for parental care (shown in Table 8) and lying (shown in Table

Table 6: Lisbon citygoers: Heterogeneity

	(1) Index	(2) Index	(3) Index	(4) Index	Index	Index
TreatF	-1.063***	-0.379	-0.639**	-0.562*	-0.537**	-0.574**
	(0.313)	(0.239)	(0.295)	(0.287)	(0.256)	(0.285)
TreatF × female respondent	0.944**					
	(0.425)					
TreatF × foreigner		-0.621				
		(0.481)				
TreatF × graduate			0.110			
			(0.424)			
TreatF × above median age				-0.090		
				(0.422)		
TreatF × has kids					-0.151	
					(0.454)	
TreatF × married						-0.036
						(0.427)
female respondent	-0.206	0.251	0.254	0.272	0.251	0.260
	(0.289)	(0.212)	(0.213)	(0.212)	(0.213)	(0.212)
age	0.020***	0.021***	0.021***	0.001	0.001	0.022***
	(0.007)	(0.007)	(0.007)	(0.012)	(0.009)	(0.007)
religiosity index	0.337***	0.335***	0.334***	0.333***	0.302***	0.333***
	(0.083)	(0.084)	(0.084)	(0.083)	(0.084)	(0.084)
foreigner	-0.765***	-0.470	-0.782***	-0.861***	-0.790***	-0.777***
	(0.246)	(0.341)	(0.247)	(0.246)	(0.246)	(0.247)
graduate	-0.059	-0.025	-0.083	-0.098	0.036	-0.018
	(0.221)	(0.220)	(0.290)	(0.223)	(0.220)	(0.223)
above median age				0.860**		
				(0.430)		
has kids					1.074***	
					(0.374)	
married						-0.075
						(0.292)
Day Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
N	967	967	967	967	966	967
Mean	12.83	12.83	12.83	12.83	12.83	12.83

Taken together with the experimental treatment effects in Table 5, these meta-norm responses suggest a sizable mismatch between perceived and actual patterns of gendered moral judgment: respondents overwhelmingly predict harsher moral judgment for women, while our measured treatment effects indicate the opposite on average in this sample.

Table 7: Lisbon citygoers: Lied to boss

	TreatM	TreatF	Total
the same	262	275	537
	48.79%	51.21%	
better	25	165	190
	13.16%	86.84%	
worse	202	37	239
	84.52%	15.48%	
Total	489	477	966

Table 8: Lisbon citygoers: Did not pick kids

	TreatM	TreatF	Total
the same	222	226	448
	49.55%	50.45%	
better	31	219	250
	12.40%	87.60%	
worse	236	32	268
	88.06%	11.94%	
Total	489	477	966

Table 9: Lisbon citygoers: Adultery

	TreatM	TreatF	Total
the same	198	210	408
	48.53%	51.47%	
better	10	243	253
	3.95%	96.05%	
worse	281	25	306
	91.83%	8.17%	
Total	489	478	967

4.3 4.3 Discussion and interpretation

The central empirical pattern we identified is not simply that gender matters, but that the direction of discrimination often runs counter to common expectations. Despite the large majority of people expecting women to be judged more harshly, we found the opposite across

all scenarios and in two different contexts, Women were never judged more harshly than men and across several scenarios women were significantly less judged than men.

In parental care, respondents sanction men more strongly for failing to pick up children from school due to work. One reading is that respondents treat the father's absence as more discretionary or less excusable, or infer a different level of responsibility from the same. Another reading is that respondents attribute higher baseline caregiving to mothers and therefore judge failures as less diagnostic of character or less blameworthy. These are not mutually exclusive, and both fit a broader pattern in which moral sanctions depend heavily on perceived choice and intent.

In adultery, we find again a greater leniency toward women's behaviour in the citygoers sample. This result sits in tension with the sexual double standard literature, in which women's sexual transgressions appear to be sanctioned more than men's (Crawford and Popp 2003). The findings from this literature, however, are more nuanced; while findings from qualitative and within-study designs - more prone to demand characteristics - support the idea that women's sexual behaviour is more harshly judged, the findings from between-subject designs are more mixed with several studies finding no evidence for double standards for sexual behaviour (e.g. Sprecher, Regan, McKinney, Maxwell, and Wazienski, 1997; Gentyr, 1998).

One way to reconcile leniency with persistent sexualized prescriptions is to consider "protective" evaluations that are subjectively positive but reinforce gender hierarchy, often discussed under ambivalent or benevolent sexism. (Glick and Fiskee 1997). In this view, leniency can reflect paternalistic protection, lower attributions of agency, or greater moral concern about harm to women, rather than egalitarian evaluation. Importantly, such protection can coexist with restrictive prescriptions about women's sexuality and role expectations in other settings or under different framing.

In dishonesty, we do not detect a robust gender difference. One interpretation is that the norm being applied is less gender typed, so category based inferences matter less. Another

is that dishonesty is evaluated through a different moral code (fairness or harm) than sex, where purity based processing is more likely to be activated. (Rozin et al. 1999).

The domain pattern is informative: gendered sanctioning appears more concentrated in scenarios connected to family and sexuality, than in a workplace honesty violation.

In contrast to the consistency of results across both contexts - Catholic pilgrims and citygoers - we find that among pilgrims, the overall effect is driven more by women being lenient toward women, while among citygoers it is driven more by men being lenient toward women. This suggests that the social meaning and enforcement of norms varies across populations even when the average direction of discrimination is similar

Finally, our finding on meta-norms is also consistent with previous research that found 93% of participants definitely or probably believed that "women who have had many sex partners are judged more harshly than men who have had many sex partners" (Milhausen and Herold's, 1999)

5 Concluding remarks

We study gender discrimination in moral judgment using a field survey experiment implemented in two distinct settings in Lisbon: a large religious event (World Youth Day 2023) and a non-religious citygoers sample. Across both settings, we find that respondents judge female norm violators less harshly than male norm violators on average. The most consistent difference concerns parental care: in both samples, men are judged more harshly than women for failing to pick up children from school due to work. In the citygoers sample, we also find a pronounced treatment effect in the adultery scenario, while the lying scenario shows no detectable gender difference in either setting.

Beyond first-order treatment effects, we examine whether enumerator gender shifts responses and find little evidence consistent with experimenter demand effects. Finally, in the citygoers sample we document a strong meta-norm misperception: respondents largely expect

women to be judged more harshly than men, despite our experimental evidence indicating the opposite on average.

These results are relevant for understanding how prescriptive gender norms operate in everyday moral evaluation, and for identifying where perceived social norms diverge from actual judgment patterns. Such divergences matter because moral judgments can shape consequential decisions in domains such as workplace evaluation, family law, and social sanctioning.

References

Asmat, R., Borowiecki, K., and Law, M. T. (2024). Competing for equality: Gender bias among juries in international piano competitions, 1890-2023. *Available at SSRN 4910483*.

Becker, S. O., Bentzen, J., and Kok, C. C. (2025). Gender and religion: a survey.

Becker, S. O., Rubin, J., and Woessmann, L. (2021). Religion in economic history: A survey. *The Handbook of Historical Economics*, pages 585–639.

Bertrand, M. and Duflo, E. (2017). Field experiments on discrimination. *Handbook of economic field experiments*, 1:309–393.

Bohren, J. A., Hull, P., and Imas, A. (2025). Systemic discrimination: Theory and measurement. *The Quarterly Journal of Economics*, 140(3):1743–1799.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, 131(4):1753–1794.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2019). Beliefs about gender. *American Economic Review*, 109(3):739–773.

Bursztyn, L., Cappelen, A. W., Tungodden, B., Voena, A., and Yanagizawa-Drott, D. H. (2023). How are gender norms perceived? Technical report, National Bureau of Economic Research.

Bursztyn, L., González, A. L., and Yanagizawa-Drott, D. (2020). Misperceived social norms: Women working outside the home in Saudi Arabia. *American Economic Review*, 110(10):2997–3029.

Bursztyn, L. and Yang, D. Y. (2022). Misperceptions about others. *Annual Review of Economics*, 14(1):425–452.

Casciaro, T. and Lobo, M. S. (2005). Competent jerks, lovable fools, and the formation of social networks. *Harvard Business Review*, 83(6):92–99.

Charness, G., Gneezy, U., and Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, 81(1):1–8.

Coffman, K. B., Exley, C. L., and Niederle, M. (2021). The role of beliefs in driving gender discrimination. *Management Science*, 67(6):3551–3569.

De Quidt, J., Haushofer, J., and Roth, C. (2018). Measuring and bounding experimenter demand. *American Economic Review*, 108(11):3266–3302.

De Quidt, J., Vesterlund, L., and Wilson, A. J. (2019). Experimenter demand effects. In *Handbook of research methods and applications in experimental economics*, pages 384–400. Edward Elgar Publishing.

Dhar, D., Jain, T., and Jayachandran, S. (2022). Reshaping adolescents' gender attitudes: Evidence from a school-based experiment in India. *American Economic Review*, 112(3):899–927.

Ellemers, N. (2018). Gender stereotypes. *Annual Review of Psychology*, 69(1):275–298.

Exley, C. L. and Nielsen, K. (2024). The gender gap in confidence: Expected but not accounted for. *American Economic Review*, 114(3):851–885.

Glick, P. and Fiske, S. T. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, 70(3):491—-512.

Gundersen, S. J. and Beck, S. V. (2020). *Religion and Gender*, pages 1–22. Springer International Publishing, Cham.

Heilman, M. E., Caleo, S., and Manzi, F. (2024). Women at work: pathways from gender stereotypes to gender bias and discrimination. *Annual Review of Organizational Psychology and Organizational Behavior*, 11(1):165–192.

Heilman, M. E., Wallen, A. S., Fuchs, D., and Tamkins, M. M. (2004). Penalties for success: reactions to women who succeed at male gender-typed tasks. *Journal of Applied Psychology*, 89(3):416.

Huber, S. and Huber, O. W. (2012). The centrality of religiosity scale (CRS). *Religions*, 3(3):710–724.

Kleven, H., Landais, C., and Leite-Mariante, G. (2024). The child penalty atlas. *Review of Economic Studies*.

Kleven, H., Landais, C., and Søgaard, J. E. (2019). Children and gender inequality: Evidence from Denmark. *American Economic Journal: Applied Economics*, 11(4):181–209.

Lee, W.-S. and Guven, C. (2013). Engaging in corruption: The influence of cultural values and contagion effects at the microlevel. *Journal of Economic Psychology*, 39:287–300.

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

Niederle, M. and Vesterlund, L. (2011). Gender and competition. *Annual Review of Economics*, 3(1):601–630.

Rudman, L. A. and Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues*, 57(4):743–762.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, 13:75–98.

Appendix

World Youth Day

Table A1: Nationalities in our sample

Nationality	Freq.	Percent
portuguese	516	28.67
spanish	223	12.39
brazilian	146	8.11
french	137	7.61
american	116	6.44
mexican	77	4.28
italian	39	2.17
polish	35	1.94
british	30	1.67
colombian	28	1.56
german	26	1.44
australian	25	1.39
indian	22	1.22
angolan, canadian	21	1.17
irish	19	1.06
phillipines	18	1.00
peruvian	16	0.89
panamanian, salvadoran	14	0.78
chilean, puerto rican	13	0.72
south african	11	0.61
dominican, venezuelan	10	0.56
argentinian, czech	9	0.50
ecuadorian, guatemalan, singaporean	8	0.44
belgian, croatian	7	0.39
austrian, honduran	6	0.33
cape verdean, cuban, dutch, east timorese, equatorial guinea, filipino, indonesian, kenyan, uruguayan	5	0.28
equatorian, ghanaian, hungarian, mozambican	4	0.22
bolivian, costa rican, lebanese, malay, nigerian, slovakian, tanzanian, zimbabwean	3	0.17
belize, cameroonese, chinese, congolese, danish, finnish, jordanian, mauritian, romanian, serbian, swiss, vietnamese	2	0.11
azerbeijan, canada, greek, icelandic, iraqi, ivorian, japanese, lithuanian, maltese, myanmar, nepalese, nicaraguan, pakistani, paraguay, saint thomas, seychelles, slovenian, solomon islanders, south korean, thai, uae, ugandan, ukrainian	1	0.06
Total	1800	100

Table A2: Balance Table: Treatment

	Control	TreatF	Difference
female respondent	0.567	0.597	0.030
	(0.496)	(0.491)	(0.023)
female assistant	0.464	0.483	0.019
	(0.499)	(0.500)	(0.024)
age	27.317	27.255	-0.062
	(10.786)	(10.308)	(0.497)
in Lisbon for WYD	0.884	0.896	0.012
	(0.320)	(0.305)	(0.015)
religiosity index	4.540	4.542	0.003
	(0.509)	(0.546)	(0.025)
foreigner	0.719	0.707	-0.012
	(0.450)	(0.455)	(0.021)
graduate	0.653	0.664	0.011
	(0.476)	(0.473)	(0.022)
women should be able to preside over the Mass	3.092	3.173	0.081
	(2.037)	(2.057)	(0.097)
same-sex couples should be able to marry in the Church	2.680	2.740	0.060
	(1.982)	(1.986)	(0.094)
Observations	905	895	1,800

Table A3: Robustness Table 3: including Assistant FE

	Index		Lied to boss		Did not pick kids		Adultery	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TreatF	-0.486***	-0.472***	-0.106	-0.097	-0.245***	-0.242***	-0.135**	-0.132**
	(0.142)	(0.139)	(0.070)	(0.069)	(0.078)	(0.076)	(0.062)	(0.062)
female respondent		-0.022		-0.168**		0.122		0.024
		(0.143)		(0.071)		(0.078)		(0.064)
age		0.043***		0.013***		0.030***		0.000
		(0.008)		(0.004)		(0.004)		(0.003)
in Lisbon for WYD		-0.266		-0.030		-0.168		-0.068
		(0.236)		(0.113)		(0.127)		(0.100)
religiosity index		0.619***		0.264***		0.106		0.249***
		(0.139)		(0.067)		(0.075)		(0.064)
foreigner		-0.072		-0.122		0.159*		-0.109
		(0.168)		(0.082)		(0.093)		(0.075)
graduate		-0.120		-0.091		0.028		-0.058
		(0.155)		(0.080)		(0.086)		(0.070)
Day Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Assistant Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
N	1,797	1,797	1,797	1,797	1,797	1,797	1,797	1,797
Mean Y	13.19	13.19	4.43	4.43	3.36	3.36	5.40	5.40

Notes: Robust standard errors in parentheses.* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Table 4 Heterogeneity: Women discriminated morally by gender while men did not

	Index		Lied to boss		Did not pick kids		Adultery	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TreatF	-0.230	-0.146	-0.108	-0.081	-0.035	-0.001	-0.087	-0.064
	(0.214)	(0.210)	(0.106)	(0.105)	(0.118)	(0.116)	(0.098)	(0.097)
female respondent	0.199	0.219	-0.182*	-0.175*	0.310***	0.321***	0.071	0.073
	(0.196)	(0.194)	(0.098)	(0.098)	(0.110)	(0.109)	(0.087)	(0.087)
TreatF × female respondent	-0.450	-0.572**	0.021	-0.020	-0.364**	-0.412***	-0.107	-0.140
	(0.284)	(0.280)	(0.141)	(0.140)	(0.156)	(0.152)	(0.127)	(0.126)
age		0.043***		0.013***		0.029***		0.001
		(0.007)		(0.004)		(0.004)		(0.003)
in Lisbon for WYD		-0.236		-0.018		-0.156		-0.062
		(0.231)		(0.112)		(0.126)		(0.098)
religiosity index		0.593***		0.261***		0.107		0.225***
		(0.137)		(0.066)		(0.074)		(0.062)
foreigner		-0.115		-0.123		0.122		-0.114
		(0.163)		(0.080)		(0.089)		(0.070)
graduate		-0.123		-0.075		0.009		-0.056
		(0.152)		(0.077)		(0.083)		(0.069)
Day Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
N	1,800	1,800	1,800	1,800	1,800	1,800	1,800	1,800
Mean Y	13.19	13.19	4.43	4.43	3.36	3.36	5.40	5.40

Notes: Robust standard errors in parentheses.* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: Robustness Table A4: including Assistant FE

	Index		Lied to boss		Did not pick kids		Adultery	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TreatF	-0.199	-0.112	-0.110	-0.082	-0.019	0.015	-0.071	-0.045
	(0.213)	(0.209)	(0.105)	(0.105)	(0.119)	(0.116)	(0.096)	(0.096)
female respondent	0.265	0.278	-0.161	-0.156	0.329***	0.337***	0.096	0.097
	(0.198)	(0.197)	(0.099)	(0.099)	(0.111)	(0.110)	(0.088)	(0.088)
TreatF × female respondent	-0.492*	-0.615**	0.014	-0.025	-0.394**	-0.440***	-0.112	-0.150
	(0.283)	(0.280)	(0.141)	(0.140)	(0.156)	(0.152)	(0.125)	(0.125)
age		0.043***		0.013***		0.030***		0.001
		(0.008)		(0.004)		(0.004)		(0.003)
in Lisbon for WYD		-0.273		-0.030		-0.173		-0.070
		(0.234)		(0.113)		(0.127)		(0.099)
religiosity index		0.632***		0.264***		0.115		0.253***
		(0.139)		(0.067)		(0.075)		(0.063)
foreigner		-0.064		-0.122		0.165*		-0.107
		(0.168)		(0.083)		(0.092)		(0.075)
graduate		-0.118		-0.091		0.030		-0.057
		(0.155)		(0.080)		(0.086)		(0.070)
Day Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Assistant Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
N	1,797	1,797	1,797	1,797	1,797	1,797	1,797	1,797
Mean Y	13.19	13.19	4.43	4.43	3.36	3.36	5.40	5.40

Notes: Robust standard errors in parentheses.* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: The gender of the assistant did not impact moral judgment

	Index		Lied to boss		Did not pick kids		Sex with stranger	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TreatF	-0.418**	-0.406**	-0.086	-0.070	-0.211**	-0.203*	-0.121	-0.133
	(0.193)	(0.190)	(0.097)	(0.096)	(0.107)	(0.104)	(0.085)	(0.084)
female assistant	0.123	0.124	0.036	0.054	0.118	0.128	-0.032	-0.059
	(0.194)	(0.193)	(0.098)	(0.097)	(0.109)	(0.109)	(0.086)	(0.086)
TreatF × female assistant	-0.158	-0.156	-0.032	-0.048	-0.071	-0.083	-0.056	-0.025
	(0.283)	(0.278)	(0.140)	(0.139)	(0.154)	(0.152)	(0.125)	(0.124)
female respondent		-0.069		-0.188***		0.111		0.009
		(0.141)		(0.070)		(0.077)		(0.063)
age		0.043***		0.013***		0.029***		0.000
		(0.007)		(0.004)		(0.004)		(0.003)
in Lisbon for WYD		-0.225		-0.017		-0.149		-0.058
		(0.233)		(0.112)		(0.127)		(0.099)
religiosity index		0.580***		0.259***		0.097		0.224***
		(0.137)		(0.066)		(0.074)		(0.062)
foreigner		-0.120		-0.122		0.120		-0.118*
		(0.164)		(0.080)		(0.090)		(0.071)
graduate		-0.130		-0.078		-0.000		-0.052
		(0.152)		(0.077)		(0.084)		(0.069)
Day Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
N	1,800	1,800	1,800	1,800	1,800	1,800	1,800	1,800
Mean Y	13.19	13.19	4.43	4.43	3.36	3.36	5.40	5.40

Notes: Robust standard errors in parentheses.* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B1: Nationalities in our sample

Nationality	Freq.	Percent
Portuguese	641	66.29
Brazilian	76	7.86
Angolan	39	4.03
São Toméan	26	2.69
German	17	1.76
American	16	1.65
Cape Verdean	15	1.55
French	14	1.45
British	12	1.24
Guinean	9	0.93
Italian	8	0.83
Indian, Mozambican	7	0.72
Canadian, Spanish	6	0.62
Pakistani, Ukrainian	5	0.52
Czech	4	0.41
Argentine, Chinese, Nepalese, Dutch, Polish, Swiss	3	0.31
Belarusian, Irish, Lithuanian, Moroccan, Romanian, Russian, Turkish, Austrian	2	0.21
Algerian, Bangladeshi, Belgian, Colombian, Congolese, Egyptian, Ecuadorian, Guatemalan, Hong Kong, Hungarian, Israeli, Japanese, Mexican, Peruvian, Sri Lankan, Swedish, Serbian, Uruguayan, Venezuelan, Yemeni	1	0.10
Total	967	

Lisbon citygoers

Table B2: Balance Table: Treatment

	(1) TreatM		(2) TreatF		(2)-(1) Pairwise t-test	
Variable	N	Mean/(SE)	N	Mean/(SE)	N	Mean difference
female respondent	489	0.517	478	0.494	967	-0.024
		(0.500)		(0.500)		
female assistant	489	0.540	478	0.550	967	0.010
		(0.499)		(0.498)		
age	489	34.650	478	35.184	967	0.534
		(15.890)		(15.409)		
religiosity index	489	2.775	478	2.668	967	-0.107
		(1.285)		(1.302)		
foreigner	489	0.325	478	0.349	967	0.024
		(0.469)		(0.477)		
graduate	489	0.440	478	0.498	967	0.058*
		(0.497)		(0.501)		

Table B3: Robustness Table 5 – Including Day FE

	Index		Lied to boss		Did not pick kids		Adultery	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TreatF	-0.567**	-0.532**	0.010	0.005	-0.232**	-0.232**	-0.344***	-0.305***
	(0.226)	(0.224)	(0.105)	(0.105)	(0.116)	(0.115)	(0.107)	(0.105)
female respondent		0.323		-0.017		0.256**		0.084
		(0.225)		(0.108)		(0.119)		(0.105)
age		0.017**		0.010***		0.016***		-0.009**
		(0.008)		(0.004)		(0.004)		(0.004)
religiosity index		0.317***		0.092**		-0.005		0.230***
		(0.092)		(0.044)		(0.049)		(0.042)
foreigner		-0.301		-0.042		0.043		-0.302**
		(0.298)		(0.139)		(0.144)		(0.139)
graduate		0.017		0.079		-0.063		0.001
Day Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	895	895	895	895	895	895	895	895
Mean Y	12.79	12.79	4.27	4.27	3.45	3.45	5.07	5.07

Table B4: Robustness Table 5 – Including Assistant FE

	Index		Lied to boss		Did not pick kids		Adultery	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TreatF	-0.642***	-0.589***	-0.035	-0.025	-0.228**	-0.221**	-0.379***	-0.343***
	(0.212)	(0.208)	(0.101)	(0.100)	(0.111)	(0.110)	(0.098)	(0.096)
female respondent		0.260		-0.029		0.193*		0.096
		(0.209)		(0.102)		(0.114)		(0.097)
age		0.021***		0.010***		0.019***		-0.008**
		(0.007)		(0.003)		(0.004)		(0.003)
religiosity index		0.350***		0.118***		0.025		0.208***
		(0.082)		(0.041)		(0.047)		(0.038)
foreigner		-0.688***		-0.297**		-0.151		-0.239*
		(0.261)		(0.132)		(0.138)		(0.124)
graduate		-0.040		0.068		-0.038		-0.069
		(0.217)		(0.106)		(0.118)		(0.101)
Assistant Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	967	967	967	967	967	967	967	967
Mean Y	12.83	12.83	4.26	4.26	3.47	3.47	5.10	5.10

Table B5: Lisbon citygoers: enumerator gender did not impact moral judgment

	Index		Lied to boss		Did not pick kids		Adultery	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
TreatF	-0.680**	-0.678**	-0.131	-0.130	-0.259	-0.272*	-0.290**	-0.276**
	(0.286)	(0.281)	(0.139)	(0.138)	(0.164)	(0.164)	(0.125)	(0.124)
female assistant	-0.797***	-0.735**	-0.394***	-0.367**	-0.071	-0.099	-0.333**	-0.270**
	(0.284)	(0.292)	(0.140)	(0.144)	(0.157)	(0.161)	(0.129)	(0.132)
TreatF×female assistant	0.090	0.168	0.219	0.227	0.010	0.038	-0.139	-0.098
	(0.423)	(0.420)	(0.200)	(0.199)	(0.224)	(0.223)	(0.196)	(0.193)
female respondent		0.293		-0.010		0.179		0.124
		(0.212)		(0.102)		(0.113)		(0.098)
age		0.023***		0.012***		0.019***		-0.008**
		(0.007)		(0.003)		(0.004)		(0.003)
religiosity index		0.313***		0.098**		0.031		0.184***
		(0.083)		(0.041)		(0.046)		(0.039)
foreigner		-0.562**		-0.216*		-0.100		-0.246**
		(0.259)		(0.126)		(0.130)		(0.121)
graduate		-0.052		0.057		-0.035		-0.074
N	967	967	967	967	967	967	967	967
Mean Y	12.83	12.83	4.26	4.26	3.47	3.47	5.10	5.10