



Subjective well-being and the gender composition of the reference group: Evidence from a survey experiment[☆]

Elena Fumagalli^{a,*}, Laura Fumagalli^b

^a Utrecht University, Kriekenpitplein 21–22, Utrecht, 3584 EC (NL)

^b ISER, University of Essex, Wivenhoe Park, Colchester, CO43SQ, (UK)



ARTICLE INFO

Article history:

Received 22 October 2020

Revised 25 October 2021

Accepted 12 December 2021

Available online 31 December 2021

JEL classification:

C99

I31

J16

Keywords:

Gender differences

Well-being

RCT

ABSTRACT

This paper tests how people's subjective well-being reacts when they compare themselves with other people of the same gender, and if this reaction differs between women and men. We implement a randomized control trial prompting some respondents to compare themselves with people of the same gender and leaving the reference group of others unconstrained. Treated women report higher income and leisure satisfaction. Evaluating satisfaction in relation to a given reference group may be cognitively demanding. When accounting for this, we find that the treatment also increases women's health satisfaction. No or small effects are found for men, suggesting that the reference group affects subjective well-being reporting of men and women differently.

© 2021 The Author(s). Published by Elsevier B.V.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

1. Introduction

In recent decades, in the US and in most European countries, women's subjective well-being (SWB) has declined relative to men's, despite women having reduced their disadvantage compared to men in education, the labor market, and civil rights (Stevenson and Wolfers, 2009; Bertrand, 2020). A possible explanation for this puzzle (Bertrand, 2011) emphasizes the role of the gender composition of the reference group. In the past women mainly compared themselves with other women, while now they increasingly compare themselves with men. As women are more disadvantaged than men in many life domains (e.g., in the labor market), women should report higher (lower) SWB when they compare themselves with women (men). We test Bertrand's hypothesis that women report higher SWB when they compare themselves with women

[☆] We are grateful to Cara Booker, Chris Bollinger, Tom Crossley, Conchita D'Ambrosio, Peter Ditto, Thomas Martin, Elaine Prentice-Lane, Stephanie Rosenkranz, Anna Salomons, and seminar participants at Utrecht University (Utrecht, Netherlands), the 2018 Lisbon Meeting on Economics and Political Science (Lisbon, Portugal), the third Dondena workshop on public policy (Milan, Italy), the 9th Alpine Population Conference (la Thuile, Italy) and the 2019 Understanding Society Scientific Conference (Colchester, UK) for useful comments and suggestions. This study can be found in the AEA RCT Registry (AEARCTR-0005177). Understanding Society is an initiative funded by the Economic and Social Research Council and various Government Departments, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by NatCen Social Research and Kantar Public. The research data are distributed by the UK Data Service. One of the authors was funded by ESRC ES/N00812/1 and ES/S012486/1.

* Corresponding author.

E-mail addresses: e.fumagalli@uu.nl (E. Fumagalli), lfumag@essex.ac.uk (L. Fumagalli).

only, at least in domains where women are disadvantaged compared to men.¹ We also test what happens to men's SWB when men compare themselves with men only. These two tests have never been performed.

We implement a randomized control trial (RCT) on a nationally representative sample of British respondents. We experimentally assign the women's reference group by randomly prompting some women to evaluate their SWB (measured as satisfaction with: health, income, amount of leisure time and life overall) by comparing themselves with women only (treatment group); we ask other women to answer the standard satisfaction questions used in most British surveys (control group). These standard satisfaction questions do not refer to an explicit reference group. To test for heterogeneity by gender, we implement the same test for men: men in the treatment group are prompted to compare themselves only with men, men in the control group are asked the standard satisfaction questions.

The estimated treatment effects are likely to differ across genders and satisfaction domains as they are likely to depend on the reference group considered by the individuals in the control group and the perceived gender gap in each domain. We can make some hypotheses on the estimated treatment effects based on what we expect the perceived gender gap to be. Consider the case of women who report their satisfaction about income: a domain where women are likely to believe they are worse-off than men. If at least some women in the control group compare themselves with men and if women's SWB decreases with the perceived income of the reference group, women in the treatment group should report higher income satisfaction than those in the control group. This is because the reference group of the women in the treatment group has lower perceived income than the reference group of the women in the control group. The estimated treatment effects for men and for satisfaction domains other than income are more difficult to predict. For example, the perceived gender health gap is uncertain (see also [Section 3](#)). Women live longer than men but there is mounting evidence suggesting that gender discrimination, health practices biased towards men, and new trends in risk behaviors are harming women's health more than men's (see [Ravindran et al., 2020](#); [Vijayasingham et al., 2020](#); [Amin et al., 2021](#); [Feeny et al., 2021](#); [Seedat and Rondon, 2021](#), for a recent discussion).

The treatment modifies women's satisfaction with income and leisure. We find that, when prompted to comprise their reference group solely of women, women report higher satisfaction with income and leisure: domains where women are likely to think they are more disadvantaged than men (see [Section 3](#)). This suggests that when answering the standard satisfaction questions (at least some) women include men in their reference group. Treatment effects on income satisfaction are larger for women who work in sectors with a large gender pay gap; treatment effects on leisure satisfaction are larger for women who do more housework than their partner.

The treatment has an ambiguous effect on women's satisfaction with health, possibly due to a combination of a positive treatment effect and attenuation bias (see [Sections 3](#) and [7](#)). Recent literature suggests that, when reporting health through Likert scales, people choose the middle outcome more often than what objective measures of their health would suggest [Greene et al. \(2015\)](#). This 'box-ticking' strategy, which arise as a more socially desirable alternative to non-response ([Malhotra et al., 2014](#); [Sturgis et al., 2014](#)), can lead to attenuation bias.

To tackle this problem, we use a Middle Inflated Ordered Probit model (MIOP): an ordered probit where the middle outcome is inflated (see [Bagozzi and Mukherjee, 2012](#)). The identification of the parameters of the MIOP is helped by exclusion restrictions: variables explaining the adoption of a 'box-ticking' strategy, but uncorrelated with satisfaction. As exclusion restrictions we use likely determinants of data quality affected or generated by other randomized survey experiments carried out on the same sample. The results from the MIOP regressions, purged of attenuation bias, suggest that the treatment also increases women's health satisfaction. These results are coherent with a situation where women on average think that women are less healthy than men.

There is little or no evidence that men report lower satisfaction when prompted to compare themselves with other men. There are three, non mutually exclusive, reasons for this. First, at least in some domains (e.g., leisure or health), men may not think they are better-off than women. Second, men may compare themselves predominantly with men even when they answer the control questions.² Third, while individuals' satisfaction is likely to be negatively influenced by being worse-off than the reference group, individuals' satisfaction is not necessarily positively influenced by being better-off than the reference group.³ All these reasons lead to 'upward' comparisons, meaning that individuals' SWB is more affected by comparisons with better-off people than by comparisons with worse-off people (see also [Duesenberry, 1949](#); [Ferrer-i Carbonell, 2005](#); [Boyce et al., 2010](#); [Card et al., 2012](#)).

This paper contributes to the understanding of how the characteristics of the reference group affect SWB. A prominent literature tests whether one's SWB (generally measured as happiness or life satisfaction) decreases with the income of one's reference group ('relative income hypothesis'). In a large part of this literature, the income of the reference group is computed by the researcher as the income of individuals similar to the respondent in a set of characteristics (for example: age, gender, education, occupation, region). These studies conclude that being exposed to better-off people decreases SWB

¹ A test of whether women compare themselves with men more than in the past is left for future research.

² In fact a comparison with men could be the default option for both men and women, as [Criado Perez \(2019\)](#) puts it, the results of our 'deeply male-dominated culture is that the male experience, the male perspective, has come to see at universal, while the female experience [...] is seen as, well, niche'.

³ This is consistent with a utility function concave in relative outcomes where the decrease in utility an individual experiences from being worse-off than their reference group is larger than the increase in utility an individual experiences from being better-off than their reference group. For a mathematical formalization, see [Appendix B](#).

when it induces well-being comparisons, but increases SWB when it suggests that improvements in status are possible (tunnel effects).⁴ Adopting an increasingly popular approach, other studies (e.g., Luttmer, 2005; Kingdon and Knight, 2007; Clark et al., 2009b; Deaton and Stone, 2013; Ifcher et al., 2018; Brodeur and Fleche, 2019; Noy and Sin, 2021) avoid arbitrary choices of the reference group and define the reference group based on one single characteristic: geographical proximity. These studies generally find that the income of proximal (distant) neighbors is positively (negatively) correlated with SWB.⁵

The identification of the effect of the characteristics of the reference group on SWB is complicated by the fact that the reference group people think of when reporting SWB is endogenous and unknown. Several authors have addressed this problem by leveraging quasi-experimental or experimental variation in the outcomes of people who are likely to be in one's reference group or in the available information about these outcomes. For example, Kuhn et al. (2011) makes use of random variation in the income of neighbors induced by a lottery win. Card et al. (2012) and Perez-Truglia (2020) exploit random or quasi-random variation in information about pay or income of coworkers or fellow citizens. McBride (2010) and Ifcher et al. (2020) test whether the SWB of the participants in a lab experiment is affected by knowing the performance of other participants. This quasi-experimental/experimental literature generally suggests that one's SWB decreases when the income of the reference group increases.⁶

Our contribution is to tests how SWB changes when the reference group – and particularly the gender of the reference group – is experimentally assigned. While a growing number of papers shed light on the characteristics of the reference groups people refer to when formulating SWB statements (see, for example: Senik, 2009; Clark and Senik, 2010; Pérez-Asenjo, 2011; Akay et al., 2012; Godechot and Senik, 2015; Goerke and Pannenberg, 2015; Senik et al., 2017; Neumann-Böhme et al., 2021), none of the existing papers has looked explicitly at the role of gender in the definition of the reference group, nor at how the role of the reference group varies across genders and satisfaction domains. Moreover, we are not aware of any other paper in the relevant literature that carries out a controlled experiment on a nationally-representative sample. Our work is a first step in filling this gap.

This paper also contributes to the literature on the differences between women and men in outcomes. A growing literature shows that these gender differences in outcomes can be partially explained by differences between women and men in behavior and psychological attributes such as: competitiveness, risk aversion, social preferences, attitude to networking, and leader effectiveness (see, for example: Niederle and Vesterlund, 2007; Croson and Gneezy, 2009; Gneezy et al., 2009; Croson et al., 2012; Booth and Nolen, 2012a; 2012b; Booth et al., 2014; Pekkarinen, 2015; Shurchkov and Eckel, 2018; Buser and Yuan, 2019; Cai et al., 2019; Iriberri and Rey-Biel, 2019; Gauriot and Page, 2019; Grossman et al., 2019; Hoyer et al., 2020; Mengel, 2020). While this literature has successfully unpacked many of the determinants of the gender gap in outcomes, new forms of gender heterogeneity need to be investigated (Blau and Kahn, 2017; Donna and Veramendi, 2021). We move this literature a step forward by investigating whether and how this gender gap in outcomes translates into SWB, and, more in general, by studying the gender determinants of SWB. We find differences in how the SWB of women and men reacts when people are prompted to compare themselves with other people of the same gender. This is a type of heterogeneity that has never been discussed in the literature.

New research is needed to explain why women and men react differently to SWB comparisons with people of the same gender. Our results suggest that this is at least partially due to the fact that women do (or perceive they do) less well than men in several domains. Other factors may be at play. For example, the cognitive process of formulating well-being statements may differ by gender. Recent literature finds some evidence that women and men's SWB does react differently to comparisons with other people. This is true even when the outcomes over which these comparisons are made do not differ by gender (Ifcher et al., 2020). More generally, new evidence suggests that women and men differ in the aspects they consider when reporting SWB (Benjamin et al., 2021). The identification of gender differences in reported well-being and their causes is a very interesting development in both the literature on SWB and in the literature on the differences between women and men.

2. Experimental design

We use experimental data from wave five of the Understanding Society Innovation Panel: a longitudinal survey, based on a household level probability sample representative of the population living in Britain south of the Caledonian Canal, collected annually since 2008 to conduct survey experiments (Burton et al., 2008, University of Essex et al., 2019). All adult members of sampled households were surveyed. Wave five (IP5), collected in 2012, has a sample of 1224 households. Our experiment was included in IP5, as a winning project of the 2011 public call (Auspurg et al., 2013).⁷

⁴ Clark and Oswald (1996), Ferrer-i Carbonell (2005) and Diriwachter and Shvartsman (2018) find support for the relative income hypothesis. Senik (2004), Clark et al. (2009a) find evidence of tunnel effects. Senik (2008), Caporale et al. (2009), Kuhn et al. (2011) Brown et al. (2015), Brown and Gray (2016), find mixed or no results. Reviews of the role of interpersonal comparison are: Clark et al. (2008) and Clark and D'Ambrosio (2015). On the interplay between interpersonal comparisons, aspirations and SWB, see Crosby (1976); Blanchflower and Oswald (2004); Stutzer (2004); Genicot and Ray (2017).

⁵ See Ifcher et al. (2018) for a careful analysis of the channels leading to the effect.

⁶ Kuhn et al. (2011) only finds effects on behavior, but not on SWB.

⁷ The experiment is part of a broader project also studying comparisons by education. The education treatment was administered to another subsample of individuals. To avoid order or priming effects, individuals studied in this papers were not administered the education treatment.

Table 1
Treatment effect on income satisfaction Δs_i .

	Gender and Income			
	Woman		Man	
	(1)	(2)	(3)	(4)
Unprompted reference group	Low-income	High-income	Low-income	High-income
1. Same gender	$\Delta s_i = 0$	$\Delta s_i = 0$	$\Delta s_i = 0$	$\Delta s_i = 0$
2. Same income	$\Delta s_i = ?$	$\Delta s_i > 0$	$\Delta s_i < 0$	$\Delta s_i = ?$
3. Same gender & income	$\Delta s_i < 0$	$\Delta s_i > 0$	$\Delta s_i < 0$	$\Delta s_i > 0$
4. Unrestricted	$\Delta s_i > 0$	$\Delta s_i > 0$	$\Delta s_i < 0$	$\Delta s_i < 0$

Alongside the other questions contained in the IP5 questionnaire, the Understanding Society respondents were asked to answer a set of questions on satisfaction with four domains: health, household income, the amount of leisure time, and life overall. For each domain, we produced two different versions of these questions and we randomly allocated them to respondents. The treatment question asks women (men) to compare themselves with other women (men) as follows:

Treatment questions. *How dissatisfied or satisfied are you with your [domain] if you compare yourself with other [(wo)men]?*

The control question is the Understanding Society standard SWB question and does not mention any comparison group:

Control question. *How dissatisfied or satisfied are you with your [domain]?*

As standard practice in Understanding Society, answers were recorded using Likert scales with the options: ‘completely dissatisfied’, ‘mostly dissatisfied’, ‘somewhat dissatisfied’, ‘neither satisfied nor dissatisfied’, ‘somewhat satisfied’, ‘mostly satisfied’, ‘completely satisfied’.

3. Perceived gender gap, reference group and treatment effect

We now discuss how exogenously assigning the gender of the reference group is likely to affect SWB. A formalization of this argument can be found in [Appendix B](#). We focus on income satisfaction, but the argument can be extended to other domains. When answering questions on income satisfaction, people are likely to take into account both their income and the income of their reference group. Following the literature on the relative income hypothesis, we assume that satisfaction increases with own income and decreases with the income the reference group. The treatment prompts a specific reference group and thus it can modify the reference group respondents would have normally thought of when unprompted.

The treatment effect is the difference in satisfaction between the case in which the reference group is prompted (‘prompted reference group’) and the case in which the reference group is not prompted (‘unprompted reference group’). As the treatment is random, this difference is identified by comparing the average satisfaction between the treatment and the control groups. The treatment effect is likely to depend on: i) the perceived gender income gap; ii) the difference in the gender composition of the unprompted and the prompted reference group, iii) the respondent's income, and iv) the respondent's gender. The expected signs of the treatment effect for different combinations of gender, income and unprompted reference group are summarized in [Table 1](#). In what follows, we will discuss the general intuition and some specific cases. Please refer to [Appendix B](#) for a more formal and comprehensive discussion.

We focus on women first. For simplicity, we assume there are only two levels of income: low and high, and that women perceive that women have lower income than men.⁸ Consider the case of a high-income woman (column 2 in [Table 1](#)). If the woman's unprompted reference group is already composed by all women (‘same gender’ reference group, column 2, row 1), then the treatment does not modify the woman's reference group and the treatment effect is zero. If the woman's unprompted reference group includes only high-income people irrespective of their gender (‘same income’ reference group, column 2 row 2), the treatment makes the woman remove high-income men from her reference group, and include low-income. As a result, the prompted reference group has lower income than the unprompted reference group and the treatment effect is positive. If the woman's unprompted reference group includes only high-income women (‘same gender and income’ reference group, column 2 row 3), the treatment makes the woman add low-income women to her reference group. Again, the prompted reference group has lower income than the unprompted reference group and thus the treatment effect is positive.⁹ Finally, if the woman's unprompted reference group is composed by all men and women (‘unrestricted’ reference group, column 2 row 4), the treatment induces the woman to remove men from her reference group, which reduces the income of the prompted reference group and leads to a positive treatment effect.

⁸ The UK has one of the largest gender pay gaps in Europe. In 2012 in the UK the average gross hourly earnings of female paid employees was 22.6% lower than that of male paid employees (Eurostat, 2020). The European average was 17.4%. More information on the gender pay gap by sectors of occupation can be found in [Fig. A.3](#).

⁹ An example conceptually similar to this case is when a woman includes both married and single women in her unprompted reference group, but only single women in her prompted reference group. If this woman considers married women to have higher income than unmarried women, then the prompted reference group has lower income than the unprompted reference group, and the treatment effect is positive. We thank an anonymous referee for pointing out this case.

The scenario for low-income women is more involved (as detailed in Column 1 of Table 1 and in the model in Appendix B). The cases where the woman has the ‘same gender’ and the ‘unrestricted’ unprompted reference group yield the same results as in the high-income woman scenario, namely a zero and a positive treatment effect (see Table 1, column 1, rows 1 and 4). The case where the woman’s unprompted reference group is the ‘same gender and income’ reference group is undetermined (see Appendix B for details). We discuss here the only case where the treatment effect for women is negative, that is the case where the unprompted reference group of a low-income woman is composed by low-income women only (Table 1, column 1, row 3). In this case, the treatment effect makes the woman add high-income women to her reference group. As a consequence, the prompted reference group has higher income than the unprompted reference group and the treatment effect is negative.

Predictions can also be made for men (see Table 1, columns 3 and 4). Consider, for example, income satisfaction of a low-income man and assume he thinks men have higher income than women. If his unprompted reference group is already composed by men only, the treatment effect would be zero (column 3, row 1 of Table 1). In all the other cases, the treatment effect will be negative as the prompted reference group – which is restricted to males – has higher income than the unprompted reference group.

Theoretically, the predictions for men mirror the predictions for women (column 3 mirrors column 2 and column 4 mirrors column 1). However, men and women are likely to differ in various aspects. For example, for many men the unprompted reference group may be already composed mainly or exclusively by men, as men may consider comparing themselves with women an unfair test, and a way of cheating on self-evaluation. This may not be the same for women, because social norms may put more pressure on women to compete with men than on men to compete with women. As a consequence, more men than women may have a zero treatment effect.

The treatment effect on other satisfaction domains depends on the perceived gender gap in the considered domain. Women may think they are worse off than men regarding leisure: women have been found to enjoy less or worse quality leisure than men (Krueger, 2007; Bertrand, 2011). In the UK, men have been found to spend more time on leisure than women. In 2015, on average, UK men spent six hours and nine minutes per day in leisure activities, while UK women spent five hours and 29 minutes per day (ONS, 2017). Men were more likely than women to spend their leisure time in sports and hobbies, while women were more likely to spend their leisure time socializing (ONS, 2017). Men may also think they are worse off than women regarding leisure, as, on average, men spend more time than women in paid work. Therefore, the perceived gender leisure gap is unclear.

The perceived gender health gap is also unclear. Women live longer than men, although in the UK this advantage is the narrowest in the WHO European region (see Figure 2.6 in: WHO, 2018). In a UK sample similar to ours, women have also been found to suffer less than men from chronic stress and to be less at risk of cardiovascular diseases (see: Davillas and Pudney, 2017). However, there is also extensive evidence highlighting several health-related aspects where women are worse off than men. Women report worse health conditions and higher pain levels than men, are less likely than men to receive analgesic treatment (Case and Paxson, 2005; Chen et al., 2008; Crimmins et al., 2010; Bartley and Fillingim, 2013). In addition, health practices biased towards men, and new trends in risk behaviors may affect women’s health more than men’s (Ravindran et al. (2020); Vijayasingham et al. (2020); Amin et al. (2021); Feeny et al. (2021); Seedat and Rondon (2021)). Data from the UK (see, for example: Carrieri and Jones, 2017; Davillas and Pudney, 2017; Chaparro et al., 2019) also suggest that, compared to men, women have higher levels of fibrinogen (a marker for inflammation), and lower iron levels.

The treatment effect on life overall is almost impossible to predict as it depends on which aspects of life respondents consider and how they aggregate them into a single answer. This process is mentally demanding and driven by the information most accessible to respondents at the time they are surveyed (Schwarz and Strack, 1999). Therefore, we focus on satisfaction with specific life domains.

4. Empirical framework

Randomization achieves treatment exogeneity. Table 2 reports balancing tests on the differences in means between the control and the treatment group of crucial characteristics: respondent’s age, whether the respondent is female, is in a registered partnership/marriage, has post-compulsory education, has a long-term illness or disability, lives in England, and is white British. These tests suggest that the treatment and the control group are comparable. Marginally statistically significant differences are only found in education (for women only) and disability (for men only). Therefore, in our estimates of the treatment effect we control for these characteristics.

Randomization cannot address the problem that the treatment question is likely to be more cognitively demanding than the control question. There are three reasons why the treatment question may be more cognitively demanding than the control question. First, the treatment question is longer and uses if clauses, and both these aspects have been linked to increased cognitive load (Krosnick, 1991; Pitler and Nenkova, 2008). Second, the control question is a standard satisfaction question. Therefore, our panel respondents are likely to find the control question easy to answer, as they might have answered it before, either in previous waves of our survey or in other surveys. Third, research suggests that respondents do not generally maximize the quality of their answers, but, rather, they minimize their response effort subject to a threshold of minimum required response quality (see, for example Krosnick, 1991; Malhotra et al., 2014). This has been called ‘satisficing’ behavior. When asked the control question, the respondents are free to pick the reference group they prefer, which is likely to be the reference group that appears the most natural to them. This minimizes response effort, while still delivering

Table 2
Balancing tests.

	(1)	(2)	(3)
	Full sample	Females	Males
Females	-0.002 (0.032)		
Registered couple	0.042 (0.032)	0.050 (0.043)	0.033 (0.048)
Highly educated	-0.050 (0.032)	-0.105 (0.043)	0.019 (0.048)
Age	1.184 (1.207)	1.509 (1.619)	0.772 (1.813)
Disability	-0.059 (0.031)	-0.025 (0.041)	-0.101 (0.046)
England	0.030 (0.022)	0.006 (0.030)	0.060 (0.034)
White British	0.008 (0.020)	0.018 (0.028)	-0.004 (0.028)
N	970	540	430

Note: IP5 data. Females: whether respondent is female; Registered couple: respondent is in a registered partnership/marriage; Highly educated: respondent has post-compulsory education; Age: respondent's age; Disability: respondent has a long-term illness or disability; England: respondent lives in England; White British: respondent is White British. Point estimates are differences between the control and the treatment group. Associated standard errors are reported in parentheses. Stars for significance levels are not reported.

a satisfactory answer. When a specific reference group is prompted, respondents may have to consider a reference group that is different from the one they find most natural and this may result in increased cognitive load.

If the treatment increases the cognitive load, the estimated treatment effect may be the combination of the real treatment effect (the change in satisfaction induced by the prompted reference group) and the treatment effect due to cognitive load. Cognitive load may affect answers recorded through Likert scales in two ways (Krosnick, 1991; Sturgis et al., 2014). First, with Likert scales with more than three options, requiring a judgment on both the direction and the intensity of the opinion expressed, respondents facing cognitively demanding questions may decide to choose less or more extreme options than what they would have chosen otherwise. Second, when the neutral option of the scale is labeled as 'neither/nor', respondents may react to cognitively demanding questions by choosing it dis-proportionally often.

We address this problem as follows. First, for all domains we collapse the seven points of the Likert scale into three: 'Satisfied', 'Neutral' and 'Dissatisfied' (see Figure A.2.).¹⁰ This is a standard practice in cases where data are collected through cognitively demanding questions and questions are recorded through Likert scales (see, for example: Chaparro et al., 2019), and ensures that results are not driven by movements within the satisfied/dissatisfied domain due to cognitive load (Krosnick and Presser, 2010). Second, we add to our main specification (ordered probit) a specification relaxing the ordinality assumption (multinomial probit). Third, we disregard individuals who chose the neutral answer and we estimate a probit model using, as dependent variable, a dummy variable indicating whether the individual is satisfied or dissatisfied. Fourth, for the case of health satisfaction, we use a Middle inflated Ordered Probit to address the problem that respondents - and particularly treated respondents - may dis-proportionally choose the neutral option.

There are two reasons why we do not use the MIOP on income and leisure satisfaction. The first reason is that, for the case of income and leisure satisfaction, the multinomial probit models do not suggest any tendency of disproportionately choosing the neutral option (see Fig. 1 and Table 4 in the next section). The second reason is that there is a growing literature that claims that answering general questions about health or health satisfaction using Likert scales is a particularly demanding process (see Jylhä, 2009; Greene et al., 2015; Chaparro et al., 2019), as respondents need to consider a very diverse range of physical and mental conditions and summarize their evaluations of all these aspects into a single answer.

¹⁰ For completeness, we also present a specification of the order probit model that considers all the seven Likert scale points (see: Fig. A.1 and Table A.1 of Appendix A). This specification suggests that movements within the satisfied/dissatisfied domain may mask the real treatment effect.

Table 3
Ordered probit: MEMs (all outcomes).

	<i>Health: Dissatisfied</i>		<i>Health: Neutral</i>		<i>Health: Satisfied</i>	
	(1)	(2)	(3)	(4)	(5)	(6)
	Females	Males	Females	Males	Females	Males
Treatment	-0.051 (0.035)	0.039 (0.039)	-0.013 (0.010)	0.010 (0.010)	0.064 (0.044)	-0.049 (0.049)
<i>Income: Dissatisfied</i>						
<i>Income: Neutral</i>						
<i>Income: Satisfied</i>						
Treatment	-0.117 (0.052)	0.030 (0.065)	-0.022 (0.013)	0.004 (0.009)	0.139 (0.065)	-0.034 (0.074)
<i>Leisure: Dissatisfied</i>						
<i>Leisure: Neutral</i>						
<i>Leisure: Satisfied</i>						
Treatment	-0.060 (0.034)	-0.032 (0.035)	-0.014 (0.009)	-0.010 (0.012)	0.074 (0.043)	0.042 (0.047)
<i>Overall: Dissatisfied</i>						
<i>Overall: Neutral</i>						
<i>Overall: Satisfied</i>						
Treatment	0.038 (0.033)	0.030 (0.034)	0.015 (0.012)	0.014 (0.015)	-0.053 (0.044)	-0.044 (0.048)

Note: IP5 data. Controls: high education and disability dummies. N(health): Women=948, Men=777; N(income): Women=433, Men=321 (respondents not in couples); N(leisure): Women=950, Men=773; N(overall): Women=951, Men=778. Stars for significance levels are not reported.

Table 4
Multinomial probit: MEMs (all outcomes).

	<i>Health: Dissatisfied</i>		<i>Health: Neutral</i>		<i>Health: Satisfied</i>	
	(1)	(2)	(3)	(4)	(5)	(6)
	Females	Males	Females	Males	Females	Males
Treatment	-0.095 (0.037)	0.028 (0.043)	0.106 (0.046)	0.023 (0.039)	-0.011 (0.050)	-0.051 (0.052)
<i>Income: Dissatisfied</i>						
<i>Income: Neutral</i>						
<i>Income: Satisfied</i>						
Treatment	-0.134 (0.058)	-0.015 (0.071)	0.024 (0.059)	0.086 (0.075)	0.111 (0.070)	-0.072 (0.080)
<i>Leisure: Dissatisfied</i>						
<i>Leisure: Neutral</i>						
<i>Leisure: Satisfied</i>						
Treatment	-0.068 (0.038)	-0.051 (0.038)	-0.000 (0.033)	0.030 (0.041)	0.068 (0.045)	0.021 (0.050)
<i>Overall: Dissatisfied</i>						
<i>Overall: Neutral</i>						
<i>Overall: Satisfied</i>						
Treatment	-0.017 (0.035)	0.029 (0.040)	0.109 (0.042)	0.007 (0.039)	-0.092 (0.048)	-0.036 (0.050)

Note: IP5 data. Controls: high education and disability dummies. N(health): Women=948, Men=777; N(income): Women=433, Men=321 (respondents not in couples); N(leisure): Women=950, Men=773; N(overall): Women=951, Men=778. Stars for significance levels are not reported.

5. Results

Fig. 1 shows marginal effects at the mean (MEMs) from ordered and multinomial probit models, by gender and satisfaction domain (see also Table 3 and Table 4).¹¹ For income satisfaction we only keep respondents not in couples, as the Understanding Society income satisfaction question considers household income. Taken together, results suggest that the treatment: i) increases women's income and leisure satisfaction; ii) increases the probability that women choose the neutral answer in health and life satisfaction questions; iii) has an ambiguous effect on women's health satisfaction, possibly due to a combination of a positive treatment effect and an attenuation effect due to cognitive load; iv) affects men minimally. We now look each set of results (ordered probit models and multinomial probit models) separately.

We first look at the results for women using ordered probit models (see Table 3 and triangular filled markers in Fig. 1). These MEMs indicate that the treatment increases women's income and leisure satisfaction: treated women are 12 percentage points (pp) less likely than untreated women to report income dissatisfaction (Table 3, second panel, column 1) and

¹¹ Our discussion focuses on marginal effects. The associated coefficients are not presented and are available on request. Differences in means between the treatment and the control group (Likert scale collapsed to a three points Likert scale) by gender and domain are reported in Table A.2

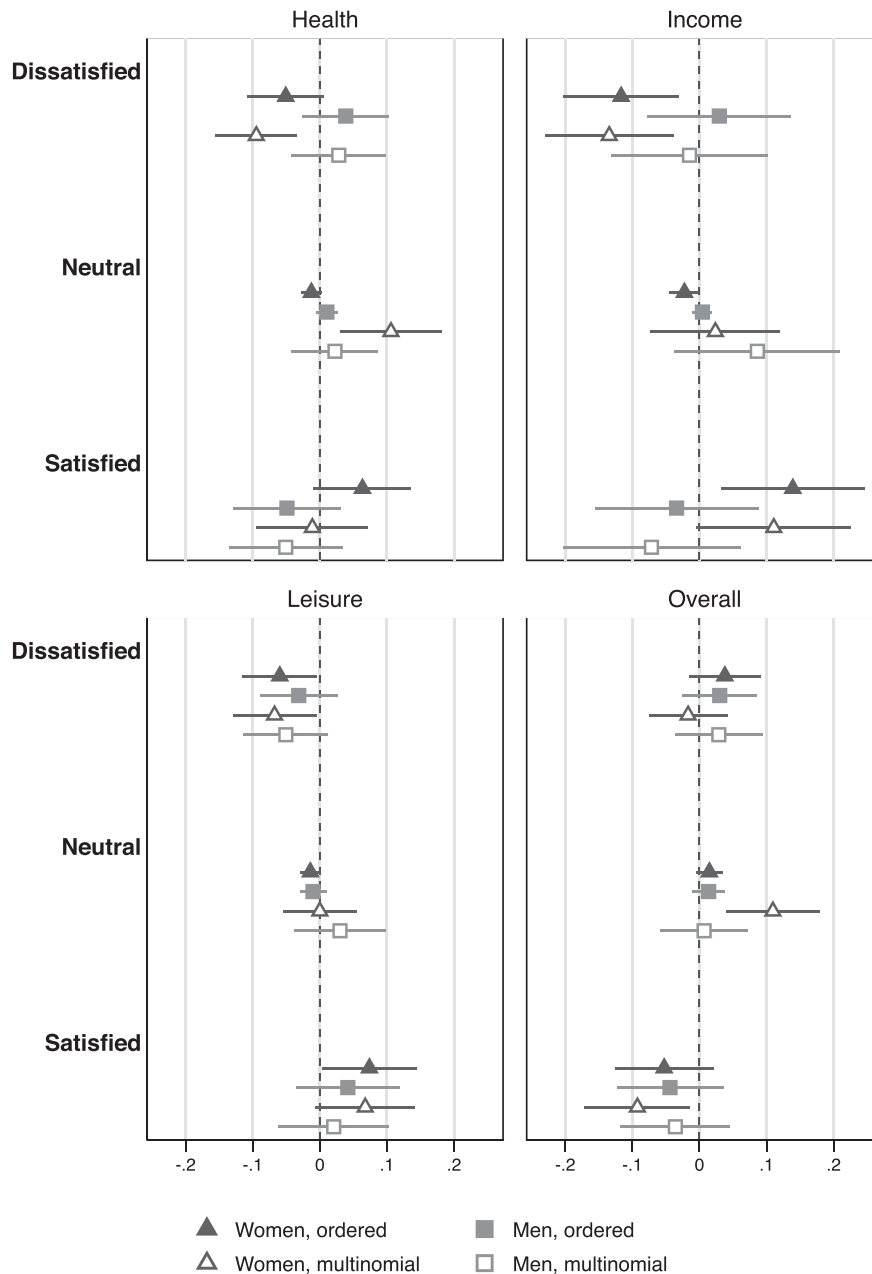


Fig. 1. MEMs from ordered and multinomial probit models (all outcomes). Note: IP5 data. 90% confidence intervals. Controls: high education and disability dummies. N(health): Women=948, Men=777; N(income): Women=433, Men=321 (respondents not in couples); N(leisure): Women=950, Men=773; N(overall): Women=951, Men=778.

14pp more likely to report income satisfaction (Table 3, second panel, column 5); treated women are 6pp less likely to report leisure dissatisfaction and 7pp more likely to report leisure satisfaction (Table 3, third panel, columns 1 and 5).¹² While imprecisely estimated, MEMs from ordered probit models suggest that the treatment may also increase women's health sat-

¹² The results for income are estimated by restricting the sample to people not in couples, and should not be generalized to people in couples. The SWB of people in couples and not in couples may react differently to well-being comparisons. For example, the "neighbors as negatives" result in Luttmer (2005) is driven by married respondents. If the results in Luttmer (2005) are driven by married people being more affected by social comparisons *in general*, then our results on income may be underestimated. If, instead, the results in Luttmer (2005) are driven by married people being more affected by comparisons *with neighbors*, for example as they are more settled than unmarried people, then Luttmer's results on the difference between married and unmarried people on the relevance of social comparison should not matter for our results where the reference group is defined by gender rather than place of residence.

isfaction: treated women are 5pp less likely than untreated women to report health dissatisfaction and 6pp more likely to report health satisfaction. MEMs for satisfaction with life overall are imprecisely estimated. Ordered probit models may not be the most appropriate models when respondents choose the middle answer as an alternative to non-response (box-ticking strategy). Multinomial probit models are better suited for situations like these and are also useful to detect cases where the treatment might have incentivized the emergence of box-ticking strategies.

We now look at the marginal effects at the mean from multinomial probit models estimated for women (see Table 4 and empty triangular markers in Fig. 1). The MEMs for income and leisure satisfaction from multinomial probit models are very similar to those from ordered probit models and indicate that the treatment increases women's income and leisure satisfaction. For income and leisure, the decrease in dissatisfaction due to the treatment translates into an increase in satisfaction: treated women are 13pp less likely than untreated women to report income dissatisfaction and 11pp more likely to report income satisfaction (Table 4, second panel, columns 1 and 5); treated women are 7pp less likely to report leisure dissatisfaction and 7pp more likely to report leisure satisfaction (Table 4, third panel, columns 1 and 5). In both domains, the probability of reporting a neutral answer is unaffected by the treatment (Table 4, second and third panels, column 3). The fact that the treatment leads to a decrease in dissatisfaction almost completely counterbalanced by an increase of satisfaction suggests that, for the case of income and leisure, the treatment leads to an increase in satisfaction rather than a to squeeze towards the middle answer.

For health and life overall, marginal effects at the mean from multinomial probit models do suggest that the treatment increases the probability of reporting a neutral answer for women (see the empty black triangular markers for the neutral outcomes in Fig. 4). The decrease in health satisfaction due to the treatment translates into an increase in the probability of choosing the neutral answer: treated women are 10pp less likely than untreated women to report health dissatisfaction (Table 4, top panel, column 1), 1pp more likely to report health satisfaction (Table 4, top panel, column 5), and 11pp more likely to report a neutral answer (Table 4, top panel, column 3). For life overall, the treatment increases the probability of reporting the neutral answer, at the expense of both the probability of reporting satisfaction and the probability of reporting dissatisfaction: treated women are 2pp less likely than untreated women to report life dissatisfaction (Table 4, bottom panel, column 1), 9pp less likely to report life satisfaction (Table 4, bottom panel, column 5), and 11pp more likely to report a neutral answer (Table 4, bottom panel, column 3). The emergence of box-ticking strategies in the case of health and life satisfactions is not surprising, as answering these questions is a complex process, requiring respondents to evaluate satisfaction with different domains and summarize this information into a single measure.¹³

Why do we find positive treatment effects for women? Section 3 suggests that positive treatment effects arise when women: i) think they are worse-off than men in the considered domain; ii) have some men in the unprompted reference group. Why are the positive treatment effects for women only found for income and leisure satisfaction? Income and leisure are domains where women are more likely to think they are worse-off than men (see discussion in Section 3): the treatment effect should be stronger in these domains.

We now look at the results for men. The marginal effects at the mean for men are shown by the square markers in Fig. 1 (empty for the ordered probit models and filled for the multinomial probit models). The treatment effects estimated for men are negligible and/or imprecisely estimated. Moreover, in the case of men, none of the MEMs estimated for the middle outcome via multinomial probit is statistically significant, suggesting no emergence of box-ticking strategy.

Why don't we find negative treatment effects for men, at least for domains where men are likely to think they are better-off than women, such as income? For income satisfaction, MEMs for men suggest negative treatment effects, but the estimated effects are small and/or imprecisely estimated. These weak results may indicate that men's default comparison group is already composed entirely or at least predominantly by other men and/or that people's satisfaction is not significantly inflated by being better-off than their reference group.

6. Income and leisure satisfaction: Heterogeneous effects

We now look at heterogeneous effects. We focus on women's income and leisure satisfaction, where the largest treatment effects were found. We estimate heterogeneous effects by proxies of the gender gap experienced by women. For income satisfaction, we estimate heterogeneous effects by the gender pay gap women experience at work; for leisure satisfaction, we estimate heterogeneous effects by the gender gap women experience in the allocation of housework within the couple.

To measure the gender pay gap experienced by women, we use data from the Office for National Statistics (ONS) on the UK level gender pay gap in the respondent's sector of employment (SIC 2007, mayor group 20 codes). Fig. 2 presents marginal effects at the mean from ordered probit models for income satisfaction for women not in couples. We present two specifications. Specification one (left panel) interacts the treatment indicator with an indicator of whether the gender pay gap in the respondent's sector of employment is above the sample median. Specification two (right panel) interacts the treatment indicator with a continuous measure of the gender pay gap in the respondent's sector of employment.

¹³ See Jylhä (2009); Greene et al. (2015); Chaparro et al. (2019) on the cognitive challenges posed by answering self-reported health questions and Benjamin et al. (2021) on the process of aggregating different domains when answering satisfaction questions.

Income

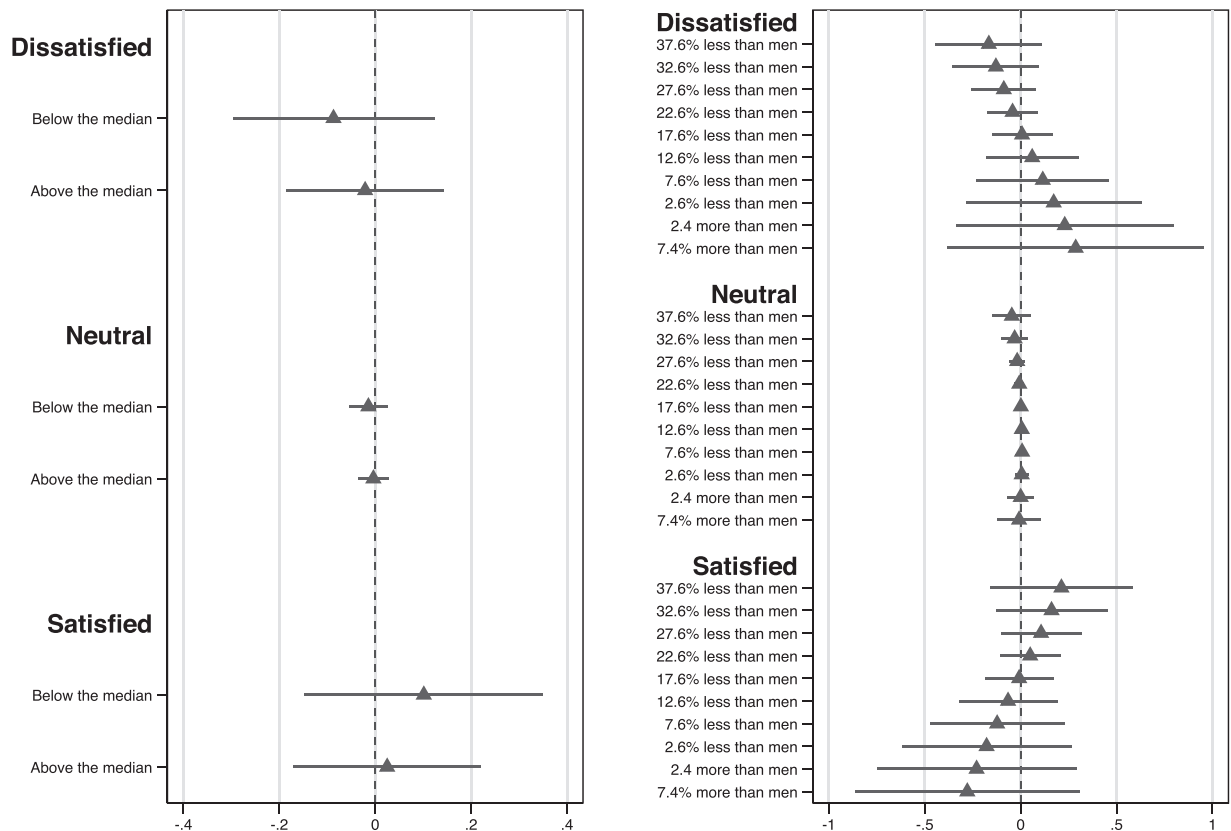


Fig. 2. MEMs from ordered probit models (income satisfaction). Treatment interacted with gender pay gap at the sector level. Note: IP5 data and ONS data on gender pay gap by sector for 2012 (SIC 2007, mayor group 20 codes), estimated using data from the Annual Survey of Hours and Earnings (ASHE). 90% confidence intervals. Controls: high education and disability dummies. N=224 (women not in couples).

Both panels of Fig. 2 suggest that women in sectors where women earn much less than men report higher income satisfaction when treated. For women in sectors where women earn the same or more than men, the treatment effect is zero or negative. Results (available on request) using data on the gender pay gap by occupation are similar.

To measure the gender gap in housework allocation, we use two measures only available for women in couples: i) the share of housework done by women over the total housework done by the couple; ii) self-reported data on women's standard of housework compared to their partner's: an indicator of the importance women attribute to housework (Auspurg et al., 2017).¹⁴ Both measures are likely to be positively correlated with the burden of housework borne by women and, thus, negatively correlated with the amount/quality of leisure women enjoy.

Fig. 3 presents MEMs from ordered probit models for leisure satisfaction. Due to the availability of data on housework allocation, we restrict the sample to women in couples. Specification one (left panel) interacts the treatment indicator with a continuous indicator of women's share of housework. MEMs are estimated at five points: all housework, 75%, 50%, 25% or none is done by the woman. Specification two (right panel) interacts the treatment indicator with indicators of whether the woman's standard for housework compared to the partner's is: much higher, higher, the same, lower or much lower.

Women who do more housework than their partners report higher leisure satisfaction when treated. The estimated treatment effect for women who do all or 75% of the housework is positive; it becomes zero or negative for women reporting a lower share of housework than their partners (Fig. 3, left panel). Equally, the treatment effect decreases with women's standard of housework; it becomes zero or negative for women reporting a lower standard of housework than their partners' (Fig. 3, right panel).

Our analysis of heterogeneous effects shows that the treatment effect increases with measures that proxy for the gender income gap (gender pay gap in the sector women belong to) and the gender leisure gap (share of housework in the couple). If women use their experience to infer the gender gap, our analysis is coherent with the hypothesis that the treatment

¹⁴ The question wording of the question is "How is your standard of housework compared with that of your spouse or partner?" The question is asked to all people living with a partner, when both members of the couple are eligible for the interview.

Leisure

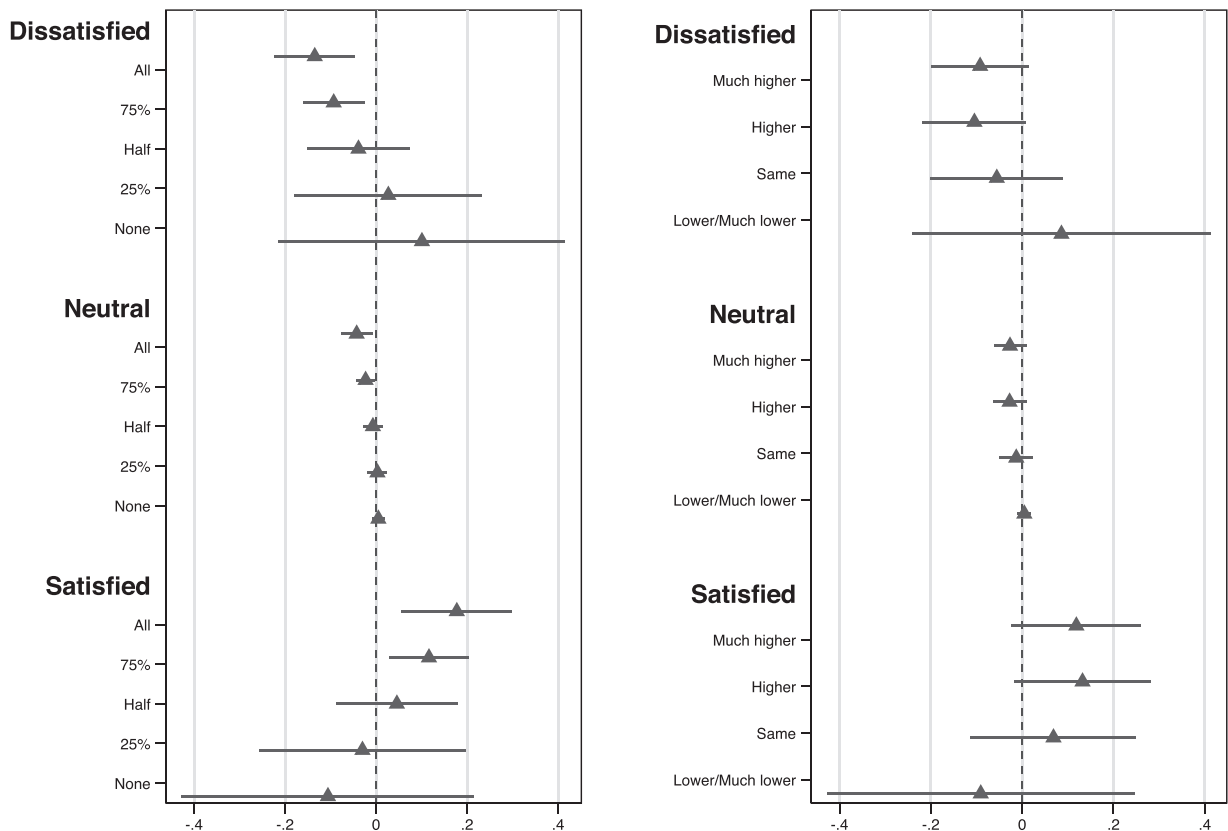


Fig. 3. MEMs from ordered probit models (leisure satisfaction). Treatment interacted with women's share and standard of housework compared to the partner's. Note: IP5 data. 90% confidence intervals. Left panel: treatment interacted with women's share of housework. N=589. Right panel: treatment interacted with women's standards of housework compared to the partner's. N=603. Sample: women in couples. Controls: high education and disability dummies.

effect increases with the perceived gender gap. However, this analysis is descriptive and does not permit to identify the effect of the perceived gender gap separately from the effect of the gender composition of the unprompted reference group, which we claim are the major drivers of the treatment effect. For example, if women experiencing high income/leisure gaps have a low share of men in their unprompted reference group, the effect of the perceived gender gap and the effect of the composition of the unprompted reference group may have opposite signs. Consider a woman working in health/social work: a sector with a high share of women, but also where women suffer a large gender pay gap.¹⁵ In this case, the perceived gender income gap is likely to be large, suggesting a large treatment effect. However, if the woman has mainly coworkers in her unprompted reference group, her unprompted reference group is likely to be already mainly composed by women. Therefore, the treatment is likely to have a small effect, despite the large gender income gap likely perceived by the women.

7. Tackling the attenuation effect problem

In this section we address the problem that the treatment may increase the probability that respondents select the neutral outcome as a socially acceptable alternative to non response and this may mask the 'real' treatment effect: the change in satisfaction induced by the prompted reference group. We perform two robustness checks. First, for all satisfaction domains, we estimate a model where the dependent variable is a dichotomous variable equal to one if the respondent reported satisfaction and equal to zero if the respondent reported dissatisfaction. Respondents who selected the neutral answer are excluded. The model is estimated via probit. Second, for health satisfaction only (that is the case for which we find the strongest evidence of a box-ticking strategy), we estimate the treatment effect using a Middle Inflated Ordered Probit (MIOP).

¹⁵ See the dot indicated with the letter Q in Fig. A.3, plotting the relationship between share of women in each employment sector and gender pay gap.

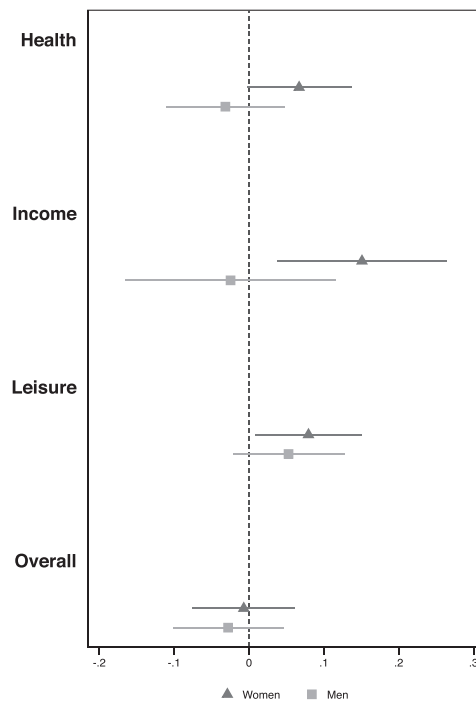


Fig. 4. MEMs from probit models excluding the neutral answers (all satisfaction domains). Note: IP5 data. 90% confidence intervals. Controls: high education and disability dummies. N(health): Women=814, Men=672; N(income): Women=347, Men=250 (respondents not in couples); N(leisure): Women=801, Men=643; N(overall): Women=807, Men=654.

Marginal effects at the mean from probit models that exclude the observations indicating a neutral answer are presented in Fig. 4 and Table A.3. For women, the estimated treatment effects on the probability of being 'satisfied' are slightly larger than those estimated through ordered probit models (in Table 3) and suggest that the treatment increases women's income, leisure and health satisfaction, although the estimated treatment effect for health satisfaction is at the margin of statistical significance. Treatment effects for men and life overall are small or null.

Removing neutral answers is particularly taxing in terms of sample size in the case of health satisfaction. For this case, the multinomial probit models in Section 5 suggest that the treatment substantially increases the probability that respondents report a neutral answer. As a consequence, excluding the observations reporting a neutral outcome means disregarding a lot of -potentially meaningful- information. Therefore, for the case of health satisfaction, we present an additional robustness check that uses the information contained in the neutral answers.

Following Bagozzi and Mukherjee (2012), we assume that respondents first decide between exerting effort to provide a genuine answer and selecting the neutral outcome as a socially acceptable alternative to non-response (spurious neutral answer). Respondents who decide to provide a genuine answer then decide whether to report dissatisfaction, genuine neutrality, or satisfaction. Therefore, the reported neutral answers are a combination of spurious and genuine neutral answers. We model this data generation process using a Middle Inflated Ordered Probit (MIOP): an ordered probit where the middle outcome is inflated.¹⁶ The parameters of the model are estimated via maximum likelihood. In theory, identification of the parameters of the model can be obtained via functional form only. In practice, identification is helped by exclusion restrictions: variables correlated with whether the respondent chooses against a spurious neutral answer, but uncorrelated with satisfaction.

Our first exclusion restriction (ER1) is a measure of the tendency of choosing spurious neutral answers: the share of neutral answers reported in other IP5 questions unrelated to the experiment and measured via Likert scale. ER1 is likely to be uncorrelated with health satisfaction, particularly in our case, as IP5 respondents are subject to multiple experiments: these experiments induce random variation in respondents' motivation and fatigue and thus in the likelihood of choosing a spurious neutral answer. Our second exclusion restriction (ER2) is an exogenous determinant of data quality: the receipt of a 'Norm' motivational letter. ER2 exploits the fact that respondents were sent different motivational letters (Auspurg et al., 2013). The 'Norm' letter was meant to motivate the respondents by informing them that most sample members responded

¹⁶ MIOP is similar to the zero-inflated ordered probit by Harris and Zhao (2007), but accommodates the cases where the middle outcome - rather than the lowest outcome - is inflated. Other applications of similar models are: Brooks et al. (2012); Greene et al. (2015).

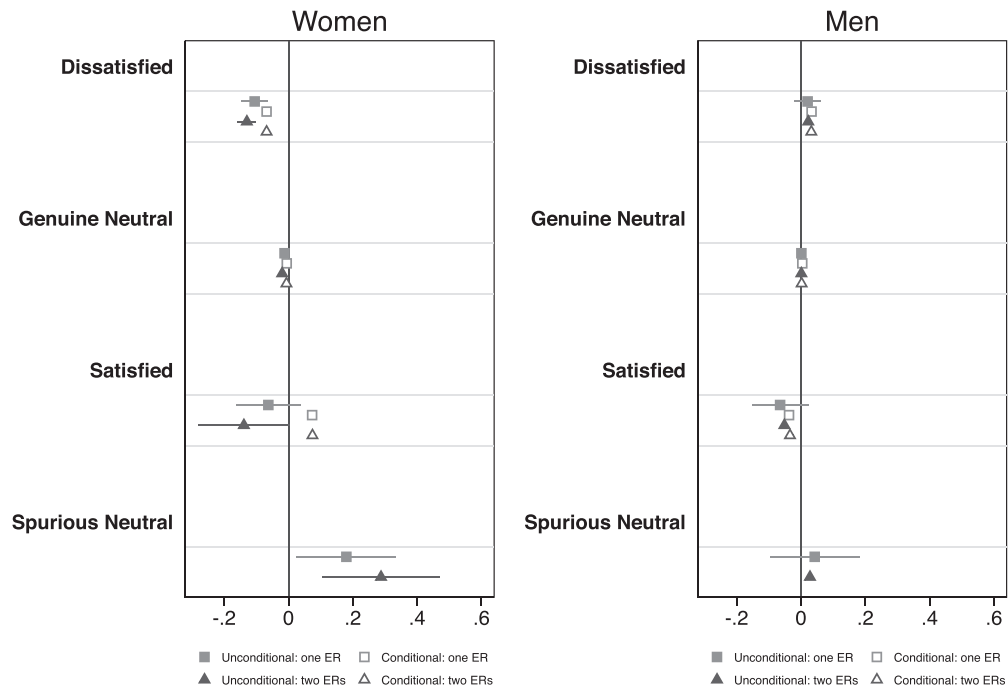


Fig. 5. MEMs from MIOP (health satisfaction). Note: IP5 data. 90% confidence intervals. Controls: high education and disability dummies. Exclusion restrictions: share of Neutral answers in questions measured via Likert scale in IP5; receipt of 'norm' letter (dummy). N(health): R1: Women=939, Men=773. R2: Women=938, Men=772.

in the previous wave of the survey: respondents' desire to comply with the norm should improve data quality. As letters were randomly allocated, ER2 should be uncorrelated with satisfaction.

We are primarily interested in how the treatment affects the probabilities of reporting genuine outcomes. For each genuine outcome we estimate conditional marginal effects at the mean (MEMs) of the treatment. Conditional MEMs are estimated as the difference between treated and untreated women in the predicted probability of choosing the outcome, conditional on choosing to report a genuine answer (dissatisfied, neutral, satisfied). Conditional MEMs capture the effect on satisfaction induced by the prompted reference group, once the attenuation effect due to the adoption of a box-ticking strategy has been removed.

For comparability with the results from the ordered probit and the multinomial probit models, we also compute unconditional MEMs as the difference between treatment and control women and men in the joint probability of choosing a genuine answer and choosing dissatisfaction, genuine neutrality, or satisfaction. These unconditional MEMs capture the total effect of the treatment on self-reported satisfaction: the combination of the effect of the prompted reference group on satisfaction and the effect of the prompted reference group on the probability of adopting a box-ticking strategy. We also present the marginal effect at the mean of the treatment on the probability of choosing a spurious neutral outcome.

Fig. 5 presents marginal effects at the mean from the MIOP.¹⁷ The MEM for the spurious neutral outcome is positive and sizable in the case of women (left graph, bottom panel), suggesting the treatment increases spurious neutral answers. The top three panels present MEMs for dissatisfaction, genuine neutrality, and satisfaction. The unconditional MEMs are similar to those from multinomial probit models (see Fig. 1) and different from the conditional MEMs. This suggests that the treatment may have increased the probability of disproportionately choosing the neutral outcome, and this may have masked the real treatment effect on health satisfaction. The conditional MEMs suggest that the treatment increases health satisfaction: treated women are 7pp less likely than untreated women to report health dissatisfaction and 7–8pp (depending on the model) more likely to report health satisfaction. MEMs for men are small.

8. Conclusions

We test the role of the gender composition of the reference group in reported SWB, measured as satisfaction across multiple domains. Using an RCT on a representative sample of British respondents, we prompt some respondents to evaluate

¹⁷ See also Table A.4. Standard errors are bootstrapped (50 repetitions). The coefficients of the MIOP show that the exclusion restrictions help identification. AIC and BIC suggest that the MIOP is preferable over ordered probit. Results are available upon request.

their satisfaction by comparing themselves with others of their own gender. We leave control respondents' reference group unprompted.

Women and men respond differently to the treatment. We find that women report higher satisfaction when prompted to compare themselves with women only, especially in domains (income and leisure) where women are likely to think they are worse-off than men. For income satisfaction, larger treatment effects are found for women experiencing larger gender pay gaps; for leisure satisfaction, larger treatment effects are found for women experiencing unfavorable allocation of housework. When we account for potential attenuation effects due to the adoption of a box-ticking strategy, we find that the treatment also increases women's health satisfaction. We find no effects for men. At least for domains where men think they are better off than women (e.g., income), this can be interpreted as evidence that, in the absence of treatment, interpersonal comparisons are upward and not downward. In our case, this could be due to the fact that women compare themselves with (at least some) men, but men do not generally compare themselves with women.

The paper contributes to the literature that identifies and analyzes differences in women's and men's outcomes. Our paper suggests a new form of gender heterogeneity: women modify their self-reported satisfaction when prompted to compare themselves with women only; men do not modify theirs when prompted to compare themselves with men only. This form of heterogeneity has never been discussed in the literature and this paper is a first step to fill that gap. More research is needed to identify the reasons of this difference. For example, we still know very little about how much importance people assign to their gender when defining their reference group, and if this importance relates to one's gender identity and experience of gender segregation in social environments (for example on the job). We also do not know how competition within and across genders affects SWB, and how people perceive the gender gap in various domains. Surveys like the Understanding Society Innovation Panel are a powerful tool to answer these questions and to design new experiments, for example to see what happens if men are asked to compare themselves only to women, and women to compare themselves only to men.

Our paper also suggests that prompting a specific reference group may, in some cases, lead to an increase in measurement error. A promising literature (e.g., Benjamin et al., 2021) has started comparing different SWB questions in a series of aspects, including how cognitively demanding these questions are. New evidence is needed to identify the consequences of cognitive burden on SWB, and how researchers can deal with these consequences in the analysis of SWB data. This would make it possible to analyze potentially very interesting, but arguably cognitively demanding questions, like what happens to SWB when people are asked to compare themselves with people of another - rather than their own - gender.

Finally, this paper contributes to the literature on SWB and particularly to the debate on why women's SWB does not seem to have followed the improvements in the women's condition. Specifically, this paper is the first study that adds experimental evidence to the claim that this can be due to women comparing themselves to men. It also indirectly suggests women consider themselves more disadvantaged than men in multiple domains.

Declaration of competing interest

None.

Appendix A. Additional figures and tables

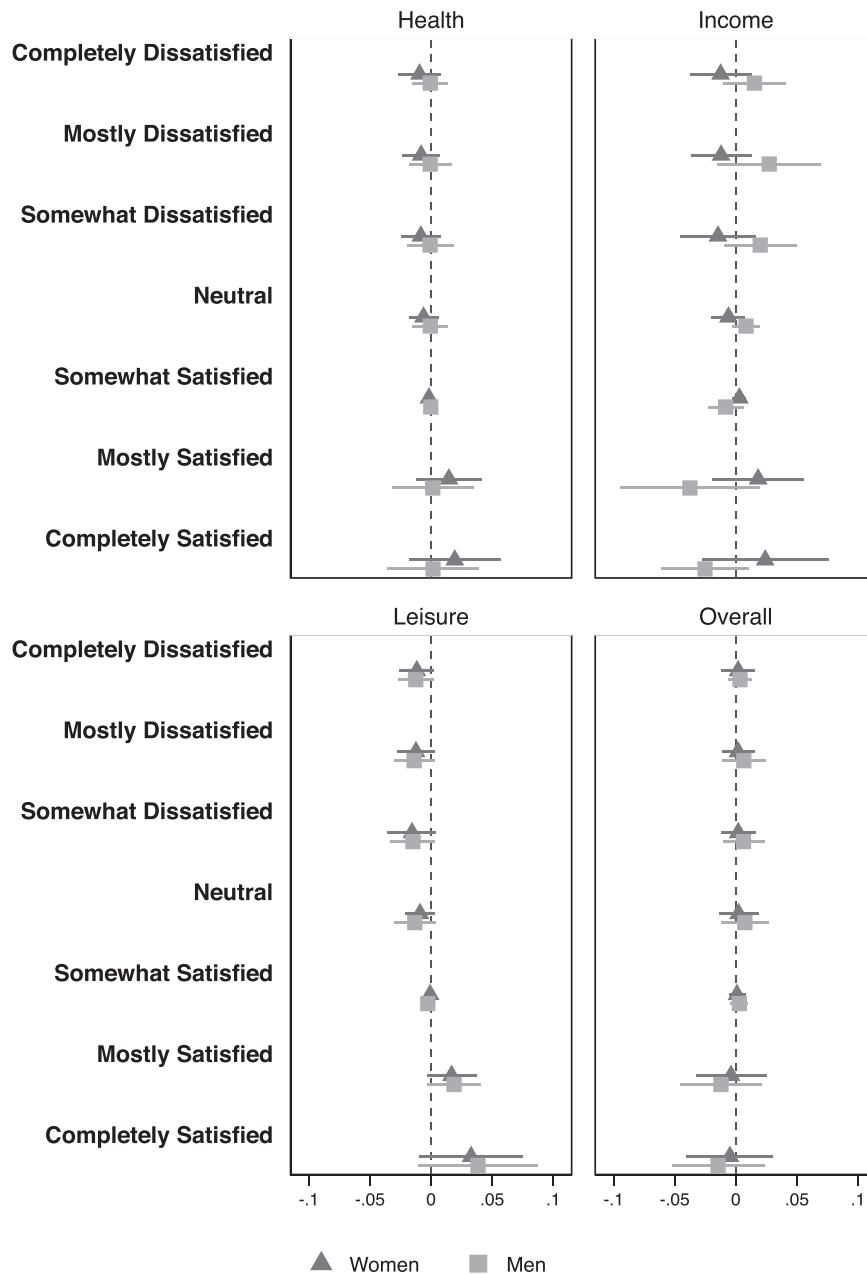


Fig. A.1. MEMs from ordered probit (all outcomes, 7 points Likert scale). Note: IP5 data. 90% confidence intervals. Controls: high education and disability dummies. N(health): Women=948, Men=777; N(income): Women=433, Men=321 (respondents not in couples); N(leisure): Women=950, Men=773; N(overall): Women=951, Men=778.

Table A.1
Ordered probit: MEMs (all outcomes, 7points Likert scale).

	<i>Health: Completely Dissatisfied</i>		<i>Health: Mostly Dissatisfied</i>		<i>Health: Somehow Dissatisfied</i>		<i>Health: Neutral</i>		<i>Health: Somehow Satisfied</i>		<i>Health: Mostly Satisfied</i>		<i>Health: Completely Satisfied</i>	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	Females	Males	Females	Males	Females	Males	Females	Males	Females	Males	Females	Males	Females	Males
Treatment	-0.009 (0.011)	-0.001 (0.009)	-0.008 (0.009)	-0.001 (0.011)	-0.008 (0.010)	-0.001 (0.011)	-0.006 (0.007)	-0.001 (0.009)	-0.002 (0.002)	-0.000 (0.003)	0.015 (0.016)	0.001 (0.020)	0.019 (0.023)	0.002 (0.023)
	<i>Income: Completely Dissatisfied</i>		<i>Income: Mostly Dissatisfied</i>		<i>Income:Somewhat dissatisfied</i>		<i>Income: Neutral</i>		<i>Income: Somewhat Satisfied</i>		<i>Income: Mostly Satisfied</i>		<i>Income: Completely Satisfied</i>	
Treatment	Females	Males	Females	Males	Females	Males	Females	Males	Females	Males	Females	Males	Females	Males
	-0.012 (0.015)	0.015 (0.015)	-0.012 (0.015)	0.027 (0.026)	-0.015 (0.019)	0.020 (0.018)	-0.006 (0.008)	0.008 (0.007)	0.003 (0.003)	-0.008 (0.009)	0.018 (0.023)	-0.038 (0.035)	0.024 (0.032)	-0.025 (0.022)
	<i>Leisure: Completely Dissatisfied</i>		<i>Leisure: Mostly Dissatisfied</i>		<i>Leisure:Somewhat dissatisfied</i>		<i>Leisure: Neutral</i>		<i>Leisure: Somewhat Satisfied</i>		<i>Leisure: Mostly Satisfied</i>		<i>Leisure: Completely Satisfied</i>	
Treatment	Females	Males	Females	Males	Females	Males	Females	Males	Females	Males	Females	Males	Females	Males
	-0.012 (0.009)	-0.013 (0.009)	-0.012 (0.009)	-0.014 (0.010)	-0.016 (0.012)	-0.015 (0.011)	-0.009 (0.007)	-0.013 (0.010)	-0.001 (0.002)	-0.003 (0.003)	0.017 (0.012)	0.019 (0.013)	0.033 (0.026)	0.038 (0.030)
	<i>Overall: Completely Dissatisfied</i>		<i>Overall: Mostly Dissatisfied</i>		<i>Overall:Somewhat dissatisfied</i>		<i>Overall: Neutral</i>		<i>Overall: Somewhat Satisfied</i>		<i>Overall: Mostly Satisfied</i>		<i>Overall: Completely Satisfied</i>	
Treatment	Females	Males	Females	Males	Females	Males	Females	Males	Females	Males	Females	Males	Females	Males
	0.002 (0.008)	0.003 (0.006)	0.002 (0.008)	0.007 (0.011)	0.002 (0.009)	0.006 (0.010)	0.002 (0.010)	0.007 (0.012)	0.001 (0.004)	0.003 (0.004)	-0.004 (0.018)	-0.012 (0.020)	-0.005 (0.021)	-0.014 (0.023)

Note: IP5 data. Controls: high education and disability dummies. N(health): Women=948, Men=777; N(income): Women=433, Men=321 (respondents not in couples); N(leisure): Women=950, Men=773; N(overall): Women=951, Men=778.

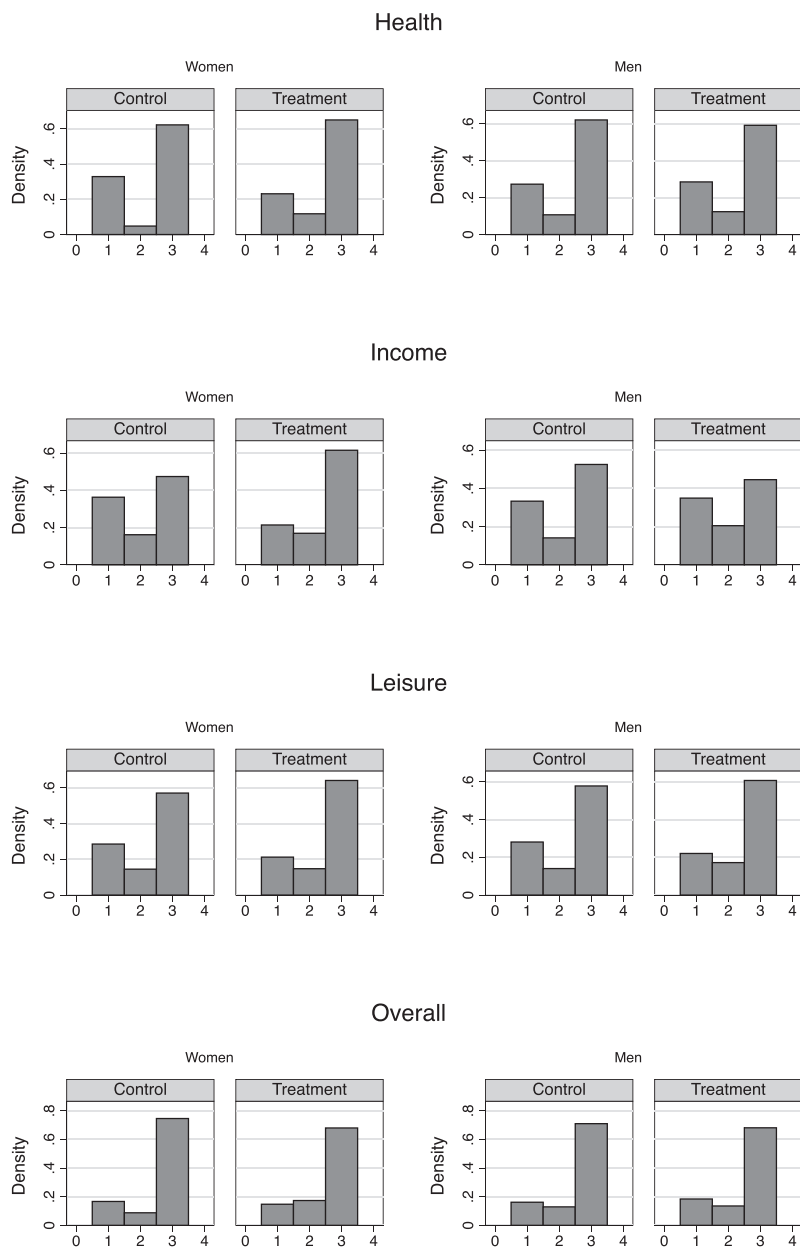


Fig. A.2. Health, income, leisure and life overall satisfaction: distributions by gender and treatment group Note: IP5 data. N(Health, Women): Control=228, Treatment=229; N(Health, Men): Control=187, Treatment=186; N(Income, Women): Control= 99 Treatment=112; N(Income, Men): Control=78, Treatment=83; N(Leisure, Women): Control=228, Treatment=231; N(Leisure, Men): Control=185, Treatment=186; N(Overall, Women): Control=228, Treatment=231; N(Overall, Men): Control=186, Treatment=185. The sample for income is only composed by respondents not in couples.

Table A.2

T-tests.

	(1)	(2)
	Females	Males
<i>Health: Dissatisfied</i>	-0.098 (0.042)	0.012 (0.047)
<i>Health: Neutral</i>	0.070 (0.026)	0.017 (0.033)
<i>Health: Satisfied</i>	0.028 (0.045)	-0.029 (0.051)
<i>N</i>	457	373
<i>Income: Dissatisfied</i>	-0.149 (0.062)	0.016 (0.075)
<i>Income: Neutral</i>	0.008 (0.052)	0.064 (0.060)
<i>Income: Satisfied</i>	0.141 (0.068)	-0.080 (0.079)
<i>N</i>	211	161
<i>Leisure: Dissatisfied</i>	-0.073 (0.040)	-0.061 (0.045)
<i>Leisure: Neutral</i>	0.002 (0.033)	0.032 (0.038)
<i>Leisure: Satisfied</i>	0.071 (0.046)	0.029 (0.051)
<i>N</i>	459	371
<i>Overall: Dissatisfied</i>	-0.019 (0.034)	0.022 (0.039)
<i>Overall: Neutral</i>	0.085 (0.031)	0.006 (0.035)
<i>Overall: Satisfied</i>	-0.066 (0.042)	-0.029 (0.048)
<i>N</i>	459	371

Note: IP5 data. The sample for income satisfaction is restricted to respondents not in couples. Differences are computed as treatment group-control group. Stars for significance levels are not reported.

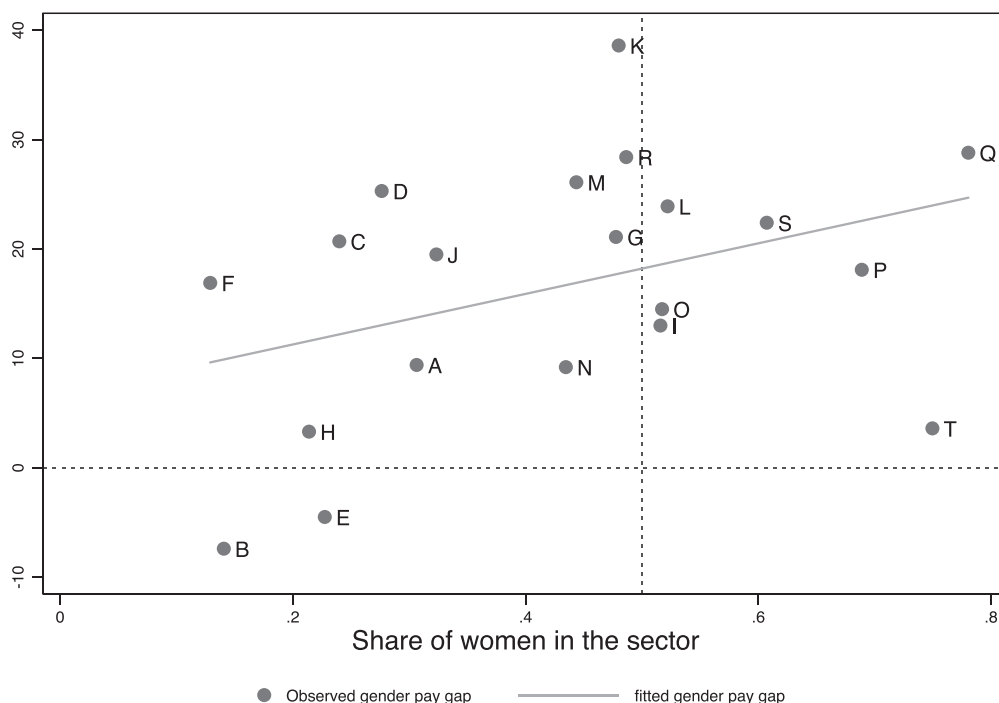


Fig. A.3. Occupational segregation and gender pay gap Note: Data: The observed gender pay gap is the gender pay gap in the respondent's sector of employment (SIC) in the UK. The share of women is share of women in the respondent's sector of employment (SIC) in the UK. Data are from the Office for National Statistics (ONS). Dotted horizontal and vertical lines represent, respectively, the null gender pay gap and the case where the share of women is equal to 50%. The fitted gender pay gap has been computed by regressing the observed pay gap on the share of women in each sector. The value of the estimated coefficient is 23.11 (p-value=0.096). The correlation between the share of women and the gender pay gap in the sector is 0.383. The sectors presented in the graph are the following: A : Agriculture, forestry and fishing; B : Mining and quarrying; C : Manufacturing; D : Electricity, gas, steam and air conditioning supply; E : Water supply; sewerage, waste management and remediation activities; F : Construction; G : Wholesale and retail trade; repair of motor vehicles and motorcycles; H : Transportation and storage; I : Accommodation and food service activities; J : Information and communication; K : Financial and insurance activities; L : Real estate activities M : Professional, scientific and technical activities; N : Administrative and support service activities; O : Public administration and defence; compulsory social security; P : Education; Q : Human health and social work activities; R : Arts, entertainment and recreation; S : Other service activities; T : Activities of households as employers; undifferentiated goods-and services-producing activities of households for own use.

Table A.3
Probit models excluding the neutral answer: MEMs.

	<i>Health: Satisfied</i>	
	(1)	(2)
	Females	Males
Treatment	0.067 (0.043)	-0.032 (0.048)
<i>Income: Satisfied</i>		
	Females	Males
Treatment	0.151 (0.069)	-0.025 (0.085)
<i>Leisure: Satisfied</i>		
	Females	Males
Treatment	0.079 (0.043)	0.053 (0.045)
<i>Overall: Satisfied</i>		
	Females	Males
Treatment	-0.007 (0.042)	-0.028 (0.045)

Note: IP5 data. Controls: high education and disability dummies. N(health): Women=814, Men=672; N(income): Women=347, Men=250 (respondents not in couples); N(leisure): Women=801, Men=643; N(overall): Women=807, Men=654. Stars for significance levels are not reported.

Table A.4
MIOP: MEMs (health satisfaction).

	<i>One ER</i>		<i>Two ERs</i>	
	(1)	(2)	(3)	(4)
	Females	Males	Females	Males
Unconditional Dissatisfied	-0.105 (0.025)	0.021 (0.025)	-0.129 (0.018)	0.023 (0.004)
Unconditional Genuine Neutral	-0.012 (0.008)	0.001 (0.005)	-0.020 (0.007)	0.001 (0.000)
Unconditional Satisfied	-0.063 (0.061)	-0.065 (0.054)	-0.139 (0.087)	-0.052 (0.003)
Conditional Dissatisfied	-0.068 (0.007)	0.033 (0.005)	-0.068 (0.003)	0.032 (0.005)
Conditional Genuine Neutral	-0.005 (0.001)	0.004 (0.000)	-0.006 (0.001)	0.002 (0.000)
Conditional Satisfied	0.074 (0.008)	-0.037 (0.005)	0.075 (0.002)	-0.034 (0.006)
Spurious Neutral	0.180 (0.094)	0.043 (0.085)	0.288 (0.111)	0.028 (0.002)
Total Neutral	0.168 (0.086)	0.044 (0.079)	0.268 (0.105)	0.029 (0.002)
N	939	773	938	772

Note: IP5 data. Controls: high education and disability dummies. Exclusion restrictions: share of neutral answers in other questions measured via Likert scale in IP5 (ER1). Dummy variable indicating the random receipt of a 'Norm' motivational letter (ER2). For each genuine outcome, unconditional MEMs are computed using the joint probabilities of deciding against a spurious neutral answer and choosing the considered outcome; conditional MEMs are computed using the probabilities of choosing the considered outcome conditional on choosing against a spurious answer. MEMs for the spurious neutral answer are defined as the difference in the probabilities of choosing against a neutral answer between treatment and control.

Appendix B. Model

In this Appendix we formalize the relationship between perceived gender gap, reference group and treatment effect described in Section 3. Assume that individuals are characterized by gender (men and women) and income (low and high). high-income women (men) are women (men) whose income is above the average income for women (men). Consider first a high-income woman and assume also that i perceives that women have lower income than men.

Consider income satisfaction in absence of the treatment (*unprompted* income satisfaction). Let i be a generic high-income woman, let x_i be i 's income and let s_i be the satisfaction i derives from x_i . Let $N = \{1, 2, 3, \dots, n\}$ be a set of agents organized in a network \mathbf{g}^U , with i and $j \in N$. The network \mathbf{g}^U (*unprompted* network) is the network defining the comparison relationships in absence of the treatment. \mathbf{g}^U can be represented by the adjacency matrix \mathbf{G}^U with generic entry g_{ij}^U . If $g_{ij}^U = 1$, i compares herself with j ; if $g_{ij}^U = 0$, i does not. $g_{ii}^U = 0$. Let $N_i^U = \{j \in N : g_{ij}^U = 1\}$ be i 's *unprompted* reference group: the set of nodes j i compares herself with in absence of the treatment, i.e., when i answers the control question. Notice that i may or may not know exactly the income of every j in her reference group.

i 's income satisfaction in absence of the treatment (*unprompted* income satisfaction) can be written as:

$$s_i^U = u_i(x_i) + v_i[x_i - \sum_{j \in N} g_{ij}^U \hat{x}_j] + \theta_i \quad (1)$$

where $u_i(\cdot)$ and $v_i(\cdot)$ are increasing concave functions. Further, $v_i(0) = 0$ and $v_i(\cdot) > 0$ if $x_i - \sum_{j \in N} g_{ij}^U \hat{x}_j > 0$, $v_i(\cdot) < 0$ if $x_i - \sum_{j \in N} g_{ij}^U \hat{x}_j < 0$.¹⁸ $u_i(x_i)$ is i 's income satisfaction x_i , when i does not compare herself with j ('non-comparative component'), and $v_i[x_i - \sum_{j \in N} g_{ij}^U \hat{x}_j]$ is the 'comparative component'. The comparative component depends on $\sum_{j \in N} g_{ij}^U \hat{x}_j$: the average income in i 's reference group N_i^U , where \hat{x}_j is the income of individual j and $\bar{g}_{ij}^U = \frac{g_{ij}^U}{\sum_{j=1}^N g_{ij}^U}$.¹⁹ Since i does not necessarily know the income of everyone in \mathbf{g}^U , the incomes \hat{x}_j are just perceived (hence the hat symbol). Finally, θ_i is an individual specific component capturing individual level heterogeneity.

The treatment modifies i 's reference group by including both low and high-income women.²⁰ We call \mathbf{g}^P is the *prompted* network: the network induced by the treatment. The treatment makes women compare themselves with other women irrespective of whether these women were in \mathbf{g}^U . The *prompted* reference group is $N_i^P = \{j \in N : g_{ij}^P = 1\}$, where $g_{ij}^P = 1$ if j is a woman, $g_{ij}^P = 0$ otherwise. Δs depends on N_i^U as follows:

1. $N_i^P = N_i^U$. N_i^U contains high and low-income women. We call this *same gender unprompted reference group*.

$$\Delta s_i = v_i[x_i - \sum_{j \in N} \bar{g}_{ij}^P \hat{x}_j] - v_i[x_i - \sum_{j \in N} \bar{g}_{ij}^U \hat{x}_j] = 0$$

i compares herself with low and high-income women both before and after the treatment: the treatment does not change her reference group. The treatment effect is null.

2. $N_i^P \neq N_i^U$ and N_i^P is not a subset/superset of N_i^U . N_i^U contains high-income women and men. We call this *same income unprompted reference group*.

$$\Delta s_i = v_i[x_i - \sum_{j \in N} \bar{g}_{ij}^P \hat{x}_j] - v_i[x_i - \sum_{j \in N} \bar{g}_{ij}^U \hat{x}_j] > 0$$

$x_i - \sum_{j \in N} \bar{g}_{ij}^P \hat{x}_j > 0$, because i is a high-income woman and thus her income is higher than the average income in the low and high-income women subsample. In absence of the treatment i compares herself with high-income women and men. The treatment makes her consider low-income women and disregard high-income men. Therefore, $x_i - \sum_{j \in N} \bar{g}_{ij}^U \hat{x}_j < x_i - \sum_{j \in N} \bar{g}_{ij}^P \hat{x}_j$: the treatment effect is positive.

3. $N_i^P \neq N_i^U$ and $N_i^P \supset N_i^U$. N_i^U contains high-income women only. We call this *same gender-income unprompted reference group*.

$$\Delta s_i = v_i[x_i - \sum_{j \in N} \bar{g}_{ij}^P \hat{x}_j] - v_i[x_i - \sum_{j \in N} \bar{g}_{ij}^U \hat{x}_j] > 0$$

$x_i - \sum_{j \in N} \bar{g}_{ij}^P \hat{x}_j > 0$, because i is a high-income woman and thus her income is higher than the average income in the low and high-income women subsample. In absence of the treatment i compares herself with high-income women only. The treatment makes her consider low-income women in addition to high-income ones. Therefore, $x_i - \sum_{j \in N} \bar{g}_{ij}^U \hat{x}_j < x_i - \sum_{j \in N} \bar{g}_{ij}^P \hat{x}_j$: the treatment effect is positive.

¹⁸ See Clark and D'Ambrosio (2015) for a survey of properties of well-being functions. Card et al. (2012) uses a similar function.

¹⁹ $\sum_{j \in N} \bar{g}_{ij}^U \hat{x}_j$ is the average income in i 's reference group because it is the sum of the j incomes \hat{x}_j weighted by \bar{g}_{ij}^U , the entries of the adjacency matrix $\tilde{\mathbf{G}}^U$. Notice that $\tilde{\mathbf{G}}^U$ is the row normalized version of \mathbf{G}^U , i.e. the matrix where the entries of row i of \mathbf{G}^U are divided by the sum of the elements of row i .

²⁰ Treated high-income women may consider also high-income women only. Results assuming it is the case are available upon request.

4. $N_i^P \neq N_i^U$ and $N_i^P \subset N_i^U$. N_i^U contains low and high-income women and men. We call this *unrestricted unprompted reference group*.

$$\Delta s_i = v_i[x_i - \sum_{j \in N} \tilde{g}_{ij}^P \hat{x}_j] - v_i[x_i - \sum_{j \in N} \tilde{g}_{ij}^U \hat{x}_j] > 0$$

$x_i - \sum_{j \in N} \tilde{g}_{ij}^P \hat{x}_j > 0$, because i is a high-income woman and thus her income is higher than the average income in the low and high-income women subsample. In absence of the treatment i compares herself with low and high-income women and men. The treatment makes her disregard all men. Therefore, if i perceives that women have lower income than men, $x_i - \sum_{j \in N} \tilde{g}_{ij}^U \hat{x}_j < x_i - \sum_{j \in N} \tilde{g}_{ij}^P \hat{x}_j$: the treatment effect is positive.

Consider the case where i is a low-income woman exposed to the treatment. Δs_i depends on N_i^U as follows:

5. $N_i^P = N_i^U$. N_i^U contains high and low-income women. We call this *same gender unprompted reference group*.

$$\Delta s_i = v_i[x_i - \sum_{j \in N} \tilde{g}_{ij}^P \hat{x}_j] - v_i[x_i - \sum_{j \in N} \tilde{g}_{ij}^U \hat{x}_j] = 0$$

i compares herself with low and high-income women both in the unprompted and in the prompted case: the treatment does not change her reference group. The treatment effect is null.

6. $N_i^P \neq N_i^U$ and N_i^P is not a subset/superset of N_i^U . N_i^U contains low-income women and men. We call this *same income unprompted reference group*.

$$\Delta s_i = v_i[x_i - \sum_{j \in N} \tilde{g}_{ij}^P \hat{x}_j] - v_i[x_i - \sum_{j \in N} \tilde{g}_{ij}^U \hat{x}_j] = ?$$

$x_i - \sum_{j \in N} \tilde{g}_{ij}^P \hat{x}_j < 0$, because i is a low-income woman and thus her income is lower than the average income in the low and high-income women subsample. In absence of the treatment i compares herself with low-income women and men. The treatment makes her consider high-income women and disregard low-income men. If i perceives that women have lower income than men, it is not *a priori* clear whether $x_i - \sum_{j \in N} \tilde{g}_{ij}^P \hat{x}_j$ is greater or less than $x_i - \sum_{j \in N} \tilde{g}_{ij}^U \hat{x}_j$: the treatment effect is undetermined.

7. $N_i^P \neq N_i^U$ and $N_i^P \supset N_i^U$. N_i^U contains low-income women only. We call this *same gender-income unprompted reference group*.

$$\Delta s_i = v_i[x_i - \sum_{j \in N} \tilde{g}_{ij}^P \hat{x}_j] - v_i[x_i - \sum_{j \in N} \tilde{g}_{ij}^A \hat{x}_j] < 0$$

$x_i - \sum_{j \in N} \tilde{g}_{ij}^P \hat{x}_j < 0$, because i is a low-income woman and thus her income is lower than the average income in the low and high-income women subsample. In absence of the treatment i compares herself with low-income women only. The treatment makes her consider high-income women in addition to low-income ones. Therefore, $x_i - \sum_{j \in N} \tilde{g}_{ij}^U \hat{x}_j > x_i - \sum_{j \in N} \tilde{g}_{ij}^P \hat{x}_j$: the treatment effect is negative.

8. $N_i^P \neq N_i^U$ and $N_i^P \subset N_i^U$. N_i^U contains low and high-income women and men. We call this *unrestricted unprompted reference group*.

$$\Delta s_i = v_i[x_i - \sum_{j \in N} \tilde{g}_{ij}^P \hat{x}_j] - v_i[x_i - \sum_{j \in N} \tilde{g}_{ij}^U \hat{x}_j] > 0$$

$x_i - \sum_{j \in N} \tilde{g}_{ij}^P \hat{x}_j < 0$, because i is a low-income woman and thus her income is lower than the average income in the low and high-income women subsample. In absence of the treatment i compares herself with low and high-income women and men. The treatment makes her disregard men. If i perceives that women have lower income than men, $x_i - \sum_{j \in N} \tilde{g}_{ij}^U \hat{x}_j > x_i - \sum_{j \in N} \tilde{g}_{ij}^P \hat{x}_j$: the treatment effect is positive.

Treatment effects for men are mirrored (see Table 1).

References

- Akay, A., Bargain, O., Zimmermann, K.F., 2012. Relative concerns of rural-to-urban migrants in china. *J. Econ. Behav. Org.* 81 (2), 421–441.
- Amin, A., Remme, M., Allotey, P., Askew, I., 2021. Gender equality by 2045: reimagining a healthier future for women and girls. *BMJ* 373. doi:10.1136/bmj.n1621. <https://www.bmj.com/content/373/bmj.n1621.full.pdf>
- Auspurg, K., Burton, J., Cullinane, C., Delavande, A., Fumagalli, L., Iacovou, M., Jackle, A., Kaminska, O., Lynn, P., Mathews, P., Nicolaas, G., Nicoletti, C., Ye, C., Zafar, B., 2013. Understanding Society Innovation Panel Wave 5: results from methodological experiments. Understanding Society Working Paper Series. Understanding Society at the Institute for Social and Economic Research. <https://ideas.repec.org/p/ese/ukhslp/2013-06.html>
- Auspurg, K., Iacovou, M., Nicoletti, C., 2017. Housework share between partners: experimental evidence on gender-specific preferences. *Soc. Sci. Res.* 66, 118–139.
- Bagozzi, B.E., Mukherjee, B., 2012. A mixture model for middle category inflation in ordered survey responses. *Politic. Anal.* 20 (3), 369–386.
- Bartley, E.J., Fillingim, R.B., 2013. Sex differences in pain: a brief review of clinical and experimental findings. *Br. J. Anaesth.* 111 (1), 52–58.
- Benjamin, D.J., Guzman, J.D., Fleurbaey, M., Heffetz, O., Kimball, M.S., 2021. What Do Happiness Data Mean? Theory and Survey Evidence. Working Paper. National Bureau of Economic Research doi:10.3386/w28438. <http://www.nber.org/papers/w28438>
- Bertrand, M., 2011. Chapter 17 - New Perspectives on Gender. In: Card, D., Ashenfelter, O. (Eds.), *Handbook of Labor Economics*, Vol. 4. Elsevier, pp. 1543–1590. doi:10.1016/S0169-7218(11)02415-4. <http://www.sciencedirect.com/science/article/pii/S0169721811024154>

- Bertrand, M., 2020. Gender in the twenty-first century. In: *AEA Papers and Proceedings*, Vol. 110, pp. 1–24.
- Blanchflower, D., Oswald, A., 2004. Well-being over time in Britain and the USA. *J. Public Econ.* 88, 1359–1386.
- Blau, F.D., Kahn, L.M., 2017. The gender wage gap: extent, trends, and explanations. *J. Econ. Lit.* 55 (3), 789–865.
- Booth, A., Cardona-Sosa, L., Nolen, P., 2014. Gender differences in risk aversion: do single-sex environments affect their development? *J. Econ. Behav. Org.* 99, 126–154.
- Booth, A., Nolen, P., 2012. Choosing to compete: how different are girls and boys? *J. Econ. Behav. Org.* 81 (2), 542–555.
- Booth, A.L., Nolen, P., 2012. Gender differences in risk behaviour: does nurture matter? *Econ. J.* 122 (558), F56–F78. doi:10.1111/j.1468-0297.2011.02480.x. <https://academic.oup.com/ej/article-pdf/122/558/F56/26412463/ej0156.pdf>
- Boyce, C.J., Brown, G.D., Moore, S.C., 2010. Money and happiness: rank of income, not income, affects life satisfaction. *Psychol. Sci.* 21 (4), 471–475. doi:10.1177/0956797610362671. PMID: 20424085
- Brodeur, A., Fleche, S., 2019. Neighbors' Income, public goods, and well-being. *Rev. Income Wealth* 65 (2), 217–238. doi:10.1111/roiw.12367. <https://onlinelibrary.wiley.com/doi/pdf/10.1111/roiw.12367>
- Brooks, R., Harris, M.N., Spencer, C., 2012. Inflated ordered outcomes. *Econ. Lett.* 117 (3), 683–686.
- Brown, S., Gray, D., 2016. Household finances and well-being in Australia: an empirical analysis of comparison effects. *J. Econ. Psychol.* 53, 17–36. doi:10.1016/j.joep.2015.12.006. <http://www.sciencedirect.com/science/article/pii/S0167487015001488>
- Brown, S., Gray, D., Roberts, J., 2015. The relative income hypothesis: a comparison of methods. *Econ. Lett.* 130, 47–50. doi:10.1016/j.econlet.2015.02.031. <http://www.sciencedirect.com/science/article/pii/S0165176515000919>
- Burton, J., Laurie, H., Uhrig, S.N., Bryan, M.L., Desousa, C., Fumagalli, L., Jackle, A., Knies, G., Lynn, P., Nandi, A., Platt, L., Pudney, S., 2008. Understanding Society. Some preliminary results from the Wave 1 Innovation Panel. Understanding Society Working Paper Series. Understanding Society at the Institute for Social and Economic Research. <https://ideas.repec.org/p/ese/ukhisp/2008-03.html>
- Buser, T., Yuan, H., 2019. Do women give up competing more easily? evidence from the lab and the Dutch math olympiad. *Am. Econ. J.: Appl. Econ.* 11 (3), 225–252.
- Cai, X., Lu, Y., Pan, J., Zhong, S., 2019. Gender gap under pressure: evidence from China's national college entrance examination. *Rev. Econ. Stat.* 101 (2), 249–263.
- Caporale, G.M., Georgellis, Y., Tsitsianis, N., Yin, Y.P., 2009. Income and happiness across Europe: do reference values matter? *J. Econ. Psychol.* 30 (1), 42–51. doi:10.1016/j.joep.2008.06.004. <http://www.sciencedirect.com/science/article/pii/S0167487008000809>
- Ferrer-i Carbonell, A., 2005. Income and well-being: an empirical analysis of the comparison income effect. *J. Public Econ.* 89 (5), 997–1019. doi:10.1016/j.jpubeco.2004.06.003. <http://www.sciencedirect.com/science/article/pii/S004727270400088X>
- Card, D., Mas, A., Moretti, E., Saez, E., 2012. Inequality at work: the effect of peer salaries on job satisfaction. *Am. Econ. Rev.* 102 (6), 2981–3003. doi:10.1257/aer.102.6.2981. <http://www.aeaweb.org/articles?id=10.1257/aer.102.6.2981>
- Carrieri, V., Jones, A.M., 2017. The income-health relationship 'beyond the mean': new evidence from biomarkers. *Health Econ.* 26 (7), 937–956. doi:10.1002/hec.3372. <https://onlinelibrary.wiley.com/doi/pdf/10.1002/hec.3372>
- Case, A., Paxson, C., 2005. Sex differences in morbidity and mortality. *Demography* 42 (2), 189–214.
- Chaparro, M., Hughes, A., Kumari, M., Benzeval, M., 2019. Is the association between self-rated health and underlying biomarker levels modified by age, gender, and household income? evidence from understanding society - the UK household longitudinal study. *SSM - Populat. Health* 8, 100406. doi:10.1016/j.ssmph.2019.100406. <https://www.sciencedirect.com/science/article/pii/S2352827318303318>
- Chen, E.H., Shofer, F.S., Dean, A.J., Hollander, J.E., Baxt, W.G., Robey, J.L., Sease, K.L., Mills, A.M., 2008. Gender disparity in analgesic treatment of emergency department patients with acute abdominal pain. *Acad. Emergency Med.* 15 (5), 414–418.
- Clark, A.E., D'Ambrosio, C., 2015. Chapter 13 - Attitudes to Income Inequality: Experimental and Survey Evidence. In: Atkinson, A.B., Bourguignon, F. (Eds.), *Handbook of Income Distribution*, Vol. 2. Elsevier, *Handbook of Income Distribution*, pp. 1147–1208. doi:10.1016/B978-0-444-59428-0.00014-X. <http://www.sciencedirect.com/science/article/pii/B978044459428000014X>
- Clark, A.E., Frijters, P., Shields, M.A., 2008. Relative income, happiness, and utility: an explanation for the Easterlin paradox and other puzzles. *J. Econ. Lit.* 46 (1), 95–144. doi:10.1257/jel.46.1.95. <http://www.aeaweb.org/articles?id=10.1257/jel.46.1.95>
- Clark, A.E., Kristensen, N., Westergaard-Nielsen, N., 2009. Job satisfaction and co-worker wages: status or signal? *Econ. J.* 119 (536), 430–447. doi:10.1111/j.1468-0297.2008.02236.x.
- Clark, A.E., Oswald, A.J., 1996. Satisfaction and comparison income. *J. Public Econ.* 61 (3), 359–381. doi:10.1016/0047-2727(95)01564-7. <http://www.sciencedirect.com/science/article/pii/S0047272795015647>
- Clark, A.E., Senik, C., 2010. Who compares to whom? the anatomy of income comparisons in Europe. *Econ. J.* 120 (544), 573–594.
- Clark, A.E., Westergaard-Nielsen, N., Kristensen, N., 2009. Economic satisfaction and income rank in small neighbourhoods. *J. Eur. Econ. Assoc.* 7 (2–3), 519–527. doi:10.1162/JEEA.2009.7.2-3.519. http://backfile/content_public/journal/jee/7/2-3/10.1162_jee.2009.7.2-3.519/1/jee0519.pdf
- Criado Perez, C., 2019. Invisible women: Exposing data bias in a world designed for men. *Random House*.
- Crimmins, E.M., Kim, J.K., Solé-Auró, A., 2010. Gender differences in health: results from SHARE, ELSA and HRS. *Eur. J. Public Health* 21 (1), 81–91.
- Crosby, F., 1976. A model of egoistical relative deprivation. *Psychol. Rev.* 83 (2), 85.
- Croson, R., Gneezy, U., 2009. Gender differences in preferences. *J. Econ. Lit.* 47 (2), 448–474.
- Croson, R., Gneezy, U., Rey-Biel, P., 2012. Gender differences in risk aversion and competition [special issue]. *Journal of Economic Behavior and Organization*, Vol. 83.
- Davillas, A., Pudney, S., 2017. Concordance of health states in couples: analysis of self-reported, nurse administered and blood-based biomarker data in the UK understanding society panel. *J. Health Econ.* 56, 87–102. doi:10.1016/j.jhealeco.2017.09.010. <https://www.sciencedirect.com/science/article/pii/S0167629617300358>
- Deaton, A., Stone, A.A., 2013. Two happiness puzzles. *Am. Econ. Rev.* 103 (3), 591–597. doi:10.1257/aer.103.3.591. <http://www.aeaweb.org/articles?id=10.1257/aer.103.3.591>
- Diriwaechter, P., Shvartsman, E., 2018. The anticipation and adaptation effects of intra- and interpersonal wage changes on job satisfaction. *J. Econ. Behav. Org.* 146, 116–140. doi:10.1016/j.jebo.2017.12.010. <https://www.sciencedirect.com/science/article/pii/S0167268117303542>
- Donna, J., Veramendi, G., 2021. Gender differences within the firm: evidence from two million travelers. *J. Human Resour.* forthcoming.
- Duesenberry, J., 1949. *Income, savings and the theory of consumer behaviour*. Harvard University Press.
- Eurostat, 2020. Gender pay gap in unadjusted form (sdg_05_20).
- Feeny, E., Dain, K., Varghese, C., Atiim, G.A., Rekve, D., Gouda, H.N., 2021. Protecting women and girls from tobacco and alcohol promotion. *BMJ* 374. doi:10.1136/bmj.n1516. <https://www.bmj.com/content/374/bmj.n1516.full.pdf>
- Gauriot, R., Page, L., 2019. Does success breed success? a quasi-Experiment on strategic momentum in dynamic contests. *Econ. J.* 129 (624), 3107–3136. doi:10.1093/ej/uezo40. <https://academic.oup.com/ej/article-pdf/129/624/3107/31271159/uezo40.pdf>
- Genicot, G., Ray, D., 2017. Aspirations and inequality. *Econometrica* 85 (2), 489–519. doi:10.3982/ECTA13865.
- Gneezy, U., Leonard, K.L., List, J.A., 2009. Gender differences in competition: evidence from a matrilineal and a patriarchal society. *Econometrica* 77 (5), 1637–1664.
- Godechot, O., Senik, C., 2015. Wage comparisons in and out of the firm. evidence from a matched employer–employee French database. *J. Econ. Behav. Org.* 117, 395–410.
- Goerke, L., Pannenberg, M., 2015. Direct evidence for income comparisons and subjective well-being across reference groups. *Econ. Lett.* 137 (C), 95–101. doi:10.1016/j.econlet.2015.10. <https://ideas.repec.org/a/eee/econlet/v137y2015icp95-101.html>
- Greene, W.H., Harris, M.N., Hollingsworth, B., 2015. Inflated responses in measures of self-assessed health. *Am. J. Health Econ.* 1 (4), 461–493.

- Grossman, P.J., Eckel, C., Komai, M., Zhan, W., 2019. It pays to be a man: rewards for leaders in a coordination game. *J. Econ. Behav. Org.* 161, 197–215. doi:10.1016/j.jebo.2019.04.002. <https://www.sciencedirect.com/science/article/pii/S0167268119301052>
- Harris, M.N., Zhao, X., 2007. A zero-inflated ordered probit model, with an application to modelling tobacco consumption. *J. Econom.* 141 (2), 1073–1099.
- Hoyer, B., van Huizen, T., Keijzer, L., Rezaei, S., Rosenkranz, S., Westbrook, B., 2020. Gender, competitiveness, and task difficulty: evidence from the field. *Labour Econ.* 101815.
- Ifcher, J., Zarghamee, H., Graham, C., 2018. Local neighbors as positives, regional neighbors as negatives: competing channels in the relationship between others' income, health, and happiness. *J. Health Econ.* 57, 263–276. doi:10.1016/j.jhealeco.2017.08.003. <https://www.sciencedirect.com/science/article/pii/S0167629616305616>
- Ifcher, J., Zarghamee, H., Houser, D., Diaz, L., 2020. The relative income effect: an experiment. *Exp. Econ.* 23 (4), 1205–1234. doi:10.1007/s10683-020-09648-. https://ideas.repec.org/a/kap/expeco/v23y2020i4d10.1007_s10683-020-09648-w.html
- Iriberry, N., Rey-Biel, P., 2019. Competitive pressure widens the gender gap in performance: evidence from a two-stage competition in mathematics. *Econ. J.* 129 (2020), 1863–1893.
- Jylhä, M., 2009. What is self-rated health and why does it predict mortality? towards a unified conceptual model. *Soc. Sci. Med.* 69 (3), 307–316.
- Kingdon, G.G., Knight, J., 2007. Community, comparisons and subjective well-being in a divided society. *J. Econ. Behav. Org.* 64 (1), 69–90. doi:10.1016/j.jebo.2007.03.004. <https://www.sciencedirect.com/science/article/pii/S0167268107000972>
- Krosnick, J.A., 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Appl. Cogn. Psychol.* 5 (3), 213–236. doi:10.1002/acp.2350050305. <https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.2350050305>
- Krosnick, J.A., Presser, S., 2010. Questionnaire Design. In: Wright, J.D., Marsden, P.V. (Eds.), *Handbook of Survey Research* (Second Edition). West Yorkshire, England: Emerald Group.
- Krueger, A., 2007. Are we having more fun yet? categorizing and evaluating changes in time allocation. *Brookings Pap. Econ. Act.* 38 (2), 193–218. <https://EconPapers.repec.org/RePEc:bin:bpeajo:v:38:y:2007:i:2007-2:p:193-218>
- Kuhn, P., Kooreman, P., Soetevent, A., Kapteyn, A., 2011. The effects of lottery prizes on winners and their neighbors: evidence from the dutch postcode lottery. *Am. Econ. Rev.* 101 (5), 2226–2247. <https://ideas.repec.org/a/aea/aecrev/v101y2011i5p2226-47.html>
- Luttmer, E.F.P., 2005. Neighbors as negatives: relative earnings and well-being. *Q. J. Econ.* 120 (3), 963–1002. doi:10.1093/qje/120.3.963. http://oup/backfile/content_public/journal/qje/120/3/10.1093/qje/120.3.963/2/120-3-963.pdf
- Malhotra, N., Miller, J.M., Wedeking, J., 2014. The relationship between nonresponse strategies and measurement error. *Online Panel Res.: Data Q. Perspect.*, A 313–336.
- McBride, M., 2010. Money, happiness, and aspirations: an experimental study. *J. Econ. Behav. Org.* 74 (3), 262–276. doi:10.1016/j.jebo.2010.03.002. <https://www.sciencedirect.com/science/article/pii/S0167268110000399>
- Mengel, F., 2020. Gender differences in networking. *Econ. J.* 130 (630), 1842–1873. doi:10.1093/ej/ueaa035. <https://academic.oup.com/ej/article-pdf/130/630/1842/33664373/ueaa035.pdf>
- Neumann-Böhme, S., Attema, A.E., Brouwer, W.B., van Exel, J.N., 2021. Life satisfaction: the role of domain-specific reference points. *Health Econ.* 30 (11), 2766–2779.
- Niederle, M., Vesterlund, L., 2007. Do women shy away from competition? do men compete too much? *Q. J. Econ.* 122 (3), 1067–1101.
- Noy, S., Sin, I., 2021. The effects of neighbourhood and workplace income comparisons on subjective wellbeing. *J. Econ. Behav. Org.* 185, 918–945. doi:10.1016/j.jebo.2020.11.008. <https://www.sciencedirect.com/science/article/pii/S0167268120304091>
- ONS, 2017. Leisure time in the UK: 2015. Technical Report. Office for National Statistics.
- Pekkarinen, T., 2015. Gender differences in behaviour under competitive pressure: evidence on omission patterns in university entrance examinations. *J. Econ. Behav. Org.* 115, 94–110.
- Pérez-Asenjo, E., 2011. If happiness is relative, against whom do we compare ourselves? implications for labour supply. *J. Popul. Econ.* 24 (4), 1411–1442.
- Perez-Truglia, R., 2020. The effects of income transparency on well-being: evidence from a natural experiment. *Am. Econ. Rev.* 110 (4), 1019–1054. doi:10.1257/aer.20160256. <https://www.aeaweb.org/articles?id=10.1257/aer.20160256>
- Pitler, E., Nenkova, A., 2008. Revisiting readability: A unified framework for predicting text quality. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 186–195. <http://dl.acm.org/citation.cfm?id=1613715.1613742>
- Ravindran, T.S., Teerawattananon, Y., Tannenbaum, C., Vijayasingham, L., 2020. Making pharmaceutical research and regulation work for women. *BMJ* 371. doi:10.1136/bmj.m3808. <https://www.bmj.com/content/371/bmj.m3808.full.pdf>
- Schwarz, N., Strack, F., 1999. Reports of subjective well-being: judgmental processes and their methodological implications. *Well-being: Foundat. Hedonic Psychol.* 7, 61–84.
- Seedat, S., Rondon, M., 2021. Women's wellbeing and the burden of unpaid work. *BMJ* 374. doi:10.1136/bmj.n1972. <https://www.bmj.com/content/374/bmj.n1972.full.pdf>
- Senik, C., 2004. When information dominates comparison: learning from russian subjective panel data. *J. Public Econ.* 88 (9), 2099–2123. doi:10.1016/S0047-2727(03)00066-5. <http://www.sciencedirect.com/science/article/pii/S0047272703000665>
- Senik, C., 2008. Ambition and jealousy: income interactions in the 'Old' europe versus the 'New' europe and the united states. *Economica* 75 (299), 495–513.
- Senik, C., 2009. Direct evidence on income comparisons and their welfare effects. *J. Econ. Behav. Org.* 72 (1), 408–424.
- Senik, C., et al., 2017. Gender gaps in subjective wellbeing: a new paradox to explore. *Rev. Behav. Econ.* 4 (4), 349–369.
- Shurchkov, O., Eckel, C.C., 2018. Gender differences in behavioral traits and labor market outcomes. *Oxford Handbook Women Econ.* 481.
- Stevenson, B., Wolfers, J., 2009. The paradox of declining female happiness. *Am. Econ. J.: Economic Policy* 1 (2), 190–225. <https://ideas.repec.org/a/aea/aejpol/v1y2009i2p190-225.html>
- Sturgis, P., Roberts, C., Smith, P., 2014. Middle alternatives revisited: how the neither/nor response acts as a way of saying 'i don't know'? *Sociol. Method. Res.* 43 (1), 15–38.
- Stutzer, A., 2004. The role of income aspirations in individual happiness. *J. Econ. Behav. Org.* 54 (1), 89–109. doi:10.1016/j.jebo.2003.04.003. <http://www.sciencedirect.com/science/article/pii/S0167268103002038>
- University of Essex, Institute for Social and Economic Research, NatCen Social Research, Kantar Public, 2019. Understanding society: Waves 1–9, 2009–2018 and harmonised bhps: Waves 1–18, 1991–2009. [data collection]. 12th edition. uk data service. sn: 6614.. 10.5255/UKDA-SN-6614-13.
- Vijayasingham, L., Govender, V., Witter, S., Remme, M., 2020. Employment based health financing does not support gender equity in universal health coverage. *BMJ* 371. doi:10.1136/bmj.m3384. <https://www.bmj.com/content/371/bmj.m3384.full.pdf>
- WHO, 2018. The Health and Well-being of Men in the WHO European Region: Better Health through a Gender Approach. World Health Organization. Regional Office for Europe.