

Mental health & gender stereotypes: Experimental evidence^{☆,☆☆}Paola Bertoli[☆], Veronica Grembi

University of Verona, Italy

Sapienza, University of Rome, Italy

ARTICLE INFO

JEL classification:

I100

I120

J160

Keywords:

Mental health

Depression

Gender stereotypes

ABSTRACT

International statistics reveal a persistent paradox: women receive depression diagnoses at rates two to three times higher than men, yet male suicide rates are two to three times higher than female rates. We investigate whether gender stereotype priming affects self-reported mental health using an experimental design embedded in a representative survey of the Italian population. We show that exposure to the treatment is associated to worse self-reported mental distress, with statistically comparable effects across genders. K-means clustering analysis reveals substantial heterogeneity: treatment effects are strongest among individuals who fail to conform to traditional gender expectations. These findings suggest that gender stereotype salience imposes psychological costs, particularly for those who fall short of traditional role expectations, offering insights into persistent gender disparities in mental health reporting and outcomes.

1. Introduction

Mental disorders have become a global public health concern which, however, affects men and women differently. Mental health differences between men and women are well established in psychiatry with women suffering more frequently from internalizing disorders (e.g., depression, anxiety and eating problems) and men reporting more externalizing disorders (e.g., substance abuse). For instance, across all EU, externalizing disorders are twice as frequent among men than women who, instead, self-report higher levels of internalizing disorders (Boyd et al., 2015).

Sources of these gaps are found in the role of sex hormones (Kuehner, 2017; Li and Graham, 2014) and in a different response of the hypothalamic–pituitary–adrenal axis (Bangasser and Valentino, 2014), as well as in female low self-esteem (Kuehner, 2017), gender-based violence and abuse (Oram et al., 2017), and social discrimination. For example, according to Astbury (2001), a driver of the gender gap in ordinary mental conditions is women's feeling of dis-empowerment and lack of control over the main socioeconomic determinants such as income, employment and social ranking. Stevenson and Wolfers link the

drop in women's self-reported well-being compared to men to a shift in the aspirations of women who no longer feel self-fulfilled by their role in the household but still face higher expectations in that domain.

Beside social, environmental and biological factors, there is also a behavioral explanation behind these mental health differences. Men lag far behind women when it comes to seeking help from health professionals (Galdas et al., 2005; Wang et al., 2007; Parent et al., 2018) and they tend to primarily report externalizing features than internalizing depression symptoms. As a result, they are less likely than women to be diagnosed with mental health issues.¹ Consequently, they more frequently turn to dangerous behaviors. In the EU, regardless of the higher prevalence of internalizing disorders among women, men die by suicide at a rate four-to-five times higher (OECD, 2014). This negative help-seeking is a consequence of men's unwillingness to talk about distressful situations (Vogel and Wester, 2003), to experience distressing feelings (Komiya et al., 2000) while avoiding social stigma (Deane and Chamberlain, 1994) and conforming to masculine norms (Mahalik et al., 2003; Lindinger-Sternart, 2015). At the same time, masculine norms, which emphasize emotional stoicism and self-reliance, may suppress the expression of internalizing symptoms while

[☆] This article is part of a Special issue entitled: 'Healthcare Trends & Future Perspectives' published in Economics and Human Biology.

^{☆☆} We thank Giorgio Coricelli, Raffaella Rumiati and participants to the XXX AIES Annual Conference, at the seminar at the Sapienza University of Rome, University of Verona, University of Padua, and University of Urbino for their comments on previous versions of this project.

^{*} Corresponding author.

E-mail address: paola.bertoli@univr.it (P. Bertoli).

¹ Another behavioral explanation looks at the supply side. Once they seek help, men are less likely to receive a diagnosis of depression than women. Then, they are less likely to be treated or to receive timely treatment.

² "Many men find there's no one around them at the end of the day. And they find they have become trapped by a series of unspoken rules. If you're a man, you're supposed to be strong and independent. If you connect deeply, especially with other men, it's probably because you're gay. If you crave deep emotional bonds, well, that's just unmanly." (from the Podcast Hidden Brain October 14, 2019 - Guys We Have a Problem: How American Masculinity Creates Lonely Men).

facilitating externalizing manifestations of distress (Addis and Mahalik, 2003; Magovcevic and Addis, 2008). For the same token, women may be more prone to seek help because they tend to take care of themselves and the family members more often and are not subject to the same social stigma of men.² In addition, women are more vulnerable to mental health problems as they are confronted with disproportionate care-giving expectations and work-family strain (Kuehner, 2017; Shira, 2014). In this perspective, gender stereotypes may play a key role in shaping both help-seeking behaviors and mental health manifestations by setting expectations on what behaviors are deemed appropriate for men and women.

The present paper studies how gender stereotypes may affect self-reported mental distress and if they may explain differences in it between men and women. Investigating the causal impact of gender stereotypes on mental health posits several empirical challenges. One possible approach would be to study how institutional changes that alter gender role expectations affect mental health differentially across contexts with varying degrees of gender norm rigidity. This strategy builds on insights from political economy regarding the complementarity between culture and institutions (Alesina and Giuliano, 2015; Bisin and Verdier, 2001). While gender stereotypes exhibit substantial persistence over generations (Nunn, 2012), formal institutions can rapidly change through policy reforms. Still, there are two substantial obstacles to this approach. First, finding valid measures of gender stereotypes that are also comparable across contexts is methodologically demanding (Seguino, 2011; Brandt, 2011). Available measures, whether survey-based indicators from cross-national studies (e.g., World Values Survey) or revealed preference measures from behavioral data, are threatened by endogeneity, measurement error, and cultural comparability (Prentice and Carranza, 2002; Ellemers, 2018). Second, a further difficulty lies in the identification of plausibly exogenous institutional changes that are both clearly dated and theoretically linked to gender role pressure. While certain policy reforms, such as changes in divorce laws (Gonzalez and Viitanen, 2009), abortion legislation (Levine et al., 1999), parental leave policies (Patnaik, 2019), or affirmative action mandates (Beaman et al., 2009), represent potentially useful sources of variation, establishing their exogeneity relative to evolving mental health trends and cultural attitudes is empirically challenging. In addition, these institutional changes may be endogenous to shifting cultural norms, creating identification concerns (Alesina and Fuchs-Schundeln, 2007).

Given these challenges, we adopted an alternative approach and designed an online experiment to elicit individual-level gender stereotypes and examine their association with self-reported mental distress in a controlled setting. This allows us to overcome several key limitations of observational studies. First, by directly measuring gender stereotypes at the individual level, we avoid the aggregation problems and measurement issues that plague cross-country or cross-regional comparisons of cultural attitudes. Second, the experimental setting enables us to establish clear temporal ordering and reduce concerns about reverse causality that complicate observational analyses. Third, by collecting rich individual-level data, we can explore mechanisms and heterogeneity that are difficult with aggregate administrative data.

Specifically, we administered an online survey to a representative sample of the Italian population whose respondents were randomly assigned to view different sets of images providing stereotype cues. All images portray household rooms (i.e., bathroom, living room, and kitchen). The control set displays empty rooms, while the two treatment sets show a non-same-sex couple within the same rooms: in the stereotyped condition, the woman is engaged in housekeeping activities, whereas in the counter-stereotyped condition the man is engaged in housekeeping activities. Exposure to these couple dynamics is meant to make traditional gender roles salient, especially under the stereotyped condition, while holding the same broader visual context. We show that making gender stereotypes salient increases self-reported

mental distress as captured by the validated Patient Health Questionnaire depression scale (PHQ-8), with treated respondents reporting, on average, levels 0.091 standard deviations higher than controls. Importantly, this effect is comparable across genders, challenging the notion that stereotypes differentially impact men and women. Still, the observed treatment effects on men suggest that making stereotypes salient genuinely worsens their perceived mental distress rather than merely discouraging disclosure. If the latter were predominant, we would observe the opposite pattern. These findings remain robust across alternative outcome specifications and when accounting for migration patterns through region of birth fixed effects, as well as for participants' response times. Both the single-dimension heterogeneity analysis and the k-means clustering analysis show that the treatment effects are strongest among individuals who fail to conform to traditional gender role expectations. In particular, the analysis based on k-means clustering shows that women who partially meet traditional female stereotypes (i.e., working mothers with low stereotypical beliefs) and men who fail to meet masculine norms (i.e., unemployed, not in couple, without children) experience the largest increases in perceived mental distress when exposed to gender role imagery. This pattern weighs in favor of the fact that stereotypes genuinely worsen the perception of one's own mental health rather than merely affecting disclosure behavior, as the psychological burden falls most heavily on those who fall short of the traditional expectations that stereotypical cues make salient.

The paper is organized as follows. Section 2 describes our survey design, explains its content and the information collected. Section 3 presents the experimental design, defining our primary outcomes of interest. Section 4 outlines our empirical strategy. Section 5 presents our baseline results and robustness checks, together with the heterogeneity analyses. Section 6 concludes.

2. Survey

2.1. Participants and general procedure

We design an online experiment to directly manipulate the salience of individual gender role stereotypes to test if, as a result of such activation, there is a change in the revealed mental health status as self-assessed by respondents.

To this end, during June 2022, a market survey company (Demetra opinioni.net srl) administered an online questionnaire (CAWI interviews) to a sample of 2611 individuals representative of the working age (18–65) Italian population by gender, age groups (younger than 35, 35–55, older than 55) and region of residence.³ On average, it took 10 min to complete the 36-question survey. The questionnaire includes a pre-treatment part collecting the basic socio-demographic information of participants (e.g., biological sex, age, level of education, province of residence, marital status), their working and financial conditions (e.g., employment status, sector of employment), individual explicit gender norms and gender roles perceptions, and the respondent's own mental health.⁴

Table 1 shows the distribution of our respondents across regions, 21 administrative units (19 regions and 2 autonomous provinces) also in charge of the Italian healthcare system, as well as information on the married respondents and age distribution. The comparison of these basic demographics with data from the Italian population as provided for year 2022 from the National Institute of Statistics (ISTAT) shows consistency. In addition, 71% of respondents are employed with the

³ The survey was approved by the Ethics Committee of the Scuola Internazionale Superiore di Studi Avanzati (SISSA) (Prot. N. 11542 - 13/05/2022) and registered at AEA-RCT registry (AEARCTR-0009422). All participants gave their informed consent at the beginning of the survey.

⁴ The full questionnaire translated in English is available in Appendix B.

Table 1
Representativeness (demographics).

Region	ISTAT 2022	Our survey
Piedmont	7.21	7.32
Valle d'Aosta	0.21	0.19
Lombardy	16.84	16.01
Trentino Alto Adige	1.82	0.96
Veneto	8.21	9.04
Friuli Venezia Giulia	2.02	1.72
Liguria	2.56	2.76
Emilia Romagna	7.50	7.39
Tuscany	6.21	5.48
Umbria	1.45	0.92
Marche	2.52	3.10
Lazio	9.68	10.38
Abruzzo	2.16	1.84
Molise	0.49	0.61
Campania	9.53	9.23
Puglia	6.65	8.08
Basilicata	0.92	1.03
Calabria	3.14	3.14
Sicily	8.19	7.62
Sardinia	2.69	3.18
Younger than 29	19.34	14.48
Age 30–50	43.99	53.39
Older than 50	36.66	32.13
Married	46.12	46.64

Notes: ISTAT = Italian National Institute of Statistics.

overwhelming majority in the private sector, 38% have a college degree or higher level of education, 63% declare to be religious even though not necessarily practicing, and 55% of them do have at least one child (see Table A1 for more on the definitions and summary statistics of our main variables of interest). We also ask questions related to satisfaction and feelings of guilt within respondents' social networks. The purpose is to capture both individual well-being and the pressure arising from relationships, social expectations, and social norms. Overall, 62% of respondents report being satisfied with their job, 58% with their work-life balance, and 47% with their salary. Feelings of guilt are less prevalent, ranging from 16% towards friends to 25% towards family members. Female respondents are more likely to feel guilt towards children and family members than male respondents, but not with respect to partners; at the same time, they report lower levels of satisfaction across all three dimensions.

2.2. Stereotypes indexes

Relying on two questions before the treatment, we construct two indexes to capture the potential gender biases of respondents, before exposing them to the treatments: a *Gender stereotypes* index and a *Profession stereotypes* index. These do not necessarily measure hidden biases, but rather what respondents are willing to reveal.⁵ This means that, as any other self-revealed measure, these indexes may be affected by a social desirability bias and they should be seen as a lower bound of the true individual stereotypes.

The first question aims at capturing the so-called explicit gender stereotype and was borrowed from the World Value Survey and the 2014 and 2018 surveys on gender stereotypes run by ISTAT. It asked respondents to indicate their level of agreement with five statements using a 4-point Likert scale ranging from strongly agree to strongly disagree (question n. 30, Appendix B). The statements concern the roles of men and women in the home and working domains highlighting the stereotyped role of women as housewives and men as more career driven. One statement is, for instance, “for men, more than for women, it

is very important to be successful at work”. The second question (question n. 33, Appendix B) presents a classifying task on genderizing occupations. Respondents were asked to classify a list of eleven jobs (from teacher to crane operator) as more suitable for men, for women, or for both genders. We moved from the assumption that ex ante there should not be any particular reason to state that any of these occupations is gender specific. Hence, we considered that there is a stereotypization of an occupation any time the respondent picked up one of the two genders rather than selecting the option “both”.

We converted responses into dummy variables coded as 1 for stereotypical responses. For the first question, this included agreeing or strongly agreeing with stereotype statements. For the second question, it captured gender-specific job assignments (i.e., selecting ‘man’ or ‘woman’ rather than ‘both’). We then summed all dummy variables for each question to create a continuous index ranging from 0 to 5 for gender stereotypes and from 0 to 11 for occupations. In both cases, higher scores capture more traditional beliefs about behaviors and roles deemed appropriate for men and women. As a last step, we standardized these indexes by computing z-scores with mean zero (corresponding to the national mean) and standard deviation equal to 1 to create our final indexes: *Gender stereotypes* and *Profession stereotypes*. Fig. 1 plots these measures by gender, revealing that men display higher than average gender stereotypes and they approach professions in a more gendered way.

We both internally and externally validate our measures. In Table 2 Panel (A), correlations between indexes within our sample show consistency. At the individual level, higher gender stereotypes are positively correlated with a more stereotypical way to think professional patterns. Panel (B) shows the results of an exercise of external validation using the 2019 Gender Equality (GE) Index at the regional level. The GE Index is a measure defined for the first time in 2005 by the European Institute for Gender Equality (EIGE) to compare gender equality across European Member States and to monitor its evolution in time.⁶ The GE Index is a synthetic measure summarizing the information on women's empowerment retrieved from six areas (i.e., domains): work, money, knowledge, time, power, health. A higher GE captures more equality between genders. After collapsing our indexes at the regional level, higher levels of GE are negatively correlated with our measures.

2.3. Mental well-being indexes

We assess the pre-treatment mental health status of respondents asking them to self-evaluate their overall health status, and then to assess three dimensions of their mental health status. First, respondents were asked how frequently (from very often to never) they experienced 5 situations (i.e., feeling depressed, anxious, afraid, having sleeping problems, and having panic attacks; question n. 17, Appendix B) in the six months prior to the survey. This approach based on self-reported symptoms is part of standard medical practice in mental health diagnosis (Chan, 2009). We then generate a dummy for each statement equal to one if the situation was experienced from *sometimes* to *very often*. We summed the dummy variables to create a score ranging from 0 to 5, which was standardized to a z-score (i.e., *Mental health status*) with mean 0 and standard deviation 1. The higher the score the poorer the mental health status. Throughout the paper, we use *Mental health status* to refer to this broader pre-treatment measure of overall mental health, while we use the term self-reported mental distress to refer to the symptom-level status captured by the post-treatment scales presented in Section 3.1. Sub-Figure (a) in Fig. 2 plots the distribution of this mental health index by gender showing a clear polarization: on average female respondents reveal worse than average mental status while

⁵ In order to measure the true hidden bias, an implicit association test should have been submitted.

⁶ For more details, see the Methodological Report available at https://eige.europa.eu/sites/default/files/documents/ti_pubpdf_mh0417333enn_pdfweb_20171030101356.pdf.

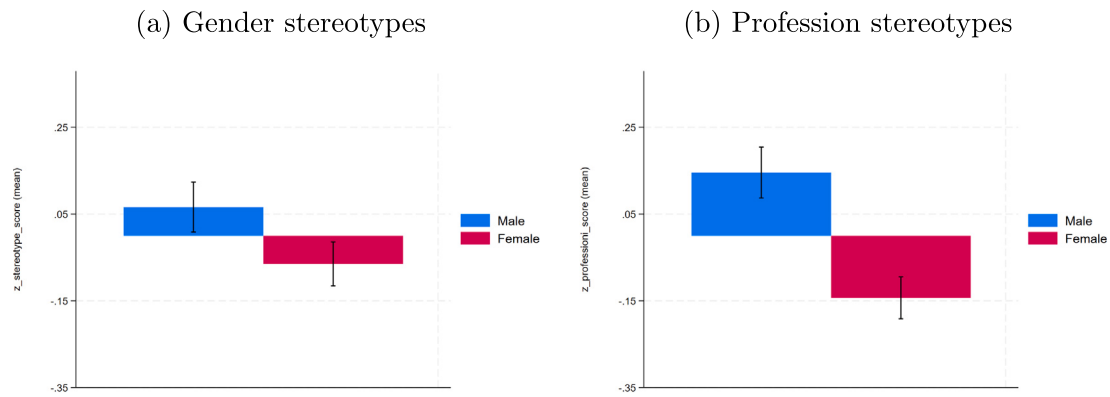


Fig. 1. Gender roles. *Notes:* *Gender stereotypes* is a standardized z-score based on the sum of five dummies taking value 1 if respondent agreed or strongly agreed with the related stereotype statement. *Profession stereotypes* is standardized z-score based on the sum of 11 dummies taking value 1 if respondent classified as gender specific the related profession.

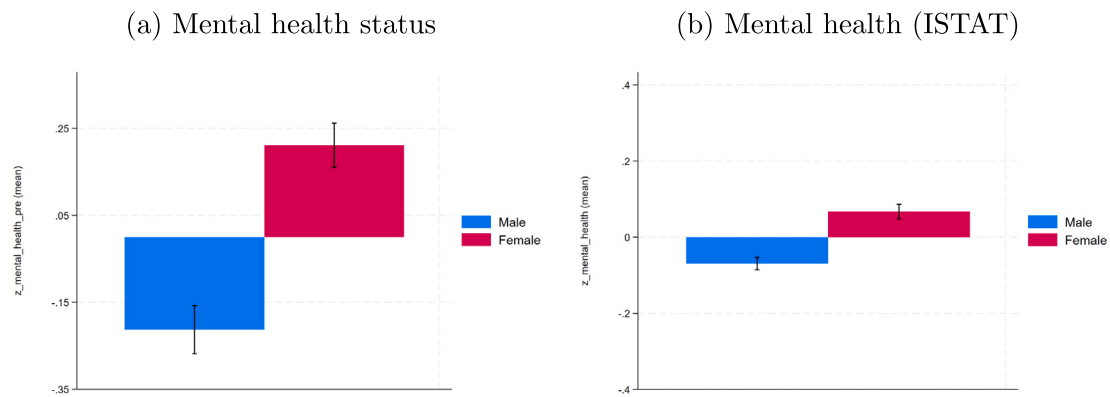


Fig. 2. Mental health status (pre-treatment): external validation. *Notes:* We use three questions from the 2022 *Aspetti di vita quotidiana* by ISTAT. We construct the z-score respect to the national average and standard deviation of the sum of three dummies which are equal to one if the respondent replies that she was (1) Often to always, sad in the last 4 weeks, (2) Often to always, anxious in the last 4 weeks, and (3) Often to always, down in the last 4 weeks. Higher values of the index mean worse mental health.

male respondents better than average mental status. This distribution is validated also using out-of-sample data, as shown in Sub-figure (b) where we draw on data from the 2022 wave of the ISTAT's annual survey of the Italian population (i.e., *Aspetti di vita quotidiana*). ISTAT poses a slightly different question to its respondents since it asks how frequently certain feelings (e.g., anxiety and sadness) had been experienced in the month prior to the survey. Yet, female respondents reveal a worse than average status versus male respondents who reveal a better than average status.

As a second and third dimensions to define the mental status of respondents, we ask if they have ever referred to a psychologist and if they have ever taken (over and not over the counter) drugs to fight anxiety and sleeping problems in the year preceding the survey. About a third of the sample replied in a positive way to these questions. However, as apparent from Fig. 3, female respondents resort more to both psychologists and medications than their male counterparts.

Table 3 shows the internal consistency of our measures of health: lower level of mental health status are positively correlated with having met a psychologist and using drugs, and negatively correlated with a self-assessed overall good health. It is interesting to notice that respondents seem to make coincide good health status with physical health, since the correlation coefficients are quite low. This seems to be more accentuated for male respondents compared to whom female ones report a negative and higher correlation. Women tend more to consider mental health as a determinant of their overall health status. In addition, Table A2 provides evidence of general response consistency,

Table 2

Internal and external validation: Gender roles indexes.

Panel A: Internal validation (individual level indexes)			
	Gender stereotypes	Profession stereotypes	
Gender stereotypes	1.000		
Profession stereotypes	0.415	1.000	
Panel B: External validation (regional level indexes)			
	GE	Profession stereotypes	Gender stereotypes
GE index (2019)	1.000		
Profession stereotypes	−0.262	1.000	
Gender stereotypes	−0.578	0.788	1.000

Notes: *GE* stands for gender equality index. *Gender stereotypes* is a standardized z-score based on the sum of five dummies taking value 1 if respondent agreed or strongly agreed with the related stereotype statement. *Profession stereotypes* is standardized z-score based on the sum of 11 dummies taking value 1 if respondent classified as gender specific the related profession.

showing coherent correlation patterns across conceptually related satisfaction and guilt-related measures, with no signals of systematic reporting inconsistencies or random answering behavior.

3. Experiment

The experiment consists in exposing randomly assigned respondents to one out of 3 sets of images presented within an attention test: a more

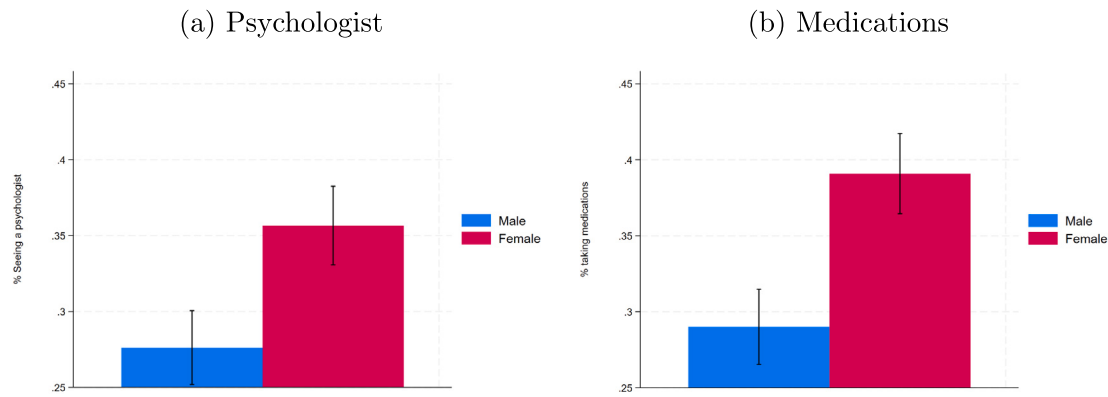


Fig. 3. Additional dimensions of mental health. *Notes:* *Psychologist* is a dummy taking value 1 if the respondent reports to have ever been referred to a psychologist and 0 otherwise. *Medications* is a dummy taking value 1 if the respondent reports to have taken (over and not over the counter) drugs to fight anxiety or sleeping problems in the year preceding the survey.

Table 3

Internal validation: Mental health status.

	<i>Mental health status</i>	<i>Psychologist</i>	<i>Drugs</i>	<i>Good health</i>
<i>Mental health status</i>	1.000			
<i>Psychologist</i>	0.235	1.000		
<i>Drugs</i>	0.380	0.289	1.000	
<i>Good Health</i>	−0.006	−0.064	−0.024	1.000
Female respondents				
<i>Mental health status</i>	1.000			
<i>Psychologist</i>	0.183	1.000		
<i>Drugs</i>	0.355	0.276	1.000	
<i>Good Health</i>	−0.023	−0.061	−0.032	1.000
Male respondents				
<i>Mental health status</i>	1.000			
<i>Psychologist</i>	0.265	1.000		
<i>Drugs</i>	0.382	0.290	1.000	
<i>Good Health</i>	0.004	−0.070	−0.020	1.000

Notes: *Mental health status* is a standardized z-score based on the sum of five dummies taking value 1 if respondent self-reports (pre-treatment) the related symptom (i.e., depression, anxiety, fear, sleeping problems, and panic attacks) from “sometimes” to “very often” in the six months prior to the survey. *Psychologist* is a dummy taking value 1 if the respondent reports to have ever been referred to a psychologist and 0 otherwise. *Medications* is a dummy taking value 1 if the respondent reports to have taken (over and not over the counter) drugs to fight anxiety or sleeping problems in the year prior to the survey. *Good Health* is a dummy taking value 1 if the respondent self-evaluates his/her overall health status as good, very good or excellent and 0 otherwise.

conservative activating set of images (i.e., Stereotyped), a set of control images (i.e., neutral), and a less conservative activating set of images (i.e., Counter-stereotyped).⁷ Tables A3 and A4 report the balance tests of the observables between the control group (871 respondents) and the overall treated group (1740 respondents), and also for each treated group separately (870 respondents each). As a further balance check, we compare response times for questions asked before treatment on potentially sensitive topics, that is, those measuring pre-treatment mental health status, professional stereotypes, and gender stereotypes. Since these questions precede treatment assignment, any systematic difference in response times would indicate pre-existing differences in cognitive engagement with potentially sensitive topics. However, as shown in Figure A1, we see no statistically significant differences in response times between treated and control respondents for any of these questions.

⁷ Before participants were randomly assigned to their set of images, they were reminded of the importance to pay attention and we anticipated that they would be asked some questions on the images that they were about to see to test their level of attention.

All images are pictures taken in household rooms (i.e., bathroom, living room, and kitchen). The control images portray the empty rooms (Fig. 4 - Panel A), while the treatment images either portray two non-same-sex fictional characters with the woman engaged in housekeeping activities (Stereotyped, Fig. 4 - Panel B), or the man engaged in housekeeping activities (Counter-stereotyped, Fig. 4 - Panel C). The exposure to couple dynamics within a household is expected to affect participants' stereotypical beliefs on gender roles but differently depending on the type of dynamic (i.e., stereotypical or counter-stereotypical).⁸

To test these salient effects of the images, immediately after the treatment, we asked respondents to complete two statements with the first words that came to mind. The statements were (i) *Carlo is really a good ...*; (ii) *Carla is really a good ...* with Carlo being a male name and Carla a female name.⁹ Then, in the spirit of Chiu et al. (2018), we submitted the complete list of words provided by respondents (without frequency information) to a panel of 27 external evaluators, heterogeneous in terms of gender and age. Since Italian words usually have feminine and masculine forms, evaluators were instructed to focus on the meaning of each word regardless of its grammatical declination and to classify it into one of three categories: *female gender*, *male gender* or *gender neutral*. We assigned numerical values to each classification: 0 for *gender neutral*, 1 for *female gender*, and −1 for *male gender*. For each word, we averaged the classifications across all evaluators to obtain a stereotype score. We then inverted the scores for Carlo (i.e., multiplied by −1) to ensure consistent interpretation. The resulting score (*Implicit Gender Bias*) ranges from −1 to 1, where 1 indicates maximum stereotype (the word chosen for Carla is consistently classified as *female gender* and the word for Carlo as *male gender*), and −1 indicates maximum counter-stereotype (the word chosen for Carla is consistently classified as *male gender* and the word for Carlo as *female gender*).

As apparent from Figure A2, images depicting non-same-sex couples effectively activate stereotypical beliefs in respondents.¹⁰ Compared to those who viewed only empty rooms, respondents exposed to couple

⁸ The use of a home-related setting might be per se evoking gender dynamics. This element should be considered in the final interpretation of the treatment effects compared to a situation in which a non-same-sex couple is watched out of a household setting.

⁹ According to ISTAT, in 1999 the name Carlo accounted for the 0.33% of the names used for baby boys (0.3% in 2020) while Carla accounted for the 0.12% (0.09% in 2020) of the names used for baby girls.

¹⁰ The same exercise was also performed by submitting a list of neutral words to another panel of 27 external evaluators with similar composition to the original panel. Words were presented in gender-neutral form by truncating gender-specific endings. For instance, the word ‘nurse’ translates to *infermiera* (feminine) and *infermiere* (masculine) in Italian, and was presented as “*inferm**”. We obtained consistent results.

(A) Neutral condition

(a) Bathroom



(b) Kitchen



(c) Living room

**(B) Stereotypes condition**

(d) Bathroom



(e) Kitchen



(f) Living room

**(C) Counter-stereotypes condition**

(g) Bathroom



(h) Kitchen



(i) Living room



Fig. 4. Experimental images. *Notes:* Respondents were randomly assigned to one of the three experimental conditions.

images displayed higher implicit gender biases, though the magnitude of the effect varied across image types. Images showing less traditional gender roles (such as a man cleaning) produced lower gender biases than images depicting more conventional gender roles. Nevertheless, even the less traditional images elicited higher biases (i.e., more stereotypical associations) than the empty room baseline. These findings align with existing literature showing that both stereotypical and counter-stereotypical images can activate implicit gender stereotypes even if to a different extent (e.g., Finnegan et al., 2015; Blair et al., 2001).¹¹

¹¹ We also previously tested the images on 170 individuals aged 20–64 living in Italy (who did not take part in the main experiment) who were recruited on MTurk between November 17 and December 11, 2021. After viewing the images, participants were asked to complete two tasks. First, they completed the same statements presented in the main experiment. Second, they were asked to “list the first six objects that come to mind” following Chiu et al. (2018). Two independent coders (one man and one woman), who were blind to the image conditions, classified each listed object and completion word as masculine (e.g., hammer, plumber), feminine (e.g., purse, housekeeper), or gender-neutral (e.g., chair, person). Participants reported the highest number

3.1. Outcomes of interest

Self-reported symptom assessment represents standard clinical practice in mental healthcare (Chan, 2009). Clinical interviews and questionnaires, as well as clinician-rated scales, all rely on patients’ ability to recall their emotional states. This approach is based on the principle that occasional mental health challenges – such as anxiety, stress, or sleep difficulties – are common, but only when they occur with a sufficient frequency over an extended period of time do they raise clinical concerns.

Consistently with clinical practice, to measure the effect of activating stereotypical beliefs on respondents’ reported mental distress, we administered two validated psychological scales after the treatment: the 4-item Perceived Stress Scale (PSS-4) and the 8-item Patient Health Questionnaire depression scale (PHQ-8). The PSS-4 as validated by

of stereotyped words and objects in the Stereotype condition, followed by the Counter-stereotype and Control conditions, respectively.

Cohen et al. (1983)¹² is one of the most commonly used measures of perceived stress. It includes two negatively stated and two positively stated items aimed at capturing the degree to which one perceives her life as unpredictable and overloaded (Wang et al., 2011). Specifically, respondents are asked to rate how often they had certain feelings or thoughts over the previous 4 weeks on a 5-point Likert scale from “never” (i.e., equal to 0) to “very often” (i.e., equal to 4). After reversing the two positively worded items, the rates given to each item are summed up to produce the PSS-4 score ranging from 0 to 16.

The PHQ-8 is a canonical depression screener known to be a strong predictor of medical diagnoses (Kroenke and Spitzer, 2002; Kroenke et al., 2009). It includes eight out of the nine domains from the diagnostic criteria for major depressive disorders (MDD) in the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) (Kroenke et al., 2009).¹³ Respondents are asked to rate how often they have been bothered by a certain problem over the previous 4 weeks on a 4-point Likert scale from “never” (i.e., equal to 0) to “nearly every day” (i.e., equal to 3). Adding together the scores for all 8 domains, one obtains the PHQ-8 score which ranges from 0 to 24.

Both scores share the same high-level interpretation: higher PHQ-8 scores identify more severe depressive symptoms and higher PSS-4 scores identify greater perceived stress. The two scores are generally highly correlated (e.g., Gold, 2015; Tafet and Nemeroff, 2016), although stress is more strongly influenced by environmental factors, particularly the work environment (Tennant, 2001). We standardized both scales to z-scores based on the control group’s mean and standard deviation obtaining our outcomes of interest: *Depression* and *Perceived stress*. The correlation coefficient between the two outcomes is 0.65.

4. Empirical strategy

We assess the effect of activating stereotypical beliefs on self-reported mental distress estimating the model in Eq. (1):

$$\begin{aligned} Mental\ health_i = & \delta Treatment_i + \lambda SES_i' + \beta Female_i + \gamma Good\ health_i' \\ & + \mu Covid19_i \\ & + \tau Mental\ health\ status_i + \pi Psychologist_i \\ & + \rho Medications_i + \nu Qcorrect_i \\ & + \phi Gender\ stereotypes_i + \omega Profession\ stereotypes_i \\ & + \alpha Region_r + \epsilon_{ir} \end{aligned} \quad (1)$$

where *Mental health_i* represents post-treatment outcomes for individual *i* (i.e., the standardized depression score *Depression* and the standardized perceived stress scores *Perceived stress*). *SES_i* captures socioeconomic characteristics including age, education, marital status, presence of children, being religious, employment status, being self-employed, working in the private sector and having experienced financial problems (see Table A1). *Female_i* is a gender indicator. Health-related controls include self-assessed overall health status (*Goodhealth_i*), pre-treatment self-assessed mental health status (*Mental health status_i*), whether the respondent ever tested positive

to Covid-19 (*Covid19_i*), whether the respondent ever consulted a psychologist (*Psychologist_i*), and used anxiety or sleep medications in the year prior to the survey (*Medications_i*). Finally, we control for respondent’s gender stereotype score (*Gender stereotypes_i*), profession stereotype score (*Profession stereotypes_i*) and for her attention test performance (i.e., number of correct answers, *Qcorrect_i*). The model also includes regional fixed effects for Italy’s 21 administrative units to account for time-invariant regional characteristics. These fixed effects capture cultural and social norms, and persistent attitudes toward gender roles in the region of residence of respondents. Standard errors are clustered at the regional level. The key explanatory variable is *Treatment_i*, which indicates exposure to images depicting non-same-sex couples in household settings.

5. Results

Since both stereotypical and counter-stereotypical images have proved to activate gender role associations – though with varying intensity – we start pooling these treatment conditions into a single treatment group, with respondents exposed only to empty rooms serving as the control group, before separately examining the effect of exposure to stereotypical and counter-stereotypical images.

Table 4 reports our baseline estimates. Panel A shows that exposure to non same-sex couple imagery increases self-reported mental distress overall and for both genders, but only as captured by the depression score. Treated respondents report a depression score 0.091 standard deviations higher than controls; this increase amounts to 0.079 standard deviations for female respondents and to 0.091 standard deviations for their male counterparts. The magnitude of these effects, however, does not differ significantly across gender, providing an important first insight: stereotype activation has comparable impacts on both men and women.

This holds also once we separate treatment conditions. The comparison between Panel B and Panel C shows that the exposure to stereotypical images (i.e., women performing housework) affects female respondents who report a 0.139 standard deviation increase in depression, while counter-stereotypical images (i.e., men performing housework) do not exert any statistically significant effect. In contrast, male respondents are affected by both sets of images. Still, the magnitude of the effect is stronger when they are exposed to stereotypical images than to counter-stereotypical ones (i.e., 0.111 vs. 0.094 increase in standard deviations). Even if men and women still do not exhibit a statistically significant difference in their reaction to the activation of stereotypical beliefs, breaking down the two treatment conditions provides an important second insight: such activation actually worsens men’s self-reported mental distress over the recent past rather than simply discouraging disclosure. Even if we cannot rule out the latter mechanism, it cannot be predominant. If it were, we would observe negative treatment effects as traditional gender role cues – which emphasize male “strength” – would discourage men from sharing their past mental distress.

In Tables A6 and A7, we test the robustness of our results to alternative definitions of our outcomes of interest. Rather than using z-score, we use the non-standardized version of both *Depression* and *Perceived stress*. In addition, we created the categorical variable *PHQ-8 Ranking* by categorizing the depression score into the standard 5 classes used in clinical practice (from 1 “Minimal depression” and 5 “Severe depression”) (Kroenke et al., 2001; Williams et al., 2009). Similarly, for perceived stress, we also created *PSS-4 Ranking*, a categorical variable ranging from 0 “Low stress” to 2 “High perceived stress”. In addition, in Table A8, we consider respondents’ region of birth rather than their region of residence to exclude that our results are driven by migration patterns which do affect the cultural background of the respondents. Hence, we apply region of birth fixed effects and standard error are clustered at the level of the region of birth. Treatment effects are confirmed in magnitude and significance across all robustness checks. Finally, Table A9 shows that our results are also robust to the inclusion of the response times to the post-treatment questions as additional controls.

¹² We use the Italian translation produced by Fossati with the authorization of the American Sociological Association (Fossati, 2010) and further tested on Italian subjects by Mondo et al. (2021). For more information, visit <https://www.cmu.edu/dietrich/psychology/stress-immunity-disease-lab/scales/index.html>.

¹³ The eight domains are (i) anhedonia, (ii) feeling down, depressed or hopeless, (iii) sleep problems, (iv) fatigue or low energy, (v) poor appetite or overeating, (vi) feelings of worthlessness or excessive guilt, (vii) trouble concentrating, and (viii) psychomotor agitation or retardation. We use the Italian version of the PHQ-8 as validated and released by the Istituto Superiore di Sanità (ISS, the National Institute of Health in Italy). For more information, see <https://www.epicentro.iss.it/ben/2013/gennaio/1>.

5.1. Heterogeneity analysis on depression

According to our baseline analysis, increasing the salience of gender stereotypes worsens self-reported mental distress for both men and women as captured by depression scores. This finding deserves further investigation as our setting does not allow us to separately assess the two potential mechanisms at play: the genuine worsening of individuals' perception of their mental health or the increased willingness to disclose their pre-existing distress. Moreover, the average treatment effects may mask important heterogeneity in how individuals respond to gender role cues. Individuals differ substantially in their pre-existing beliefs, demographic characteristics, and socioeconomic circumstances, all factors that may mediate responses to stereotype priming. We therefore perform a heterogeneity analysis, as described by Eq. (2).

$$\begin{aligned}
 \text{Mental health}_i = & \delta \text{Treatment}_i + \psi \text{Treatment}_i * D_i + \lambda \text{SES}'_i + \beta \text{Female}_i \\
 & + \gamma \text{Good health}'_i \\
 & + \mu \text{Covid19}_i + \tau \text{Mental health status}_i \\
 & + \pi \text{Psychologist}_i + \rho \text{Medications}_i \\
 & + \nu \text{Qcorrect}_i + \phi \text{Gender stereotypes}_i \\
 & + \omega \text{Profession stereotypes}_i \\
 & + \alpha \text{Region}_r + \epsilon_{ir}
 \end{aligned} \quad (2)$$

The dummy variable D_i alternatively identifies 9 potentially relevant individual characteristics: being a female, being 44 or older, having a college degree, being in a couple, having children, being religious, being employed, having high gender stereotypes and having a poor pre-treatment mental health status (for a definition of each dummy, see Table A5). Using these dummies, we estimate Eq. (2), which includes the interaction term $\text{Treatment}_i * D_i$. Thus, δ represents the effect of the treatment on those not possessing the given individual characteristic and $\delta + \psi$ represents the effect on those possessing it, while ψ indicates whether the difference between the two groups is statistically significant. For instance, δ captures the effect of being employed, $\delta + \psi$ captures the effect on those unemployed, and ψ captures the difference between the two groups.

As shown in Table 5, we find evidence of different responses to the treatment only for the following dimensions: religiosity, employment status, pre-existing gender stereotypes, and pre-treatment mental health status. When female and male respondents are pooled together (Panel A), respondents who do not self-assess as religious react more strongly than religious ones and respondents with poor pre-treatment mental health show a larger response than those reporting a good pre-treatment mental status. According to panels B and C, these results are primarily driven by male respondents. Among women (Panel B), only religiosity yields a marginally significant difference with non-religious women reacting more than religious ones to the activation of gender role cues. Religiosity offers interpretive structures and coping mechanisms that may buffer the psychological tension triggered by gender role cues, therefore non-religious individuals end up experiencing the salience of traditional norms more strongly. Beyond religiosity, the treatment effect is concentrated on women with stronger pre-existing stereotypical beliefs, who report a 0.115 standard deviation increase in self-reported distress (significant at the 5% level), while the effect is essentially null for those with weaker beliefs. Although the difference between the two groups is not statistically significant, this dimension is worth noting because it mirrors, in the opposite direction, the pattern observed among men. For women the response is driven by those who most strongly endorse traditional roles, for whom making such roles salient may intensify the pressure to meet standards they themselves embrace.

Among men (Panel C), unemployed respond more than employed ones, while men holding weaker pre-existing stereotypical beliefs react more than those holding stronger ones, and men with poor pre-treatment mental health react more than those with good mental

Table 4

Baseline results.

	Depression			Perceived stress		
	All	Female	Male	All	Female	Male
<i>Panel A: Treatment vs. Control</i>						
Treatment	0.091*** (0.025)	0.079* (0.045)	0.091*** (0.029)	0.088* (0.042)	0.086 (0.066)	0.087 (0.059)
Obs.	2611	1315	1296	2611	1315	1296
<i>Panel B: Stereotype vs. Control</i>						
Treatment	0.125*** (0.040)	0.139** (0.057)	0.111** (0.050)	0.071 (0.054)	0.057 (0.092)	0.090 (0.064)
Obs.	1741	877	864	1741	877	864
<i>Panel C: Counter vs. Control</i>						
Treatment	0.068** (0.031)	0.032 (0.061)	0.094*** (0.029)	0.106* (0.057)	0.117 (0.071)	0.100 (0.082)
Obs.	1741	877	864	1741	877	864

Notes: All estimates account for the sets of covariates detailed in Table A1, as well as for region fixed effects. Standard errors clustered at the regional level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

health. The differential response by employment status can be explained by the fact that the salience of traditional gender roles is particularly burdensome for those who fail to meet them and employment is a core dimension of the traditional role of men. Then, when unemployed men are confronted with household imagery, they may suffer from their distance from this norm showing a worsening of their self-reported mental status. The asymmetry along pre-existing gender stereotypes is consistent with a form of cognitive dissonance according to which men with weaker stereotypical beliefs do not conform to the traditional masculine model that the treatment makes salient. As consequence, their stronger response may be the result of the psychological tension triggered when an explicitly rejected model is reactivated by external cues. By contrast, men embracing more traditional beliefs can be expected to be less destabilized by cues that are coherent with their own view.

Finally, the fact that respondents already in poor mental health, especially men, report a stronger response is consistent with two non-mutually-exclusive interpretations, that is, a greater willingness to disclose pre-existing distress once stereotype cues lower the threshold for self-reporting, or a greater sensitivity to the psychological pressure caused by these cues. However, the treatment effect is also positive and statistically significant among respondents reporting good pre-existing mental health, both in the pooled sample and among men. This finding is difficult to reconcile with a disclosure-based mechanism. Individuals reporting a poor mental health status pre-treatment are already willing to report their distress, whereas those reporting good mental health may also include respondents experiencing distress but reluctant to disclose it. If stereotype cues mainly affected reporting behavior, the strongest effects should be observed among the latter group and would be expected to decrease or leave unchanged reported distress rather than increase it. The positive treatment effect observed among respondents with good mental health status therefore supports an actual worsening in perceived mental distress caused by stereotype salience, rather than merely a change in reporting behavior.

5.2. Heterogeneity by clusters

The heterogeneity analysis by single-characteristic suggests that the treatment effect is not uniformly distributed across the population, but concentrates among individuals who, for different reasons, face greater tension when traditional gender roles are made salient: unemployed men, men holding weaker stereotypical beliefs, non-religious respondents, and individuals with poor pre-existing mental status. Still, the notion of conformity to traditional gender roles is inherently multidimensional. For example, an unemployed man without a partner

Table 5
Heterogeneities on depression.

	Female	44 years & older	College degree	In a couple	With kids	Religious	Employed	Stereotypes level	Mental status
<i>Panel A: ALL</i>									
	No	No	No	No	No	No	No	Low	Good
Treatment	0.101*** (0.031)	0.092* (0.047)	0.085** (0.040)	0.122** (0.058)	0.076** (0.035)	0.167*** (0.051)	0.110*** (0.038)	0.123*** (0.040)	0.057** (0.024)
	Yes	Yes	Yes	Yes	Yes	Yes	Yes	High	Poor
Treatment	0.080** (0.035)	0.092** (0.042)	0.100 (0.965)	0.071** (0.030)	0.103** (0.049)	0.047*** (0.030)	0.081** (0.034)	0.081** (0.031)	0.287** (0.112)
Difference	−0.021 (0.043)	0.000 (0.074)	0.015 (0.081)	−0.051 (0.072)	0.026 (0.069)	−0.119* (0.061)	−0.030 (0.051)	−0.043 (0.050)	0.229* (0.126)
Obs.	2611	2611	2611	2611	2611	2611	2611	2611	2611
<i>Panel B: FEMALE</i>									
	No	No	No	No	No	No	No	Low	Good
Treatment	0.125** (0.056)	0.089 (0.057)	0.115 (0.096)	0.050 (0.075)	0.162** (0.065)	0.007 (0.055)	0.007 (0.055)	0.007 (0.074)	0.040 (0.039)
	Yes	Yes	Yes	Yes	Yes	Yes	Yes	High	Poor
Treatment	0.030 (0.062)	0.061 (0.101)	0.059 (0.049)	0.099 (0.060)	0.034 (0.049)	0.134* (0.062)	0.115** (0.054)	0.115** (0.054)	0.213 (0.133)
Difference	−0.095 (0.077)	−0.028 (0.128)	−0.056 (0.111)	0.049 (0.100)	−0.128* (0.068)	0.127 (0.077)	0.107 (0.089)	0.107 (0.089)	0.173 (0.142)
Obs.	1315	1315	1315	1315	1315	1315	1315	1315	1315
<i>Panel C: MALE</i>									
	No	No	No	No	No	No	No	Low	Good
Treatment	0.030 (0.059)	0.075 (0.047)	0.123* (0.060)	0.083* (0.040)	0.169** (0.062)	0.274*** (0.048)	0.269*** (0.060)	0.269*** (0.060)	0.051* (0.029)
	Yes	Yes	Yes	Yes	Yes	Yes	Yes	High	Poor
Treatment	0.159*** (0.044)	0.114* (0.063)	0.067 (0.053)	0.098 (0.069)	0.044 (0.049)	0.026 (0.043)	0.020 (0.039)	0.020 (0.039)	0.484*** (0.157)
Difference	0.130 (0.089)	0.038 (0.094)	−0.056 (0.097)	0.015 (0.096)	−0.125 (0.095)	−0.248*** (0.079)	−0.250*** (0.083)	−0.250*** (0.083)	0.433** (0.165)
Obs.	1296	1296	1296	1296	1296	1296	1296	1296	1296

Notes: All estimates account for the sets of covariates detailed in Table A1, as well as for region fixed effects. Standard errors clustered at the regional level in parentheses. *** p < 0.01, ** p < 0.05, * p < 0.1.

and children fails to meet traditional masculine expectations along multiple reinforcing dimensions. By examining one characteristic at a time, the model in Eq. (2) disregards these combinations and may not correctly identify the subset of respondents on whom stereotype activation impacts the most. For instance, the observed employment status effect may actually be driven by a specific subgroup of unemployed respondents who simultaneously have other characteristics associated with greater sensitivity to gender-role salience, such as older age or more traditional beliefs. In general, effects attributed to a single characteristic may capture the joint influence of multiple correlated dimensions. Since responsiveness to gender-role salience is fundamentally multidimensional, a heterogeneity analysis by single-characteristic may fail to capture the combinations of characteristics that drive treatment responses. We therefore apply a multidimensional approach and examine more complex forms of heterogeneity by performing a clustering analysis to group respondents into different types.

To perform this heterogeneity without imposing arbitrary subgroup definitions, we employ k-means clustering analysis on the controls described in Table A1. This data-driven approach allows us to identify naturally occurring clusters of respondents with similar profiles across multiple dimensions simultaneously, revealing whether certain combinations of individual characteristics make respondents more or less responsive to the exposure of stereotype cues. Following Riley (2024), clusters have been identified using the k-means command in STATA and examining the sum of within-cluster distances by number of groups to select the natural break-point. We double checked this approach with the Calinski–Harabasz pseudo-F index stopping rule, with larger pseudo-F index indicating more distinct clustering. Both methods identified 4 groups for women and 3 for men. K-means clustering is by construction sensitive to the relative variance of the input variables because variables with larger variance tend to dominate the distance computation and therefore the cluster assignment. However,

our analysis is not affected by the risk of individual variables disproportionately driving cluster assignment, as all our clustering variables are expressed on comparable scales. Specifically, socio-demographic characteristics are coded as dummy variables, while the stereotype indices are standardized as z-scores.

Table 6 reports the descriptive statistics for the four female groups. Group 1 is composed of women without children, not in a couple, who, on average, have a job and are younger and less likely to define themselves as religious than women in the other groups. By contrast, Groups 2–4 are mainly mothers, in a couple, and quite religious. Still, they differ in terms of employment status. If all women in group 2 are unemployed, those in Group 4 are all employed while 54% of women in Group 3 do have a job. In terms of stereotypes, Groups 1, 2, and 4 are comparable with women characterized by both gender and profession stereotypes below average and much lower than women in Group 3.

We examine heterogeneity across these four groups in Table 7, Column (1). The reference group is Group 4, composed of women who partially conform to the traditional female stereotype (i.e., married working mothers) but hold low stereotypical beliefs. The treatment significantly increases their self-reported mental distress. This finding is consistent with the two possible mechanisms previously discussed. First, stereotypical cues may confront these women with their incomplete conformity to traditional expectations. Second, such cues may encourage greater disclosure, as complaining behavior is traditionally accepted among women. The two other groups show statistically different treatment effects from Group 4. Specifically, Groups 2 and 3 exhibit smaller responses to stereotypical cues. Women in Group 2 hold low stereotypical beliefs but fully conform to the traditional female stereotype (i.e., married stay-at-home mothers). For this group, stereotype activation likely increases disclosure of past mental distress rather than worsening their perception of it. Conversely, women in Group 3 hold strong stereotypical beliefs, but fail to meet traditional expectations

Table 6
Descriptive statistics of female clusters.

	Group 1	Group 2	Group 3	Group 4
Age	34.72 (11.18)	49.35 (11.00)	45.06 (11.02)	45.15 (10.20)
College degree or higher	0.49 (0.50)	0.15 (0.36)	0.30 (0.46)	0.42 (0.49)
Couple	0.00 (0.00)	0.92 (0.27)	0.81 (0.40)	0.86 (0.35)
Kids	0.02 (0.14)	0.83 (0.38)	0.73 (0.45)	0.77 (0.42)
Religious	0.48 (0.50)	0.71 (0.45)	0.77 (0.42)	0.67 (0.47)
Employed	0.59 (0.49)	0.00 (0.00)	0.54 (0.50)	1.00 (0.00)
Financial distress	0.67 (0.47)	0.63 (0.48)	0.69 (0.47)	0.63 (0.48)
Private sector	0.73 (0.44)	0.70 (0.41)	0.73 (0.47)	0.74 (0.44)
Self-employed	0.11 (0.32)	0.00 (0.00)	0.14 (0.34)	0.18 (0.39)
Medication	0.37 (0.48)	0.35 (0.48)	0.46 (0.50)	0.41 (0.49)
Psychologist	0.38 (0.49)	0.33 (0.47)	0.24 (0.43)	0.40 (0.49)
Good health	0.57 (0.50)	0.65 (0.48)	0.55 (0.50)	0.68 (0.47)
Mental health status	0.35 (0.88)	0.14 (0.96)	0.16 (1.04)	0.24 (0.93)
Covid-19	0.34 (0.47)	0.31 (0.46)	0.32 (0.47)	0.36 (0.48)
Stereotype score	-0.37 (0.57)	-0.38 (0.45)	1.95 (0.59)	-0.37 (0.46)
Profession score	-0.38 (0.64)	-0.24 (0.76)	0.66 (1.27)	-0.21 (0.80)
Obs	320	302	175	518

Notes: Groups selected from k-mean clustering on baseline covariates for female respondents.

Table 7
Heterogeneity on depression – K-mean clustering for female.

	Female
Treatment	0.278*** (0.094)
Group1*Treatment	-0.148 (0.153)
Group2*Treatment	-0.295** (0.112)
Group3*Treatment	-0.289** (0.124)
Group 1	0.365*** (0.110)
Group 2	0.129 (0.114)
Group 3	0.445*** (0.125)
Obs	1315
Mean group 1	0.392
Mean group 2	0.044
Mean group 3	0.389
Mean group 4	0.113
p-value group1*Treatment = group2*Treatment	0.351
p-value group1*Treatment = group3*Treatment	0.515
p-value group2*Treatment = group3*Treatment	0.971

Notes: Groups selected from k-mean clustering on baseline covariates. Means are shown overall and for each group at baseline. The bottom panel shows p-values from comparing each treatment interaction against the others. Standard errors clustered at the regional level in parentheses. *** p < 0.01, ** p < 0.05, * p < 0.1.

(i.e., they are working mothers rather than stay-at-home mothers). Hence, these women likely experience genuinely worsened mental health perceptions when confronted with traditional expectations they are not satisfying.

Table 8
Descriptive statistics of male clusters.

	Group 1	Group 2	Group 3
Age	50.71 (10.59)	41.33 (10.66)	41.58 (14.24)
College degree or higher	0.00 (0.00)	0.65 (0.48)	0.33 (0.47)
Couple	0.89 (0.31)	0.53 (0.50)	0.17 (0.37)
Kids	1.00 (0.00)	0.33 (0.47)	0.13 (0.34)
Religious	0.74 (0.44)	0.58 (0.49)	0.52 (0.50)
Employed	0.85 (0.36)	1.00 (0.00)	0.00 (0.00)
Financial distress	0.63 (0.48)	0.50 (0.50)	0.65 (0.48)
Private sector	0.75 (0.43)	0.72 (0.34)	0.72 (0.45)
Self-employed	0.11 (0.31)	0.21 (0.41)	0.00 (0.00)
Medication	0.25 (0.43)	0.30 (0.46)	0.28 (0.45)
Psychologist	0.23 (0.42)	0.29 (0.45)	0.29 (0.46)
Good health	0.67 (0.47)	0.58 (0.48)	0.57 (0.50)
Mental health status	-0.35 (0.99)	-0.22 (1.01)	-0.03 (1.00)
Covid-19	0.37 (0.48)	0.35 (0.48)	0.26 (0.44)
Stereotype score	1.18 (1.51)	1.02 (1.51)	1.04 (1.60)
Profession score	0.23 (1.08)	0.07 (1.00)	0.08 (1.06)
Obs	407	641	156

Notes: Groups selected from k-mean clustering on baseline covariates for male respondents.

We repeated the same analysis for male respondents, who were clustered into three groups. As shown in Table 8, these groups are comparable in terms of their own stereotypical beliefs, but differ in demographic characteristics. Group 1 consists primarily of in couple, religious men with children and employed — men who largely conform to the traditional male stereotype. Group 2 comprises employed men who are typically in couple and religious, but only a minority have children, thus partially conforming to stereotype given the age range. Group 3, in contrast, consists of men who fail to conform: they are predominantly unemployed, not in couple, and with no children.

Table 9, Column (2) reports the heterogeneity analysis for male respondents, using Group 3 as the reference group. The treatment significantly increases perceived mental distress for this group. This finding is consistent with the interpretation that stereotypical cues confront these men with traditional masculine expectations they fail to meet, leading them to perceive their recent mental health status more negatively. This effect differs significantly from the effects on the other two groups. Groups 1 and 2 – comprising men who largely or partially conform to male stereotypes, respectively – are less affected by the treatment. This pattern again supports the interpretation that stereotype confrontation does not merely reduce disclosure propensity but genuinely affects how men perceive their mental health status. Men in Groups 1 and 2 report somewhat lower mental distress because they suffer less from stereotype confrontation, as they more closely align with traditional masculine expectations than Group 3.

Overall, the clustering analysis highlights distinct patterns that shed further light on the mechanisms underlying our main findings. For women, the strongest treatment effects emerge among those who partially conform to traditional female stereotypes (i.e., married working mothers) but hold low stereotypical beliefs. These women appear particularly vulnerable when confronted with gender role cues that highlight their incomplete conformity to traditional expectations. In

Table 9

Heterogeneity on depression – K-mean clustering for male.

	Male
Treatment	0.295*** (0.072)
Group1*Treatment	−0.235* (0.132)
Group2*Treatment	−0.247** (0.099)
Group 1	−0.181 (0.178)
Group 2	−0.071 (0.137)
Obs	1296
Mean group 1	−0.244
Mean group 2	−0.126
Mean group 3	0.129
p-value group1*Treatment = group2*Treatment	0.931

Notes: Groups selected from k-mean clustering on baseline covariates. Means are shown overall and for each group at baseline. The bottom panel shows p-values from comparing each treatment interaction against the others. Standard errors clustered at the regional level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

contrast, women who either fully conform to traditional roles (i.e., married stay-at-home mothers) or hold strong stereotypical beliefs show smaller responses, suggesting different underlying mechanisms at play. For men, the pattern is even more striking and confirms our baseline results. Men who fail to meet traditional masculine expectations (i.e., unemployed, not in couple, without children) experience the largest increases in self-reported mental distress when exposed to stereotypical imagery. Men who conform more closely to traditional male roles show significantly weaker responses. This asymmetry across conformity status provides further evidence that stereotype activation does not merely discourage disclosure among men but genuinely worsens their perception of their mental health when they fall short of traditional expectations. Taken together, these findings suggest that the psychological burden of stereotypes falls most heavily on individuals who fail to meet – or only partially meet – the traditional gender role expectations that stereotypical cues make salient.

6. Conclusion

While gender stereotypes are extensively documented as detrimental to women across multiple domains, including education, labor market outcomes, income, and health, structured evidence on their impact on male mental health remains limited. The main challenge in studying the effects of cultural norms on behavior and perceptions stems from the fact that culture does not change in the short run. To address this constraint, we designed an experiment that exogenously increases the salience of gender stereotypes among survey respondents.

Our findings show that making stereotypes more salient is associated with worse self-reported mental distress, with comparable effects for men and women. Treated respondents report a depression score 0.091 standard deviations higher than controls, an effect of similar size for women (0.079) and men (0.091); it is largest under the stereotypical condition, where it reaches 0.125 standard deviations overall (0.139 for women and 0.111 for men). The heterogeneity analysis sheds further light on these results. When we examine respondents one dimension at a time, the strongest response is reported among men who fail to meet core dimensions of traditional masculinity (such as the unemployed) and among men holding lower pre-existing stereotypical beliefs, suggesting that the psychological tension is most acute when external cues reactivate a model that respondents either fail to satisfy or explicitly reject. Among women, instead, the response is concentrated on those holding stronger pre-existing stereotypical beliefs. A multidimensional clustering analysis reinforces these findings. For men, the largest treatment effect (0.295 standard deviations) is observed among

those who fail to meet traditional masculine expectations along multiple reinforcing dimensions (i.e., unemployed, not in a couple, without children). For women, the largest effect (0.278 standard deviations) concerns those who only partially conform to the traditional female role (i.e., working mothers) while holding low stereotypical beliefs.

Overall, our findings suggest that stereotypes impose genuine psychological costs rather than merely affecting disclosure behavior. The burden falls most heavily on individuals who fall short of the gender norms made salient by the treatment, which is difficult to reconcile with a pure disclosure mechanism. Consistently, we observe a positive and statistically significant treatment effect even among respondents reporting good pre-treatment mental health status, that is, among those with relatively little prior distress to disclose and who may also include individuals unwilling to report their distress. If stereotype cues mainly affected reporting behavior, effects in this group would be expected to be null or even negative rather than positive. While we cannot rule out that disclosure also plays a role, the evidence suggests that it is unlikely to be the predominant mechanism.

CRedit authorship contribution statement

Paola Bertoli: Writing – original draft, Visualization, Investigation, Formal analysis, Data curation, Conceptualization. **Veronica Grembi:** Writing – original draft, Methodology, Investigation, Conceptualization.

Ethics declaration

Written informed consent to take part in the study and to publish the article has been obtained from all participants or their legal representatives. The privacy rights of participants have been observed.

This study was performed in compliance with relevant laws, regulatory frameworks and guidelines where the research took place. This study was approved by the Ethics Committee of the Scuola Inter-nazionale Superiore di Studi Avanzati (SISSA). (Approval No. 11542 - 13/05/2022)

Acknowledgment

The project was funded through research funds made available by our respective universities at the time (the University of Verona for Paola Bertoli and the University of Milan for Veronica Grembi).

Appendix A. Supplementary data

Appendix A and Appendix B available as supplementary material to this article can be found online at <https://doi.org/10.1016/j.ehb.2026.101638>.

References

- Addis, M.E., Mahalik, J.R., 2003. Men, masculinity, and the contexts of help seeking. *Am. Psychol.* 58 (1), 5–14.
- Alesina, A., Fuchs-Schundeln, N., 2007. Goodbye Lenin (or not?): The effect of communism on people's preferences. *Am. Econ. Rev.* 97 (4), 1507–1528.
- Alesina, A., Giuliano, P., 2015. Culture and institutions. *J. Econ. Lit.* 53 (4), 898–944.
- Astbury, J., 2001. Gender disparities in mental health. In: *Mental health. Ministerial Round Tables 2001*, 54th World Health Assembly.
- Bangasser, D., Valentino, R., 2014. Sex differences in stress-related psychiatric disorders: neurobiological perspectives. *Front. Neuroendocrinol.* 35, 303–319.
- Beaman, L., Chattopadhyay, R., Duflo, E., Pande, R., Topalova, P., 2009. Powerful women: Does exposure reduce bias? *Q. J. Econ.* 124 (4), 1497–1540.
- Bisin, A., Verdier, T., 2001. The economics of cultural transmission and the dynamics of preferences. *J. Econom. Theory* 97 (2), 298–319.
- Blair, I.V., Ma, J.E., Lenton, A.P., 2001. Imagining stereotypes away: the moderation of implicit stereotypes through mental imagery. *J. Pers. Soc. Psychol.* 81, 828–841.
- Boyd, A., Van de Velde, S., Vilagut, G., de Graaf, R., O'Neill, S., Florescu, S., Kovess-Masfety, V., 2015. Gender differences in mental disorders and suicidality in Europe: Results from a large cross-sectional population-based study. *J. Affect. Disord.* (173), 245–254.

- Brandt, M.J., 2011. Sexism and gender inequality across 57 societies. *Psychol. Sci.* 22 (11), 1413–1418.
- Chan, D., 2009. So why ask me? Are self-report data really that bad? In: Lance, C.E., Vandenberg, R.J. (Eds.), *Statistical and Methodological Myths and Urban Legends*. Routledge/Taylor, Greenwich, Connecticut, pp. 309–336.
- Chiu, C., Hong, Y., Ching-man, I., Fu, J., Yuk, J., Sau, V., 2018. Stereotyping and Self-Presentation: Effects of gender stereotype activation. *Group Process. Intergroup Relat.* 1 (1), 81–96.
- Cohen, S., Kamarck, T., Mermelstein, R., 1983. A global measure of perceived stress. *J. Health Soc. Behav.* 24, 386–396.
- Deane, F., Chamberlain, K., 1994. Treatment fearfulness and distress as predictors of professional psychological help-seeking intentions. *Br. J. Guid. Couns.* 22, 207–217.
- Ellemers, N., 2018. Gender stereotypes. *Annu. Rev. Psychol.* 69, 275–298.
- Finnegan, E., Oakhill, J., Garnham, A., 2015. Counter-stereotypical pictures as a strategy for overcoming spontaneous gender stereotypes. *Front. Psychol.* 50 (1291).
- Galdas, P., Cheater, F., Marshall, P., 2005. Men and health help-seeking behavior: Literature review. *J. Adv. Nurs.* 49, 616–622.
- Gold, P.W., 2015. The organization of the stress system and its dysregulation in depressive illness. *Mol. Psychiatry* 20 (1), 32–47.
- Gonzalez, L., Viitanen, T.K., 2009. The effect of divorce laws on divorce rates in Europe. *Eur. Econ. Rev.* 53 (2), 127–138.
- Komiya, N., Good, G., Sherrod, N., 2000. Emotional openness as a predictor of college students' attitudes toward seeking psychological help. *J. Couns. Psychol.* 47, 138–143.
- Kroenke, K., Spitzer, R.L., 2002. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatry Ann.* 32, 509–521.
- Kroenke, K., Spitzer, R., Williams, J., 2001. The PHQ-9: validity of a brief depression severity measure. *J. Gen. Intern. Med.* 16, 606–613.
- Kroenke, K., Strine, T.W., Spitzer, R.L., Williams, J.B.W., Berry, J.T., Mokdad, A.H., 2009. The PHQ-8 as a measure of current depression in the general population. *J. Affect. Disord.* 114, 163–173.
- Kuehner, C., 2017. Why is depression more common among women than among men? *Lancet Psychiatry* 4, 146–158.
- Levine, P.B., Staiger, D., Kane, T.J., Zimmerman, D.J., 1999. Roe v Wade and American fertility. *Am. J. Public Health* 89 (2), 199–203.
- Li, S., Graham, B., 2014. Why are women so vulnerable to anxiety, trauma-related and stress-related disorders? The potential role of sex hormones. *Lancet Psychiatry* 4, 73–82.
- Lindinger-Sternart, S., 2015. Help-seeking behaviors of men for mental health and the impact of diverse cultural backgrounds. *Int. J. Soc. Sci. Stud.* 3 (1).
- Magovcevic, M., Addis, M.E., 2008. The masculine depression scale: Development and psychometric evaluation. *Psychol. Men Masculinity* 9 (3), 117–132.
- Mahalik, J., Locke, B., Ludlow, L., Diemer, M., Scott, R., 2003. Development of the conformity to masculine norms inventory. *J. Clin. Psychol.* 6, 661–674.
- Mondo, M., Sechi, C., Cabras, C., 2021. Psychological distress, resilience, and COVID-19: Coping strategies among the general population. *Curr. Psychol.* <http://dx.doi.org/10.1007/s12144-021-01666-1>.
- Nunn, N., 2012. Culture and the historical process. *Econ. Hist. Dev. Reg.* 27 (sup1), S108–S126.
- Oram, S., Khalifeh, H., Howard, L., 2017. Violence against women and mental health. *Lancet Psychiatry* 4, 159–170.
- Parent, M.C., Hammer, J.H., Bradstreet, T.C., Schwartz, E.N., Jobe, T., 2018. Men's mental health Help-Seeking behaviors: An intersectional analysis. *Am. J. Men's Health* 12 (1), 64–73.
- Patnaik, A., 2019. Reserving time for daddy: The consequences of fathers' quotas. *J. Labor Econ.* 37 (4), 1009–1059.
- Prentice, D.A., Carranza, E., 2002. What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychol. Women Q.* 26 (4), 269–281.
- Riley, E., 2024. Resisting social pressure in the household using mobile money: Experimental evidence on microenterprise investment in Uganda. *Am. Econ. Rev.* 114 (5), 1415–1447.
- Seguino, S., 2011. Help or hindrance? Religion's impact on gender inequality in attitudes and outcomes. *World Dev.* 39 (8), 1308–1321.
- Shira, O., 2014. Time with children and employed parents' emotional well-being. *Soc. Sci. Res.* 47, 192–203.
- Stevenson, B., Wolfers, J., The paradox of declining female happiness. *Am. Econ. J.: Econ. Policy* 1 (2), 190–225. <http://dx.doi.org/10.1257/pol.1.2.190>.
- Tafet, G.E., Nemeroff, C.B., 2016. The links between stress and depression: Psychoneuroendocrinological, genetic, and environmental interactions. *J. Neuropsychiatry Clin. Neurosci.* 28 (2), 77–88.
- Tennant, C., 2001. Work-related stress and depressive disorders. *J. Psychosom. Res.* 51 (5), 697–704.
- Vogel, D., Wester, S., 2003. To seek help or not to seek help: The risk of self-disclosure. *J. Couns. Psychol.* 50, 351–361.
- Wang, P., Aguilar-Gaxiola, S., Alonso, J., Angermeyer, M., Borges, G., Bromet, E., Bruffaerts, R., de Girolamo, G., de Graaf, R., Gureje, O., Haro, J.M., Karam, E., Kessler, R., Kovess, V., Lane, M., Lee, S., Levinson, D., Ono, Y., Petukhova, M., Posada-Villa, J., Seedat, S., Wells, E., 2007. Worldwide use of mental health services for anxiety, mood, and substance disorders: Results from 17 countries in the WHO World Mental Health (WMH) surveys. *Lancet* 370, 841–850.
- Wang, Z., Chen, J., Boyd, J., Zhang, H., Jia, X., Qiu, J., Xiao, Z., 2011. Psychometric properties of the Chinese version of the perceived stress scale in policewomen. *PLoS One* 6 (12), e28610.
- Williams, R., Heinemann, A., Bode, R., Wilson, C., Fann, J., 2009. Improving measurement properties of the Patient Health Questionnaire-9 with rating scale analysis. *Rehabil. Psychol.* 54 (2), 198–203.