Recognition Thresholds and the Gender Gap in Depression*

Catalina Franco† Akshay Moorthy‡ Sara Abrahamsson§

February 27, 2026

Abstract

Worldwide, depression is more prevalent among women than men, but it is unclear whether this reflects differences in distress or in who is recorded as depressed. We find that the gap is smaller when depression is measured in representative surveys rather than epidemiological data, consistent with selection into measurement. We then conduct a pre-registered vignette experiment focusing on symptom interpretation and help-seeking. For equivalent symptoms, men are less likely than women to recognize depression and to seek help. The results indicate a higher self-classification threshold for men, implying that measured prevalence partly reflects behavior rather than underlying mental health.

JEL codes: I12, I14, I18, J16, C91
Keywords: Mental health; Gender differences; Depression recognition; Help-seeking behavior; Selection

*We acknowledge financial support from the Norwegian School of Economics (NHH), and by the Research Council of Norway through its Centres of Excellence funding scheme, project number 343454. We received invaluable feedback from audiences at FAIR, the University of Lausanne, the Workshop on Recent Advances in Behavioral Economics in Lyon and the Bogotá Experimental Economics Conference 2026, and individually from Marcella Alsan, Afsane Bjorvatn, Aline Bütikofer, Francesco Capozza, Janet Currie, Fabio Galeotti, Uri Gneezy, Anna Hochleitner, Amanda Kowalski, Synnøve Nesse, Petra Persson, Mattie Toma and Egon Tripodi. We are grateful to Egon Tripodi who generously shared data with us.

†Center for Applied Research (SNF) and FAIR at NHH – Norwegian School of Economics.
‡Faculty of Business and Economics, University of Lausanne.
§Division of Health Services at the Norwegian Institute of Public Health in Oslo, Norway.

1 Introduction

Mental health disorders are a leading cause of disability worldwide, affecting over one billion people and costing about USD 5 trillion annually (Rehm and Shield, 2019; Arias et al., 2022). One of the most persistent empirical patterns in mental health is the gender gap in depression: women are diagnosed at roughly twice the rate of men, a disparity documented since the 1970s and consistently observed across settings (Weissman and Klerman, 1977; Hyde et al., 2008; Van de Velde et al., 2010). This gap matters economically because mental health affects labor supply, absenteeism, and earnings (Goodman et al., 2011; Ridley et al., 2020; Biasi et al., 2021; Currie, 2024; Carvalho et al., 2026), potentially contributing to gender gaps in labor market outcomes. However, the diagnostic gap is difficult to interpret because it may capture differences in underlying distress as well as in the recognition and measurement of depression.

In this paper, we hypothesize that the gender gap partly arises from differences in the threshold at which symptoms are interpreted as depression. To appear in healthcare data, individuals must first recognize their symptoms as depression and act on that belief. If men require a higher threshold—more severe symptoms—to classify their distress as depression, they will be systematically underrepresented in diagnosed cases even at comparable symptom levels. Measured prevalence would then reflect both true underlying distress and selection into measurement.

Empirically testing this mechanism is challenging because standard data observe individuals only after recognition and healthcare contact. However, entry into measurement is nonrandom and affected by healthcare utilization, differential screening (Corredor-Waldron and Currie, 2024), insurance coverage, social norms (Seidler et al., 2016), perceived stigma (Roth et al., 2024a) and perceptions about treatment effectiveness (Roth et al., 2024b; Batmanov et al., 2026).

To address this challenge, we first examine whether the gender gap varies with the degree of selection into measurement by comparing prevalence ratios across sources that differ in how cases are recorded—after healthcare contact versus population screening. We then implement a pre-registered online experiment to test whether men and women differ in

the symptom severity required to recognize depression and seek help. Our main contribution is to provide evidence that part of the gender gap arises because men are less likely to appear in depression statistics and that differential recognition and help-seeking plausibly generate this selection.

We compare female-to-male prevalence ratios in the United States from the Global Burden of Disease Collaborative Network (2025), which synthesizes epidemiological and clinical data, to equivalent ratios constructed from population screening data in the National Health and Nutrition Examination Survey (CDC and NCHS, 2024) and an online screening U.S. sample from Prolific (Roth et al., 2024a,b). Under our hypothesis, the gender gap should be smaller in screening samples which measure symptoms directly rather than relying on recognition and healthcare contact.

Our first main result is that the female-to-male depression prevalence ratio declines when measured using survey data rather than aggregated epidemiological and clinical data. The ratio decreases from 1.7 in the Global Burden of Disease data to 1.4 in the NHANES sample and 1.3 in the Prolific sample. This pattern is consistent with lower selection into measurement in survey samples and suggests that a substantial share of men experiencing depressive symptoms may not appear in recorded prevalence statistics.

Next, we design an experiment to bypass underrecognition and underreporting potentially present in real-life depression cases, and implement it through an online survey with a representative sample of U.S. participants.[1] We create hypothetical scenarios, in which the severity and symptoms of depression, and the subject of the scenario (whether it is the participants themselves or a hypothetical female or male who experiences the symptoms) are randomized.

Specifically, each scenario is defined by a randomly selected PHQ-9 score and a corresponding set of symptoms and frequencies randomly drawn to match that score.[2] After reviewing each scenario, participants assess whether the symptoms indicate depression (recog-

[1]We recruit participants through Prolific, an online research platform commonly used in social science research (Peer et al., 2022).

[2]The PHQ-9 (Patient Health Questionnaire-9) is a nine-item depression screening tool commonly used in clinical and self-assessment settings (Kroenke et al., 2001). It is recommended as an initial step in evaluating depressive symptoms. Accessing and using the instrument, however, requires a degree of recognition that depression may exist—i.e., selecting into screening.

nition), its perceived severity, and the likelihood of seeking help. Comparing how men and women respond to randomly assigned symptom profiles allows for the identification of gender differences in recognition, accurate severity classification, and the likelihood of seeking help, and how this varies by depression severity.

Our second main result is that men are less likely than women to recognize depression and to seek specialist help, particularly at milder severity levels. Across hypothetical scenarios, each depicting varying degrees of depression, men are 5.3 percentage points (pp) less likely to classify symptoms as depression overall and 14.5 pp less likely in mild cases. Women, in turn, are more likely to overestimate symptom severity. Men are also less likely to seek help from specialist sources (6.1 pp overall and 10.6 pp in mild scenarios), a gap that persists even at higher severity levels.[3] The recognition gap increases monotonically with age, whereas we find no comparable age gradient in overestimation or help-seeking.

Turning to mechanisms, we find that psychic costs—the emotional and mental burden associated with admitting a mental health issue—appear to be an important driver of gender differences in help-seeking. When evaluating scenarios about themselves, men are 4.2 pp less likely than women to report willingness to seek help, whereas no gender difference arises when the scenario concerns others, indicating that men view treatment as appropriate but are less willing to apply it to themselves. Additional analyses examine whether masculinity norms shape how depressive symptoms are evaluated. Although estimates are imprecise, responses vary with the gender of the vignette subject in ways consistent with gendered interpretations. We do not find any evidence of gender differences in perceived social expectations, or perceptions regarding different aspects of depression.

Our paper revisits the long-standing literature documenting gender differences in depression as one of the most robust findings in psychopathology research (Hyde et al., 2008). Prior work attributes the gap to biological, psychological, and environmental factors (Hyde et al., 2008; Girgus and Yang, 2015; Call and Shafer, 2018).[4] Although this literature has

[3] In general, men are less than or equally likely as women to seek help from different specialist and non-specialist sources, with the exception of AI-enabled mental health chatbots, where men report a greater likelihood of seeking help from that source than women, which is consistent with previously documented gender gaps in AI adoption and use (Carvajal et al., 2024; Chatterji et al., 2025).

[4] These include biological differences (such as hormonal variation and genetic predisposition), coping mechanisms (e.g., rumination and need for approval), gendered manifestation of symptoms, and differential expo-

considered the possibility of differential recognition and reporting, such "artefactual" reasons have largely been dismissed (Girgus and Yang, 2015). We provide new evidence that selection into measurement and behavioral differences in recognition and help-seeking contribute meaningfully to observed prevalence differences. These findings highlight a behavioral margin that has received limited attention in the literature and point to new and actionable levers for public health policy.

To the best of our knowledge, no prior work in economics has directly examined behavioral drivers of *gender gaps* in mental health. Much of the literature studies the effectiveness of treatment (e.g., Baranov et al., 2020; Ridley et al., 2020; Bhat et al., 2022; Angelucci and Bennett, 2024), which by design presumes that screening or diagnosis has already occurred. More recent work considers how misinformation, perceived treatment effectiveness, and stigma affect treatment demand (Acampora et al., 2022; Roth et al., 2024a,b; Batmanov et al., 2026). We complement this literature by showing that measured prevalence and treatment uptake depend on how individuals interpret symptoms, highlighting a behavioral margin and a gender angle relevant for both clinical and informational interventions.

Our paper also contributes to the literature on selection in health behaviors (Oster, 2020), particularly work documenting selection into mental illness diagnoses (Nicoletti and Vidiella-Martin, 2025). Prior research shows that measured mental health conditions depend on contact with clinicians, diagnostic practices, and institutional incentives rather than solely on underlying health (e.g., Dalsgaard et al., 2014; Cuddy and Currie, 2020a,b; Currie and MacLeod, 2020), and recent studies examine interventions that increase treatment take-up (Breza et al., 2026). However, this literature typically cannot observe why individuals fail to be diagnosed or seek care. We provide evidence on symptom recognition and self-directed help-seeking as plausible behavioral mechanisms behind this margin.

The paper is structured as follows: Section 2 presents a comparison of the female-male prevalence gap from different data sources. Section 3 describes the experiment design and procedures. Section 4 presents the main analysis, complemented by an exploration of mechanisms in Section 5. Section 6 discusses implications of our findings.

sure to stressors (e.g., sexual abuse and social image concerns).

2 Selection into Measurement

Our premise is that recorded depression prevalence reflects not only underlying symptoms but also selection into measurement. Epidemiological and administrative data record cases only after individuals recognize symptoms and seek care, whereas large-scale screening surveys measure symptoms directly, regardless of prior recognition or healthcare contact. As a result, these sources differ in the extent to which cases require prior recognition and reporting, with representative screening surveys plausibly involving the least selection and non-survey data relying more heavily on recognized and diagnosed cases. If men apply a higher threshold when interpreting symptoms as depression, then conditional on comparable distress they will be less likely to enter measurement.

To assess whether selection contributes to the gender gap in measured depression, we compare female-to-male prevalence ratios across three sources: (i) estimates provided by the Global Burden of Disease Collaborative Network (2025), which synthesizes epidemiological and clinical information,[5] (ii) survey data from the 2021-2023 National Health and Nutrition Examination Survey (CDC and NCHS, 2024), which administers standardized screening instruments to a representative U.S. sample (N=6,337), and (iii) a large online screening sample collected on Prolific in 2022 (N=18,982) for the research reported in Roth et al. (2024a) and Roth et al. (2024b).[6]

Figure 1a presents the fraction of individuals classified as depressed in the NHANES 2021-2023. Two patterns emerge. First, the overall prevalence is higher for women (13.39%) than for men (9.34%). Second, the share classified as depressed declines sharply with age for both genders based on PHQ-9 scores.

Figure 1b plots female-to-male depression prevalence ratios across data sources and definitions of depression. Using GBD estimates (U.S. only), prevalence is 6.53% for women and 3.90% for men, equivalent to a ratio of 1.7, consistent with long-standing epidemiological evidence and depression rates based on administrative data on diagnoses. Using symptom-based screening with the standard PHQ-9 threshold for major depression (score $\geq$ 10), the

[5]The GBD combines multiple inputs including epidemiological surveys, clinical records, administrative sources, and scientific studies (Zhao et al., 2025).

[6]We thank the authors for their generosity in sharing their dataset.

ratio falls to 1.4 in the NHANES sample and 1.3 in the Prolific sample. Expanding the definition to include mild symptoms (PHQ-9 ≥ 4) further reduces the ratio to 1.3 in the NHANES sample and 1.1 in the Prolific sample, bringing the ratio close to gender parity. The gender rations shrink monotonically as measurement relies less on recognition and healthcare contact, consistent with men being less likely to enter recorded prevalence when identification requires self-recognition or care-seeking.[7]

Result 1. *Female-to-male depression prevalence ratios are larger in epidemiological estimates that rely on recognized cases than in symptom-based screening surveys, and they shrink further when mild symptoms are included, consistent with selection into measurement.*

3 Experimental Design

Motivated by these patterns, we design an experiment to test whether men and women respond differently to equivalent sets of depressive symptoms. We implement a within-subjects online study in which all participants evaluate four hypothetical scenarios. Because each respondent is exposed to comparable symptom profiles and context, differences in responses reflect recognition, interpretation, and intended behavior rather than variation in underlying distress. This design overcomes a key limitation of observational data, where men and women may experience or report different symptoms, preventing direct comparisons of recognition. By varying vignette features across scenarios, we characterize gender differences in classification thresholds and the role of contextual cues in help-seeking. Although responses to hypothetical scenarios may not perfectly predict real behavior (Loomis, 2011), the design avoids confounding factors such as selection into measurement and diagnosis, material constraints, and differential self-reporting.

We randomly generate depression scenarios, consisting of a set of depression symptoms and the frequency with which they occurred over a two-week period. The scenarios are based on the PHQ-9 instrument (Kroenke et al., 2001), a widely-used screening tool for depres-

[7]The prevalence rates by gender from NHANES and Prolific are reported in Appendix D and Figures A.1a-A.2b. As expected, they are higher than the GBD estimates because these surveys capture symptoms in the general population rather than only among individuals who seek care or screening. We focus on ratios rather than variation in overall prevalence because prevalence levels vary substantially across sources.

sion. The PHQ-9 assessment is comprised of 9 depression symptoms and their occurrence frequencies over the previous two weeks. The assessment results in a score ranging from 0 to 27, which indicates depression severity (higher scores are more severe) and whether further action is required. The scenarios are constructed using a randomization procedure that selects symptom items and frequencies to match a randomly selected "target" PHQ-9 score between 4 (minimal depression) and 21 (severe depression).[8] This generates quantitative variation in the severity of the scenario because of differences in the PHQ-9 scores, and qualitative variation based on differences in the symptom-frequency combinations.

Participants evaluate four scenarios in two randomized blocks, *Self* and *Other*. The blocks differ only in the subject of the hypothetical scenario. In *Self*, the subject of the scenario is the participant themselves. In *Other*, the subject is a hypothetical individual who is either male or female, randomized between-subjects. We signal the individual's gender using a name and provide basic background information that is held constant (see details in Appendix E). The symptoms shown in the second scenario of the *Other* block are identical for all participants, and is always presented at the end of the survey.

3.1 Main outcomes

After reviewing each hypothetical scenario, participants make three evaluations (Appendix F presents a complete list of survey questions). First, they state whether they think that the subject of the scenario is suffering from depression on a 4-point scale ranging from "Definitely yes" to "Definitely no" (no neutral option). We construct an indicator, *Recognition*, equal to 1 if the participant selects "Definitely yes" or "Probably yes", and 0 otherwise.

Second, participants assess symptom severity by choosing one of five categories: none or minimal, mild, moderate, moderately severe, or severe. These correspond to the diagnostic categories of the PHQ-9 instrument. We construct an indicator variable, *Accuracy*, equal to 1 if the participant's assessment matches the PHQ-9 severity category, and 0 otherwise.

Finally, participants assess how likely the subject of the scenario would be to seek help

[8] We exclude the extreme ends of the scale to focus statistical power on scenarios that are more ambiguous. See Appendix Section C for the complete symptom list and severity classification and Appendix Section G for the randomization procedure.

from each of six sources.[9] These assessments are aggregated into an indicator, *Seek help*, which is equal to 1 if the participant thinks that the subject of the scenario would seek help from at least one of the six sources. We also construct separate indicators for seeking help from (i) specialist, and (ii) non-specialist sources.

Together, these outcomes provide a comprehensive overview of the different stages of recognition and help-seeking behavior that may shape recorded depression prevalence. We complement these outcomes with additional survey questions to study potential mechanisms, including psychic costs, perceptions, and norms (see Section 5).

3.2 Study procedures and sample descriptives

The experiment was conducted with a sample of 401 U.S. participants recruited through the online survey platform Prolific. Each participant evaluated four scenarios for a total sample size of 1,604 scenario evaluations. The sample is representative of the U.S. population along gender, age, and ethnicity. The experiments were built using oTree (Chen et al., 2016). The survey was conducted in October 2024, and we pre-registered the design and hypotheses (AEARCTR-0014621) before data collection. The study was reviewed and approved by the IRB at the Norwegian School of Economics. Our analysis largely follows the pre-analysis plan (see Appendix I); any deviations or exploratory analyses are indicated in the text.

We collected participants' background characteristics and their prior experience with mental health screening tools. Of the 1,604 scenario evaluations we analyze,[10] 820 (51.1%) were completed by women. The mean age for women (men) in the sample is 46 (45) years old, 43% of women (34.7% of men) have less than a bachelor's degree, 59.5% of women (75% of men) are employed, 53.2% of women (62.8% of men) have an annual household income of at least US $50,000, and 21% of women (40% of men) report not being familiar with depression screening tools at the time of the survey.

Figure A.3 shows a histogram of the PHQ-9 scores across the 1,604 scenarios. A quarter of the scenarios have a score of 4, corresponding to the scenario that is presented to all

[9]The sources consist of both specialist (General practitioner (GP) for this purpose, GP for another purpose, therapy) and non-specialist options (counselor at workplace or school, friend or relative, AI-enabled mental health chatbot).

[10]We exclude 44 evaluations from 11 individuals who reported non-binary gender.

participants. The distribution of scenarios is relatively uniform between scores of 5 and 21, consistent with the randomization of target PHQ-9 scores.

4 Analysis Of Gender Differences

We begin the analysis by examining the scenario-level responses. We first present binned scatter plots of the three main outcomes (depression recognition, accurate severity and help seeking) against the scenario PHQ-9 score, separately by the respondent gender, and controlling for order, scenario type, and treatment assignment.[11]

We complement these results with regression estimates for each outcome on (i) a gender indicator (*Male* equals 1 if the participant is male; see Equation 1) and (ii) the gender indicator interacted with scenario severity categories (see Equation 2). All regressions include demographic controls (age, education, employment, income level, and familiarity with depression screening tools), indicators for mild, moderate, and moderately severe symptoms, with minimal symptoms as the omitted category, and treatment and order controls. In the second specification, we interact each severity indicator with *Male*.[12] Standard errors are heteroscedasticity-robust and clustered at the participant level.

$$y_i = \alpha_0 + \alpha_1 \text{Male}_i + \gamma' X_i + \varepsilon_i \tag{1}$$

$$\begin{aligned} y_i = &\beta_0 + \beta_1 \text{Male}_i + \beta_2 \text{Mild}_i + \beta_3 \text{Male}_i \times \text{Mild}_i + \beta_4 \text{Moderate}_i + \\ &\beta_5 \text{Male}_i \times \text{Moderate}_i + \beta_6 \text{Modsevere}_i + \beta_7 \text{Male}_i \times \text{Modsevere}_i + \gamma' X_i + \varepsilon_i \end{aligned} \tag{2}$$

4.1 Depression recognition

Our first finding from the experiment is that women are more likely than men to classify a given symptom profile as depression. Figure 2a shows a binned scatterplot of depression recognition against the PHQ-9 score, separately by gender. Two patterns emerge. First,

[11] Specifically, Self vs. Other scenarios, and whether the subject in the *Other* scenario is a male or a female. Since the fourth scenario is the same for all participants, we control for whether *Self* or *Other* scenarios were presented first instead of the within-survey sequence.

[12] In the PAP, we intended to use indicators for Moderate, Moderately severe and Severe, with excluded indicator Mild. However, since we also have a scenario that depicts minimal depression (PHQ-9 score equal to 4) that all participants evaluated, we decided to include it in the analysis and changed the omitted category to Mild. We also combined the categories Moderately severe and Severe for expositional clarity.

recognition is generally high, as expected given that all scenarios depict depression: over 90% of women and 70% of men label the symptoms as depression. Second, recognition increases with symptom severity for both genders, but the gradient is steeper for men. At moderate to severe levels (PHQ-9 scores of 15–21), nearly all respondents, regardless of gender, recognize depression. At minimal to mild levels (PHQ-9 scores of 4–10), a gender gap emerges, with women significantly more likely to recognize depression than men.

Table 1 quantifies these differences. Column (1) shows an average gender gap of 5.3 pp, relative to a base of 95% of women recognizing depression. The gap is substantially larger at minimal depression severity (14.5 pp, or 17% relative to women), where 87% of women identify the scenario as depression (column (2)). Recognition rates increase sharply with severity for both genders, and are close to universal at moderate and higher severity levels. Men show a larger increase in recognition across these categories, as reflected in the positive interaction terms between male and depression severity.

Exploratory analyses show that the gender gap in recognition also varies by age: men under 30 exhibit similar recognition rates as women, whereas the gap is substantially larger at older ages (Figure 2b).

Result 2. *Women are more likely than men to classify equivalent symptom profiles as depression, consistent with a lower recognition threshold.*

Overall, the recognition gap is driven by men at lower symptom severity and by older men. The age gradient patterns are consistent with previous evidence on older men's lower mental-health literacy (Farrer et al., 2008) and greater stigma toward depression (Möller-Leimkühler, 2002).

4.2 Accuracy in recognizing depression severity

Our second pre-registered outcome measures whether individuals accurately classify the severity of depression indicated by the symptoms in a given scenario. Misclassification adds an additional margin to recognition failures: systematically underestimating severity may reduce help-seeking, while overestimating severity may contribute to excessive concern or overdiagnosis.

Figure 3 presents binned scatterplots of severity classification accuracy against scenario PHQ-9 score. Panel 3a shows that accurate classification is generally low, particularly at lower severity levels: fewer than 20% of assessments match the PHQ-9 severity category for scenarios with scores up to 10 (minimal or mild depression). Accuracy improves with severity but remains limited: only ≈50% of assessments are correct even for the most severe scenarios. Panels 3b and 3c decompose misclassification into over- and underestimation. Most errors stem from overestimating severity, especially at lower PHQ-9 scores.

A second pattern is that men are significantly more accurate than women in classifying severity throughout the distribution. Overall, men are 6.8 pp more accurate in their depression severity assessments than women, a 40% difference relative to a baseline accuracy rate of 18% among women (Table 1, column (3)). This gender gap in severity classification persists across the full range of PHQ-9 scores (Table 1, Col. (4)). We do not find an age gradient in accuracy: men are consistently more accurate than women across all ages (Figure A.4a).

Result 3. *Men classify depression severity more accurately than women for identical symptom profiles, and this accuracy gap persists across the PHQ-9 severity range.*

The fact that men classify severity more accurately but are less likely to recognize depression appears contradictory. Figure A.5 shows that overestimation is concentrated among participants who recognize the vignette as depression. Because men are less likely to recognize depression—especially at lower PHQ-9 levels where overestimation is most common—they are mechanically less likely to overestimate severity. Thus, men's higher average accuracy reflects both differences in recognition and better calibration of symptom severity.

4.3 Willingness to seek help

Our third pre-specified outcome measures the perceived likelihood of help-seeking after viewing the symptom list. This question was asked regardless of whether respondents correctly recognized the symptoms as depression, since individuals may be inclined to seek help even without labeling the condition. Participants considered six potential sources of help, which we classify as specialist or non-specialist. The measure is a proxy for active help-seeking rather than routine screening by healthcare providers.

Figure 4 presents binned scatterplots of the perceived likelihood of seeking help against the scenario PHQ-9 scores, pooling specialist and non-specialist sources in Panel 4a, and shown separately in Panels 4b and 4c. Overall, participants think help-seeking is likely in over 90% of the scenarios, with little variation by severity. Men and women exhibit similar patterns, though men report slightly lower help-seeking rates. However, confidence intervals for the two groups overlap at several points across the PHQ-9 distribution (Panel 4a). Columns (5) and (6) in Table 1 show that the point estimates are negative but only significant at the 10% level, suggesting no evidence of gender differences in help-seeking behavior.

While overall help-seeking is similar by gender, the sources to which men and women prefer to turn for help differ. Panels 4b and 4c show substantially higher likelihoods of help-seeking from specialists (general practitioners and therapists) than non-specialist sources (counselors, friends, or AI chatbots), particularly among women. Striking gender differences appear in the likelihood of seeking help from specialist sources. Women report being likely to seek specialist help in 95% of scenarios, regardless of depression severity. In contrast, men's likelihood is lower at mild severity (around 80%), and increases with severity, reaching parity with women only at the most severe depression levels. For non-specialist sources, the only source from which men would be more likely than women to seek help are AI-enabled mental health chatbots (see Figure A.6). We do not detect an age gradient in the likelihood of seeking help (Figure A.4b).

The gender difference in seeking help from specialist sources is quantified in Table B.1. Column (1) shows that the overall gender gap is 6.1 pp or 6.4% relative to a base of 95% of women reporting that the subject of the scenario would seek help. In Column (2), we see that the gender gap in seeking help from specialist sources is driven by the mild scenarios, where men are 10.6 pp or 11% less likely than women to report seeking help from specialist sources in scenarios with PHQ-9 scores between 5 and 10. However, the large and positive interaction coefficient on Male and Moderately severe scenario indicates that the gender gap in seeking help from specialist sources would close when depression is severe enough. There are no significant coefficients on Columns (3) and (4) corresponding to seeking help from non-specialist sources.

Result 4. *The likelihood of help-seeking is high for both genders, but men report a lower likelihood of seeking specialist help at lower and moderate levels of symptom severity.*

Our findings suggest that while the likelihood of help-seeking is high on average, gender differences emerge in the type of help sought. Moreover, the age gradient appears in recognition but not in help-seeking, suggesting that the primary barrier arises at the stage of identifying symptoms as depression rather than in seeking care conditional on recognition. These patterns suggest that gender differences in recorded prevalence are more plausibly driven by recognition at milder severity levels than by reluctance to obtain treatment.

4.4 Robustness checks

We conduct three robustness exercises to address concerns that factors such as survey fatigue, overattention to certain symptoms and failure to recognize depression might affect participant responses. First, we restrict the sample to the first two scenarios that participants evaluate, which limits survey fatigue and the scope for earlier scenarios to influence later responses. Second, we drop scenarios that include the suicidal ideation item (Q9 in the PHQ), to ensure that recognition is not disproportionately driven by this particularly salient item. Third, we restrict the sample to scenarios in which participants classify the scenario as depression, to assess whether subsequent responses are sensitive to recognition failures. We briefly discuss concerns related to social desirability bias in Subsection H.1.

Appendix Tables B.2 - B.4 report the results of these exercises, and show that the estimates in Table 1 are robust to these concerns. Two details are worth mentioning. First, restricting the estimation sample to the first two scenarios (Table B.2) excludes the scenario with a PHQ-9 score of 4; consequently, the minimal category is not a part of the specification and the mean recognition rate for women is 98%. The gender gap in recognition is 8.8 pp (9%), which is smaller than the 16.6% gap observed when the minimal scenario is included. Second, the gender gaps in recognition (accuracy) across severity levels increase (decrease) slightly when scenarios including the PHQ-9 item on suicidal ideation are excluded (Table B.3). Together, these analyses support the main results and suggest that gender differences are more pronounced when symptoms are more ambiguous.

5 Mechanisms

We investigate three plausible mechanisms: perceptions, gender norms, and psychic costs.

5.1 Psychic costs of recognition and help-seeking

Recognizing one's own mental health problems and seeking treatment can impose psychic costs (Cronin et al., 2024), driven by emotions such as shame, guilt, or denial, and amplified by social image concerns (Smith, 2023). We hypothesize that these costs are higher when evaluating oneself than an unknown other, and that they differ by gender. We use within-subject variation between *Self* and *Other* scenarios to test whether gender gaps in recognition reflect men identifying depression more readily in others than in themselves.

Columns (1), (3), and (5) of Appendix Table B.5 present estimates of Equation 1 including an indicator for the *Other* treatment and its interaction with the Male indicator, for the three main outcomes. For recognition and accuracy (Columns (1) and (3)), there is no evidence that the gender gap changes between *Self* and *Other* scenarios. For help-seeking, a pattern consistent with psychic costs emerges: in *Self* scenarios, men are 4.2 pp less likely to seek help than women, while in the *Other* scenarios the interaction term offsets this gap, implying no gender difference in the likelihood of seeking help. Additional analyses (Appendix Table B.6) suggest that men may be more likely to seek help from specialist sources in *Other* than in *Self* scenarios.

5.2 Norms

We hypothesize that gender norms shape men's recognition of depression and likelihood of seeking help by increasing the perceived cost of acknowledging weakness or vulnerability. Norms that emphasize toughness, self-reliance, and emotional restraint may discourage men from recognizing or seeking help for depression (De Haas et al., 2024).

We test the role of norms in two ways. First, we examine whether responses to *Other* vignettes vary with the gender of the vignette subject (male vs. female), holding symptoms constant. This captures differential evaluation of equivalent sets of symptoms across gendered targets. Second, we elicit incentivized second-order beliefs about whether other par-

ticipants would classify a vignette as depression, which measures perceived social expectations within the study sample.

In Columns (2), (4), and (6) of Appendix Table B.5, we report regressions for the three main outcomes using the *Other* scenarios only, with indicators for participant and subject gender, and their interaction. If normative expectations about how men should feel or behave shape men's lower recognition, we would expect to see lower recognition when the subject in *Other* is male, particularly among male respondents. Across outcomes, men are more likely to recognize depression, less accurate, and more likely to indicate help-seeking when the scenario subject is a man than when it is a woman. If anything, these patterns would suggest that men are less sensitive to depression symptoms for women relative to men.

The second test uses an incentivized second-order belief question, in which participants guess the fraction of participants who recognized a given scenario as depression.[13] The scenario was held constant across all participants, while the gender of the subject was randomized. If perceived social expectations drive men's lower recognition of depression, we would expect male participants to make lower estimates than females. We find no gender difference in second-order beliefs – the distribution of male and female participants' beliefs are very similar (Figure A.7; KS test: $p = 0.27$).

To summarize, we do not find any evidence of a difference in perceived social expectations between men and women. However, we do find evidence that the gender of the scenario subject shifts men's evaluations of identical symptoms.

5.3 Perceptions

Finally, we examine whether men and women differ in their perceptions of six factors related to: (i) the consequences of depressive symptoms, (ii) social image concerns and stigma, or (iii) the effectiveness of treatment options and their side effects. We use responses to survey questions asked after each of the first three scenarios (see Appendix F). We generate binary outcomes for each of the six sub-items, coded as 1 if the respondent somewhat or strongly

[13]The scenario corresponds to a PHQ-9 score of 4 (mild depression). Details are provided in Appendix F. We deviated from our pre-registration where we stated that this scenario would have a PHQ-9 score of 11.

agrees with a statement.

We find no evidence of gender differences in perceptions about the scenario. Both graphical evidence (presented in Figure A.8) and regression estimates (Figure A.9) fail to find any significant differences. Although gender differences in perceptions are largely absent, the overall levels of agreement with the statements reveal some informative patterns. Nearly half of the participants believed that the issues would resolve on their own, a third view them as not requiring treatment, and large majorities express concerns about stigma, treatment effectiveness, and medication side effects. These patterns align with recent evidence on these issues (Acampora et al., 2022; Roth et al., 2024b; Cronin et al., 2024). We discuss these patterns in more detail in Appendix H.

6 Discussion

Gender differences in depression diagnoses are well documented, yet little is known about the behavioral mechanisms underlying these gaps. We provide evidence on two such mechanisms: differences in recognition and in specialist help-seeking thresholds. Interpreted through the lens of diagnostic errors, men appear more prone to Type II errors (under-recognition), whereas women appear more prone to Type I errors, interpreting mild symptoms as severe.

Clinical guidelines recommend monitoring or low-intensity interventions for mild depression and structured treatment for moderate or severe cases (National Collaborating Centre for Mental Health, 2010). Within this framework, the two error types have distinct implications: Type II errors may delay early intervention and allow symptoms to escalate into more severe and costly episodes, whereas Type I errors may lead to premature use of specialist services, congestion, and overtreatment. The welfare implications depend on how society and policymakers trade off underdiagnosis against overtreatment.

A central implication of our findings is that part of the observed gender disparity reflects differences in recognition and help-seeking rather than underlying distress. Because our vignette design abstracts from insurance, access, material constraints, and variation in actual

mental health conditions, the margins we identify are behavioral rather than institutional or rooted in deeper determinants of mental health disparities. Our findings therefore highlight interpretation and response to symptoms as an independent source of gender disparities in measured mental health.

References

Acampora, Michelle, Francesco Capozza, and Vahid Moghani, "Mental Health Literacy, Beliefs and Demand for Mental Health Support among University Students," Technical Report, Tinbergen Institute Discussion Paper 2022.

Angelucci, Manuela and Daniel Bennett, "The economic impact of depression treatment in india: Evidence from community-based provision of pharmacotherapy," *American economic review*, 2024, *114* (1), 169–198.

Arias, Daniel, Shekhar Saxena, and Stéphane Verguet, "Quantifying the global burden of mental disorders and their economic value," *EClinicalMedicine*, 2022, *54*.

Baranov, Victoria, Sonia Bhalotra, Pietro Biroli, and Joanna Maselko, "Maternal depression, women's empowerment, and parental investment: Evidence from a randomized controlled trial," *American economic review*, 2020, *110* (3), 824–859.

Batmanov, Alisher, Idaliya Grigoryeva, Bruno Calderón-Hernández, Roberto González-Téllez, and Alejandro Guardiola Ramírez, "Beliefs, information sharing, and mental health care use among university students," *Journal of Development Economics*, 2026, *180*, 103646.

Bhat, Bhargav, Jonathan De Quidt, Johannes Haushofer, Vikram H Patel, Gautam Rao, Frank Schilbach, and Pierre-Luc P Vautrey, "The long-run effects of psychotherapy on depression, beliefs, and economic outcomes," Technical Report, National Bureau of Economic Research 2022.

Biasi, Barbara, Michael S Dahl, and Petra Moser, "Career effects of mental health," Technical Report, National Bureau of Economic Research 2021.

Breza, Emily, Kevin Carney, Vijaya Raghavan, Kailash Rajah, Thara Rangaswamy, Gautam Rao, Frank Schilbach, Sobia Shadbar, and James Stratton, "Financial Incentives, Health Screening, and Selection into Mental Health Care: Experimental Evidence from

College Students in India," Technical Report, National Bureau of Economic Research 2026.

Bütikofer, Aline, Rita Ginja, Krzysztof Karbownik, and Fanny Landaud, "(Breaking) intergenerational transmission of mental health," *Journal of Human Resources*, 2024, *59* (S), S108–S151.

Call, Jarrod B and Kevin Shafer, "Gendered manifestations of depression and help seeking among men," *American Journal of Men's Health*, 2018, *12* (1), 41–51.

Carvajal, Daniel, Catalina Franco, and Siri Isaksson, "Will Artificial Intelligence get in the way of achieving gender equality?," *NHH Dept. of Economics Discussion Paper*, 2024, (03).

Carvalho, Leandro, Damien de Walque, Crick Lund, Heather Schofield, Vincent Somville, and Jingyao Wei, "Psychological barriers to participation in the labor market: Evidence from rural Ghana," *Journal of Development Economics*, 2026, p. 103734.

CDC and NCHS, "National Health and Nutrition Examination Survey Data," U.S. Department of Health and Human Services, Centers for Disease Control and Prevention 2024. Accessed: 2026-01-15.

Chatterji, Aaron, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman, "How People Use ChatGPT," Technical Report, National Bureau of Economic Research 2025.

Chen, Daniel L, Martin Schonger, and Chris Wickens, "oTree—An open-source platform for laboratory, online, and field experiments," *Journal of Behavioral and Experimental Finance*, 2016, *9*, 88–97.

Choi, Han, Adriana Corredor-Waldron, Janet Currie, and Chris Felton, "What Can Trends in Emergency Department Visits Tell Us About Child Mental Health?," *Journal of Human Resources*, 2025.

Corredor-Waldron, Adriana and Janet Currie, "To what extent are trends in teen mental health driven by changes in reporting?: The example of suicide-related hospital visits," *Journal of Human Resources*, 2024, *59* (S), S14–S40.

Cronin, Christopher J, Matthew P Forsstrom, and Nicholas W Papageorge, "What good are treatment effects without treatment? Mental health and the reluctance to use talk therapy," *Review of Economic Studies*, 2024.

Cuddy, Emily and Janet Currie, "Rules vs. discretion: Treatment of mental illness in us adolescents," Technical Report, National Bureau of Economic Research 2020.

_ **and** _ , "Treatment of mental illness in American adolescents varies widely within and across areas," *Proceedings of the National Academy of Sciences*, 2020, *117* (39), 24039–24046.

Currie, Janet, "The Economics of Child Mental Health: Introducing the Causes and Consequences of Child Mental Health Special Issue," *Journal of Human Resources*, 2024, *59* (S), S1–S13.

_ **and Mark Stabile**, "Child mental health and human capital accumulation: the case of ADHD," *Journal of Health Economics*, 2006, *25* (6), 1094–1118.

_ **and** _ , "Mental Health in Childhood and Human Capital," 2009.

Currie, Janet M and W Bentley MacLeod, "Understanding doctor decision making: The case of depression treatment," *Econometrica*, 2020, *88* (3), 847–878.

Dalsgaard, Søren, Maria Knoth Humlum, Helena Skyt Nielsen, and Marianne Simonsen, "Common Danish standards in prescribing medication for children and adolescents with ADHD," *European child & adolescent psychiatry*, 2014, *23* (9), 841–844.

de Velde, Sarah Van, Piet Bracke, and Katia Levecque, "Gender differences in depression in 23 European countries. Cross-national variation in the gender gap in depression," *Social Science & Medicine*, 2010, *71* (2), 305–313.

Farrer, Louise, Liana Leach, Kathleen M Griffiths, Helen Christensen, and Anthony F Jorm, "Age differences in mental health literacy," *BMC public health*, 2008, *8* (1), 125.

Fletcher, Jason M, "Adolescent depression and educational attainment: results using sibling fixed effects," *Health economics*, 2010, *19* (7), 855–871.

Girgus, Joan S and Kaite Yang, "Gender and depression," *Current Opinion in Psychology*, 2015, *4*, 53–60.

Global Burden of Disease Collaborative Network, "Global Burden of Disease Study 2023 (GBD 2023)," 2025. Accessed: 2026-02-10.

Goodman, Alissa, Robert Joyce, and James P Smith, "The long shadow cast by childhood physical and mental problems on adult life," *Proceedings of the National Academy of Sciences*, 2011, *108* (15), 6032–6037.

Haas, Ralph De, Victoria Baranov, Ieda Matavelli, and Pauline Grosjean, "Masculinity around the world," *Work. Pap.*, 2024.

Hyde, Janet Shibley, Amy H Mezulis, and Lyn Y Abramson, "The ABCs of depression: integrating affective, biological, and cognitive models to explain the emergence of the gender difference in depression.," *Psychological review*, 2008, *115* (2), 291.

Kroenke, Kurt, Robert L Spitzer, and Janet BW Williams, "The PHQ-9: validity of a brief depression severity measure," *Journal of general internal medicine*, 2001, *16* (9), 606–613.

Loomis, John, "What's to know about hypothetical bias in stated preference valuation studies?," *Journal of Economic Surveys*, 2011, *25* (2), 363–370.

Möller-Leimkühler, Anne Maria, "Barriers to help-seeking by men: a review of sociocultural and clinical literature with particular reference to depression," *Journal of affective disorders*, 2002, *71* (1-3), 1–9.

National Collaborating Centre for Mental Health, "Depression: the treatment and management of depression in adults (updated edition)," in "in" British Psychological Society 2010.

Nicoletti, Cheti and Joaquim Vidiella-Martin, "ADHD, school performance, and economic outcomes," *Oxford Research Encyclopedia of Economics and Finance*, 2025.

Oster, Emily, "Health recommendations and selection in health behaviors," *American Economic Review: Insights*, 2020, *2* (2), 143–160.

Peer, Eyal, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer, "Data quality of platforms and panels for online behavioral research," *Behavior research methods*, 2022, *54* (4), 1643–1662.

Rehm, Jürgen and Kevin D Shield, "Global burden of disease and the impact of mental and addictive disorders," *Current psychiatry reports*, 2019, *21*, 1–7.

Ridley, Matthew, Gautam Rao, Frank Schilbach, and Vikram Patel, "Poverty, depression, and anxiety: Causal evidence and mechanisms," *Science*, 2020, *370* (6522), eaay0214.

Roth, Christopher, Peter Schwardmann, and Egon Tripodi, "Depression stigma," 2024.

_ , _ , and _ , "Misperceived effectiveness and the demand for psychotherapy," *Journal of Public Economics*, 2024, *240*, 105254.

Seidler, Zac E, Alexei J Dawes, Simon M Rice, John L Oliffe, and Haryana M Dhillon, "The role of masculinity in men's help-seeking for depression: a systematic review," *Clinical psychology review*, 2016, *49*, 106–118.

Smith, Emma C, "Stigma and Social Cover: A Mental Health Care Experiment in Refugee Networks," Technical Report, Working Paper 2023.

Smith, James Patrick and Gillian C Smith, "Long-term economic costs of psychological problems during childhood," *Social science & medicine*, 2010, *71* (1), 110–115.

Weissman, Myrna M and Gerald L Klerman, "Sex differences and the epidemiology of depression," *Archives of General Psychiatry*, 1977, *34* (1), 98–111.

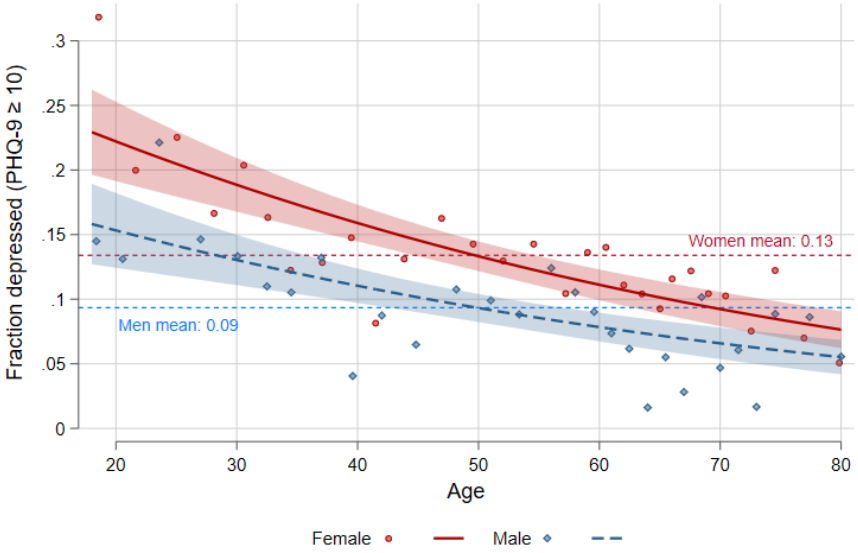
Zhao, Le, Yan Lou, Yuexian Tao, Hangsai Wang, and Nan Xu, "Global, regional and national burden of depressive disorders in adolescents and young adults, 1990–2021:

systematic analysis of the global burden of disease study 2021," *Frontiers in Public Health*, 2025, *13*, 1599602.

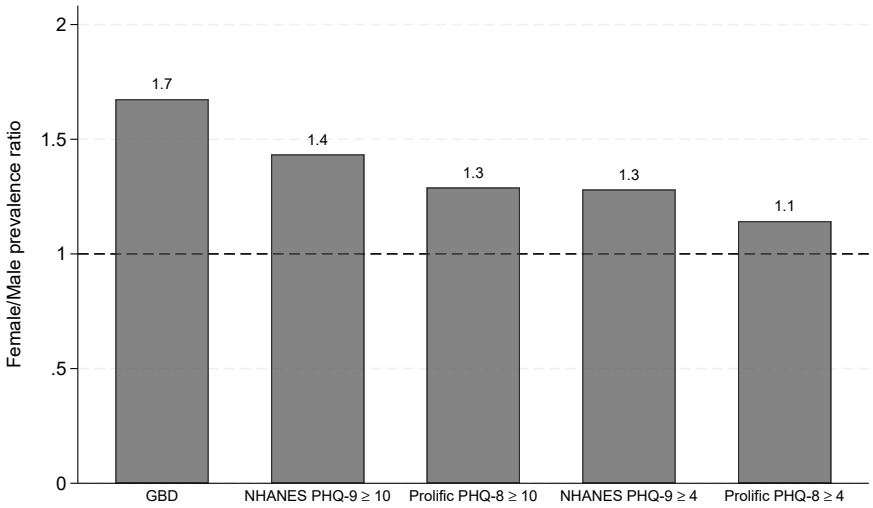
7 Figures

Figure 1: Depression prevalence by age and gender and female/male prevalence ratio from different sources

(a) Depression prevalence in NHANES



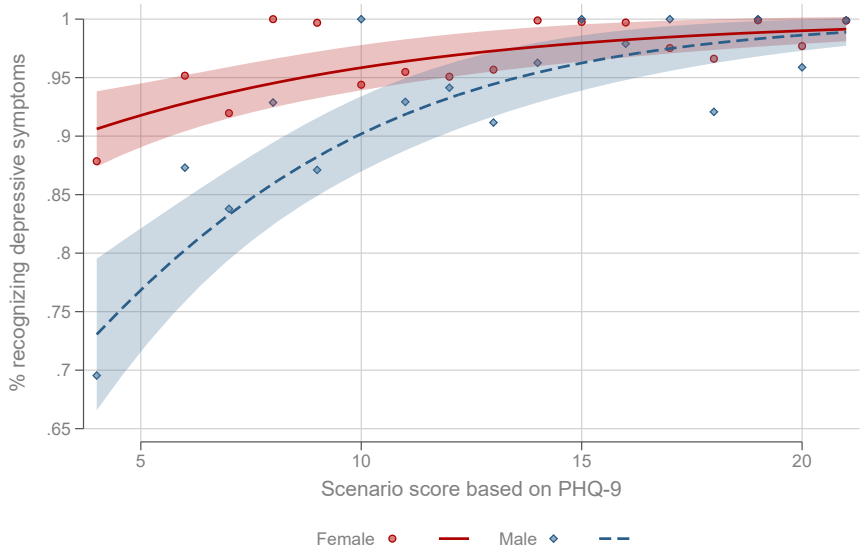
(b) Prevalence ratios across samples



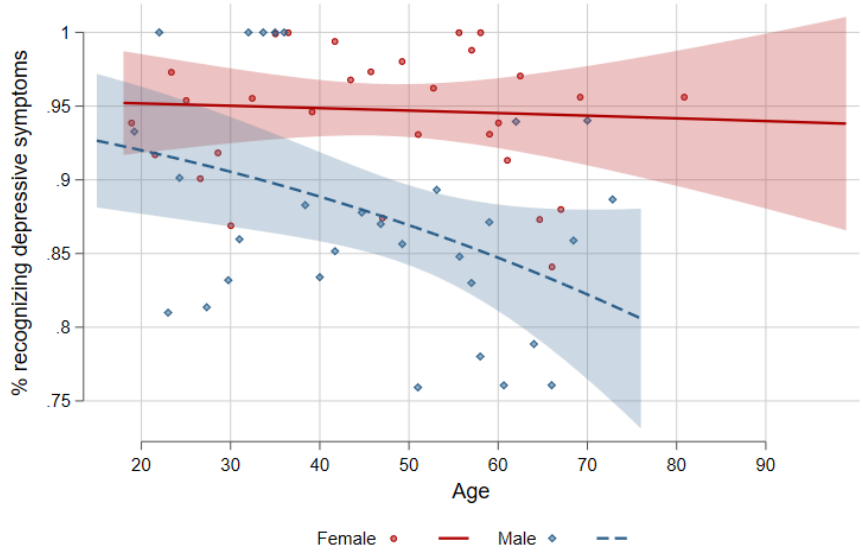
Notes. Panel A plots the prevalence of depressive symptoms by age and gender in the National Health and Nutrition Examination Survey (NHANES 2021–2023, N=6,337). Depression is measured using the PHQ-9 two-week symptom scale. Points denote binned scatterplot means and bands indicate 95% confidence intervals. Panel B reports the female-to-male prevalence ratio across three data sources: (i) Global Burden of Disease estimates commonly used in cross-country comparisons (Global Burden of Disease Collaborative Network, 2025); (ii) NHANES 2021–2023, N=6,337 (CDC and NCHS, 2024); and (iii) a screened online sample from Prolific used in Roth et al. (2024a,b), N=18,982. Ratios above one indicate higher prevalence among women. For NHANES and the Prolific samples, prevalence is constructed using two PHQ-9 cutoffs: ≥ 10 (standard clinical threshold for major depression) and ≥ (including mild symptoms).

Figure 2: Gender differences in depression recognition

(a) Recognition gap



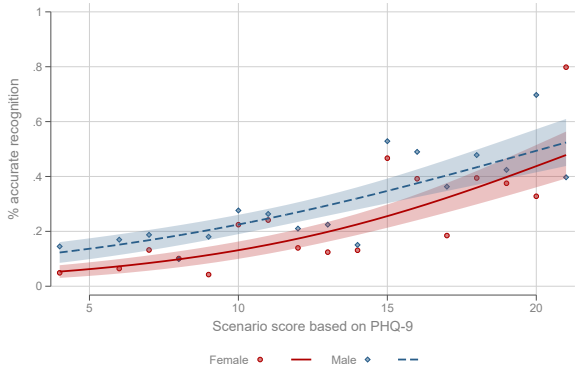
(b) Age gradient in recognition



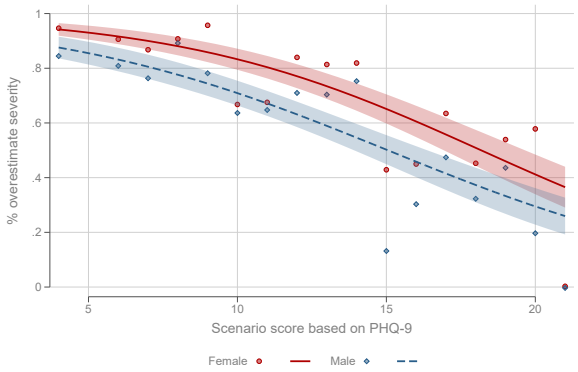
Notes. Panel A plots a binned scatterplot of depression recognition and the severity (PHQ-9) score of the hypothetical scenario, shown separately for male and female participants. Panel B plots the recognition measure by gender and age. The recognition of depressive symptoms binary variable is based on the question: Suppose that [you/NAME] were/was experiencing the hypothetical issues at the frequencies listed above, do you think [you/NAME] would have depression? *[Definitely yes, probably yes, probably no, definitely no]*. Each point in the plots represents the regression coefficient from a regression of recognition on the scenario severity, controlling for treatment assignment (self vs. other), whether the subject in the "other" scenario is a male or a female and the order in which the scenarios were presented. The bands show 95% confidence intervals, with standard errors clustered by participant.

Figure 3: Gender differences in identifying the severity of depressive symptoms

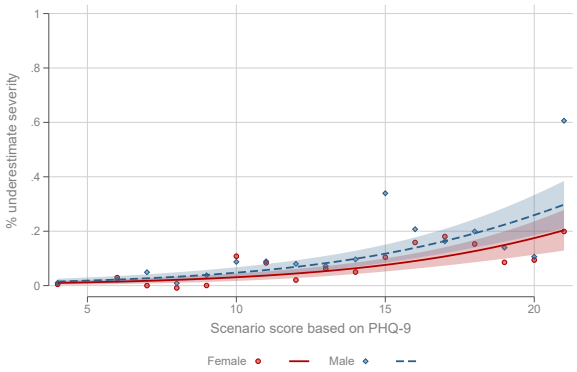
(a) Accurate severity



(b) Overestimate severity



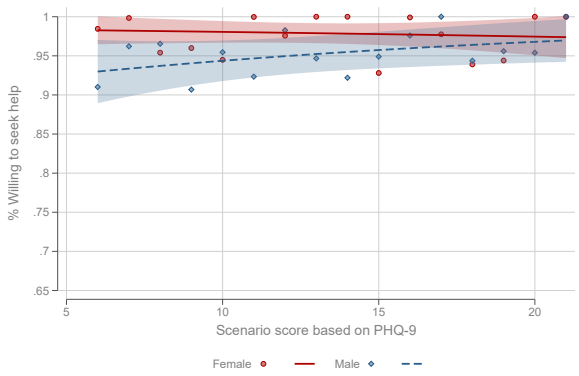
(c) Underestimate severity



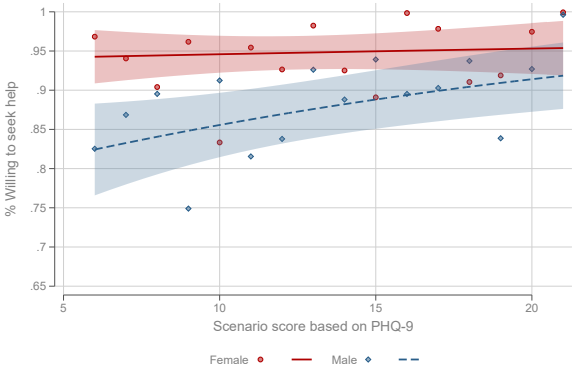
Notes. Binned scatter plot of accuracy in recognizing the severity of depression and the severity (PHQ-9) score of the hypothetical scenario, shown separately for male and female participants. The binary variables are created based on the question: How severe do you think the depression would be if [you/NAME] were experiencing these issues in real life? *[None or minimal, Mild, Moderate, Moderately Severe, Severe]*. Over-(Under-)estimate is defined as classifying the symptoms in a more (less) severe category than they actually belong to. Each point in the plot represents the regression coefficient from a regression of recognition on the scenario severity, controlling for treatment assignment (self vs. other), whether the subject in the "other" scenario is a male or a female and the order in which the scenarios were presented. The bands show 95% confidence intervals, with standard errors clustered by participant.

Figure 4: Gender differences in likelihood of seeking help

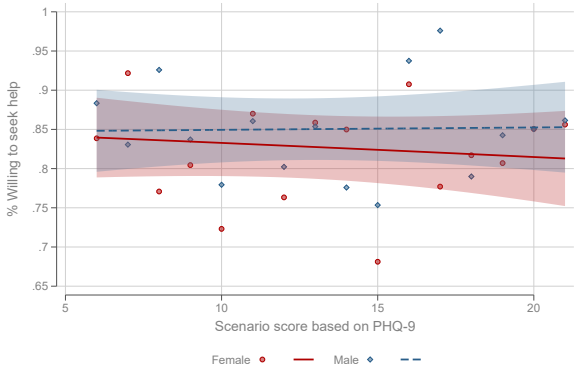
(a) Would seek help



(b) Would seek help from a specialist



(c) Would seek help from a non-specialist



Notes. Binned scatter plot of the likelihood of seeking help and the severity (PHQ-9) score of the hypothetical scenario, shown separately for male and female participants. The binary variables are created based on the question: If [you/NAME] were experiencing these hypothetical issues, how likely do you think it is that [you/she,he] will seek help from the following sources? For these questions, imagine that there are no constraints on the time or money that has to be spent, and no problems relating to health insurance coverage for these options. *[Very unlikely, Somewhat unlikely, Somewhat likely, Very likely]*. Seeking help from a specialist includes answering somewhat likely or very likely to any of the following: General Practitioner solely for this purpose, a GP during a visit for another purpose or a psychologist or a therapist. Seeking help from a non-specialist includes answering somewhat likely or very likely to any of the following: a counselor at your workplace or university, a close friend or relative or an AI-enabled mental health chatbot. Each point in the plot represents the regression coefficient from a regression of the outcomes on the scenario severity, controlling for treatment assignment (self vs. other), whether the subject in the "other" scenario is a male or a female and the order in which the scenarios were presented. The bands show 95% confidence intervals, with standard errors clustered by participant.

8 Tables

Table 1: Gender differences in depression recognition, accuracy and willingness to seek help

	Recognition		Accurate severity		Seek help	
	(1)	(2)	(3)	(4)	(5)	(6)
Male	-0.053***	-0.145***	0.068***	0.074**	-0.026*	-0.041*
	(0.018)	(0.040)	(0.022)	(0.030)	(0.014)	(0.022)
Mild	0.127***	0.095***	0.029	0.031	0.000	0.000
	(0.023)	(0.027)	(0.024)	(0.027)	(.)	(.)
Moderate	0.162***	0.095***	0.101***	0.119***	0.011	0.012
	(0.021)	(0.026)	(0.029)	(0.033)	(0.012)	(0.013)
Moderately Severe	0.191***	0.119***	0.329***	0.323***	0.010	-0.010
	(0.021)	(0.024)	(0.032)	(0.039)	(0.012)	(0.014)
Male × Mild		0.068		-0.004		0.000
		(0.047)		(0.039)		(.)
Male × Moderate		0.138***		-0.034		0.000
		(0.042)		(0.050)		(0.024)
Male × Moderately Severe		0.153***		0.014		0.042*
		(0.041)		(0.057)		(0.025)
Constant	0.774***	0.818***	-0.128	-0.129	0.974***	0.983***
	(0.065)	(0.064)	(0.085)	(0.085)	(0.047)	(0.048)
Demog. controls	Yes	Yes	Yes	Yes	Yes	Yes
Treat/Order FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean women	0.95	0.87	0.18	0.05	0.98	0.98
Observations	1604	1604	1604	1604	1203	1203

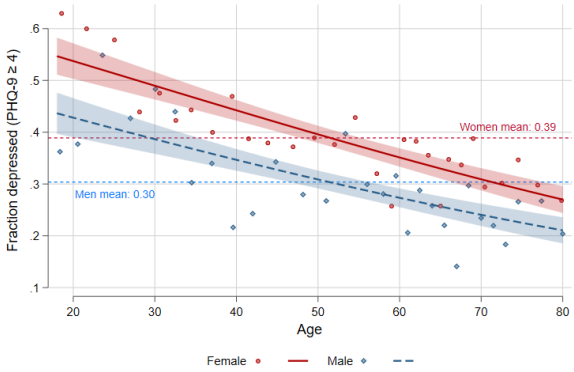
Notes. OLS regressions of depression recognition (Col. (1)–(2)), severity accuracy (Col. (3)–(4)) and likelihood of seeking help (Col. (5)–(6)) on a male indicator, and interactions with the severity category of the hypothetical scenario. Demographic controls are participant age, education, employment, income level and familiarity with depression screening tools. Order controls include which of the treatment blocks was presented first. Standard errors are heteroscedasticity-robust and clustered at the participant level. The mean of women at the bottom of the table corresponds to the mean outcome for women across all severity levels in Col. (1), (3) and (5) and the mean for women in the minimal severity level in Col. (2), (4) and (6). The seeking help question was not asked in the minimal depression scenarios so the constant and Male coefficient correspond to the Mild severity scenario. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

ONLINE APPENDIX

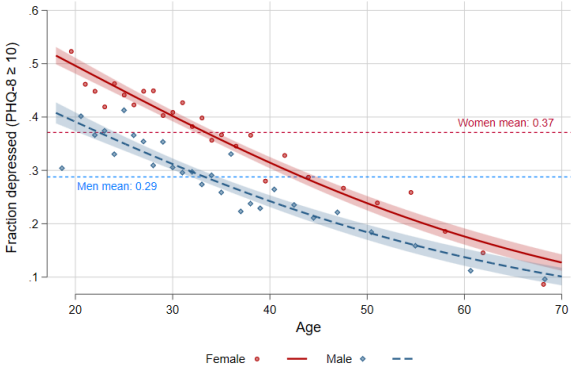
A Additional Figures

Figure A.1: Fraction classified as depressed by gender and age, across surveys

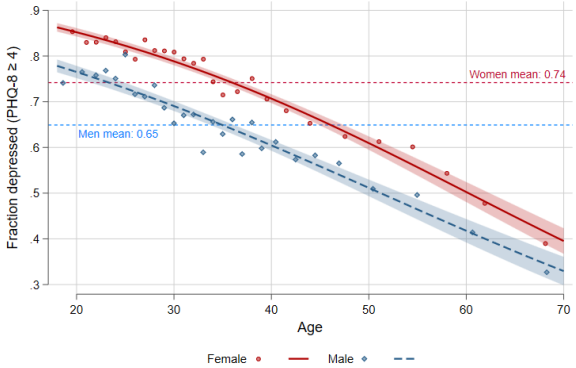
(a) Depression classification in NHANES, including mild symptoms



(b) Depression classification in Prolific, standard threshold



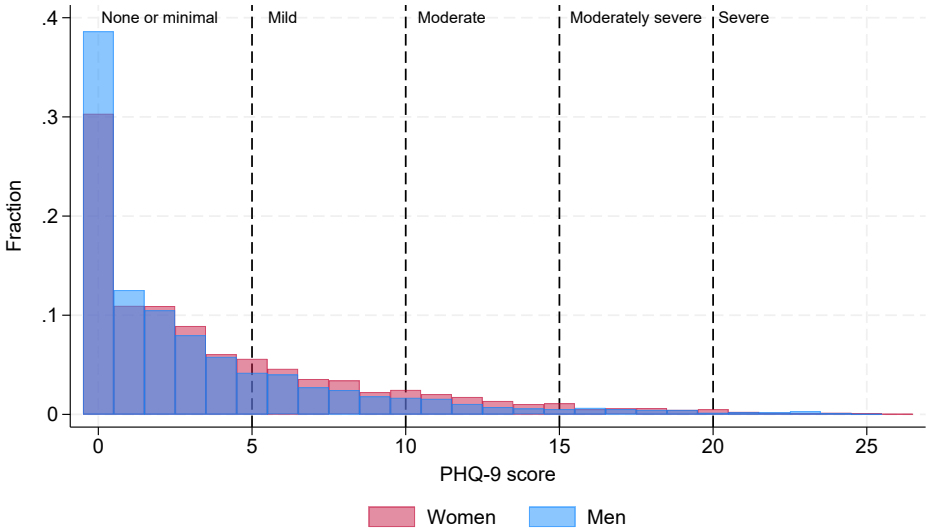
(c) Depression classification in Prolific, including mild symptoms



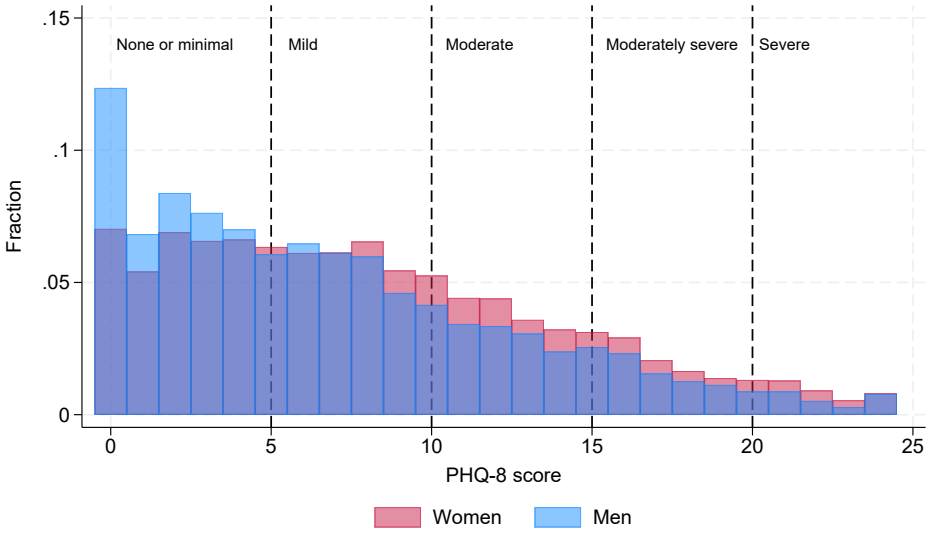
Notes. Panel A plots binned scatterplots of the prevalence of depressive symptoms by age and gender in the National Health and Nutrition Examination Survey (NHANES 2021–2023, N=6,337). Depression is measured using the PHQ-9 two-week symptom scale with the threshold of 10 points. Panels B and C plot binned scatterplots using the Prolific sample (N=18,982), prevalence is constructed using two PHQ-9 cutoffs: ≥ 10 (standard clinical threshold for major depression in panel B) and ≥ (including mild symptoms in Panel C). Points denote binned scatterplot means and vertical bars indicate 95% confidence intervals.

Figure A.2: Histograms of PHQ scores across samples

(a) Histogram of PHQ-9 scores in NHANES

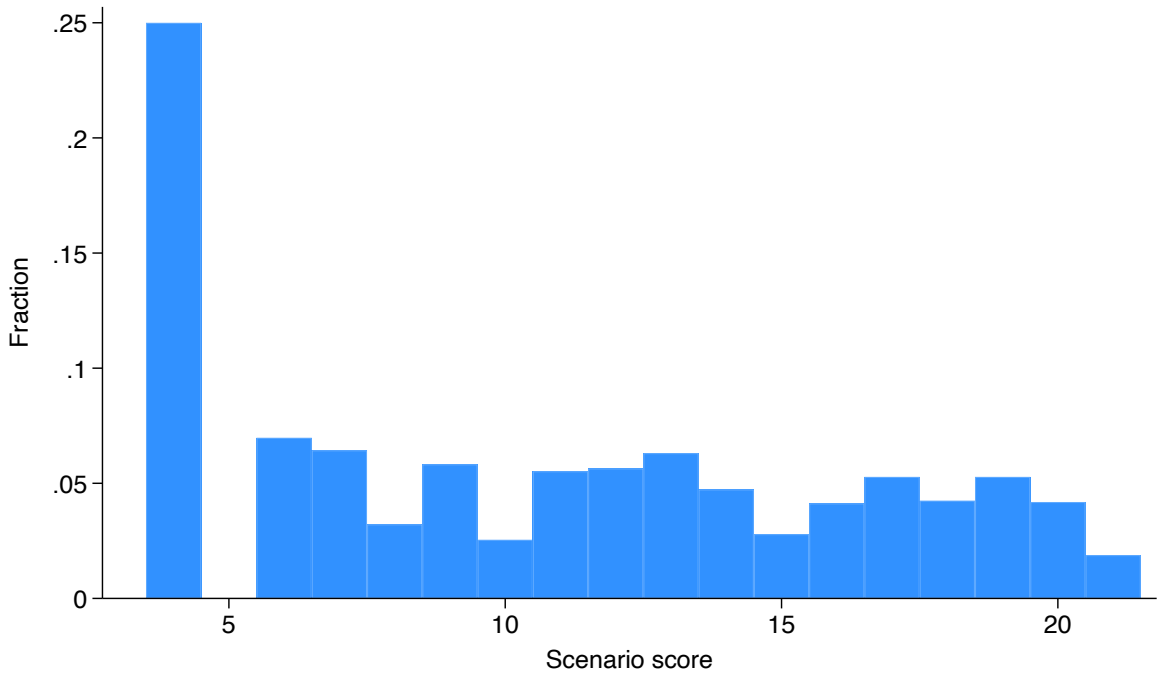


(b) Histogram of PHQ-8 scores in Prolific



Notes. Panel A presents the raw histogram of PHQ-9 scores by gender in NHANES (N=6,337). Panel B plots the raw histogram of PHQ-8 scores by gender in Prolific (N=18,982).

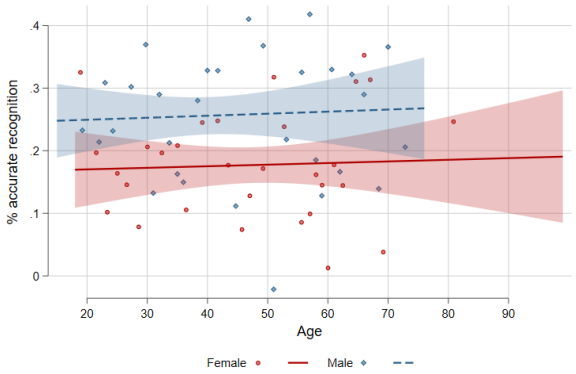
Figure A.3: Distribution of scenarios seen by participants



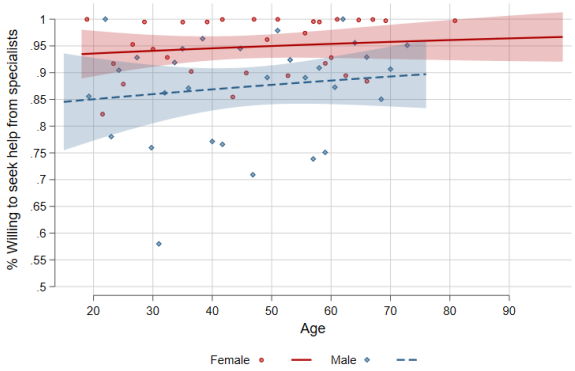
Notes. Fractions of scenarios presented to participants at every point of the support of the PHQ-9 score. The spike at 4 corresponds to the final scenario that is the same for all participants (PHQ-9 = 4.)

Figure A.4: Age gradient in accuracy and seeking help

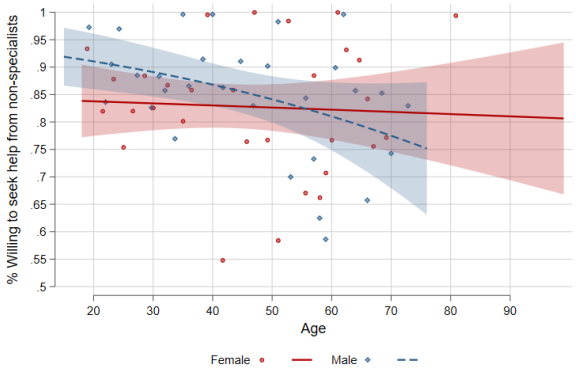
(a) Severity accuracy



(b) Seek help from specialist sources

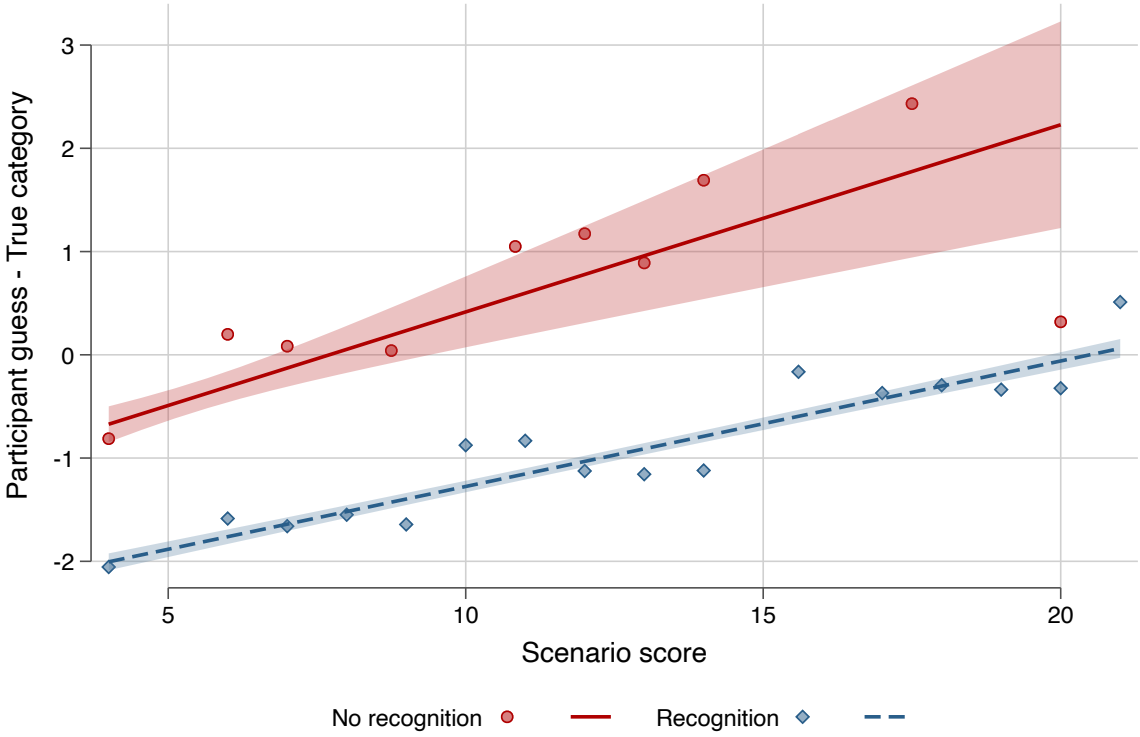


(c) Seek help from non-specialist sources



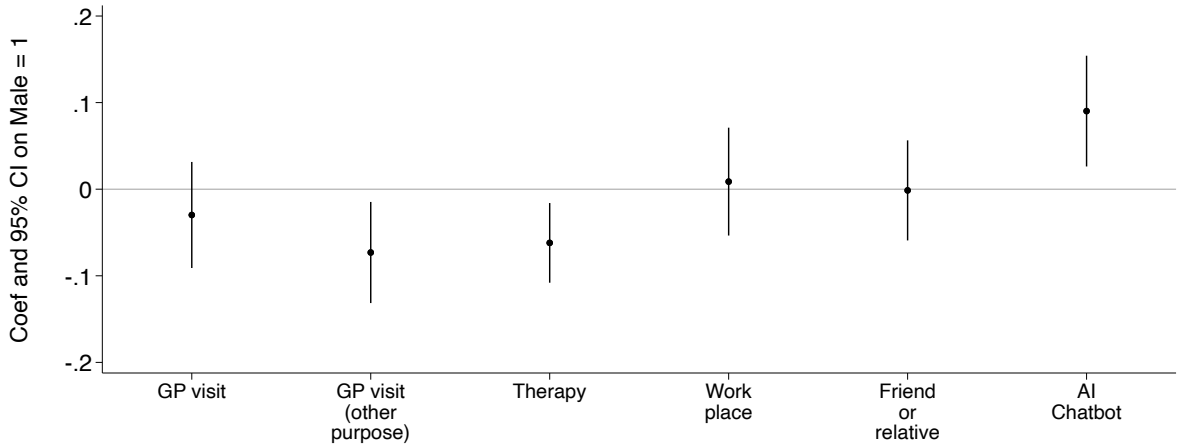
Notes. Binned scatter plots of the severity accuracy and seeking help outcomes on the age of the respondent, shown separately for male and female participants. Each point in the plot represents the regression coefficient from a regression of the outcome on the Panel labels on the scenario severity, controlling for treatment assignment (self vs. other), whether the subject in the "other" scenario is a male or a female and the order in which the scenarios were presented. The bands show 95% confidence intervals, with standard errors clustered by participant.

Figure A.5: Classification accuracy by recognition



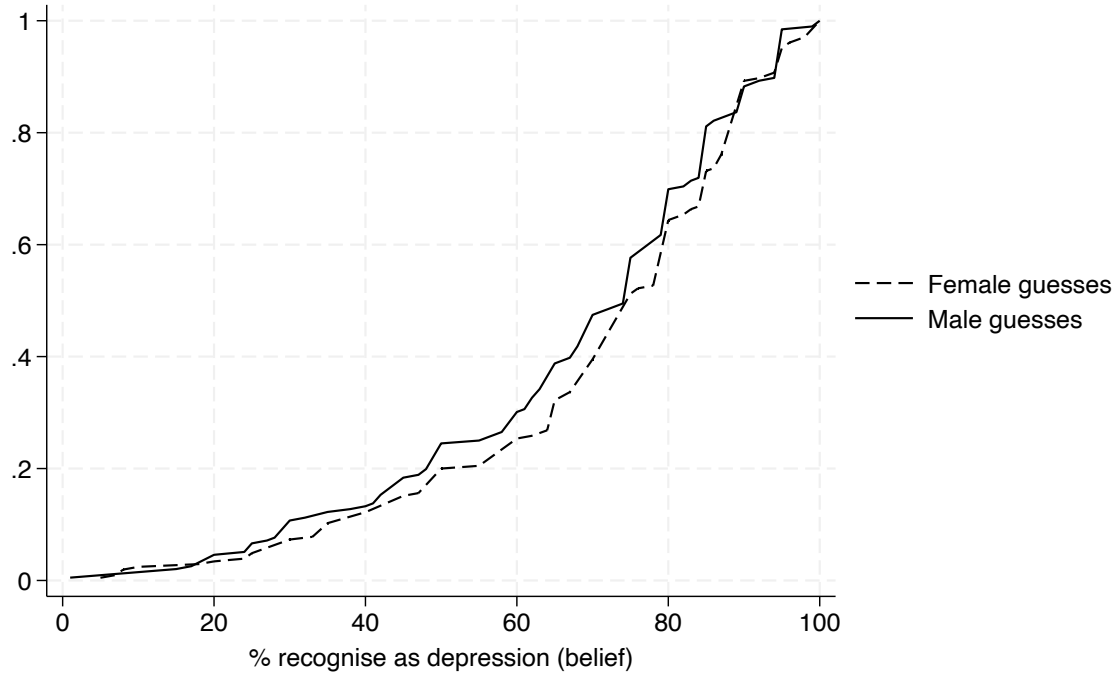
Notes. Binned scatter plots of the difference between estimated severity and actual severity of a given scenario on the scenario PHQ-9 score, shown separately for participants who recognize and do not recognize the scenario as a case of depression. We pool men and women together in this plot. Each point in the plot represents the regression coefficient from a regression of the outcome on the scenario PHQ-9 score, controlling for treatment assignment (self vs. other), whether the subject in the "other" scenario is a male or a female and the order in which the scenarios were presented. The bands show 95% confidence intervals, with standard errors clustered by participant.

Figure A.6: Gender differences in sources of seeking help



Notes. Each point in the plot represents the coefficient from separate regressions based on regressing each of the sources on the male indicator following Equation 1. The binary outcomes are equal to 1 when the respondent responds somewhat or very likely to seek help from that source in the x-axis. The estimates show gender gaps in sources of seeking help, where a positive point estimate indicates that men are more likely to women to seek help from that source. We plot 95% confidence intervals along with the point estimates of the gender gaps, with standard errors clustered by participant.

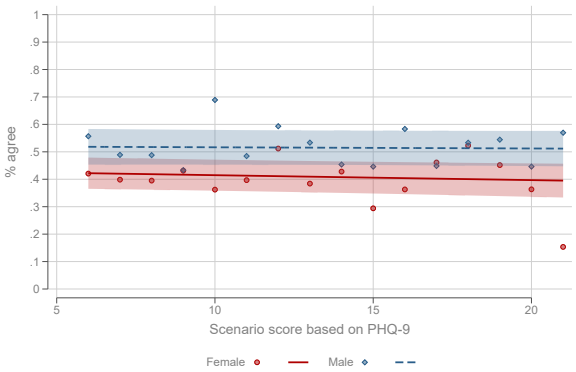
Figure A.7: Second-order beliefs



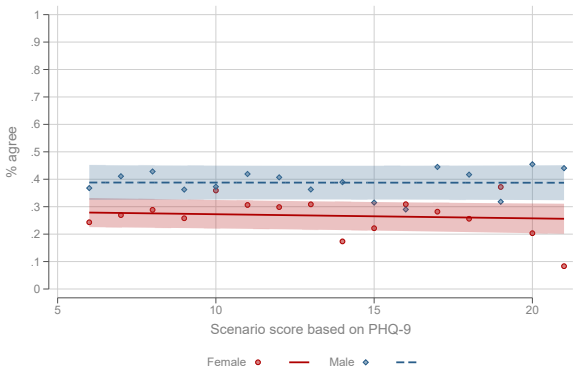
Notes. Empirical CDF of the guesses on the fraction of Americans recognizing depression in the scenario presented. CDFs of guesses are plotted separately for female and male participants. The wording of the question used to create this plot is: We conducted a similar survey with a sample of 100 Americans. The composition of respondents in this survey was broadly representative of the American population. Participants in that survey evaluated the following exact hypothetical scenario and answered whether or not they thought [NAME] would have depression if he/she were experiencing these symptoms. *List of symptoms*. Of the 100 Americans who participated in that survey, how many do you think answered Definitely yes or Probably yes to the question: Suppose that [NAME] were experiencing the hypothetical issues at the frequencies listed above, do you think [NAME] would have depression? [*Number between 0 and 100*]

Figure A.8: Fraction agreeing with the perceptions statements in each panel heading

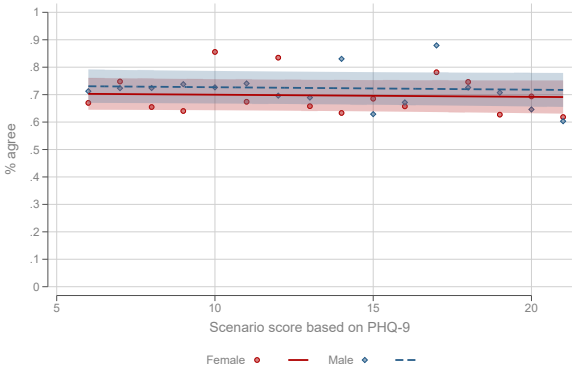
(a) The issues would go away by themselves in some time



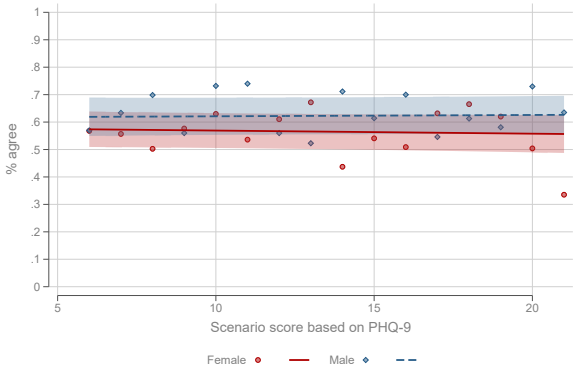
(b) The issues are not very serious and do not require treatment



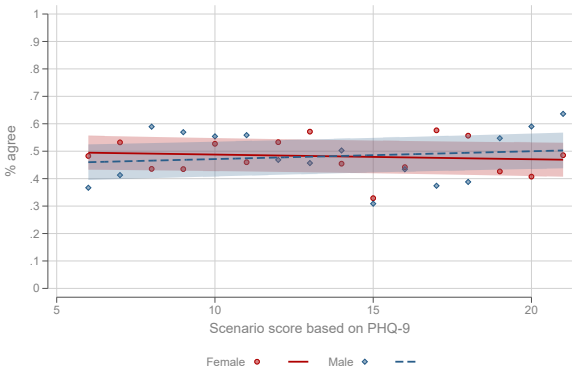
(c) Worry about what others would think



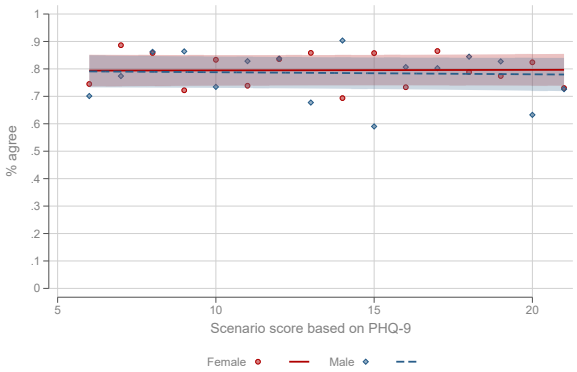
(d) Would rather deal with these issues oneself



(e) Do not think available treatments are effective

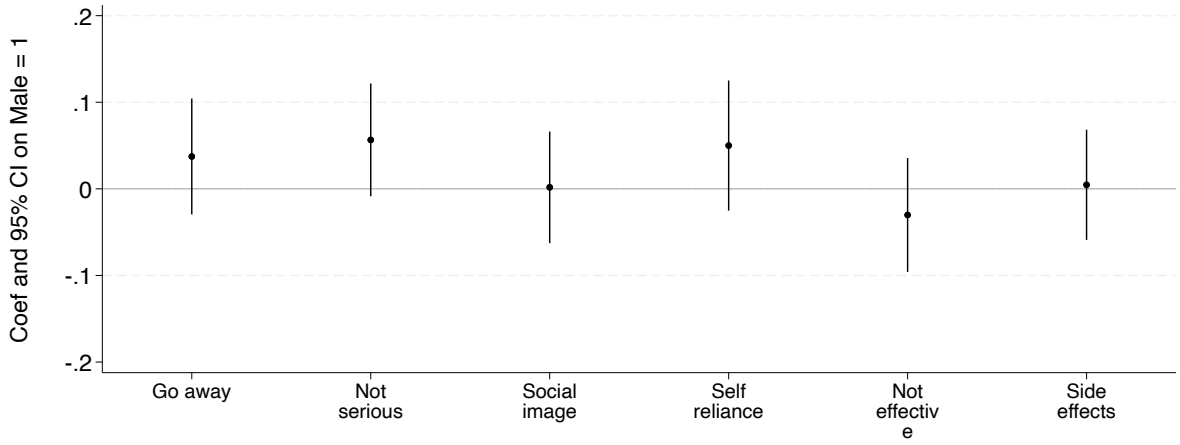


(f) Worry about the side effects of medication for depression



Notes. Binned scatter plots of the perceptions variables and the severity (PHQ-9) score of the hypothetical scenario, shown separately for male and female participants. The binary variables are equal to 1 when the respondent somewhat or strongly agrees with the statement. Each point in the plot represents the regression coefficient from a regression of the perception variable on the scenario severity, controlling for treatment assignment (self vs. other), whether the subject in the "other" scenario is a male or a female and the order in which the scenarios were presented. The bands show 95% confidence intervals, with standard errors clustered by participant.

Figure A.9: Gender differences in perceptions about depression symptoms



Notes. Each point in the plot represents the coefficient from separate regressions based on regressing each of the binary perceptions outcomes on the male indicator following Equation 1. The binary outcomes are equal to 1 when the respondent somewhat or strongly agrees with the statement in the x-axis. The estimates show gender gaps in perceptions, where a positive point estimate indicates that men are more likely to women to agree with the perceptions statement. The baseline levels that each outcome takes are in Figure A.8. We plot 95% confidence intervals along with the point estimates of the gender gaps, with standard errors clustered by participant.

B Additional Tables

Table B.1: Seeking help from specialist vs. non-specialist sources

	Specialist		Non-specialist	
	(1)	(2)	(3)	(4)
Male	-0.061***	-0.106***	0.020	0.015
	(0.021)	(0.033)	(0.030)	(0.039)
Moderate	0.015	-0.010	-0.028	-0.014
	(0.018)	(0.022)	(0.023)	(0.033)
Moderately Severe	0.035*	-0.000	-0.015	-0.033
	(0.018)	(0.022)	(0.021)	(0.030)
Male × Moderate		0.052		-0.028
		(0.037)		(0.047)
Male × Moderately Severe		0.074**		0.038
		(0.037)		(0.042)
Constant	0.902***	0.923***	1.183***	1.187***
	(0.061)	(0.062)	(0.088)	(0.089)
Demog. controls	Yes	Yes	Yes	Yes
Treat/Order FE	Yes	Yes	Yes	Yes
Mean women	0.95	0.95	0.83	0.84
Observations	1203	1203	1203	1203

Notes. OLS regressions of the likelihood of seeking help from specialist sources (Columns (1)–(2)) and from non-specialist sources (Columns (3)–(4)) on a male indicator, and interactions with the severity category of the hypothetical scenario. Demographic controls are participant age, education, employment, income level and familiarity with depression screening tools. Order controls include which of the treatment blocks was presented first. Standard errors are heteroscedasticity robust and clustered at the participant level. The seeking help question was not asked for the minimal depression scenario so the constant and Male coefficient correspond to the Mild severity scenario. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.2: Robustness: Main results, excluding last two scenarios.

	Recognition		Accurate severity		Seek help	
	(1)	(2)	(3)	(4)	(5)	(6)
Male	-0.043**	-0.088***	0.064**	0.053	-0.030*	-0.045*
	(0.018)	(0.032)	(0.031)	(0.037)	(0.016)	(0.026)
Moderate	0.024	-0.018	0.106***	0.114***	0.012	0.010
	(0.018)	(0.016)	(0.029)	(0.036)	(0.013)	(0.016)
Moderately Severe	0.042**	0.020	0.379***	0.356***	0.009	-0.009
	(0.017)	(0.014)	(0.035)	(0.046)	(0.015)	(0.015)
Male × Moderate		0.083**		-0.017		0.003
		(0.034)		(0.057)		(0.028)
Male × Moderately Severe		0.047		0.050		0.039
		(0.033)		(0.070)		(0.032)
Constant	0.866***	0.894***	-0.077	-0.070	0.994***	1.003***
	(0.102)	(0.100)	(0.134)	(0.135)	(0.051)	(0.053)
Demog. controls	Yes	Yes	Yes	Yes	Yes	Yes
Treat/Order FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean women	0.98	0.98	0.22	0.22	0.98	0.98
Observations	802	802	802	802	802	802

Notes. OLS regressions of depression recognition (Col. (1)–(2)), severity accuracy (Col. (3)–(4)) and the likelihood of seeking help (Col. (5)–(6)) on a male indicator, and interactions with the severity category of the hypothetical scenario. The sample is restricted to the first two (of four) scenarios evaluated by participants. Demographic controls are participant age, education, employment, income level and familiarity with depression screening tools. Order controls include which of the treatment blocks was presented first. Standard errors are heteroscedasticity robust and clustered at the participant level. The mean of women at the bottom of the table corresponds to the mean outcome for women across all severity levels. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.3: Robustness: Main results, excluding scenarios with suicidal ideation.

	Recognition		Accurate severity		Seek help	
	(1)	(2)	(3)	(4)	(5)	(6)
Male	-0.083***	-0.132***	0.044*	0.067**	-0.046**	-0.048*
	(0.025)	(0.040)	(0.026)	(0.030)	(0.021)	(0.025)
Mild	0.109***	0.087***	0.041	0.036	0.000	0.000
	(0.027)	(0.032)	(0.030)	(0.036)	(.)	(.)
Moderate	0.144***	0.089***	0.136***	0.173***	0.006	0.013
	(0.026)	(0.032)	(0.039)	(0.050)	(0.016)	(0.016)
Moderately Severe	0.158***	0.094***	0.351***	0.399***	-0.006	-0.028
	(0.032)	(0.036)	(0.060)	(0.077)	(0.024)	(0.033)
Male × Mild		0.047		0.006		0.000
		(0.051)		(0.044)		(.)
Male × Moderate		0.113**		-0.076		-0.014
		(0.049)		(0.066)		(0.034)
Male × Moderately Severe		0.135**		-0.105		0.046
		(0.058)		(0.112)		(0.047)
Constant	0.793***	0.817***	-0.204***	-0.214***	1.053***	1.052***
	(0.101)	(0.100)	(0.076)	(0.078)	(0.066)	(0.066)
Demog. controls	Yes	Yes	Yes	Yes	Yes	Yes
Treat/Order FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean women	0.92	0.92	0.14	0.14	0.98	0.98
Observations	991	991	991	991	590	590

Notes. OLS regressions of depression recognition (Col. (1)–(2)), severity accuracy (Col. (3)–(4)) and the likelihood of seeking help (Col. (5)–(6)) on a male indicator, and interactions with the severity category of the hypothetical scenario. The sample is restricted to scenarios which do not include the suicidal ideation question from the PHQ-9. Demographic controls are participant age, education, employment, income level and familiarity with depression screening tools. Order controls include which of the treatment blocks was presented first. Standard errors are heteroscedasticity robust and clustered at the participant level. The mean of women at the bottom of the table corresponds to the mean outcome for women. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.4: Robustness: Main results, excluding participants who do not recognize depression.

	Accurate severity		Seek help	
	(1)	(2)	(3)	(4)
Male	0.048**	-0.013	-0.017	-0.025
	(0.022)	(0.015)	(0.014)	(0.021)
Mild	0.074***	0.053**	0.000	0.000
	(0.022)	(0.025)	(.)	(.)
Moderate	0.181***	0.154***	0.014	0.016
	(0.027)	(0.033)	(0.010)	(0.013)
Moderately Severe	0.415***	0.365***	0.001	-0.010
	(0.030)	(0.037)	(0.012)	(0.015)
Male × Mild		0.050		0.000
		(0.033)		(.)
Male × Moderate		0.064		-0.004
		(0.044)		(0.021)
Male × Moderately Severe		0.112**		0.025
		(0.049)		(0.024)
Constant	-0.160*	-0.133	0.990***	0.996***
	(0.090)	(0.090)	(0.045)	(0.046)
Demog. controls	Yes	Yes	Yes	Yes
Treat/Order FE	Yes	Yes	Yes	Yes
Mean women	0.17	0.17	0.98	0.98
Observations	1463	1463	1145	1145

Notes. OLS regressions of severity accuracy (Col. (1)–(2)) and likelihood of seeking help (Col. (3)–(4)) on a male indicator, and interactions with the severity category of the hypothetical scenario. The sample is restricted to scenarios where participants recognise the presented scenario as depression. Demographic controls are participant age, education, employment, income level and familiarity with depression screening tools. Order controls include which of the treatment blocks was presented first. Standard errors are heteroscedasticity robust and clustered at the participant level. The mean of women at the bottom of the table corresponds to the mean outcome for women across all severity levels. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.5: Gender differences in depression recognition, accuracy and likelihood of seeking help: *Self* vs. *Other*

	Recognition		Accurate severity		Seek help	
	(1)	(2)	(3)	(4)	(5)	(6)
Male	-0.035*	-0.103***	0.071**	0.114***	-0.042**	-0.020
	(0.019)	(0.034)	(0.031)	(0.038)	(0.018)	(0.019)
Other	0.024		0.010		-0.005	
	(0.015)		(0.030)		(0.014)	
Male × Other	-0.038		-0.005		0.049**	
	(0.027)		(0.039)		(0.021)	
Other is male		0.002		0.029		-0.042
		(0.031)		(0.034)		(0.027)
Male × Other is male		0.111**		-0.130**		0.065*
		(0.050)		(0.053)		(0.035)
Constant	0.754***	0.809***	-0.125	-0.213***	0.959***	1.018***
	(0.065)	(0.104)	(0.082)	(0.073)	(0.046)	(0.051)
Demog. controls	Yes	Yes	Yes	Yes	Yes	Yes
Order FE	Yes	Yes	Yes	Yes	Yes	Yes
Category FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean	0.98	0.96	0.21	0.19	0.98	0.99
Observations	1604	802	1604	802	1203	401

Notes. OLS regressions of depression recognition (Col. (1)–(2)), severity accuracy (Col. (3)–(4)) and likelihood of seeking help (Col. (5)–(6)) on a male indicator, and interactions with the "Other" treatment in all scenarios in Col. (1), (3) and (5) and with the "Other is male" treatment within the "Other" scenarios only in Col. (2), (4) and (6). Demographic controls are participant age, education, employment, income level and familiarity with depression screening tools. Order controls include which of the treatment blocks was presented first. Standard errors are heteroscedasticity robust and clustered at the participant level. The mean of women at the bottom of the table corresponds to the mean outcome for women evaluating the "Self" treatment in Col. (1), (3) and (5) and women evaluating the "Other is female" treatment in Col. (2), (4) and (6). The seeking help question was not asked for the minimal depression scenario so the constant and Male coefficient correspond to the Mild severity scenario. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.6: Seeking help from specialist vs. non-specialist sources in Self vs. Other scenarios

	Overall		Specialist		Non-specialist	
	(1)	(2)	(3)	(4)	(5)	(6)
Male	-0.042**	-0.037**	-0.080***	-0.068***	0.003	0.008
	(0.018)	(0.018)	(0.025)	(0.024)	(0.036)	(0.036)
Other	-0.005	-0.004	-0.018	-0.015	0.039	0.041
	(0.014)	(0.014)	(0.021)	(0.021)	(0.027)	(0.027)
Male × Other	0.049**	0.045**	0.058*	0.048	0.054	0.050
	(0.021)	(0.021)	(0.031)	(0.031)	(0.038)	(0.038)
Recognized depression		0.119**		0.285***		0.114*
		(0.048)		(0.070)		(0.062)
Constant	0.959***	0.866***	0.909***	0.687***	1.139***	1.050***
	(0.046)	(0.063)	(0.058)	(0.085)	(0.086)	(0.105)
Demog. controls	Yes	Yes	Yes	Yes	Yes	Yes
Order FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean women	0.98	0.98	0.95	0.95	0.81	0.81
Observations	1203	1203	1203	1203	1203	1203

*Notes.*OLS regressions of seeking help from specialist and non-specialists sources combined (Columns (1)–(2)) and separate (Columns (3)–(6)) on a male indicator, and interactions with *Other* treatment assignment of the hypothetical scenario. Demographic controls are participant age, education, employment, income level and familiarity with depression screening tools. Order controls include which of the treatment blocks was presented first. Columns (2), (4) and (6) add a control for whether the participant recognized depression at an earlier stage right after evaluating the scenario. Standard errors are heteroscedasticity robust and clustered at the participant level. The seeking help question was not asked for the minimal depression scenario so the constant and Male coefficient correspond to the Mild severity scenario. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

C PHQ-9 questionnaire and severity classification

Figure C.1: The PHQ-9 questionnaire (Kroenke et al., 2001)

PATIENT HEALTH QUESTIONNAIRE-9 (PHQ-9)

Over the last 2 weeks, how often have you been bothered by any of the following problems? *(Use "✔" to indicate your answer)*	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself — or that you are a failure or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed? Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
9. Thoughts that you would be better off dead or of hurting yourself in some way	0	1	2	3

The PHQ-9 score is calculated by the simple addition of the frequencies for each symptom, with no weight for how "serious" the symptom is. Each score corresponds to a severity category:

- 0-4: None or minimal
- 5-9: Mild
- 10-14: Moderate
- 15-19: Moderately severe
- 20-27: Severe

D Background Information on GBD Data and Survey Data from NHANES and Prolific

Obtaining consistent measures of mental health prevalence in the United States is challenging because no centralized administrative registry exists. Available data instead come from fragmented sources, such as healthcare utilization records and surveys, each capturing different populations and affected by changes in screening practices, coding conventions, and reporting incentives (Choi et al., 2025). For example, increases in recorded mental health events may reflect shifts in diagnosis or reporting rather than underlying prevalence. These limitations make single-source estimates difficult to interpret and particularly problematic when studying differences across demographic groups. Aggregated estimates such as those from the Global Burden of Disease project synthesize multiple data inputs to provide a harmonized measure of prevalence and are therefore commonly used when comprehensive population-level data are unavailable.

Below we describe the three sources used in the paper and highlight differences in representativeness, measurement, and potential selection into observation.

Global Burden of Disease (GBD)

We use depression prevalence estimates from the Global Burden of Disease Study 2021 (Global Burden of Disease Collaborative Network, 2025). The GBD database collects data through various sources into a harmonized population estimate using statistical modeling. The data sources include official statistical systems (population censuses, vital statistics, and civil registration), health surveys (national health surveys and disease-specific surveys), medical records (hospital statistics, disease surveillance reports, and insurance claims), scientific research (epidemiological studies and clinical trials), and environmental monitoring (satellite imagery and air quality measurements), and other sources (Zhao et al., 2025). Because these estimates incorporate diagnosed cases and studies that often rely on contact with healthcare systems or structured diagnostic instruments, appearing in the measured prevalence may depend on symptom recognition, reporting, and access to care.

The GBD does not provide individual-level microdata and therefore does not allow analysis of symptom distributions or attrition. Instead, it provides population prevalence rates by gender and age groups. The female-to-male prevalence ratio used in the paper is constructed directly from these aggregate prevalence estimates.

National Health and Nutrition Examination Survey (NHANES)

We use the 2021–2023 wave of the National Health and Nutrition Examination Survey (NHANES), a nationally representative survey conducted by the National Center for Health

Statistics (CDC and NCHS, 2024). NHANES employs a stratified, multistage probability sampling design targeting the non-institutionalized U.S. population and provides survey weights to recover population representativeness.

The analytic sample contains $N = 6{,}337$ individuals, of which 51.4% are women and 48.6% are men after applying survey weights (unweighted: 54.9% women, 45.1% men). Depression is measured using the Patient Health Questionnaire (PHQ-9) two-week symptom scale administered as part of the mental health questionnaire module.

We construct depression prevalence using two cutoffs:

- PHQ-9 ≥ 10: standard clinical threshold for major depressive disorder
- PHQ-9 ≥ 4: inclusive threshold capturing mild symptoms

NHANES measures symptoms directly rather than relying on prior diagnosis, and participation does not depend on mental health concerns. As a result, selection into measurement is expected to be lower than in administrative or clinical sources.

Figure A.1a shows the fraction of men and women classified as depressed using the threshold at 4 points in the PHQ-9. The full distribution of PHQ-9 scores by gender is presented in Figure A.2a.

Prolific Screening Sample

We use a screened online sample collected on the Prolific platform in 2022 and originally used in Roth et al. (2024a) and Roth et al. (2024b). The screening sample contains $N = 18{,}982$ U.S. participants recruited prior to participation in the main studies.

Prolific is not a probability sample but is commonly used in social science research because it allows targeting demographic quotas and produces samples that closely match U.S. census distributions along observable characteristics relative to other online panels (Roth et al., 2024a). Participants opt into the platform but do not select into the study based on mental health conditions.

Depression is measured using the PHQ-8 questionnaire (excludes the suicidal ideation question) and we construct prevalence using the same thresholds, as is standard in practice:

- PHQ-8 ≥ 10
- PHQ-8 ≥ 4

The sample consists of 55.5% women and 44.5% men. We exclude 228 observations with missing or non-binary gender. The fraction of people classified as depressed by age is shown in Figures A.1b and A.1c. The histogram of PHQ-8 symptom distributions by gender appears in Figure A.2b.

Because the PHQ screening was administered as part of an online questionnaire with no clinical follow-up, responses carry no medical consequences and therefore do not require participants to interpret symptoms as a medical condition. For this reason, selection into measurement is expected to be low in this dataset.

We have no access to the dataset including respondents who did not pass the attention check, but selective attrition appears to be minimal.

E Vignette scenarios

E.1 Treatment *Self*

First, participants are introduced to the context of the vignette:

> Please imagine that you have been experiencing the following issues. Note that this is a *hypothetical scenario*. Review the list of hypothetical issues along with how often you would have experienced them **over the last two weeks**:

Three or more of the randomly selected symptoms were shown to the participant in a tabular form, along with a frequency which is one of *Several days, More than half the days, Nearly every day*.

E.2 Treatment *Other*

The scenario was introduced in the same way as in treatment *Self*. The key difference is that the subject of the scenario is a hypothetical male or female individual. We use first names to make gender identity salient. The chosen names – Michael and Jessica – are among the most popular names in the birth cohort in Texas.

> For these questions, imagine a hypothetical individual, [NAME]. [NAME] is 34 years old, lives in [TOWN in the US], and works as a marketing professional.
>
> Please imagine that [NAME] has been experiencing the following issues. Note that this is a *hypothetical scenario*. Review the list of hypothetical issues along with how often [NAME] would have experienced them **over the last two weeks**:

F Survey questionnaire

Q1. Suppose that [you/NAME] were/was experiencing the hypothetical issues at the frequencies listed above, do you think [you/NAME] would have depression? *[Definitely yes, probably yes, probably no, definitely no]*

Q2. How severe do you think the depression would be if [you/NAME] were experiencing these issues in real life? *[None or minimal, Mild, Moderate, Moderately Severe, Severe]*

Q3. If [you/NAME] were experiencing these hypothetical issues, how likely do you think it is that [you/she,he] will seek help from the following sources? For these questions, imagine that there are no constraints on the time or money that has to be spent, and no problems relating to health insurance coverage for these options. *[Very unlikely, Somewhat unlikely, Somewhat likely, Very likely]*

1. a General Practitioner solely for this purpose
2. a General Practitioner during a visit for another purpose
3. a Psychologist or a therapist
4. a counselor at your workplace or university
5. a close friend or relative
6. an AI-enabled mental health chatbot

Q4. Based on the hypothetical issues and their frequencies experienced by [you/NAME], please indicate the extent to which you agree or disagree with the following statements. *[Strongly disagree, Somewhat disagree, Somewhat agree, Strongly agree]*

1. The issues would go away by themselves in some time
2. The issues are not very serious and do not require treatment
3. I/[NAME] worry/worries about what others would think of me/her/him if they became aware that [I/she,he] had these issues
4. I/[NAME] would rather deal with these issues [myself/herself,himself] than rely on help from others
5. I/[NAME] do/does not think that the available treatments for these issues are effective
6. I/[NAME] am/is worried about the side effects of medication for depression

Q5. We conducted a similar survey with a sample of 100 Americans. The composition of respondents in this survey was broadly representative of the American population. Participants in that survey evaluated the following exact hypothetical scenario and answered whether or not they thought [NAME] would have depression if he/she were experiencing these symptoms.

List of symptoms

Of the 100 Americans who participated in that survey, how many do you think answered Definitely yes or Probably yes to the question: Suppose that [NAME] were experiencing the hypothetical issues at the frequencies listed above, do you think [NAME] would have depression? [*Number between 0 and 100*]

G Randomization procedure

Each vignette always includes the first two questions of the PHQ-9. Additional questions and symptom severities are randomly generated. The procedure to randomly generate a list of symptoms was as follows:

1. Randomly select one of the four categories. The likelihood with which categories are picked are: Mild - 1/6, Moderate - 1/3, Mod. Severe - 1/3, Severe - 1/6. Each category has an upper and a lower score bound.[14]

2. For the first two questions of the PHQ-9, select severities such that the score adds up to 4 (so one of (1,3),(2,2),(3,1)).

3. Next, generate a random sequence of numbers from 3 to 9, and pick them one by one. These correspond to questions in the scenario.

4. For each item in the sequence, pick a score from 1 to 3. This represents the severity of the symptom.

5. Check whether the total score exceeds the minimum threshold for the category. If not, repeat step (4). If so:

    - If all questions have been iterated through, stop
    - If score exceeds the max score for that category as well, stop
    - Else, flip a coin. If Heads, pick another question (step (4)). If Tails, stop.

[14]According to the PHQ-9 scoring, there are 5 severity levels: 0-4 None or minimal, 5-9 Mild, 10-14 Moderate, 15-19 Moderately severe, 20-27 Severe. We do not focus on the lowest severity level to ensure that we are presenting scenarios where some level of depression is expected.

H Mechanisms: Additional details and results

H.1 Psychic costs

To attribute our findings entirely to social desirability bias, women would have to systematically overstate depression recognition at mild severity and men systematically understate it, with both groups then shifting behavior at higher-severity scenarios in ways that cancel out the bias. Likewise, men would need to underreport willingness to seek help for themselves while simultaneously overstating it for others. This set of coordinated reporting patterns seems implausible, and the heterogeneity we observe across *Self* and *Other* scenarios is inconsistent with what a uniform social desirability bias would predict.

H.2 Perceptions

Here, we quantify the patterns discussed in the manuscript: The overall agreement with the statements regardless of gender or scenario severity is 46.1% for "the issues would go away by themselves in some time," 32.7% for "the issues are not very serious and do not require treatment," 71.1% for "I/[NAME] worry/worries about what others would think of me/her/him if they became aware that [I/she,he] had these issues," 59.4% for " I/[NAME] would rather deal with these issues myself/herself,himself] than rely on help from others," 48.1% for "I/[NAME] do/does not think that the available treatments for these issues are effective," and 79.1% for "I/[NAME] am/is worried about the side effects of medication for depression." These overall averages are indicative of the pattern across severity levels, as the scatterplots remain relatively flat across the full range of PHQ-9 scores (see Figure A.8). In short, nearly half of participants believe the issues would resolve on their own, a third view them as not requiring treatment, and large majorities express concern about stigma, self-reliance, treatment effectiveness, and medication side effects.

The only perceptions where men and women appear to slightly differ are in the beliefs that "the issues would go away on their own over time" and "the issues are not very serious and do not require treatment", both of which are related to help-seeking behavior. For the latter, the confidence bands do not overlap starting at moderate depression levels, suggesting that men may be more likely than women to dismiss the symptoms and believe that treatment is unnecessary, particularly as severity increases.

Although gender differences in perceptions are largely absent, the overall levels of agreement with the statements reveal striking patterns. In over 70% of scenario evaluations, participants report concerns about what others would think and about the side effects of depression medication. We interpret these findings in light of recent work in economics on perceptions of stigma and depression. Roth et al. (2024a) document widespread misperceptions about stigma: individuals believe that 38% of Americans hold stigmatizing beliefs, while the actual rate is only 16%. The high share of respondents expressing concern about

others' opinions may reflect anticipated stigma, but could also reflect general discomfort with disclosure or fears of burdening others. Concerns about medication side effects build on evidence of misperceptions about the effectiveness of online therapy (Roth et al., 2024b), suggesting that reluctance to seek treatment Cronin et al. (2024) may be even greater when pharmacological options are involved.

I Pre-Analysis Plan

The section below correspond to the pre-analysis plan we submitted to the AER RCT registry with ID AEARCTR-0014621 and fist registered on October 21, 2024 and and first published on October 28, 2024.

I.1 Introduction

Mental health disorders are a leading cause of disability worldwide, affecting over 1 billion people (Rehm and Shield, 2019). The prevalence of these conditions contributes to significant economic burdens, with annual costs of approximately $201 billion in the U.S. and $3.7 billion in Norway (Bütikofer et al., 2024). Research has shown that mental health issues negatively impact educational outcomes, such as grade repetition, test scores (Currie and Stabile, 2006, 2009), and dropout (Fletcher, 2010), as well as labor market outcomes, including lost working days (Ridley et al., 2020; Currie, 2024) and income (Smith and Smith, 2010; Goodman et al., 2011; Biasi et al., 2021).

Undetected mental health issues can also generate substantial costs as untreated individuals continue obtaining poor health and economic outcomes. Detecting and treating mental health issues is crucial, yet global statistics indicate that many individuals suffering from these conditions remain undiagnosed and untreated. Worldwide, there are substantial gender disparities in depression and suicide rates. Women are more frequently diagnosed with depression, yet men have a significantly higher rate of completed suicides. These patterns suggest that a larger proportion of men than women may not be diagnosed in time, raising important questions about how gender influences both the recognition of mental health issues and the likelihood of seeking help.

This study investigates gender differences in the recognition of mental health symptoms, particularly depression, and the likelihood of seeking help. Through an analysis of administrative data on mental health diagnosis and suicide from Norway, combined with an online experiment using a representative U.S. sample, the project aims to explore whether gender-based differences in mental health symptom recognition and help-seeking behaviors contribute to the disparities in suicide and depression rates. The study also provides insights into potential factors such as perceptions, psychic costs, and social norms that may drive these gender differences.

I.2 Design

The study uses a within-subjects design wherein all participants are exposed to two treatment blocks. In each of these blocks, participants will be shown one or more *hypothetical scenarios*. A scenario consists of a list of depression symptoms and the frequency of their occurrence over the past two weeks. The displayed symptoms are chosen from the PHQ-9

questionnaire (Kroenke et al., 2001), which is a widely used (and often self-administered) instrument that screens for depressive disorders. The list of symptoms and frequencies are randomly chosen to meet certain thresholds which indicate different levels of depression severity.

After reviewing the hypothetical scenario, participants are asked to state their views on: (i) whether the symptoms indicate that the subject of the scenario suffers from depression (*recognition*), (ii) perceptions of depression *severity*, (iii) the likelihood of *seeking help* from different sources, and (iv) beliefs and attitudes regarding depression and mental health.

Treatment blocks. The two treatment blocks, *Self* and *Other*, differ only in the subject of the hypothetical scenario. In *Self*, the subject of the hypothetical scenario is the participant. In *Other*, the subject of the scenario is a hypothetical individual (who is either Male or Female). Participants will evaluate 2 scenarios in each treatment block and the order of the blocks will be randomized. The symptoms in the second scenario in the *Other* treatment is fixed and the subject of that scenario is the same as the subject of the first scenario in that treatment. This common scenario will be used to elicit second order beliefs by asking participants to provide their best guess of the percentage of participants in the study who thought that the scenario described a person with depression. This guess will be incentivized based on the actual responses from participants in the study.

Participants will also be presented with a set of standard demographic questions and a question about their past experience with treatment for mental health issues such as depression or anxiety.

I.3 Treatment *Self*

Vignette. First, participants are introduced to the context of the vignette:

> Please imagine that you have been experiencing the following issues. Note that this is a *hypothetical scenario*. Review the list of hypothetical issues along with how often you would have experienced them **over the last two weeks**:

List of symptoms. Three or more of these symptoms are shown to the participant in a tabular form, along with a frequency which is one of *Several days, More than half the days, Nearly every day*. The list contains the questions exactly as stated in the PHQ-9 instrument:

1. Little interest or pleasure in doing things.
2. Feeling down, depressed, or hopeless.
3. Trouble falling or staying asleep, or sleeping too much.

4. Feeling tired or having little energy.

5. Poor appetite or overeating.

6. Feeling bad about yourself —or that you are a failure or have let yourself or your family down.

7. Trouble concentrating on things, such as reading the newspaper or watching television.

8. Moving or speaking so slowly that other people could have noticed. Or the opposite —being so fidgety or restless that you have been moving around a lot more than usual.

9. Thoughts that you would be better off dead, or thoughts of hurting yourself in some way.

Each frequency corresponds with a score in the PHQ-9 scale as follows:

- Several days = 1
- More than half the days = 2
- Nearly every day = 3

The PHQ-9 score is calculated by the simple addition of the frequencies for each symptom, with no weight for how "serious" the symptom is. For example, to obtain the maximum score of 27 in the PHQ-9 scale, all nine items must have a frequency equal to 3 (nearly every day). The scores map into a suggested interpretation of the severity of depression as follows: Minimal Depression (1−4), Mild (5−9), Moderate (10−14), Moderately Severe (15−19), and Severe (>= 20). We consider all categories except Minimal in this experiment.

I.4 Treatment *Other*

Vignette context. The scenario is introduced in the same way as in treatment *Self*. The key difference is that the subject of the scenario is a hypothetical male or female individual. We use first names to make gender identity salient. The chosen names – Michael and Jessica – are among the most popular names in the birth cohort in Texas.

> For these questions, imagine a hypothetical individual, [NAME]. [NAME] is 34 years old, lives in [TOWN in the US], and works as a marketing professional.
>
> Please imagine that [NAME] has been experiencing the following issues. Note that this is a *hypothetical scenario*. Review the list of hypothetical issues along with how often [NAME] would have experienced them **over the last two weeks**:

List of symptoms. Presented in the same way as in the *Self* block.

I.5 Generating the list of symptoms

Each vignette always includes the first two questions of the PHQ-9. Additional questions and symptom severities are randomly generated. The procedure to randomly generate a list of symptoms is as follows:

1. Randomly select one of the four categories. The likelihood with which categories are picked are: Mild - 1/6, Moderate - 1/3, Mod. Severe - 1/3, Severe - 1/6. Each category has an upper and a lower score bound.[15]

2. For the first two questions of the PHQ-9, select severities such that the score adds up to 4 (so one of (1,3),(2,2),(3,1)).

3. Next, generate a random sequence of numbers from 3 to 9, and pick them one by one. These correspond to questions in the scenario.

4. For each item in the sequence, pick a score from 1 to 3. This represents the severity of the symptom.

5. Check whether the total score exceeds the minimum threshold for the category. If not, repeat step (4). If so:

    - If all questions have been iterated through, stop
    - If score exceeds the max score for that category as well, stop
    - Else, flip a coin. If Heads, pick another question (step (4)). If Tails, stop.

After the scenario, participants are asked to respond to a few questions.

I.6 Outcomes of interest

We divide our outcomes of interest in two groups. The first group of outcomes aims to document gender differences in the recognition of depression symptoms, the severity of depression, along with the willingness to seek help. The second group of outcomes relate to three candidate mechanisms that may drive any observed gender differences in the first group of outcomes.

For the main outcomes related to recognition of depression, we use the answers to the following questions in the *Self* and the *Other* treatments:

[15]According to the PHQ-9 scoring, there are 5 severity levels: 0-4 None or minimal, 5-9 Mild, 10-14 Moderate, 15-19 Moderately severe, 20-27 Severe. We do not focus on the lowest severity level to ensure that we are presenting scenarios where some level of depression is expected.

Q1. Suppose that [you/NAME] were/was experiencing the hypothetical issues at the frequencies listed above, do you think [you/NAME] would have depression? *[Definitely yes, probably yes, probably no, definitely no]*

Q2. How severe do you think the depression would be if [you/NAME] were experiencing these issues in real life? *[None or minimal, Mild, Moderate, Moderately Severe, Severe]*

The main outcome is a binary variable equal to 1 if the answer to Q1 is definitely yes or probably yes, and 0 otherwise. Another primary outcome is whether participants identify the severity of depression correctly, where the variable equals to 1 if the severity level is identified correctly and 0 otherwise.[16]

The third main outcome captures the willingness to seek help after seeing the list of symptoms and answering Q1 and Q2. The questions in the survey are:

Q3. If [you/NAME] were experiencing these hypothetical issues, how likely do you think it is that [you/she,he] will seek help from the following sources? For these questions, imagine that there are no constraints on the time or money that has to be spent, and no problems relating to health insurance coverage for these options. *[Very unlikely, Somewhat unlikely, Somewhat likely, Very likely]*

1. a General Practitioner solely for this purpose
2. a General Practitioner during a visit for another purpose
3. a Psychologist or a therapist
4. a counselor at your workplace or university
5. a close friend or relative
6. an AI-enabled mental health chatbot

We will use the responses to the set in Q3 to generate an outcome variable which equals 1 if the response is somewhat likely or very likely for one or more of the options presented in Q3 and 0 otherwise. As secondary outcomes we have whether help would be sought from a mental health specialist or from non-specialists. The specialist variable will be equal to 1 if the response is somewhat likely or very likely in at least one of options 1-3 in Q3 and 0 otherwise. The non-specialist variable will be equal to 1 if the response is somewhat likely or very likely in at least one of options 4-6 in Q3 and 0 otherwise.

[16]The range of values from 5 to 27 that we generate in our scenarios all correspond to some level of depression from mild to severe.

In the second group of outcomes we propose three different factors that could underlie gender differences in recognizing depression and seeking help. We refer to these three factors as perceptions, psychic costs and norms.

To capture gender differences in perceptions we will generate binary outcomes from each of the answer options in Q4. Each of these variables would be equal to one if the response is somewhat or strongly agree and 0 otherwise. We will also generate an index summarizing the perceptions in a single variable.

Q4. Based on the hypothetical issues and their frequencies experienced by [you/NAME], please indicate the extent to which you agree or disagree with the following statements. *[Strongly disagree, Somewhat disagree, Somewhat agree, Strongly agree]*

1. The issues would go away by themselves in some time
2. The issues are not very serious and do not require treatment
3. I/[NAME] worry/worries about what others would think of me/her/him if they became aware that [I/she,he] had these issues
4. I/[NAME] would rather deal with these issues [myself/herself,himself] than rely on help from others
5. I/[NAME] do/does not think that the available treatments for these issues are effective
6. I/[NAME] am/is worried about the side effects of medication for depression

To study psychic costs, we will use the *Self* and *Other* scenarios and create a binary variable equal to 1 if the scenario being evaluated corresponds to *Other*.

To study norms, we will survey responses to Q5.

Q5. We conducted a similar survey with a sample of 100 Americans. The composition of respondents in this survey was broadly representative of the American population. Participants in that survey evaluated the following exact hypothetical scenario and answered whether or not they thought [NAME] would have depression if he/she were experiencing these symptoms.

List of symptoms

Of the 100 Americans who participated in that survey, how many do you think answered Definitely yes or Probably yes to the question: Suppose that [NAME] were experiencing the hypothetical issues at the frequencies listed above, do you think [NAME] would have depression? [*Number between 0 and 100*]

Everyone in the study will see the same list of symptoms and frequencies. The symptoms are 1, 2, 5, 6, and 7 from the PHQ. The frequencies corresponding to these symptoms (in terms of score levels) are 2, 2, 2, 3, and 2 respectively.

Participants respond with a number from 0 to 100, scaled to a fraction from 0 to 1. For example, if the participant believes that 50% of the Americans participating in the study thought that the scenario represented a depression case, the outcome will take the value of 0.5.

This question is incentivized—participants are told the following: If the difference between your guess and the number of people (out of a 100) who answered Definitely or Probably yes is within +/- 5 of the actual number, you will earn a bonus reward of $ 1.

I.7 Gender differences in depression recognition and seeking help

We are interested in whether men and women differ in how likely they are to recognize depression-related symptoms and to seek help given the symptoms. Recognizing symptoms of depression is crucial, as it is the first step toward seeking help. When symptoms go unrecognized, a condition may remain untreated, potentially leading to severe outcomes, including suicide. Differences in the ability to recognize depression may impose barriers that prevent people from seeking help when dealing with mental health issues.

I.8 Proposed analyses to assess gender differences in recognition and seeking help

Our design involves randomly choosing symptoms and frequencies from the PHQ-9 instrument and showing them in a tabular form to participants. As explained above, we choose symptoms (i.e., items form the PHQ-9 scale) and frequencies so that the combinations of symptoms and frequencies gives an overall score between 5 and 27.

We first propose to analyze the raw data by plotting the three main outcome variables (recognize, correct and seek help) against the PHQ-9 score, and separately by the gender of the respondent.

Second, we will conduct a regression analysis of the three main outcomes on a gender indicator ($Male = 1$ if the participant is male) (see Equation 3), and *Male* along with indicators for severity levels of the scenario (Moderate, Moderately severe and Severe, with excluded indicator Mild) and the interactions between Male and the severity levels (see Equation 4). The regressions will include respondent fixed effects (γ_i). In robustness checks we will explore PHQ-9 item fixed effects and adding controls such as other demographics and familiarity with screening tools for depression and anxiety.

$$y_i = \alpha_0 + \alpha_1 \text{Male}_i + \gamma_i + \varepsilon_i \tag{3}$$

$$y_i = \beta_0 + \beta_1 \text{Male}_i + \beta_2 \text{Moderate}_i + \beta_3 \text{Male}_i \times \text{Moderate}_i + \beta_4 \text{Modsevere}_i + \beta_5 \text{Male}_i \times \text{Modsevere}_i + \beta_6 \text{Severe}_i + \beta_7 \text{Male}_i \times \text{Severe}_i + \gamma_i + \varepsilon_i \quad (4)$$

I.9 Hypotheses

We formulate the following hypotheses regarding the main outcomes (recognize, correct and seek help):

1. Conditional on a severity level, men are less likely to recognize the symptoms as depression and to correctly classify the severity of depression than women.

2. Men are less likely to state that they will seek help than women if the scenario is about *Self*.

3. In the case of scenarios referring to *Other*, it may be the case that the differences in recognition and seeking help are smaller than in *Self* depending on whether the subject of the scenario is a man or a woman. We explore this in Section I.10 and do not formulate a specific hypothesis for now.

I.10 Potential factors underlying gender differences in depression recognition and seeking help

If we find that the hypotheses in Section I.7 are validated, we propose three factors that could be behind these gender differences in depression recognition and seeking help.

Perceptions. Men and women may differ in how they perceive depression symptoms, what others would think of them, how effective treatments options are, or whether they have side effects. Differences in these dimensions are potential candidates for why they decide whether or not to seek help.

Psychic costs. Recognizing that one may be experiencing issues can be difficult for some people because it imposes a emotional and mental burden that can include feelings of shame, guilt, fear, or denial. These psychic costs would be lower when evaluation scenarios about someone else instead of oneself.

Norms. Gender or masculinity norms may affect how men are able to recognize or seek help when they are experiencing mental health issues. Traditional norms often emphasize traits like toughness, self-reliance, and emotional restraint in men, which can create barriers to acknowledging mental health struggles.

I.11 Proposed analyses to provide evidence on potential underlying factors

For perceptions, we will provide visual evidence of gender differences of the response items obtained from Q4 and t-tests for whether the differences are statistically significant. We will also conduct a regression analysis using Equations 3 and 5.

For psychic costs we will present regressions of the three main outcomes on the Male indicator, an indicator for treatment Other, and the interaction between the two:

$$y_i = \delta_0 + \delta_1 \text{Male}_i + \delta_2 \text{Other}_i + \delta_3 \text{Male}_i \times \text{Other}_i + \gamma_i + \varepsilon_i \tag{5}$$

For norms, we will provide regression results of the second order belief (SOB) when the subject of the scenario being evaluated is a man or a woman. We will regress the percent value from Q5 on the Male indicator as in Equation 3.

I.12 Hypotheses

We formulate the following hypotheses regarding perceptions, psychic costs and norms:

1. Men are more likely to respond somewhat or strongly agree to most/all of the six items in Q4 about perceptions than women. In Equation 3, we expect the coefficient α_1 to be larger for men, indicating that the value of the index is higher for men than for women.

2. Psychic costs are larger for men than for women. We hypothesise that δ_3 in Equation 5 is positive and significant, indicating that men are more likely to recognize depression in others than in themselves. We do not have a specific hypothesis on whether this is the case among women.

3. Men have more traditional norms than women regarding mental health issues. Men report similar SOB than women when the subject of the scenario is a woman but lower SOB then women when the subject of the scenario is a man.