Stigma and Social Cover: A Mental Health Care Experiment in Refugee Networks

Emma C. Smith*

Latest Version Here

This Version: July 1, 2025

Abstract

People may withhold useful information from others to avoid becoming associated with a stigmatized service. A field experiment with 849 Syrian refugee friend groups shows, first, low willingness to share information about mental health services, despite largely accurate knowledge of friends' need. But giving social cover, by encouraging individuals to disclose they are paid to share information, raises sharing rates by 34%. Effects are strongest for senders who used mental health services previously. Without social cover, senders ration messages to their highest need friends. Social cover led to more sharing from experienced users, generating greater social connectedness and noisy increases in services use.

*Georgetown University. Email: emma.smith@georgetown.edu. I am grateful to Emily Breza, Rema Hanna, Edward Miguel and Gautam Rao for guidance and support. For helpful comments I thank Layane Alhorr, Jishnu Das, Eliana La Ferrara, Edward Glaeser, Ronak Jain, Asim Khwaja, Guillermo Palacios-Diaz, I. Bailey Palmer, Benjamin Roth, Simone Schaner, Abdulrazzak Tamim, Lise Vesterlund, and seminar, workshop and conference participants at Georgetown University, Harvard University, the University of Pittsburgh, NEUDC, UNHCR-WB Joint Data Center Research Conference on Forced Displacement, and Y-Rise. The project would not have been possible without support from Rachit Shah, Eva Kaplan and the International Rescue Committee, Laura Rodriguez Takeuchi and Matthew Wai-Poi, Noor Daoud, Majd Masannat, and other members of Mindset. This study received generous funding and support from the Innovations for Poverty Action Peace and Recovery Fund, Weiss Fund, SurveyCTO, Weatherhead Institute, Institute for Humane Studies, and Institute for Quantitative Social Sciences. The study received IRB approval from Harvard University protocol number IRB20-1023. The experiments were pre-registered under RCT IDs AEARCTR-0010702 and AEARCTR-0011928.

1 Introduction

Often, services meant to help vulnerable people are associated with stigma. In settings where people rely on personal networks and word-of-mouth (Banerjee et al., 2019, Beaman et al., 2021), this can present an important information friction. While service users' experience makes them a potentially valuable source of firsthand knowledge, they may withhold information in order to hide their user status. And, even non-users may worry that if they share information, others will assume they have used the stigmatized services. This can stifle broader awareness of available services, and prevent others from learning about quality.

I experimentally investigate this in the context of mental health services for Syrian refugees in Jordan. Despite being the leading cause of disability, mental health remains under-treated and stigmatized across the globe (Bloom et al., 2012, WHO et al., 2004 and Pescosolido et al., 2013). Refugees experience particularly high rates of mental health problems – estimates suggest one in three refugees has depression, anxiety or post-traumatic stress disorder (PTSD) (Blackmore et al., 2020). Rates are even higher among Syrian refugees in Jordan, where roughly half the adult population likely has depression or anxiety. But, I document high rates of mental health stigma and little care-seeking. Individuals recognize that they are psychologically distressed and self-report that they believe mental health services could lower their distress, yet only a third can name an organization providing mental health care. Moreover, at baseline I test individuals' knowledge of their friends' mental health need and find they have accurate knowledge of who is more likely to be depressed, above and beyond what can be explained by observable covariates. This together motivates the study hypothesis: fear of being labeled a mental health care user prevents information from circulating.

The main experiment measures Syrian refugees' willingness to share information about a free phone counseling service with their friend group, while subtly varying only the introduction used. A key insight of the paper is that, while the counseling might help friends' mental health, choosing to share the information might hurt the sender's or friend's social image.[1] For one, senders may worry that sharing the information will signal that they used the service themselves. Second, the sender may also worry that sharing information will insult her friend by insinuating that the friend needs mental health services. To identify these concerns I randomly vary two dimensions of the senders' introduction. First, the study encourages some senders to reveal they are being paid to share the campaign. In fact all senders are paid, so disclosing this in some friend groups and not others creates variation only in the "social cover" that senders have. The

[1]Social image is defined following the intuition laid out in Bursztyn and Jensen, 2017: "an individual exhibits social image concerns when her utility depends on the posterior expectations of her type held by others, conditional on observing her behavior".

second source of variation comes from encouraging senders in their introduction either to say that they are trying to share with all of their friends, or say they are trying to share with friends who they think will benefit from the information. To the recipients, this creates variation in whether they think they were targeted. Privately the study asks senders to share information with all of their friends (listed by name) since everyone in this context can likely benefit in some way. 849 refugee friend groups participate in the experiment, consisting of one pre-selected "sender" and their elicited network of close friends who form the "recipients".

The first stage results from the experiment are consistent with low willingness in general to share information about mental health, despite all senders agreeing to share when enrolled. Over half of recipients likely have depression or anxiety based on standard screenings, and the results show that senders often know their friends' need, yet they share the campaign with only about a quarter of recipients.

However, encouraging senders to disclose that they are paid led to significantly more sending. Consistent with the hypothesis that disclosing payment creates a stigma-alleviating excuse, senders are 7.7 percentage points (34%) more likely to share under the two introductions that disclose they are paid, relative to the introduction saying they "want to share" but are still compensated privately (p-value= 0.013). On the other hand I find very little evidence that sending behavior is different when the introduction suggests the recipient was targeted. The increase when disclosing that the sender is paid indicates that sharing mental health information carries a costly signal, and revealing that the sender is paid dampens this signal.

Prior mental health care users, who likely have valuable firsthand information about accessing services and service quality, are the most sensitive to these image concerns. Machine learning heterogeneity points to whether the sender is a prior user as the one covariate most predictive by far of the strongly heterogeneous treatment effects (Chernozhukov et al., 2018). These individuals more than double their sharing rates when encouraged to use the "disclosed" framing relative to the "non-disclosed" framing (p-value = 0.001).

Furthermore, consistent with binding image costs, when senders do not have social cover they ration messages to friends who are most in need. When they do have social cover, in the "disclosed" framing, they send at equally high rates to recipients who do and do not likely have depression or anxiety.

Turning to the impacts of the campaign, there is noisy evidence that recipients take-up mental health services, but not the advertised helpline, after six months. In the most-shared treatment arm (disclosed, non-targeted) there is a marginally significant 7.2 percentage point (31%) increase in having ever used mental health services (p-value= 0.082, ITT estimate). The

pooled IV estimate is statistically insignificant but represents a 51% increase in service use (p-value= 0.222). But for the phone counseling service specifically, the results rule out even low take-up (including any possibility of spillovers) since less than 2% of both the treatment and control groups used the service. Even lowering the fixed cost of take-up by offering to have the helpline contact the person directly did not show any treatment effects for those exposed to the campaign, though lowering this fixed cost did increase overall interest.

The most significant effects of the campaign were on activation of informal social support. Treated recipients experienced a 0.37 standard deviation increase in an index of social connectedness, driven by a 0.45 standard deviation increase in the number of times the recipient spent time helping or being helped by a network member, both significant at the 95 percent confidence level. Treated recipients also engaged in face-to-face or phone conversations about mental health 16 percentage points more (excluding the campaign messages themselves), more than doubling the rate relative to control (p-value = 0.011). This deeper engagement was driven by the "disclosed, targeted" framing, suggesting that recipients responded positively to the sender noticing their need.

Using a follow-up experiment I show that senders can use the excuse of being paid without causing recipients to de-value the services. The follow-up experiment shuts down endogenous selection in who is exposed to which messages by having the enumerator share the information with new recipients on behalf of the new senders. Promisingly, the average effect of disclosing that the sender is paid is insignificant, as is the average effect of targeted phrasing. This means encouraging senders to use an excuse may increase sending without negatively affecting how recipients receive the information. But, there is a significant negative interaction of the two framings, and this is strongly driven by the recipients who are more in need of mental health services. While targeted phrasing used with the disclosure was associated with more positive effects when the message was tailored by the sender in the main experiment, this phrasing can backfire when the delivery is impersonal.

Together these results show, first, that even when individuals know their friends' need for services, in this case mental health care, they may choose to withhold information due to image concerns. And, those with more firsthand experience, whose information may be most useful, are most sensitive. Generalizing the results suggest that, from food banks and unemployment assistance, to HIV and substance use disorder treatment, users of stigmatized services may withhold referrals and recommendations to avoid signaling their own use. Providing social cover alleviates this. Even though revealing their financial incentive could cause senders to be regarded as less prosocial, here the stigma costs dominate and senders demand social cover (Bénabou

and Tirole, 2006, Gneezy et al., 2011). Using social cover does not exacerbate perceptions of stigma, and actually broadens the pool of people who get the messages, which may help normalizing the content. Second, there is evidence of demand by recipients for vulnerability and connection, and a combination of targeted phrasing and social cover can sometimes achieve the dual goal of addressing senders' image concerns while fostering connection with recipients. When senders are encouraged to use targeted phrasing along with the payment disclosure it leads to the deepest engagement between recipients and senders. But, if recipients' vulnerability is highlighted without a show of goodwill it can backfire, as it did when enumerators delivered more impersonal messages. Third and finally, the low-cost campaign showed moderately promising results, leading to both greater social connection and noisy increases in care-seeking.

This paper contributes to three literatures. First, the study advances the literature on social learning, and particularly the role of image and reputational concerns. A central contribution of this paper is that firsthand experience can become a double-edged sword that hinders learning about services that users do not want to be associated with. Experience is an important, possibly necessary, input for social learning (Conley and Udry, 2010). Accounting for sending-side constraints changes how information is expected to flow and the optimal design of outreach programs in the presence of stigma. These results are especially relevant to low and middle income countries where weaker institutional capacity and frequent misinformation lead people to rely on their networks to learn about opportunities (Walsha, 2024). While the prior literature documents reputational concerns in networks in general (Breza and Chandrasekhar, 2019 and Karing, 2018), and in information seeking specifically (Chandrasekhar et al., 2018 and Banerjee et al., 2018), this is one of the first studies to consider both the *sharing* party's side and the recipient's potential benefit, and do so in the context of a stigmatized service. This combination of features is key to the study's contribution, as it this captures a context where sender experience is simultaneously valuable to recipients and costly to senders' image. Additionally the study sheds new evidence on information rationing, to the highest-benefit individuals, under binding image constraints.

Second, the paper contributes to the literature on mitigating the negative consequences of social image costs, by directly dampening signals that lead to negative inferences with social cover (Raisaro, 2023). Closely related work by Bursztyn and co-authors identifies social cover in online experiments, where participants were more willing to express dissenting views on social media, and were judged less harshly by their in-group, when provided with the social cover (Bursztyn et al., 2023). The results of this paper are complementary to those of Bursztyn et al., 2023, and extend the findings to a field experiment setting where participants made naturalistic choices

over sharing information with their close friends about a real and stigmatized service. The fact that a social cover mechanism appeared even in this study, where participants made their sharing choices independently over a one week period, points to the strength of the mechanism identified in both studies. This sheds light on how to incentivize behavior in the presence of real, rather than misperceived, social stigma. In doing so this work builds significantly on a foundational prior literature that, in establishing the importance of social image and perceptions in decision-making, largely focused on either correcting inflated perceptions of stigma (Bursztyn et al., 2020) or avoiding social stigma concerns entirely by providing privacy (Bursztyn et al., 2019, Bursztyn and Jensen, 2017). In many contexts, such as mental health and safety net programs, stigma is widespread and notoriously slow to change. This paper poses an approach to mitigating the immediate frictions posed by those stigmas.

Third, this work contributes to the literature on demand-side constraints in the take-up of mental health services in general, and among refugees in particular. With mental illness the leading cause of disability globally (Bloom et al., 2012), and concentrated among vulnerable populations such as those in low-income countries and displaced people (Banerjee et al., 2023, Stillman et al., 2022), there is an urgent need to increase both supply and demand of mental health services. This paper contributes evidence on the (low) demand for phone counseling among a high-need population, and the potential for facilitating peer-to-peer awareness raising to improve informal support. Furthermore the results point to demand for higher-touch services than phone counseling, and indicate that, when information does circulate from friends, it may increase take-up of those services. This focus on demand-side constraints is motivated by prior work that has established the negative relationship between mental health with economic well-being (Ridley et al., 2020), the efficacy of mental health treatment in low-resource settings including with forcibly displaced populations (Bhat et al., 2022, Harker Roa et al., 2023, Islam et al., 2021), and additional costs of poor mental health such as workplace discrimination (Ridley, 2022). In focusing on stigma and information as barriers to accessing services, this paper also contributes to a rich literature on stigma in other areas of public health, perhaps most notably regarding HIV (Yang et al., 2023). Finally, the paper contributes to a small but growing body of randomized controlled trials in humanitarian settings and with forcibly displaced populations (Alan et al., 2021, Baseler et al., 2023, Hussam et al., 2022, Tamim et al., 2025).

The remainder of the paper is organized as follows: I describe the context in Section 2. Section 3 presents the data and experimental design. Section 4 outlines the conceptual framework. Section 5 describes the empirical strategy, Section 6 presents the results, and Section 7 concludes.

2 Context and Motivation

I conduct the study with a sample of Syrian refugees living in Jordan. Most of the roughly 660,000 Syrian refugees living in Jordan at the time of the study live outside of camps among the host population, and a majority have been in Jordan for over decade, having been predominantly displaced after the Syrian Civil War began in 2011 (UNHCR, 2024). In this setting there is a large mental health burden, with representative surveys suggesting almost half of the adult population likely has depression (Stillman et al., 2022).

Yet I document that use of mental health services remains very low. Prior to the experiments I collect nationally representative data on depressive symptoms and care-seeking among 1516 Syrian refugees in Jordan. Like prior studies, I find that roughly half of adults may likely have depression based on the PHQ-2 scale. But fewer than 7% of households had someone seek mental health services in the previous 2 weeks. These large treatment gaps are mirrored by substantial knowledge gaps as well. Later in the experiments I see that, in the control group, 70% of recipients cannot name a single organization that provides mental health services (even though several humanitarian and local organizations do so in this context), and likewise over 70% have not spoken to anyone outside their household about mental health in the past 6 months.

A first-order question to some may be whether it is reasonable to believe that mental health services can help refugees whose external circumstances are so difficult and outside their control. A majority of Syrian refugees outside of camps in Jordan live below the poverty line and are largely banned from accessing formal employment even after being displaced for over a decade (Portection and Operations, 2022, Erik et al., 2021). As such children have limited employment pathways to aspire to, although school enrollment rates did recover to pre-displacement levels (Krafft et al., 2022). Few refugees expected these conditions to change, with only 20% saying in 2023 that the war in Syrian would likely end in the next two years. In addition to the difficult circumstances while being hosted as refugees, many individuals left Syria in distress and may have experienced trauma before or during their displacement.

Despite these extremely difficult circumstances that Syrian refugees face, a variety of mental health interventions have shown positive effects for this group and similar populations. For example in several refugee and post-conflict settings in-person psychosocial support by non-specialist para-professionals has yielded positive mental health impacts (Rahman et al., 2019, Islam et al., 2021, De Graaff et al., 2023), perhaps most notably through the WHO's widely implemented Problem Management+ intervention. Evidence on lighter-touch interventions (closer to the phone counseling studied in this paper) have yielded smaller but still positive impacts on refugee mental health, often at lower cost. For example different self-guided resources succeeded

in reducing distress among Syrian and Ukrainian refugees (Burchert et al., 2024, Khedari, 2020, Asanov et al., 2024).

It could be that few people use services because of beliefs – they do not believe they need mental health help, or that services are effective. Yet, a majority of control group recipients state that they likely have depression or anxiety, and in fact on average over-report experiencing this relative to validated assessments. Additionally, recipients at baseline report they believe their distress levels would be significantly lower if they used mental health services such as calling a helpline, visiting a specialist, or receiving medication (Figure D.1). These self-reported measures are only suggestive, because they may be biased by social desirability or demand effects, but nonetheless the responses do not point to perceived lack of need as a driving reason for low service take-up. Instead, stigma toward care-seekers emerges as a likely culprit. Over 40% of a representative sample surveyed says that they would not marry someone who once sought professional mental health services. Strikingly, the rates remain high across different segments of the population, such as male or female, above or under the age of 30, married or unmarried (Figure D.2). People are broadly aware that others hold stigmatizing views, as shown by the fact that at baseline half of respondents say they worry their friends would consider them unreliable if they used mental health services.

3 Data and Experimental Design

In order to study the spread of stigmatized information within friend groups, I conduct two rounds of peer referrals, first to construct a sample of potential "senders", and then to collect data on each sender's' friend group and construct the intended "recipient" sample. I then implement the main information sharing experiment in which I "seed" information about the mental health phone counseling service via the senders. Within that experiment I study stigma barriers to sharing the information, by randomizing the framing senders are asked to use when sending information to their friends. From the experiment I study senders' differential willingness to spread information, and recipients' interest in taking up a free phone counseling helpline after exposure to information from their friend.

3.1 Sample and Recruitment

3.1.1 Sender Nominations

The enrolled "sender" sample comprised 849 individuals who agreed to participate after being nominated by peers in an otherwise-unrelated representative survey of Syrian refugees across

Jordan in late 2021 through early 2022. The representative sample that provided the nominations was drawn from the UNHCR universe of registered Syrian refugee households in Jordan. Overlap in networks was minimized by the sampling strategy – the nominating sample is dispersed across the country and accounts for 1% of the registered Syrian refugee households, or less than 0.5% of individuals. Out of 1516 surveys of the representative sample, 726 respondents agreed to nominate individuals and the average number of nominations was 2.5. Potential senders could be nominated if the nominator felt the person fell into any of the following three categories: being "well-regarded or well-know", or "community-minded", or "good at spreading news", with the final category informed by the literature on identifying individuals with high diffusion centrality (Banerjee et al., 2019). Respondents were presented with the three nomination types in random order, and were not made aware of the mental health focus on the intended study.

3.1.2 Sender Eligibility

The sender sample was surveyed by phone in January and February 2023. The enumerator first asked about demographics and attitudes around mental health, but did not mention an awareness campaign or a mental health focus of the study. The sender next completed a social network elicitation focused on the sender's close social network outside her household, such as people the senders socialize with frequently, borrow from or lend to, go to for advice or give advice to, spend time helping or being helped. The median number of friends named was 3. Next, the senders were asked to share the phone numbers for their friends. Conditional on sharing any phone numbers, senders were informed for the first time of the WhatsApp mental health awareness intervention. Senders were asked if they were willing to share mental health awareness information with their friends over WhatsApp, as part of an NGO campaign. Conditional on saying yes, the sender was enrolled in the study to be randomized. The final sender sample consisted of 849 senders who listed friends, provided the phone numbers for their friends, and expressed willingness to participate in the campaign.

3.1.3 Targeting Data

The sender survey also collected information the senders' perceptions of their friends' mental health need. After completing the social network elicitation, and within the same survey, the sender was asked to rank his or her friends according to their benefit from mental health resources. Though this question is sensitive, less than five percent of the sender sample declined to do this ranking, suggestive of high rates of trust in the survey's confidentiality. The respondent was asked:

> "Existing research shows that over 50% of people in Jordan are living in distress, including ongoing sadness, helplessness, stress, or having trouble sleeping. If we go back and think of the [number of friends] friends who you listed, which of them do you think suffer from sadness and stress in their lives, and who would benefit the most from receiving information about identifying and managing psychological distress? Please help me list them in order of who will benefit the most and who will benefit the least."

3.2 Recipient Sample and Baseline

The recipient sample comprised of the senders' friends elicited in the sender survey, and and consisted of 2668 individuals. The final friend networks have little overlap, with only 5% of recipients appearing in more than 1 friend group. This is due in large part to the initial sampling strategy described above, which drew from the refugee population across Jordan.

Short baseline phone surveys were attempted with the new recipients immediately following each sender survey. The timing of recipient baselines and the campaign roll-out were scheduled so that recipient baselines were only attempted before those recipients' senders received the campaign. Of the 2668 recipients, 1422 were reached for a baseline survey. In the survey recipients were assessed for likely depression or anxiety, using the 9-item Patient Health Questionnaire (PHQ-9) (a standard screening tool for depression) and the 2-item Generalized Anxiety Disorder tool to screen for anxiety. Recipients were not informed at that stage of the broader mental health campaign.

3.3 Sender Randomization

Randomization was at the sender level and stratified on gender and the sender's original nominator if the nominator identified multiple senders. Senders were randomized either to treatment (N=642) or control (207), with the treatment group senders asked to share the campaign over WhatsApp with the friends listed in the elicitation. Within the treatment group senders there was additional random variation in how the messages were introduced, which is discussed below. Control group senders were not contacted again for the study and did not receive the awareness content during the experiment period.

Treatment and control are balanced on baseline covariates but some imbalances arose between individual framing arms (F-statistics 1.54 and 1.93 respectively, see appendix). The primary results are highly robust to forcing the inclusion of imbalanced covariates (above and beyond covariates selected by lasso), shown in Appendix Table D.7.

3.4 WhatsApp Messaging Intervention

The content treatment group senders were asked to share was designed by the International Rescue Committee (IRC) in Jordan to increase awareness of mental health need and direct individuals to a free phone counseling service. Specifically the study informed participants of the Jordan River Foundation's free phone counseling helpline.

The Jordan River Foundation (JRF) is a local Jordanian non-profit that is part of the Queen Rania Foundation – a highly respected organization in Jordan. JRF focuses largely on women and children, and operates a free counseling helpline that is available to all. The helpline is operated by professional counselors, and serves as an entry-point to multiple potential services, including one-time assistance, routine counseling by phone, and, for more acute cases, in-person services with a licensed psychiatrist. Additionally the helpline conducts referrals to other providers. JRF partnered with the study to increase awareness and take-up of these services.

The content itself consisted of awareness messages written in text, infographic-type content such as a comic strip, and links to YouTube videos of a Jordanian psychologist discussing how to recognize and manage common symptoms of distress (see appendix for examples). These were developed by the IRC with behavioral science and human-centered design principals in mind, and received extensive input from Syrian refugee community members and the professional and cultural expertise of a Jordanian psychologist. The campaign aimed to help individuals identify whether they are experiencing distress, learn about self-care approaches, introduce the helpline, and highlight that many people in Jordan have used the helpline.

The content was sent in 3 batches over 8 days, and senders additionally received 3 reminders, one each day after a batch of content was shared.[2] The campaign was administered on a rolling basis in weekly batches, such that senders surveyed in a given week typically began receiving the campaign the following week.

Senders were instructed to copy the content and send it to all of their friends who they had named in the original survey. To remind the sender who to message, the recipient friends' names were listed in the instructions each time the sender received new content to share. Senders were incentivized to share screenshots confirming that they sent the campaign to their friends, and could receive $1.40 if they shared documentation of sending at least one piece of content to at least one person. Sender incentives were delivered as e-wallet transfers or phone credit transfers,

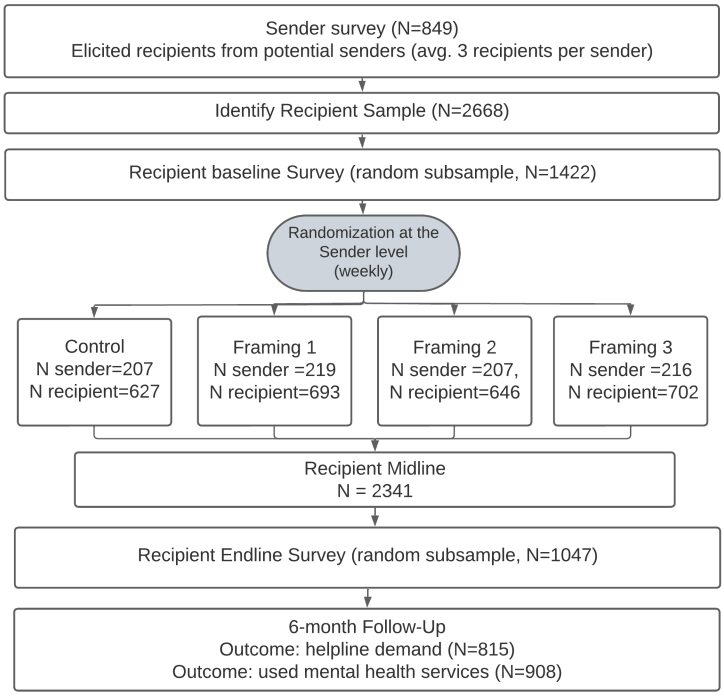
[2] An implementation error caused a random subset of non-disclosed+non-targeted senders to not receive one of the 3 batches of campaign content, and an indicator for this is included in the covariates considered by lasso double selection. The results are qualitatively robust and remain statistically significant if controlling for this or dropping the full affected week, see appendix Table D.7

depending on the respondent's preference.

Follow-up data on recipients collected after the intervention is discussed further below.

The study design is summarized in Figure 1 below.

Figure 1: Experimental Design



3.5 Message Framing Experiment

Senders could worry about what will be signaled by sharing the campaign. Sending the campaign to their friend could signal that they have used the services before themselves, and it could also signal that they think their friend needs mental health services, both of which are stigmatized. Variation in the introduction that senders were asked to share identifies these concerns.

Within the sender treatment group one sentence in the WhatsApp message that introduced the content to the recipient was randomized. The content that the sender was instructed to share began with "Here is some mental health information I received from the International Rescue Committee." This was followed by one of the three sentences below.

> Disclosed Compensation + Non-targeted: An NGO is compensating me to share this with all of my close friends.
>
> Disclosed Compensation + Targeted: An NGO is compensating me to share this with friends who I think can benefit from the information.
>
> Non-Disclosed Compensation + Non-targeted: I want to try to share this with all of my close friend.

In all three conditions the senders were offered the same incentives for participation, and reminded of the incentive each time they received a batch of content or a reminder. And, in all three conditions the sender was told to share the content with all of the friends who she mentioned in the baseline survey, since everyone in this context can benefit from the information, even if only to be able to share it with others. To be clear that all the friends should get the information, all the sender's recipients' names were listed in the instructions that she received with every batch of content. Because the actual compensation and intended recipient group were held constant across senders, the framing conditions vary only the recipient's perception of why the sender shared information.

The randomization achieved balance between treatment and control on 15 of 16 covariates, with an insignificant F-statistic of 1.02. By framing arm, some imbalances arose, with the test of joint significance across covariates for framings 1 and 2 relative to framing 3 being 1.93 and 1.54 respectively, both statistically significant at traditional confidence levels. The imbalanced covariates are included in the list of covariates considered in the pre-specified lasso double selection procedure, and as an additional robustness test I present the primary results when forcing the inclusion of those imbalanced covariates, regardless of whether they are selected by lasso. The robustness test shows that the primary results are not qualitatively different and remain statistically significant when forcing the inclusion of these covariates.

The framings enable me to test for image concerns when sharing this stigmatized information. First, the "disclosed compensation" versus "non-disclosed compensation" comparison tests whether recipients' knowledge of the financial incentive provides social cover that increases senders' sharing. The comparison tests for specifically a social image signaling effect of financial incentives, because the compensation itself is constant across treatments while only visibility of the incentives varies, through disclosure.

The second comparison provided by the framings is that of the "targeted" versus the "non-targeted" framings. This comparison tests whether senders withhold messages that carry a more negative social image signal about the recipient. When the sender tells the recipient that she is trying to send messages to people who may especially benefit, the sender reveals that she think the recipient may be in need. This could be good if it helps the recipient identify that she is good fit for the program. But the sender may worry that her friend will feel insulted or uncomfortable from having her vulnerability revealed. If in fact senders believe and internalize that recipients could feel uncomfortable having their need exposed then senders will be less likely to send the "targeted" framing.

One ex-ante concern could be that revealing that she is paid will be awkward and feel

unnatural for the sender. But, on the contrary, the design for the framing arms was borne out of piloting during which, unprompted, many senders informed their friends that an NGO was paying them to share the WhatsApp information. When the behavior persisted even after the research team repeatedly clarified to pilot participants that they did not need to share this information with their friends, it prompted the hypothesis that senders were seeking social cover. Given that some senders were already using the excuse of being paid, the estimated impacts of the framing arms presented in this paper may be underestimates of the effect of social cover.

3.6 Measuring Sender Sharing

Whether the sender shared the campaign with the recipients is measured using all available data collected through sender screenshots and recipient self-reports at midline and endline surveys described below.[3] The primary analysis uses the measure constructed at the recipient-level, where a recipient is recorded to have received a message if they report this in the midline or endline survey, or if their name shows as the message recipient in a screenshot shared by their sender.

The appendix also includes the primary results at the sender-level, where follow-through which is coded as 1 (relative to 0) if the senders shared any screenshots indicating that they shared the content, or any recipients in the sender's friend group report that they received the content. Note that rates of sending and receiving are not perfectly equivalent since, first, a sender might not share with everyone in her friend group, and second, in some sender screenshots it was impossible to conclusively determine the recipient of the WhatsApp message pictured. In these cases the sender was coded as having shared, but no recipient was coded as "1" in return.

Personalized trackable links provide a third way to measure sender sharing. Each of the three batches of content included a personalized trackable link to a YouTube video with mental health awareness content. The link tracking data does not reveal the user's identity, but indicates how many times the link was clicked on by unique devices. Each sender received unique links allowing me to measure which senders' content was engaged with more regardless of whether the sender or recipient self-reported sending or receiving the content.

[3] In an earlier version of the paper recipients were only counted as having received the messages if they shared the name of the person who sent them the content. That more conservative measure was not pre-specified. The results using that measure are included in the appendix.

3.7 Follow-Up Data Collection for Recipients

3.7.1 Midline

Recipients were contacted the week after their sender received the campaign and were asked whether they received the campaign and had used the advertised helpline. 2,341 recipients were reached for this short midline check-in, representing 88% of the total recipient sample.

3.7.2 Endline

An endline phone survey was conducted with a random sample of recipients three weeks after the recipient's last message was scheduled to be received. 1,046 endline surveys were completed. The endline survey collected the recipients' self-reported use of the advertised helpline, as well as their mental health, stigma attitudes and perceptions, and interactions with their social network.

In addition to the endline survey, the helpline conducted a short survey with 98% all first-time callers to the helpline during the study period. With the caller's consent the helpline recorded the caller's phone number in order for it to be matched to the study sample in the analysis.

3.7.3 6 Month Follow-Up

Two distinct samples of data are collected after 6 months. First, the study attempted to reach female recipients and asked if they would like the helpline to contact them directly to receive free phone counseling. 815 female recipients were reached for this question, representing 55% of the female recipient sample. Only female recipients were included due to the helpline's programmatic priorities. Using this data I construct and analyze an indicator variable for helpline demand.

Additionally, the study attempted to survey all recipient households on a series of mental health questions, including whether anyone in the household had ever used mental health services. Because of the helpline's focus on recruiting female users at that time, in households of male recipients the study surveyed a female member of the same household, rather than surveying the original male recipient himself. Because the question on mental health services asked about all household members, the female respondent can be thought of as a proxy respondent for the original male recipient. 908, or approximately a third, of the recipient households were reached for this survey.

3.7.4 Recipient Attrition

Across the recipient surveys attrition was balanced by overall treatment status and by treatment framing arm, with the exception only of the six month mental health follow-up. There, recipients from one of the framing arms were significantly less likely to be found, as shown in the appendix. The results indicate no significant 6-month impacts of that framing arm, and therefore the differential attrition does not meaningfully influence the interpretation of results.

3.8 Secondary Follow-Up Experiment

Recipients' demand for phone counseling could be affected by the framings in multiple ways. While the "disclosed compensation" framing was effective at increasing sending rates, one might be concerned that this framing will lead recipients to infer that the sender does not think the helpline is useful, and is only sharing the information because she is being paid. Or, recipients may infer from this framing that stigma is high (hence the sender's desire to use an excuse) and be deterred from using the helpline. The "targeted" framing might encourage or discourage recipient take-up. On one hand it may increase take-up if recipients learn from the framing that they are a good fit for the service. But if it causes the recipient to feel singled out and exposed there could be a backlash effect.

Though the main experiment provides evidence on the impact of different message framings on sending rates, it does not identify whether the framing that senders use affects recipients' demand for services. This is because in the main experiment the sender decides whether the recipient gets a message and therefore the estimated effects of messages on recipients cannot be compared to one another. Therefore six months after the main experiment I conduct a smaller follow-up experiment to identify whether differential recipient impacts are at play.

The follow-up experiment sample consists once again of senders and recipient and is summarized in Figure 2. New senders were recruited from among the recipients of the main experiment focusing exclusively on women due to helpline priorities. 908 surveys were completed with former female recipients or women in the household of a former male recipient, and in those surveys 598 named at least 1 female friend, yielding a sample of 1385 potential new recipients. After eliciting the sender's female social network, the enumerator asked the respondent for permission for the study to contact her friends and inform them about the helpline, while mentioning the focus respondent's name. The enumerator explicitly obtained the sender's consent to use each of 4 different framings when introducing the helpline to the focus respondent's friends. Senders received a small financial incentive of on average $2.20 in exchange for agreeing to this.

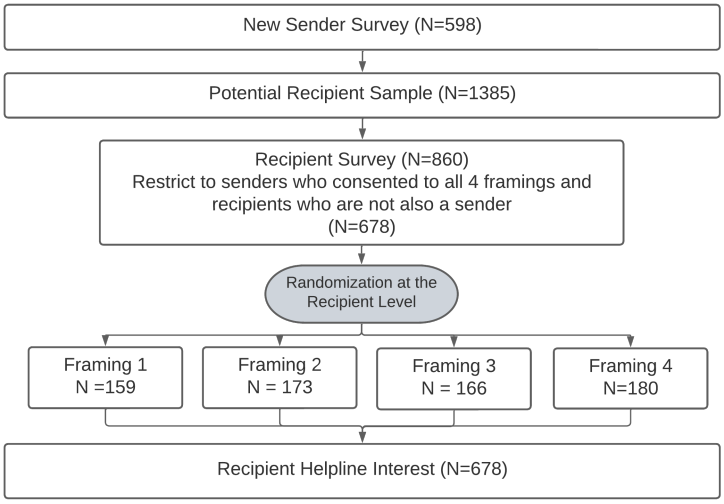
Immediately after the survey the enumerator attempted to contact the friends by phone,

introduced the helpline using a randomly assigned introduction, and elicited their interest in the helpline. Introduction framings were randomized at the final recipient level and stratified on the sender's treatment status in the main experiment and the first framing the sender was exposed to in the follow-up experiment. 860 recipients were reached, and the analysis is restricted to the 678 recipients whose associated senders agreed that the enumerator could use any of the four introductions, and recipients who were not identified to have already participated as a sender in the follow-on experiment.

The four framing conditions reflect the same framings used in the main experiment with the addition of the framing "non-disclosure + targeted".

> Disclosed Compensation + Non-targeted: Our project offered to pay your friend [sender name] to help us check if her friends are interested in this helpline, and she mentioned that you are one of the people she knows.
>
> Disclosed Compensation + Targeted: Our project offered to pay your friend [sender name] to help us check if her friends are interested in this helpline, and she indicated that you or people you know might benefit from the information.
>
> Non-Disclosed Compensation + Non-targeted: Your friend [sender name] wanted us to check if her friends are interested in this helpline, and she mentioned that you are one of the people she knows.
>
> Non-Disclosed Compensation + Targeted: Your friend [sender name] wanted us to check if her friends are interested in this helpline, and she indicated that you or people you know might benefit from the information.

Figure 2: Follow-Up Experiment Design



4 Conceptual Framework

The predicted effects of the message framings on sender's sharing decisions can be fleshed out further through a signaling framework. The conceptual framework closely follows the model proposed by Chandrasekhar et al., 2018, adapted here for sending rather than seeking information and focusing exclusively on the social image costs of sending information rather than including an "interaction" cost as those authors consider.

Senders decide whether to share information given an expected health benefit to the recipient and social image costs to both the sender and recipient. The sender's image cost arises if people associate her with mental health services or, more simply, people think that she is a mental health care user. Likewise recipients face a social image cost if people perceive them to be vulnerable or, "the type of person who needs mental health care".

In the stylized model, a fraction π of the potential sender population are prior users of mental health services (type A) and the remainder are non-users (type B). There are two types of recipients, the vulnerable type V and the unvulnerable type UV, with ω representing the proportion that are vulnerable. Senders choose a binary sending action $S \in 0, 1$ to maximize their utility which is increasing in the benefit to the recipient and decreasing in the social image loss from sending.

The sender's utility from sending action S is given by

$$U(S) = \underbrace{\xi(h\mathbf{1}_{s=1})}_{\substack{\text{Health benefit} \\ \text{to recipient} \\ \text{(sender's belief)}}} - \underbrace{\pi_{post}}_{\substack{\text{social image cost} \\ \text{to sender}}} - \underbrace{\omega_{post}}_{\substack{\text{social image cost} \\ \text{to recipient}}} - C\mathbf{1}_{s=1} \tag{1}$$

where $\xi(h)$ are the sender's beliefs about the benefit for the recipient, given a normally distributed true health benefit h to the recipient, and a monotonic function $\xi(.)$. C is the fixed cost of sending, where $C_{s=0} = 0$. The sender's utility is decreasing in the observer's posterior belief that the sender is a mental health care user or that the recipient is vulnerable, which are given by $\pi_{post} = P(Type_s = A|S)$ and $\omega_{post} = P(Type_r = V|S)$, respectively. This can result in an equilibrium in which there is a cutoff level α^* of expected recipient health benefit above which the senders share and below which they do not. When defining $R(S)$ as the change in image cost from sending, such that $R(S) \equiv \pi_{post,s=1} - \pi_{post,s=0} + \omega_{post,s=1} - \omega_{post,s=0}$, then $\alpha^* = R(S) + C$. If expected benefit is unrelated to sender's and recipient's type then $R(S) = 0$ and senders will share with recipients whose health benefit exceeds the fixed cost C, with no inferences made about sender's or recipient's type.

But, introducing two key assumptions makes it the case that sharing information conveys a

signal about the sender's or recipient's type.

Assumption 1. Prior users have higher efficacy beliefs, such that $\xi_A(h) = \xi_B(h) + q$, $q > 0$, for all h. Let α denote the non-prior user expectation of benefit to recipients, such that $\alpha \equiv \xi_B(h)$.

Assumption 2. Vulnerable types have higher health benefit from services than non-vulnerable types, such that $F_V(h) < F_{NV}(h)$ for all h, where $F(h)$ is the cumulative distribution of h.

Because prior users think that people benefit more from mental health care than non-prior users do (A1), they would, all else equal, always be more likely to share information than non-prior users. Likewise because vulnerable types can benefit more from care than non-vulnerable types (A2), again all else equal, people would be more likely to send them information about services. This however means that sharing information will reveal something about the sender's or recipient's type, creating a social image cost that senders internalize in equilibrium.

Specifically this image cost will increase $R(S)$ and drive up the cutoff of recipient benefit (α_A^*) needed for the sender to justify sharing the message. In equilibrium prior users will continue to share more than non-prior users (since for the same h they believe the recipient benefits more), but both groups will face a higher threshold for being willing to share and so less information will be shared overall. This is detailed further in the appendix.

4.1 Mitigating Social Image Costs

In this situation introducing an observable incentive M can not only directly incentivize sending, but additionally crowd in sending through a secondary effect on the image cost. The sender's utility is then given by

$$U(S) = \underbrace{\xi(h\mathbf{1}_{s=1})}_{\substack{\text{Health benefit} \\ \text{to recipient} \\ \text{(sender's belief)}}} - \underbrace{\pi_{post}}_{\substack{\text{social image cost} \\ \text{to sender}}} - \underbrace{\omega_{post}}_{\substack{\text{social image cost} \\ \text{to recipient}}} - C\mathbf{1}_{s=1} + \underbrace{M\mathbf{1}_{s=1}}_{\substack{\text{monetary incentive} \\ \text{if sends}}} \tag{2}$$

and the corresponding cutoff for prior users will be

$$\alpha_A^* = R(s) + C - q - M \tag{3}$$

while for non-prior users it will be denoted $\alpha_B^* = R(s) + C - M$.[4]

First, the incentivize encourages prior users and non-prior users alike to send more, assuming

[4]Recall that both types will face the same threshold but prior users will believe that recipients receive larger health benefit.

both types have the same preference for money. This can be seen by the fact that the threshold is decreasing in M:

$$\frac{\partial \alpha_A^*}{\partial M} = \frac{\partial R(s)}{\partial M} - 1 \tag{4}$$

As long as the image cost does not increase with M, as confirmed next, then the threshold will go down with an increase in the monetary incentive.

Cash indirectly increases sending further by dampening the negative inferences associated with sharing. Focus for simplicity on the case of only one type of recipient but two sender types, such that the sending cutoff is given by:

$$\alpha_A^* = \varphi(\pi_{post,s=1}) - \varphi(\pi_{post,s=0}) + C - q - M \tag{5}$$

Furthermore assume that the recipient has no way of becoming aware an incentive was offered unless sending occurs, therefore $\frac{\partial \varphi(\pi_{post,s=0})}{\partial M} = 0$. With these assumptions the change in the threshold for a change in M simplifies to

$$\frac{\partial \alpha_A^*}{\partial M} = \frac{\partial \varphi(\pi_{post,s=1})}{\partial M} - 1 \tag{6}$$

Given that α is normally distributed with $\alpha_A = \alpha + q$ and $q > 0$, the change in the image cost $\frac{\partial \varphi(\pi_{post,s=1})}{\partial M}$ will be negative, as further detailed in the appendix. Cash induces both users and non-users to share more, which, critically, reduces the proportion of prior users among the people who share. Even when more users than non-users are induced to share, the proportion of non-users will be higher among the marginal senders than those already sharing, and this will dilute the negative signal. Put another way, cash makes it less "telling" that someone who shares is a prior user, by offering a credible excuse for why someone might share while not being a prior-user. The larger the incentive, the more plausible that excuse. Returning to the scenario with two recipient types, the same mechanisms lead those who receive when there is a known financial incentive to be less vulnerable in expectation than when there is no incentive.

This closely parallels the familiar intuition of how financial incentives may crowd out prosocial actions. Financial incentives might crowd out prosocial actions by dampening the inference about the actor's prosociality (Bénabou and Tirole, 2006). In this setting, the inference is about a stigmatized trait, and so the financial incentive can crowd *in* sharing by dampening the inference about being a prior user.

4.2 Experimentally Identifying Social Image Costs

The secondary effect of the incentive operates purely through the social image channel, and identifying it is the focus of the main experiment. The direct effect of the financial incentive, via a preference for money, does not change when the incentive is public or private and rates of sharing are private. But the secondary effect, by which the incentive dampens the signal conveyed by sharing, comes from its observability. Therefore by varying whether others know about the incentive the experiment identifies the social cover effect, $\frac{\partial \varphi(\pi_{post})}{\partial M}$.

Thus far the framework assumes the sender cares not only about their own image but also the recipient's. The experiment test for evidence of this, by varying the recipient's perception of how targeted sending was. If recipients credibly know there was no targeting, then there is nothing to infer about the recipient's type from receiving the message. In the model this is equivalent to $\omega^{NS} = \omega^{S}$, and this could arise for example if the incentives were very large. Conversely, if recipients are told that they are targeted, then the recipient has greater certainty about the sender's decision criteria, she will now update her beliefs more conditional on the the sender sharing.

This yields two key hypotheses that guide the analysis:

H1: If senders share more when encouraged to disclose that they are paid, while not changing the monetary incentive itself, then image concerns are binding.

H2: If senders share less when assigned to the "targeted" phrasing then recipients' image concerns are binding to the sender.

5 Empirical Strategy

The effect of each framing on sending rates is estimated using first a fully flexible specification and then a specification in which the two disclosure arms are pooled, both at the recipient-level. In all the estimated specifications standard errors are clustered at the sender level.

$$p_r = \alpha_0 + \alpha_1 F_{1s} + \alpha_2 F_{2s} + \alpha F_{3s} + X_s'\beta_1 + X_r'\beta_2 + \Gamma_r + \varepsilon_r \tag{7}$$

$$p_r = \alpha_0 + \psi_1 F_{1or2s} + \psi_2 F_{3s} + X_s'\beta_1 + X_r'\beta_2 + \Gamma_r + \varepsilon_r \tag{8}$$

p_r is a binary indicator of whether the recipient received a message. F_{1s} is an indicator for the "disclosed compensation, non-targeted" framing (Framing 1: "An NGO is compensating me to share this information with all my close friends"). F_{2s} is an indicator for the "disclosed compensation, targeted" framing (Framing 2 : "An NGO is compensating me to share this

information with my friends who I think will benefit from the information.") F_{3s} is an indicator for the "non-disclosed, non-targeted" framing (Framing 3: "I want to try to share this information with all my close friends"). F_{1or2s} is an indicator for the sender being assigned to framing 1 or 2. 5% of the recipients appeared in more than 1 sender friend group, and so, because framings were randomized at the sender level, these recipients could receive multiple treatments which were assigned randomly and independently at the sender level. (The results are robust to dropping these recipients with degree greater than 1, shown in the appendix.) X'_s and X'_r are covariate vectors for the sender and recipient respectively and Γ_r are week of survey fixed effects. The final covariates and fixed effects included in each estimation are selected using the lasso double-selection procedure as was pre-specified. Note that the full pre-specified specification tested the same hypotheses using a less easily digestible specification[5], the results of reflect the same pattern of results and are provided in the appendix (see Table D.2). Additionally 8 of the 51 pre-specified covariates and fixed effects were inadvertently dropped from the baseline data collection and therefore not included.

The same specifications are also estimated using the sender-level outcome of link clicks. In those instances the vector of recipient-level covariates takes the median of the recipient outcomes in the sender's friend group, and the error term is at the sender level (ε_s). The three framings were randomized mutually exclusively over the sender sample and together comprise the complete treatment group.

When it comes to recipient impacts, I consider both the pooled effect of receiving the campaign, and the effect of specific message framing arms as pre-specified. The pooled effect estimated using the following two stage least squares specification:

$$T_s = \gamma_0 + \gamma_1 A_s^T + \gamma_2 A_r^T + X'_r\lambda_1 + X'_s\lambda_2 + \Gamma_r + \nu_r \tag{9}$$

$$y_r = \pi_0 + \pi_1 \widehat{T_s} + + X'_r\phi_1 + X'_s\phi_2 + \Gamma_r + \eta_r \tag{10}$$

where, given the low follow-through rates, T_s is an indicator taking 1 if the sender shared any messages (rather than only if the specific recipient was known to receive a message as was originally pre-specified) and y_r are recipient outcomes such as mental health care take-up and social connectedness. Treatment is instrumented using both the initial sender's assignment to treatment A_s^T and assignment for any sender linked to the recipient (A_r^T), to account for the 5%

[5] $log\left(\frac{p_r}{1-p_r}\right) = \alpha_0 + \alpha_1 I_s^S + \alpha_2 I_s^R + \delta I_s^S I_s^R + X'_s\beta_1 + X'_r\beta_2 + \Gamma_r + \varepsilon_r$ where I_s^S is an indicator for the sender being assigned to a framing that alleviates the "sender social image concern" (ie. F_1 or F_2 framing), and I_s^R is an indicator for the sender being assigned a framing that alleviates the "recipient social image concern" (ie. F_1 or F_3), and the interaction of the two takes one when both concerns are alleviated (corresponding to F_1 only).

of recipients who appeared in more than 1 sender friend group. As before X_s', X_r', and Γ_r are the vectors of sender covariates and recipient covariates, and survey week fixed effects, respectively. The specific covariates and fixed effects included in each estimation are selected using the lasso double-selection procedure.

For the effect of the campaign depending on message framing, I focus on the intent to treat estimates, since the different framings have differential sending rates:

$$y_r = \alpha_0 + \alpha_1 F_{1s} + \alpha_2 F_{2s} + \alpha F_{3s} + X_s'\beta_1 + X_r'\beta_2 + \Gamma_r + \varepsilon_r \tag{11}$$

with all variables following the definitions given above and standard errors are clustered at the sender level.

Lastly, for the follow-up experiment I estimate the effects of the message framings themselves, using the intent to treat estimates. Recall from Section 3.8 that the follow-up experiment used roughly the same three introductions as the main experiment, with the addition of a "non-disclosed, targeted" phrasing. And, in order to avoid endogenous selection in who the sender chose to message, in the follow-up experiment respondents gave the study enumerators permission to contact their friends and introduce the helpline using any of four introductions, referencing the sender's name. The primary outcome is whether the (new) recipients agree to be contacted by the helpline. I obtain the intention to treat estimates using the pre-specified specification below where k_r is a binary indicator for whether the recipient consents to be contacted.

$$k_r = \beta_0 + \beta_1 Disclose_r + \beta_2 Target_r + \delta DiscloseXTarget_r + X_s'\phi_1 + X_r'\phi_2 + \Gamma_r + \varepsilon_r \tag{12}$$

$Dislose_r$ takes 1 if the recipient was assigned to the "disclosed, targeted" or "disclosed, non-targeted" framing and 0 otherwise; $Target_r$ takes 1 if the recipient was assigned to the "disclosed, targeted" or non-disclosed, targeted" framing and 0 otherwise, and $DiscloseTarg_r$ takes 1 if the recipient was assigned to the "disclosed, targeted" framing and 0 otherwise. X_s and X_r are vectors of sender-level and recipient-level covariates, Γ_r are survey week fixed effects, and standard errors are clustered at the sender level. Note that because there is no pure control group, the "non-disclosed compensation, non-targeted" group is the omitted reference category.

6 Main Results

The main analysis first establishes individuals' private knowledge of their friends' mental health status. Second I test how social image concerns affect senders' willingness to share information with their friends. I explore mechanisms using heterogeneity analysis, including whether senders utilize their knowledge of who will benefit most from mental health information when deciding whether to share socially uncomfortable information. Lastly I present impacts on recipients' demand for mental health services and other related outcomes.

6.1 Sender Knowledge

A key rationale for involving community members in outreach efforts is that they have private knowledge about who will benefit most from programs. This could be particularly valuable when fit for a program is not easily observable, as in the case of mental health.

Comparing senders' ranking of who would benefit the most from mental health care to recipients' mental health outcomes shows that senders have quite accurate knowledge. Column 1 of Table 1 regresses a recipient-level indicator for being the recipient the sender thinks would benefit the most on an indicator for whether the recipient likely has depression at baseline, given their PHQ-9 score. This predictive analysis reveals that the friend identified as more in need is 10.7 percentage points, or 24% more likely to have depression (p-value= 0.002). The relationship drops only to 9.4 when controlling for demographics, indicating that senders have information above and beyond observable characteristics (p-value= 0.001). The same analysis with anxiety instead of depression shows that senders' rankings are less strongly predictive of anxiety, but still have a positive and marginally significant correlation (point estimate 5.1 percentage points, p-value= 0.073). Once controlling for demographics this relationship is insignificant. The difference in targeting accuracy may be related to the fact that anxiety was measured using the 2-item GAD-2 and therefore is less precise than the depression indicator which was measured using the 9-item PHQ-9.

6.2 Experimental Effect of Message Framing on Sending Rates

Despite senders agreeing at baseline to share the campaign, knowing their friends are in need, receiving frequent reminders, and being financially incentivized, most recipients never receive the campaign. As shown in Table 2, assignment to treatment led to only a 21 percentage point increase in recipients receiving the campaign, and only 16% of senders had any links clicked on. This first stage is statistically strongly significant, but the low rates of sharing are consistent with the baseline evidence that mental health is a sensitive topic that people rarely discuss.

Table 2 also shows that 6.3% of the control group recipients reporting that they saw the campaign. When only counting recipients who also told the study the name of the person who shared the content the spillovers drop to 1.2% (see appendix). Given these small spillovers the estimated impacts on recipients may be under-estimated.

Table 3 presents the main result of the effect of message framing on sharing rates. Consistent with social cover increasing sharing, encouraging senders to disclose that they were paid led to a 7.2 percentage point increase, equivalent to a 33% increase relative to the non-disclosed group (p-value= 0.047). This estimate comes from comparing the "disclosed, non-targeted" and "non-disclosed, non-targeted" rates of sending presented in column 1. There is a similar 7 percentage point increase in click rates, equivalent an even larger 57% increase, given the low click rates overall (p-value = 0.048).

But I find little evidence that senders withhold messages that signaled that the recipient was targeted on need. The difference in sending rates for the "targeted" and "non-targeted" framings (holding disclosed compensation constant) is less than 1 percentage point and has a p-value of 0.85. The difference in click rates is marginally greater at 1.9 percentage points, but still statistically insignificant. This may mean that senders internalize their own image concerns more than those of the recipient, but could also mean that the encouragement design did not induce first-stage differences in which framing the sender used. The recipient results presented further below point to the first explanation being more likely.

Given the negligible differences between the "targeted" and "non-targeted" version of the "disclosed compensation" framing, I combine the two into one pooled "disclosed compensation" framing that I compare to the "non-disclosed, non-targeted" framing. The pooled point estimates shows a similar 7.7 percentage points increase in sharing when providing social cover through the disclosed compensation (p-value = 0.013), and a 6.5 percentage point increase in click rates (p-value = 0.029).

6.3 Drivers of senders' responsiveness to disclosure

Hereogeneity analysis sheds further light on the connection between the impacts and the social image mechanisms. Using machine learning heterogeneity analysis following Chernozhukov et al., 2018 I first test whether in fact the disclosure effect has heterogeneous effects on the sample, and if yes, then which of 49 covariates are associated with the difference in treatment effects. Figure 4 shows the group average treatment effects of disclosure compared to non-disclosure from the least to most affected quantiles of the sample. The comparison of G5-G1 shows that indeed there are large significant differences between the treatment effects for the least affected and most

affected quantiles. To understand what characteristics are associated with these heterogeneous treatment effects I follow Chernozhukov et al., 2018 in comparing the average characteristics of participants in the most affected quantile to the characteristics of participants in the least affected quantile after retricting to characteristics that vary significantly between the first and fifth quantiles at the 99% confidence level.

One characteristic dominates any other: whether the sender is a prior mental health care user herself. Senders who are prior users are 25 percentage points (125%) more likely to send the disclosed compensation framing than those who have not used mental health services before (Table 4, p-value$<$ 0.001). If responsiveness to the disclosure compensation were driven by something besides social image concerns, such as salience of the financial incentive, it is unlikely that it would generate this pattern of heterogeneity. Instead, the pattern is consistent with disclosure providing social cover that prior users highly demand, perhaps because having in fact used services, they are more sensitive to the risk of "outing" themselves.

Interestingly the effect of disclosure is actually negative for one quantile of the sender sample, and this quantile is the least likely to have used mental health services (see Figure 5). This highlights the competing signal effects that come with using the excuse of monetary compensation. Using the social cover excuse has the benefit of dampening the signal that the sender may be a mental health care user, but comes with the cost of appearing less prosocial. For senders who are less worried about others thinking they used mental health services, the cost of the negative signal dominates, and they become less likely to share. While the paper's conceptual framework does not assume some senders are more sensitive to being perceived as mental health care users, the empirical results point to this being the case.

The prior-user heterogeneity also offers evidence that some senders may actually be sensitive to the more targeted framing that could cause the recipient to feel singled out. Among prior-users, senders are 15.8 percentage points (32%) less likely to send the more targeted phrasing, holding constant the disclosure framing (p-value = 0.079). This is only marginally significant, but points to these prior users being more sensitive to image concerns for not only themselves but also their friends.

In addition to the machine learning heterogeneity I test 5 pre-specified sender characteristics for heterogeneity: mental health service efficacy beliefs, own stigma views, altruism, gender, and social desirability. Of these the effect of disclosure varies only with stigma views, with lower stigma senders being more responsive to the disclosure treatment, whereas senders with high stigma are unaffected (see appendix Table D.8). This aligns with the reduction in image costs only mattering for senders who are near the threshold of being willing to share.

6.4 Do senders target those most in need?

The model assumes that senders choose to share based on the tradeoff between social image costs and potential health benefits to recipients. The interaction of recipient need with message framing points to senders making this trade-off. When senders do not have social cover, in the "non-disclosed" framing, they share the campaign less overall but ration the messages they do share to their higher-need friends. Specifically, without social cover the senders share the "non-disclosed, non-targeted" framing with only 22% of the recipients, as shown above. Breaking this down by recipient need, those senders are almost twice as likely to send those messages to friends who are likely depressed or anxious at baseline compared to friends who are not (11.5 percentage points more on a base of 13.9 percent, p-value = 0.047, Table 5).

But when the social image constraint is alleviated, this rationing disappears: in the "disclosed" framings senders share with both high and lower need recipients at similarly high rates (see Figure 7). Specifically there are no significant differences between rates of sharing with likely versus unlikely depressed or anxious recipients when the sender is assigned to either of the "disclosed compensation" framings, and the point estimates are small (3.7 and 1.5 percentage points).

The responsiveness to encouraging people to disclose that they were being paid, and the associated heterogeneity, is difficult to explain except through demand for social cover. Just like in other contexts, senders may be viewed as less prosocial once their friends know they are paid to share. The fact that most people are still more likely to share speaks to the likely strength of the stigma associated with the mental health content in this setting, and particularly when sharing it on an individual level. It seems that the social cover offered by this excuse is more valuable than being seen as prosocial.

6.5 Recipient Impacts

The main effects of the campaign came through increased social connectedness and conversations with friends about mental health. Receiving the campaign led to a more than doubling of the probability that the recipient had any conversations about mental health (not counting the campaign itself), with a treatment effect of 15.9 percentage points relative to the control mean of 11% (p-value = 0.011, Table 6). An index of social connectedness also increases by 0.37 standard deviations (p-value = 0.046). This index is based on actual interactions (rather than perceptions or attitudes) such as how many times the respondent socialized with, spoke on the phone with or helped or was helped by people outside their household. One of the largest impacts contributing to the index comes through time spent helping others or being helped,

which increases by 0.45 standard deviations (p-value = 0.026). The are the only estimated recipient impacts that survive an FDR multiple hypothesis testing adjustment. The full set of outcomes and q-values can be found in the appendix.

Interestingly, the ITT estimates broken out by framing arm show that the positive effects are predominantly driven by the "disclosed, targeted" framing. These differences are not statistically significant, but the pattern of results is notable. If recipients did react more to more targeted phrasing, it could be because of the framing itself or because of differences in who the sender chose to share with depending on the framing. Given that there is no evidence of targeting on baseline depression or anxiety between this study arm and the "disclosed, non-targeted", the most likely explanation is that recipients reacted positively to feeling their need was noticed by their friend.

The evidence on recipient take-up of mental health services is noisy but still offers some insights. First, I can reject a sharp null hypothesis that the campaign led to take-up of the phone counseling helpline. Receiving the campaign led to an insignificant 1.7 percentage point decrease in the probability of take-up, relative to a control mean of 2 percent. Second, I find no significant effect 6 months later of treatment on the recipient's willingness to have the helpline contact them directly, though rates of agreement are high, with 54% of the control group agreeing.

However, also 6 months after the intervention, there is the suggestion of noisy positive impacts on having used any mental health services. The ITT estimates by framing arm show a noisy significant effect of the disclosed, non-targeted framing (which was shared the most). Assignment to this framing led to a 7.2 percentage point (31%) increase in the probability that the recipient ever used any mental health services (p-value= 0.082, ITT estimate). This estimate, which is driven by therapy use as shown in the appendix, does not survive a multiple hypothesis testing adjustment, but suggests at potential positive impacts on care-seeking. The TOT estimate is insignificant at traditional confidence levels but is 11.8 percentage points (equivalent to a 51% increase). This points to the possibility that awareness campaigns in this context may increase refugees' demand for high touch services more so than low-touch services such as a phone helpline.

6.6 Results of Follow-Up Experiment

The final part of the analysis investigates the impact of the message framings themselves on recipients' demand for services. While the ITT results above broke out impacts by framing arm, those estimates are the result of both the message framing *and* selection in who the sender chose to message. As discussed in Section 3.8, the wording of the framings themselves could cause

recipients to react differently to the information. For example, knowing the sender was paid could lead the recipient to disregard the information. Believing they were targeted could cause the recipients to take the information more seriously, or even prompt backlash and decrease demand if recipients react defensively to feeling targeted.

The results show that neither disclosing compensation nor a more targeted phrasing on their own has a significant impact on willingness to use the helpline, but that the two conditions interact negatively. Column 1 in Table 8 shows that the effect of each framing condition individually is positive and statistically insignificant. But the interaction of the two leads to a marginally significant 12.2 percentage point (15%) decrease in willingness to accept the helpline (p-value = 0.081).

Column 2 tests for heterogeneity by whether the sender ranked the recipient as the most distressed friend[6], and shows that the negative interaction is strongly driven by the most distressed recipients. Exposure to the disclosure and targeted framings together leads to a 44.7 percentage point (64%) decrease in this group's willingness to be contacted by the helpline, relative to the effect of the framings individually (p-value = 0.002). This is not due to more distressed recipients being less willing in general to take up the helpline, and instead is driven by specifically the interaction with the two framings together.

One explanation for this could be emotional: knowing that the sender is paid might place the interaction in a transactional and impersonal domain in which the recipient finds it feels emotionally discordant and inappropriate to have her vulnerability revealed. Or the recipient may even feel that her vulnerability is being exposed for the sender's financial gain. Alternatively the response could be based on beliefs: from the excuse of being paid, the recipient might infer that there is a lot of stigma (hence the sender's motivation to use an excuse). When then told that she is believed to be in need, the recipient could think that the best way to protect her sullied reputation is by denying the services. These interpretations are purely speculative, but point to a cost arising when the messaging is not delivered with care.

Taking stock across the main and follow-up experiments, a few patterns appear. First the excuse of being paid, which increased sending rates greatly, did not on its own decrease recipients' interest in mental health services. This arm in fact led to a marginally significant increase in mental health care use 6 months after the intervention, and short-run increases in conversations with friends about mental health.

When it comes to combining framings, the results become mixed. On one hand in the follow-up experiment the causal effect of telling recipients that their friends were paid and thought they

[6]Actual mental health need is not used because the design did not allow an opportunity to collect recipient baseline data.

were in need is negative, and provokes particularly strong backlash from precisely the recipients who need mental health services the most. Yet in the main experiment, where senders had leeway to adapt the messages further and have follow-on conversations, the largest positive impacts on recipients' social connectedness came from encouraging senders to use a similar framing. One possibility is that the differences in the main experiment are driven by senders targeting different people, however there is no evidence of differential targeting by recipient need in those arms. Instead a more likely explanation is that the more targeted phrasing, when managed by the sender, created greater connection and facilitated follow-on interaction.

7 Conclusion

Together the results of this paper present a consistent picture of how individuals choose to share potentially stigmatizing information, while underscoring the complexity of how others receive that information.

Concern for their social image led many participants to withhold mental health information from their friends, even in this setting where people largely already believed that their friends were in need, and were provided with content pre-tailored to the culture and context. This raises the alarm that even when both information and need are known, stigma may prevent people from not just seeking out information (Banerjee et al., 2018), but even sharing it. Moreover, people closest to the issue can be the most sensitive, making social learning that much more difficult: More distressed recipients prove to be the most sensitive to backlash concerns. And, prior mental health care users presumably have far more knowledge of which services are high or low quality and how best to access them, yet are no more likely to share information unless provided with social cover.

Providing social cover can increase the circulation of information, spark conversations, and even foster social connection. The study finds no evidence that using social cover on its own increases perceptions of stigma. Instead, social cover induced participants with firsthand mental health care experience to share information more often, and to be less inclined to ration messages to only their highest need friends. By increasing the range of people receiving information, social cover may over time help normalize and reduce the stigma associated with receiving mental health content.

How people interpret being potentially singled out appears to influence how they receive information, leading to divergent effects. In the main experiment, where senders could tailor their messages and have follow-on interactions, when senders said the recipient was targeted it led to forging more meaningful connection. Having social cover may have created the entry point

for a deeper conversation. A majority of depressed or anxious recipients at baseline said their friends underestimated their distress, and these people may have valued seeing the sender notice their need. But, a targeted approach also comes with the risk of backlash. When the information delivery was more impersonal and delivered by an enumerator on behalf of the friend, the same framing decreased interest in the services.

The modest intervention impacts, on social connectedness and noisy increases in mental health service use, are promising and indicate that low-cost peer awareness raising may be worth exploring further in similar settings. A focus on connecting to higher-touch services such as in-person or group therapy may generate more interest, given that participants sought out those services. For lower-touch services like phone counseling, more emphasis could also be placed on demand generation for people with only moderate levels of distress. Most people in this setting recognize they are experiencing distress, yet in the endline survey respondents said the main reason they were not using mental health services was that they "do not need them". Further research would be valuable on whether these individuals would benefit from lighter touch services and what could motivate using them.

The paper offers some takeaways for designing outreach for stigmatized services. First, whether to use targeted messaging with potential beneficiaries likely depends on the outreach method and follow-up actions available. Targeted messaging may deepen engagement when outreach is personal and has opportunities for further interaction. But if resource or logistical constraints make personal connection difficult, then a targeted approach can backfire.

Second, leveraging a social cover mechanism can increase outreach effectiveness, and may be critical for engaging people with the most firsthand experience. Many different types of social cover can be used, such as offering stigmatized services in the same location as routine services (Young and Bendavid, 2010), making contract stipulations public to suggest one's "hands are tied" (Raisaro, 2023), and providing other types of observable incentives people can point to to justify their actions.

The demand for social cover indicates the clear presence of mental health stigma and the urgency for additional approaches at direct stigma reduction. In tandem, increasing social cover can circumvent existing stigma and improve the chances of people accessing sometimes critical care.

References

Alan, S., Baysan, C., Gumren, M., & Kubilay, E. (2021). Building social cohesion in ethnically mixed schools: An intervention on perspective taking. *The Quarterly Journal of Economics*, *136*(4), 2147–2194.

Asanov, A.-M., Asanov, I., & Buenstorf, G. (2024). A low-cost digital first aid tool to reduce psychological distress in refugees: A multi-country randomized controlled trial of self-help online in the first months after the invasion of ukraine. *Social Science & Medicine*, *362*, 117442.

Banerjee, A., Breza, E., Chandrasekhar, A. G., & Golub, B. (2018). *When less is more: Experimental evidence on information delivery during india's demonetization* (tech. rep.). National Bureau of Economic Research.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., & Jackson, M. O. (2019). Using gossips to spread information: Theory and evidence from two randomized controlled trials. *The Review of Economic Studies*, *86*(6), 2453–2490.

Banerjee, A., Duflo, E., Grela, E., McKelway, M., Schilbach, F., Sharma, G., & Vaidyanathan, G. (2023). Depression and loneliness among the elderly in low-and middle-income countries. *Journal of Economic Perspectives*, *37*(2), 179–202.

Baseler, T., Ginn, T., Hakiza, R., & Woldemikael, O. (2023). Can redistribution change policy views? aid and attitudes toward refugees in uganda.

Beaman, L., BenYishay, A., Magruder, J., & Mobarak, A. M. (2021). Can network theory-based targeting increase technology adoption? *American Economic Review*, *111*(6), 1918–1943.

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American economic review*, *96*(5), 1652–1678.

Bhat, B., De Quidt, J., Haushofer, J., Patel, V. H., Rao, G., Schilbach, F., & Vautrey, P.-L. P. (2022). *The long-run effects of psychotherapy on depression, beliefs, and economic outcomes* (tech. rep.). National Bureau of Economic Research.

Blackmore, R., Boyle, J. A., Fazel, M., Ranasinha, S., Gray, K. M., Fitzgerald, G., Misso, M., & Gibson-Helm, M. (2020). The prevalence of mental illness in refugees and asylum seekers: A systematic review and meta-analysis. *PLoS medicine*, *17*(9), e1003337.

Bloom, D. E., Cafiero, E., Jané-Llopis, E., Abrahams-Gessel, S., Bloom, L. R., Fathima, S., Feigl, A. B., Gaziano, T., Hamandi, A., Mowafi, M., et al. (2012). *The global economic burden of noncommunicable diseases* (tech. rep.). Program on the Global Demography of Aging.

Breza, E., & Chandrasekhar, A. G. (2019). Social networks, reputation, and commitment: Evidence from a savings monitors experiment. *Econometrica*, *87*(1), 175–216.

Burchert, S., Alkneme, M. S., Alsaod, A., Cuijpers, P., Heim, E., Hessling, J., Hosny, N., Sijbrandij, M., van't Hof, E., Ventevogel, P., et al. (2024). Effects of a self-guided digital mental health self-help intervention for syrian refugees in egypt: A pragmatic randomized controlled trial. *PLoS Medicine*, *21*(9), e1004460.

Bursztyn, L., Egorov, G., Haaland, I., Rao, A., & Roth, C. (2023). Justifying dissent. *The Quarterly Journal of Economics*, *138*(3), 1403–1451.

Bursztyn, L., Egorov, G., & Jensen, R. (2019). Cool to be smart or smart to be cool? understanding peer pressure in education. *The Review of Economic Studies*, *86*(4), 1487–1526.

Bursztyn, L., González, A. L., & Yanagizawa-Drott, D. (2020). Misperceived social norms: Women working outside the home in saudi arabia. *American economic review*, *110*(10), 2997–3029.

Bursztyn, L., & Jensen, R. (2017). Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure. *Annual Review of Economics*, *9*, 131–153.

Chandrasekhar, A. G., Golub, B., & Yang, H. (2018). *Signaling, shame, and silence in social learning* (tech. rep.). National Bureau of Economic Research.

Chernozhukov, V., Demirer, M., Duflo, E., & Fernandez-Val, I. (2018). *Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india* (tech. rep.). National Bureau of Economic Research.

Conley, T. G., & Udry, C. R. (2010). Learning about a new technology: Pineapple in ghana. *American economic review*, *100*(1), 35–69.

De Graaff, A. M., Cuijpers, P., Twisk, J. W., Kieft, B., Hunaidy, S., Elsawy, M., Gorgis, N., Bouman, T. K., Lommen, M. J., Acarturk, C., et al. (2023). Peer-provided psychological intervention for syrian refugees: Results of a randomised controlled trial on the effectiveness of problem management plus. *BMJ Ment Health*, *26*(1).

Erik, S. S., Aragie, K. T., & Maha, K. (2021). Impact of work permits on decent work for syrians in jordan, reliefweb. *URL: https://reliefweb. int/report/jordan/impact-work-permits-decent-work-syrians-jordan-september-2021-enar*.

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of economic perspectives*, *25*(4), 191–210.

Harker Roa, A., Córdoba Flechas, N., Moya, A., & Pineros-Leano, M. (2023). Implementing psychosocial support models in contexts of extreme adversity: Lessons from a process evaluation in colombia. *Frontiers in psychology*, *14*, 1134094.

Hussam, R., Kelley, E. M., Lane, G., & Zahra, F. (2022). The psychosocial value of employment: Evidence from a refugee camp. *American Economic Review*, *112*(11), 3694–3724.

Islam, A., Mozumder, T. A., Rahman, T., Shatil, T., & Siddique, A. (2021). Forced displacement, mental health, and child development: Evidence from the rohingya refugees. *Working Paper*.

Karing, A. (2018). Social signaling and childhood immunization: A field experiment in sierra leone. *University of California, Berkeley*, *2*.

Khedari, V. K. (2020). *Out of ivory towers and into refugee camps: Providing refugees with accessible resources about the biological and psychological effects of forced migration* [Doctoral dissertation, The New School].

Krafft, C., Sieverding, M., Berri, N., Keo, C., & Sharpless, M. (2022). Education interrupted: Enrollment, attainment, and dropout of syrian refugees in jordan. *The Journal of Development Studies*, *58*(9), 1874–1892.

Pescosolido, B. A., Medina, T. R., Martin, J. K., & Long, J. S. (2013). The "backbone" of stigma: Identifying the global core of public prejudice associated with mental illness. *American journal of public health*, *103*(5), 853–860.

Portection, E. C., & Operations, H. A. (2022). *Echo factsheet - jordan (19/06/2022)*. Retrieved January 5, 2024, from https://reliefweb.int/report/jordan/echo-factsheet-jordan-19062022

Rahman, A., Khan, M. N., Hamdani, S. U., Chiumento, A., Akhtar, P., Nazir, H., Nisar, A., Masood, A., Din, I. U., Khan, N. A., et al. (2019). Effectiveness of a brief group psychological intervention for women in a post-conflict setting in pakistan: A single-blind, cluster, randomised controlled trial. *The Lancet*, *393*(10182), 1733–1744.

Raisaro, C. (2023). Incentives justifying nonconformity: Experimental evidence from motortaxi organizations in uganda. *Working Paper*.

Ridley, M. (2022). *Essays on the economics of mental illness and belief formation* [Doctoral dissertation, Massachusetts Institute of Technology].

Ridley, M., Rao, G., Schilbach, F., & Patel, V. (2020). Poverty, depression, and anxiety: Causal evidence and mechanisms. *Science*, *370*(6522), eaay0214.

Stillman, S., Rozo, S. V., Tamim, A., Palmer, I. B., Smith, E., & Miguel, E. (2022). The syrian refugee life study: First glance. *Oxford Review of Economic Policy*, *38*(3), 625–653.

Tamim, A., Smith, E. C., Palmer, B., Miguel, E., Leone, S., Rozo, S. V., & Stillman, S. (2025). *Housing subsidies for refugees: Experimental evidence on life outcomes and social integration in jordan* (tech. rep.). National Bureau of Economic Research.

UNHCR. (2024). *Syrian refugee regional response*. Retrieved January 5, 2024, from https://data2.unhcr.org/en/situations/syria/location/36

Walsha, M. (2024). Do social image concerns affect technology diffusion? evidence from mobile banking in pakistan.

WHO, W. M. H. S. C., et al. (2004). Prevalence, severity, and unmet need for treatment of mental disorders in the world health organization world mental health surveys. *Jama*, *291*(21), 2581–2590.

Yang, D., Allen Iv, J., Mahumane, A., Riddell Iv, J., & Yu, H. (2023). Knowledge, stigma, and hiv testing: An analysis of a widespread hiv/aids program. *Journal of Development Economics*, *160*, 102958.

Young, S. D., & Bendavid, E. (2010). The relationship between hiv testing, stigma, and health service usage. *AIDS care*, *22*(3), 373–380.

8 Tables and Figures

Figure 3: Prevalence of Depression among Recipients at Baseline

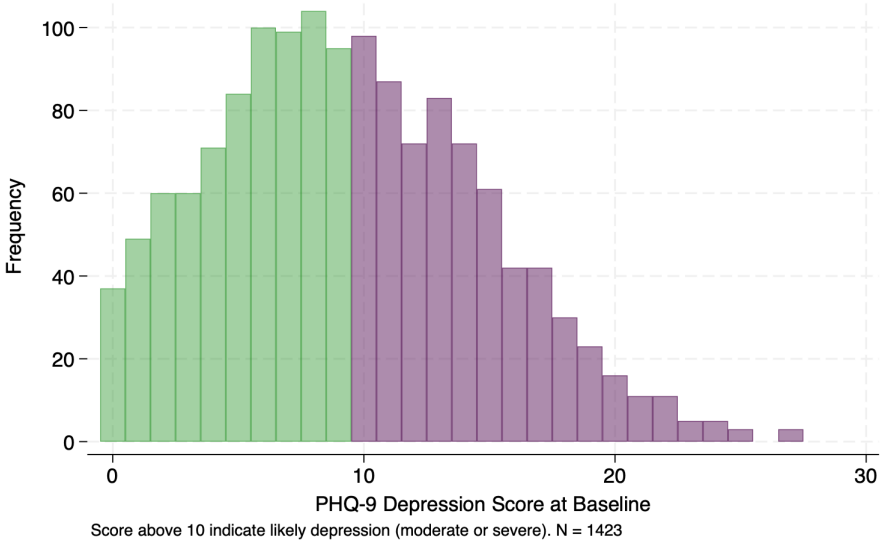


Figure 4: Prevalence of Likely Anxiety Among Recipients at Baseline

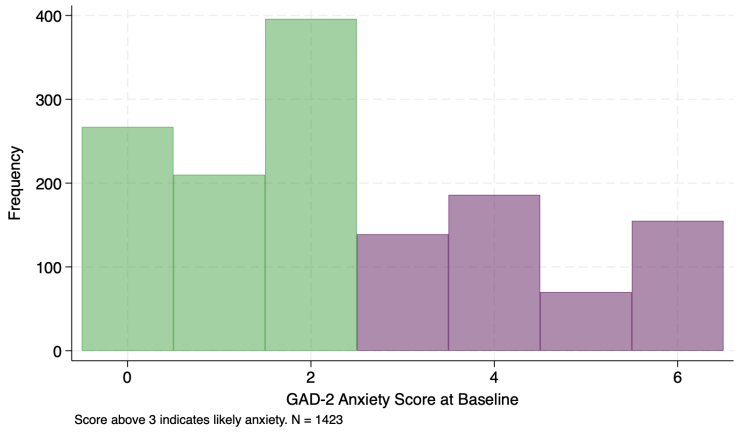


Table 1: Sender ability to target

	Recipient Depressed at Baseline		Recipient has Anxiety at Baseline	
	No Covariates	With Covariates	No Covariates	With Covariates
	(1)	(2)	(3)	(4)
Highest need recipient in friend group	0.107***	0.094***	0.051*	0.042
	(0.029)	(0.028)	(0.028)	(0.028)
Control Mean	0.436	0.436	0.375	0.375
Covariates	No controls	Demographics Selected by lasso	No controls	Demographics Selected by lasso
N	1330	1330	1330	1330

This table shows the association between senders' indication that a friend (recipient) is or is not the most in need of mental health services, and that recipient's baseline propensity to be depressed or have anxiety. Observations are at the recipient level. The sample is restricted to instances when the sender has more than 1 friend and includes only the recipients that were reached for the baseline survey. The independent variable is a binary variable of the sender having indicated that the recipient would benefit the most from mental health information. The dependent variable in columns 1 and 2 is an indicator for whether the recipient's PHQ-9 score at baseline indicates that the recipient likely has moderate to severe depression (10 or higher). The dependent variable in columns 3 and 4 is an indicator variable for whether the recipient's GAD-2 score at baseline indicates that the respondent likely has anxiety (score 3 or higher). It should be noted that the GAD-2 is only a 2-question screening and thus more imprecise than the depression measure. Columns 2 and 4 includes recipient demographic controls that are selected using the lasso double selection procedure. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2: Sender Compliance

	(1) Received campaign (recipient-level)	(2) Any clicks (0/1) (sender-level)
Treatment (sender asked to share)	0.210***	0.161***
	(0.020)	(0.015)
Control Mean	0.063	–
Covariates	Lasso Double Selection	Lasso Double Selection
N	2668	849

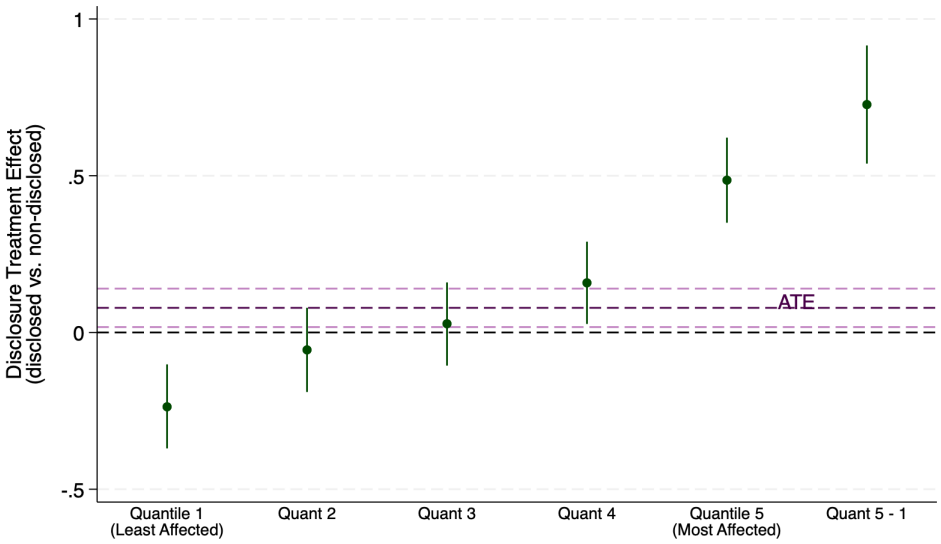
This table shows the rate at which treated senders participated by sending messages to recipients. The dependent variable in column 1 is whether the given recipient received a message from the sender. The dependent variable in column 2 is whether the sender's links were ever clicked on. A recipient is recorded to have received a message if they report this in the midline or endline survey, or if their name shows as the message recipient in a screenshot shared by their sender. Standard errors clustered at the sender level. Covariates are selected using lasso double-selection from a list of sender and recipient covariates following Belloni et al. 2014. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3: Effect of Message Framing on Sender Sharing

	(1) Received campaign (recipient-level)	(2) Received campaign (recipient-level)	(3) Any clicks (0/1) (sender-level)	(4) Any clicks (0/1) (sender-level)
Disclosed Compensation framing, *non-targeted*	0.230***		0.191***	
	(0.029)		(0.028)	
Disclosed Compensation framing, *targeted*	0.223***		0.172***	
	(0.030)		(0.027)	
Disclosed Compensation framing, *pooled*		0.240***		0.188***
		(0.023)		(0.020)
Non-Disclosed Compensation framing, *non-targeted*	0.158***	0.163***	0.121***	0.123***
	(0.028)	(0.027)	(0.025)	(0.025)
p-values				
$\text{Disclosed}_{non-targeted} - \text{Non-Disclosed}_{non-targeted}$	[.047]		[.048]	
$\text{Disclosed}_{targeted} - \text{Non-Disclosed}_{non-targeted}$	[.071]		[.138]	
$\text{Disclosed}_{targeted} - \text{Disclosed}_{non-targeted}$	[.851]		[.618]	
$\text{Disclosed}_{pooled} - \text{Non-Disclosed}_{non-targeted}$		[.013]		[.029]
Control Mean	0.063	0.063	–	–
Covariates	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection
N	2660	2660	849	849

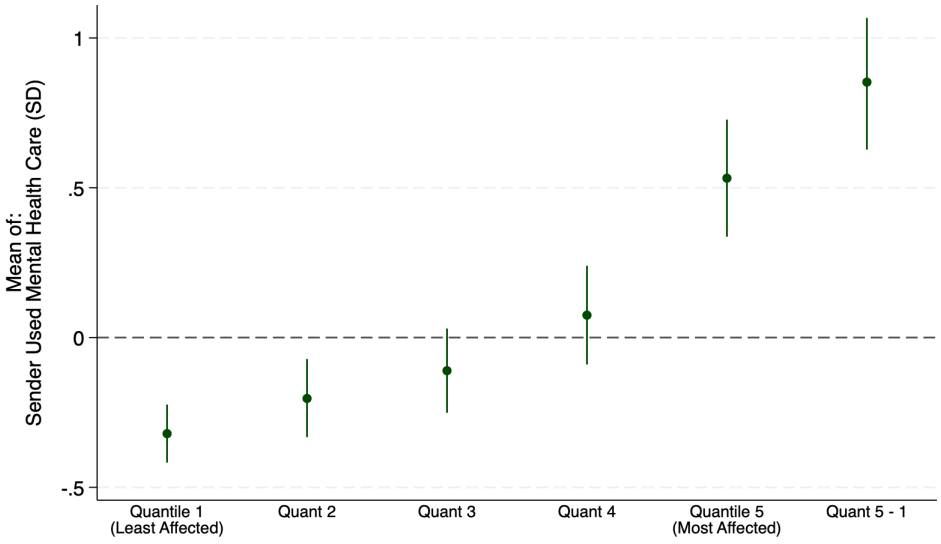
This table shows the rates of sending associated with assignment to each of the framing arms within treatment, relative to the control group which never received the campaign to share. Note that selected control covariates can vary by column. The pooled disclosed compensation framing group comprises the "disclosed compensation, non-targeted" and "disclosed compensation, targeted" groups, which were "An NGO is compensating me to share this *with all of my close friends /friends who I think can benefit from the information.* The non-disclosed compensation framing was always non-targeted, and was "I want to try to share this with all of my close friend." The framing arm coefficients are not additive. P-values are reported in brackets for the differences in point estimates. The dependent variable in columns 1 and 2 is an indicator for whether the recipient received a message from the sender. A recipient is recorded to have received a message if they report this in the midline or endline survey, or if their name shows as the message recipient in a screenshot shared by their sender. The dependent variable in columns 3 and 4 is an indicator for whether there were any clicks to links that were included in the senders' content to the recipients. These clicks may have been by anyone. In the appendix I restrict the variable to take 1 only for instances of more than 1 click from different devices and find a similar pattern of results. Standard errors of the recipient-level analysis in columns 1 and 2 are clustered at the sender level, and standard errors for the sender-level analysis in columns 3 and 4 and are robust to heteroskedasticity. Covariates are selected using the lasso double-selection procedure from a list of sender and recipient covariates following Belloni et al. 2014. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 5: Sorted Group Average Treatment Effects (Effect of Disclosure)



This figure shows the group average treatment effects of disclosure, from the least affect quantile to the most affected quantile, obtained from the machine learning heterogeneity analysis following Chernozhukov et al., 2018's "generic ML" procedure. The figure shows that there is statistically significant heterogeneity in the effect of the disclosure treatment, because the difference between the first and fifth quantiles is strongly statistically significant. The purple lines depicts the estimate and confidence interval for the average treatment effect of the disclosure treatment, pooled across quantiles.

Figure 6: Mean of "Sender Used Mental Health Care" by Quantile of Impact of Disclosure



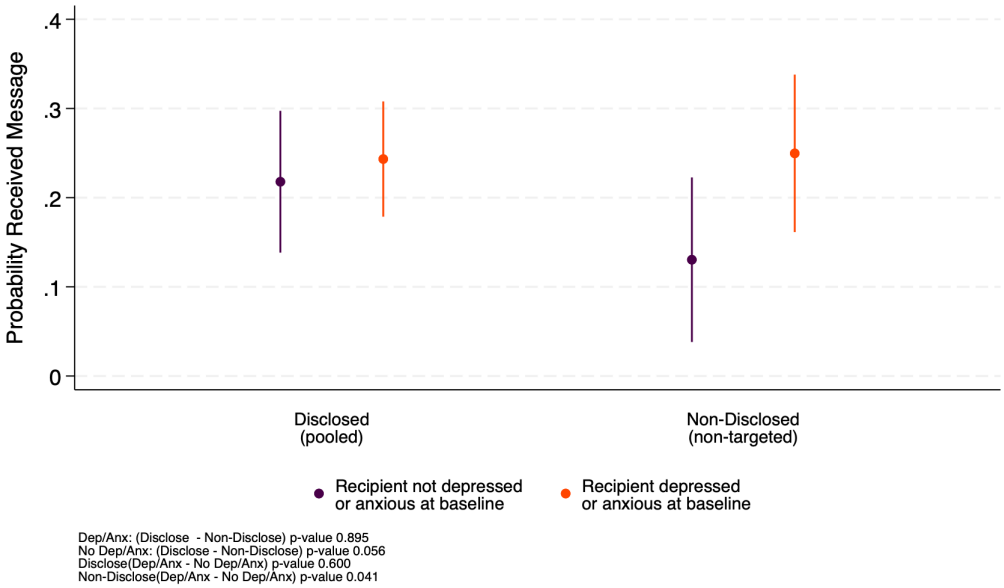
This figure shows that the quantile most affected by the disclosure treatment was much more likely to have used mental health services before compared to the least affected quantile. The figure plots the average value of the standardized variable of whether the the sender has ever used mental health services, for each quantile of the sender distribution. Out of all covariates tested, this variable is by far the most strongly associated with whether the sender is in the top or bottom quantiles of the group average treatment effects of disclosure (shown in Figure 5 above).

Table 4: Message Framing Heterogeneity by Sender Use of Mental Health Services

	(1) Received campaign (recipient-level)	(2) Received campaign (recipient-level)	(3) Received campaign (recipient-level)
Treatment (sender asked to share)	0.173***		
	(0.021)		
Sender has used mental health services X Treatment (sender asked to share)	0.204***		
	(0.056)		
Sender has used mental health services X Disclosed Compensation framing, *non-targeted*		0.314***	
		(0.077)	
X Disclosed Compensation framing, *targeted*		0.143**	
		(0.070)	
X Non-Disclosed Compensation framing, *non-targeted*		-0.015	-0.002
		(0.080)	(0.078)
Sender used mental health services X Compensation framing, *pooled*			0.246***
			(0.061)
Disclosed Compensation framing, *non-targeted*		0.185***	
		(0.027)	
Disclosed Compensation framing, *targeted*		0.197***	
		(0.032)	
Non-Disclosed Compensation framing, *non-targeted*		0.179***	0.182***
		(0.032)	(0.032)
Disclosed Compensation framing, *pooled*			0.198***
			(0.024)
Used mental health services previously	-0.013	0.013	0.001
	(0.033)	(0.035)	(0.034)
p-values for differences of means for senders who have not used mental health services			
$\text{Disclosed}_{non-targeted} - \text{Non-Disclosed}_{non-targeted}$		[.88]	
$\text{Disclosed}_{targeted} - \text{Non-Disclosed}_{non-targeted}$		[.657]	
$\text{Disclosed}_{targeted} - \text{Disclosed}_{non-targeted}$		[.745]	
$\text{Disclosed}_{pooled} - \text{Non-Disclosed}_{non-targeted}$			[.645]
p-value for senders who used services			
$\text{Disclosed}_{targeted} - \textit{Disclosed}_{non-targeted}$		[.081]	
Control Mean	0.063	0.063	0.063
Covariates	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection
N	2662	2662	2662

This table shows the interaction of whether the sender has ever used mental health services with assignment to treatment. The first column shows the interaction with assignment to the pooled treatment. Column 2 shows the interaction with each of the 3 framings, and column 3 shows the interaction when pooling the compensation framings. The dependent variable in column 1 is an indicator for whether the sender of the recipient sent any message to anyone in the friend group. A sender is recorded to have sent any message if the sender shared a screenshot with the study documenting having shared the message, or any of the sender's recipients said in the midline survey or the endline survey that they received messages. Robust standard errors clustered at the sender level. P-values for the difference in means are reported in brackets in the bottom panel. The sample includes all recipients in the experiment. Covariates are selected using lasso double-selection from a list of sender and recipient covariates following Belloni et al. 2014. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 7: Heterogeneous Framing Effects by Recipient Mental Health Need



This figure summarizes the results detailed in Table **??** below. When senders are encouraged to disclose they are paid, they send to high-need and lower-need recipients at similar rates. When senders are not encouraged to disclose, they send to high-need recipients at similar rates to the other sender group, but withhold messages from the lower-need recipients. This is consistent with senders rationing messages to the highest need recipients only when they do not have social cover.

Table 5: Interaction of Recipient Need and Message Framing

	(1) Received campaign (recipient-level)	(2) Received campaign (recipient-level)	(3) Received campaign (recipient-level)
Treatment			
(sender asked to share)	0.189***		
	(0.037)		
Recipient baseline depression or anxiety			
X Treatment (sender asked to share)	0.048		
	(0.044)		
Recipient Depressed/Anxious			
X Disclosed Compensation framing, *non-targeted*		0.037	
		(0.057)	
X Disclosed Compensation framing, *targeted*		0.015	
		(0.059)	
X Non-Disclosed Compensation framing, *non-targeted*		0.115**	0.117**
		(0.058)	(0.059)
X Disclosed Compensation framing, *pooled*			0.027
			(0.049)
Disclosed Compensation framing, *non-targeted*		0.200***	
		(0.047)	
Disclosed Compensation framing, *targeted*		0.229***	
		(0.049)	
Non-Disclosed Compensation framing, *non-targeted*		0.139***	0.133***
		(0.047)	(0.047)
Disclosed Compensation framing, *pooled*			0.218***
			(0.041)
Recipient Depression			
or Anxiety at Baseline (0/1)	-0.033	-0.043	-0.040
	(0.037)	(0.037)	(0.037)
p-values for differences of means for non-depressed recipients			
$\text{Disclosed}_{non-targeted} - \text{Non-Disclosed}_{non-targeted}$		[.235]	
$\text{Disclosed}_{targeted} - \text{Non-Disclosed}_{non-targeted}$		[.092]	
$\text{Disclosed}_{targeted} - \text{Disclosed}_{non-targeted}$		[.594]	
$\text{Disclosed}_{pooled} - \text{Non-Disclosed}_{non-targeted}$			[.064]
Control Mean	0.063	0.063	0.063
Covariates	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection
N	1423	1417	1417

See Figure 7 above for a visual summary of these results. This table shows how the interaction of recipient mental health need with assignment to treatment and message framing impacts the sending decision. The dependent variable in columns 1-3 is an indicator for whether the given recipient received any campaign message. A recipient is recorded to have received a message if they report this in the midline or endline survey, or if their name shows as the message recipient in a screenshot shared by their sender. Column 1 shows that on average senders did not target on recipient need. Column 2 tests for targeting by recipient across the three framing arms. We see that the "non-compensation, non-targeted" framing was targeted to more depressed or anxious recipients, and this targeting closes the gap in sending between the compensation and non-compensation groups. Column 3 repeats this analysis when pooling the two compensation framings. The second panel shows the differences in rates of sharing between framing arms for the base group, which is those without depression or anxiety. P-values for the difference in means are reported in brackets in the bottom panel. Robust standard errors clustered at the sender level. The sample includes only recipients in the subsample reached to be surveyed at baseline. Covariates are selected using lasso double-selection from a list of sender and recipient covariates following Belloni et al. 2014. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Main Estimated Impacts on Recipients

	(1) Any conversations about mental health	(2) Social Connectedness Index (SD)	(3) Labor Assistance Freq. (SD)
Panel A: Pooled IV Estimates			
Sender shared (to anyone)	0.159**	0.372**	0.448**
	(0.063)	(0.186)	(0.202)
Control Mean	0.110	0.000	-0.010
Covariates	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection
N	1038	1042	1042
Panel B: ITT Estimates by Message Framing Arm			
Disclosed Compensation framing, *non-targeted*	0.050*	0.107	0.092
	(0.030)	(0.091)	(0.095)
Disclosed Compensation framing, *targeted*	0.070**	0.205**	0.201**
	(0.034)	(0.092)	(0.102)
Non-Disclosed Compensation framing, *non-targeted*	0.041	0.075	0.085
	(0.031)	(0.092)	(0.098)
Control Mean	0.11	0.00	-0.01
Covariates	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection
N	1038	1042	1042

Notes: This table contains the recipient-level significant impacts that survive a multiple hypothesis testing adjustment. (See appendix for other outcomes and FDR q-values.) Panel A presents the IV estimates pooling the three treatment arms and comparing to control. Panel B presents the ITT estimates for each message framing arm. (IV estimates are not provided by message framing due to experimentally induced differential compliance by arm, per the sender-level results.) The dependent variable in column 1 is an indicator of whether the respondent had conversations about mental health with anyone outside her household in the past 6 months. The dependent variable in column 2 is a standardized index consisting of times that the respondent has socialized with, spoken on the phone with, or helped or been helped by someone in his/her network. The dependent variable in column 3 is a standardized measure of the frequency that the respondent has taken time to help someone outside his/her household with tasks such as childcare, accompanying someone to an appointment, etc., or been helped in similar ways. The sample includes only recipients in the subsample reached to be surveyed at endline. Robust standard errors clustered at the sender level. Covariates are selected using lasso double-selection from a list of sender and recipient covariates following Belloni et al. 2014. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Estimated Impacts on Recipient Take-Up of Mental Health Services

	(1) Called Helpline	(2) Willing to accept call from helpline (after 6 months)	(3) Ever used mental health services (after 6 months)
Panel A: Pooled IV Estimates			
Sender shared (to anyone)	-0.017	-0.045	0.118
	(0.021)	(0.126)	(0.096)
Control Mean	0.020	0.540	0.230
Covariates	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection
N	1041	812	905
Panel B: ITT Estimates by Message Framing Arm			
Disclosed Compensation framing, *non-targeted*	-0.009	0.036	0.072*
	(0.009)	(0.050)	(0.041)
Disclosed Compensation framing, *targeted*	-0.004	-0.078	0.062
	(0.010)	(0.049)	(0.044)
Non-Disclosed Compensation framing, *non-targeted*	-0.008	0.002	0.042
	(0.009)	(0.051)	(0.044)
Control Mean	0.02	0.54	0.23
Covariates	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection
N	1041	812	905

Called helpline (column 1) measured at endline roughly 3 weeks after the intervention. Willing to accept call from helpline (column 2) is measured 6 months after implementation among only original female recipients (due to helpline programmatic priorities). Ever used mental health services is an indicator taking 1 if the respondent reports in the endline or 6 month follow-up that anyone in the household has ever used mental health services. In column 3 the "disclosed compensation, targeted" arm, which has no significant treatment effect, exhibited differential attrition; see appendix Tables D.15 and D.16. All specifications include covariates selected by lasso and standard errors are clustered at the sender level. No results in this table survive a multiple hypothesis testing adjustment; see appendix. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 8: Experiment 2: Impact of Exogenous Messaging on Demand for Phone Counseling

	(1) Willing to accept call from helpline	(2) Willing to accept call from helpline
Disclosed Compensation framing, *pooled*	0.034	-0.019
	(0.051)	(0.077)
Targeted framing, *pooled*	0.048	0.021
	(0.048)	(0.079)
Compensation X Targeted framing	-0.122*	0.068
	(0.070)	(0.105)
Recipient Ranked Most Distressed		
X Disclosed Compensation framing, *pooled*		0.096
		(0.103)
X Targeted framing, *pooled*		0.106
		(0.107)
X Disclosed X Targeted framing		-0.447***
		(0.146)
Recipient Ranked Most Distressed		0.053
		(0.080)
Reference category mean:		
Non-Disclosed compensation, *non-targeted*	0.694	0.694
Double selection	Yes	Yes
N	676	539

This table shows that neither disclosing compensation nor targeted framing individually impact recipients' demand for the helpline, but that, like in the main experiment, the two framings interact negatively. The negative interaction is driven by the most in-need recipients. Given the design of this second experiment there is no selection on receives which message, nor any scope for attrition since the outcome is collected at the time of treatment. The dependent variable is the willingness of the new recipients to be contacted by the helpline to receive phone counseling. The measure of recipient being the most distressed is based on the sender's ranking, because this experimental design did not allow for a recipient baseline survey. Sample sizes differ in the two specifications because not all senders were willing to rank their friends' need. In column 1, missing values of the highest need indicator are imputed and an indicator is included for imputation, while column 2 includes only observations from senders who ranked their friends' need. Appendix Table D.24 shows that the column 1 result is robust to restricting to the column 2 sample. Robust standard errors clustered at the experiment 2 sender level. Covariates are selected using lasso double-selection from a list of sender and recipient covariates following Belloni et al. 2014. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Appendix A. Intervention Content

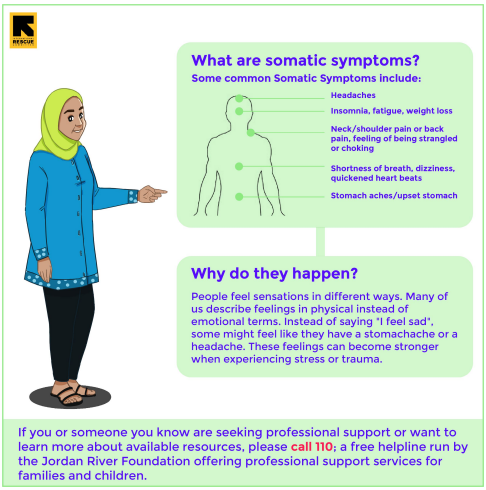


Figure A.1: Example of Campaign Content

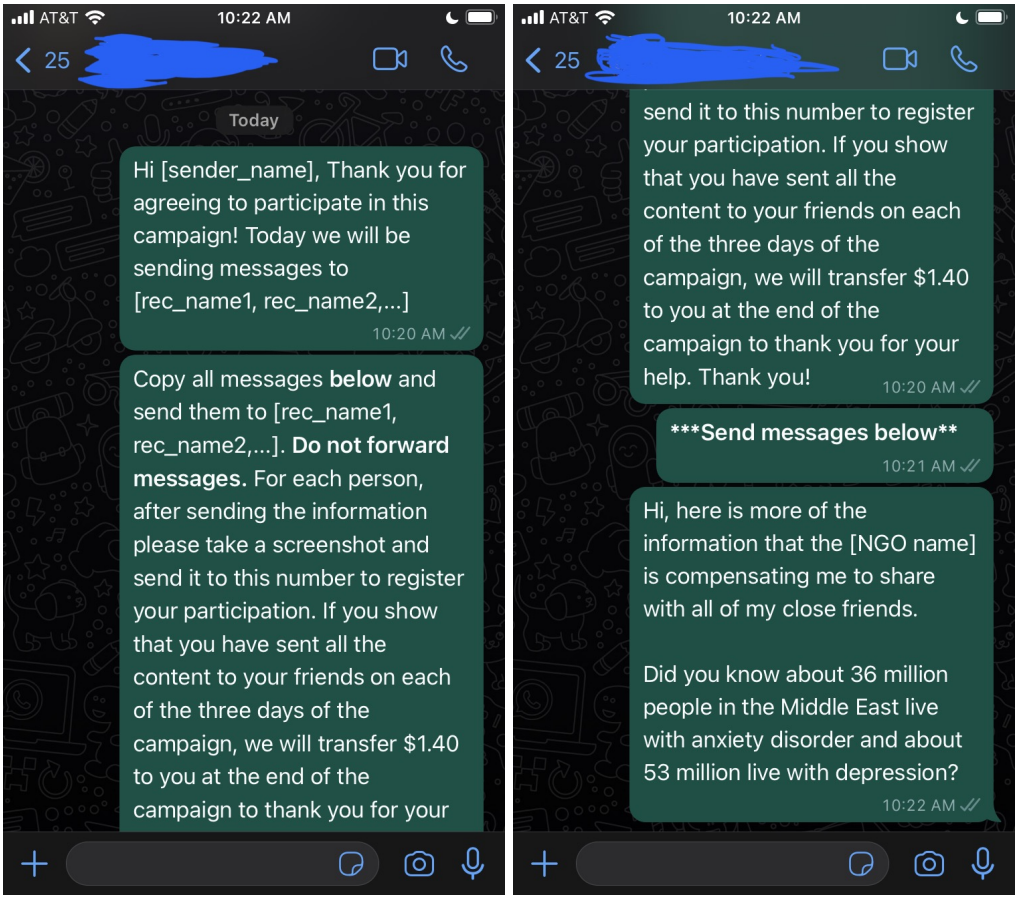


Figure A.2: Example of Campaign Instructions

Appendix B. Sender Elicitation Scripts

- **Well-known or well-regarded**: Think of the people who you know in your community, or the network of people you interact with. From among those people, tell me the name and phone number of one or two people who you know of in your community who are well-known and thought of highly. This could be because their opinions are respected, or simply because they are well-liked.

- **Community-minded**: Now, please tell me the name and phone number of one or two people you know who you believe are community-minded. This could be because they volunteer in an organized way, or they're simply very helpful to others.

- **Good at spreading information**: Now tell me the names and phone numbers of one or two people who, when they share information, many people get to know about it. For example, if they share information about job opportunities, news about Syria, or a wedding, many people would learn about it.

- **Random sample**: Identified through random digit dialing

Appendix C.

Assume in the sender population a fraction π are prior users of mental health services (type A) and the remainder are non-users (type B). The model intuition can be extended to assume two types of recipients, the vulnerable type V and the unvulnerable type UV who differ in their mental health need, with ω representing the proportion of recipients who are the vulnerable type, but here I consider only sender type. Senders choose a binary sending action $S \in 0, 1$ to maximize their utility which is increasing in the benefit to the recipient and decreasing in the social image loss from sending.

$$U(S) = \underbrace{\xi(h\mathbf{1}_{s=1})}_{\substack{\text{Health benefit} \\ \text{to recipient} \\ \text{(sender's belief)}}} - \underbrace{\pi_{post}}_{\substack{\text{social image cost} \\ \text{to sender}}} - C(S) \tag{13}$$

where h is the recipient's health benefit from the information and $\xi(h)$ are the sender's beliefs about the benefit for the recipient where $\xi(.)$ is a monotonic function of h and h is distributed normally. Let α denote non-prior users' beliefs of the health benefit, such at that $\alpha \equiv \xi_B(h)$. $\pi_{post} = P(Type_s = A|S)$ is the posterior belief that the sender is a prior user given the observed sending decision.

This can result in an equilibrium in which there is a cutoff level α^* of expected recipient health benefit above which the senders share and below which they do not. When defining $R(S)$ as the change in image cost from sending, such that $R(S) \equiv \pi_{post,s=1} - \pi_{post,s=0}$, then $\alpha^* = R(S) + C(1)$. Assume that $\pi_{post,s=0} = \pi$, meaning the posteriors are unchanged if no sending occurs (consistent with the experiment in which recipients do not know the sender faced an invitation to share). If expected benefit is unrelated to sender's and recipient's type then $R(S) = 0$ and senders will share with recipients whose health benefit exceeds the fixed cost $C(1)$, with no inferences made about sender's type.

But, introducing a key assumption makes it the case that sharing information conveys a signal about the sender's type.

Assumption 1. Prior users have higher efficacy beliefs, such that $\xi_A(h) = \xi_B(h) + q$, $q > 0$, for all h. Let α denote the non-prior user expectation of benefit to recipients, such that $\alpha \equiv \xi_B(h)$.

When $\xi_A(h) = \xi_B(h) + q$, $q > 0$, for all h, then the resulting equilibrium cutoff $\widetilde{\alpha_B^*}$ is greater than the cutoff α_B^* resulting when $q = 0$.

To see that the differences in type lead to a higher cutoff, consider that the cutoff is given by $\alpha_B^* = \pi_{post,s=1} + C(S)$ and so the cutoff is increasing in q if $\pi_{post,s=1}$ increases with q. The

posterior probability of being a user conditional on sending is given by

$$\pi_{post,s=1}|(q>0) = \frac{(1-F(\alpha-q))\pi}{(1-F(\alpha-q))\pi + (1-F(\alpha))(1-\pi)}$$
$$\pi_{post,s=1}|(q=0) = \pi$$

WTS: $\pi_{post,s=1}|(q>0) > \pi_{post,s=1}|(q=0)$

$$\frac{(1-F(\alpha-q))\pi}{(1-F(\alpha-q))\pi + (1-F(\alpha))(1-\pi)} > \pi$$
$$1-F(\alpha-q) > 1-F(\alpha)$$

This is true for all $q > 0$.

Next, an observable monetary incentive will decrease the cutoff. The sender's utility is now

$$U(S) = \underbrace{\xi(h\mathbf{1}_{s=1})}_{\substack{\text{Health benefit}\\ \text{to recipient}\\ \text{(sender's belief)}}} - \underbrace{\pi_{post}}_{\substack{\text{social image cost}\\ \text{to sender}}} - C(S) + \underbrace{M\mathbf{1}_{s=1}}_{\substack{\text{monetary incentive}\\ \text{if sends}}} \tag{14}$$

and the corresponding cutoff for prior users will be

$$\alpha_A^* = R(s) + C(s) - q - M \tag{15}$$

while for non-prior users it will be denoted $\alpha_B^* = R(s) + C(s) - M$.

Assume that the recipient has no way of becoming aware an incentive was offered unless sending occurs, therefore $\frac{\partial\varphi(\pi_{post,s=0})}{\partial M} = 0$. Then the change in the threshold for a change in M is given by

$$\frac{\partial\alpha_A^*}{\partial M} = \frac{\partial\pi_{post,s=1}}{\partial M} - 1 \tag{16}$$

Therefore the observable incentive will decrease the threshold as long as $\frac{\partial\pi_{post,s=1}}{\partial M} < 1$. Noting that $\frac{\partial\pi^S_{post,s=1}}{\partial M} = \frac{\partial\pi^S_{post,s=1}}{\partial\alpha_A^*}\frac{\partial\alpha_A^*}{\partial M}$, this can be re-stated as:

$$\frac{\partial\alpha_A^*}{\partial M} = \frac{-1}{1 - \frac{\partial\pi_{post,s=1}}{\partial\alpha_A^*}} \tag{17}$$

This shows that as long as $\frac{\partial\pi_{post,s=1}}{\partial\alpha_A^*} < 1$ the monetary incentive will decrease the threshold, and if $0 < \frac{\partial\pi_{post,s=1}}{\partial\alpha_A^*} < 1$ then there will additionally be a crowd-in effect, from the image cost decreasing.

The posterior $\pi_{post,s=1}$ will be increasing in α_A^* if, for any $\alpha < \alpha_A^*$ and $q > 0$, $\frac{f(\alpha-q)}{f(\alpha)} < \frac{f(\alpha_A^*-q)}{f(\alpha_A^*)}$. This holds for normally distributed α. As α_A^* decreases, the proportion of prior users is smaller among the marginal senders induced to participate than among the inframarginal senders.

Appendix D. Supplementary Tables and Figures

List of Tables

List of Figures

Table D.1: Robustness: Recipient only considered to receive message if shared the name of who they got the message from

	(1) Received campaign (recipient-level)	(2) Received campaign (recipient-level)
Disclosed Compensation framing, *non-targeted*	0.217***	
	(0.025)	
Disclosed Compensation framing, *targeted*	0.219***	
	(0.026)	
Disclosed Compensation framing, *pooled*		0.229***
		(0.019)
Non-Disclosed Compensation framing, *non-targeted*	0.163***	0.166***
	(0.024)	(0.024)
p-values		
$\text{Disclosed}_{non-targeted} - \text{Non-Disclosed}_{non-targeted}$	[.12]	
$\text{Disclosed}_{targeted} - \text{Non-Disclosed}_{non-targeted}$	[.099]	
$\text{Disclosed}_{targeted} - \text{Disclosed}_{non-targeted}$	[.828]	
$\text{Disclosed}_{pooled} - \text{Non-Disclosed}_{non-targeted}$		[.033]
Control Mean	0.012	0.012
Covariates	Lasso Double Selection	Lasso Double Selection
N	2668	2668

This table shows the main treatment effects when restricting the outcome variable "received campaign (recipient-level)" to 0 if the recipient does not share the name of who they received the message from. This is a more conservative measure of whether the recipient really got the campaign, and was not pre-specified. The table shows the rates of sending associated with assignment to treatment and assignment to each of the framing arms within treatment, relative to the control group which never received the campaign to share. The framing arms are mutually exclusive and together comprise the complete treatment group, therefore coefficients on the framing arms are not additive. P-values are reported for the differences in point estimates. The dependent variable is an indicator for whether the sender of the recipient sent any message to anyone in the friend group, taking 1 if the recipient says they received the campaign and gives the name of the person who shared the campaign with them. The dependent variable also takes 1 if the recipient is identifiable from screenshots shared by the sender documenting their participation. Robust standard errors clustered at the sender level and reported in parentheses. Covariates are selected using the lasso double-selection procedure from a list of sender and recipient covariates following Belloni et al. 2014. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.2: Pre-Specified Specification for Main Analysis

	(1) Received campaign (recipient-level)	(2) Any clicks (0/1) (sender-level)
Sender's image concerns alleviated	1.813***	7.041***
	(0.238)	(2.055)
Recipient's image concerns alleviated	1.387***	6.538***
	(0.245)	(2.045)
Both sender's and recipients image concerns alleviated	-1.329***	-6.292***
	(0.304)	(2.048)
Estimated effect of sender image concerns $(\alpha_1 + \alpha_2 + \delta) - \alpha_1$	0.484	0.749
p-value	[.013]	[.012]
Estimated effect of recipient image concerns $(\alpha_1 + \alpha_2 + \delta) - \alpha_2$	0.058	0.246
p-value	[.757]	[.342]
Control Mean	0.063	–
Covariates	Lasso Double Selection	Lasso Double Selection
N	2660	848

This table presents the main results using the pre-specified specification. The results are very similar in magnitude and significance, but are less easily digestible. The specification is:

$$log\left(\frac{p_r}{1-p_r}\right) = \alpha_0 + \alpha_1 I_s^S + \alpha_2 I_s^R + \delta I_s^S I_s^R + X_s'\beta_1 + X_r'\beta_2 + \Gamma + \varepsilon_r \tag{18}$$

where I_s^S is an indicator for the sender being assigned to a framing that alleviates the "sender social image concern" (ie. m_1 or m_2 framing), and I_s^R is an indicator for the sender being assigned a framing that alleviates the "recipient social image concern" (ie. m_1 or m_3), and the interaction of the two takes one when both concerns are alleviated (corresponding to m_1 only). Recall that the regression coefficients are rates of sending relative to a pure control group. The reference group for calculating treatment effects is when both concerns are alleviated, ie. $I^S = 1$ and $I^R = 1$, captured by the sum of the three regression coefficients $\alpha_1 + \alpha_2 + \delta$. The estimated effect of alleviating the sender image concern is given by the difference between the rate when both concerns are alleviated versus only recipient image concerns are alleviated: $(\alpha_1 + \alpha_2 + \delta) - \alpha_2$. The estimated effect of alleviating the recipient image concern is given by the difference between the rate when both concerns are alleviated versus only sender image concerns are alleviated: $(\alpha_1 + \alpha_2 + \delta) - \alpha_1$. As in all the analysis, X_s is a vector of sender-level covariates, X_r is a vector of recipient baseline covariates, Γ are week of survey fixed effects, and covariates are selected from the list below using the double post lasso method following Belloni et al. 2014. Differences in rates of sharing between the conditions are reported in the second panel, and p-values are reported in brackets below the associated difference in point estimates. The dependent variable in column 1 is an indicator for whether the sender of the recipient sent any message to anyone in the friend group. In column 2 the dependent variable is a sender-level indicator for whether anyone clicked on any of the sender's links. Robust standard errors clustered at the sender level and reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.3: Robustness: Sender follow-through with unique click data

	(1) Received campaign (recipient-level)	(2) Received campaign (recipient-level)	(3) Any clicks (0/1) (sender-level)	(4) Any clicks (0/1) (sender-level)	(5) More than 1 unique click (0/1)	(6) More than 1 unique click (0/1)
Disclosed Compensation framing, *non-targeted*	0.230***		0.191***		0.088***	
	(0.029)		(0.028)		(0.020)	
Disclosed Compensation framing, *targeted*	0.223***		0.172***		0.065***	
	(0.030)		(0.027)		(0.018)	
Non-Disclosed Compensation framing, *non-targeted*	0.158***	0.163***	0.121***	0.123***	0.040**	0.045***
	(0.028)	(0.027)	(0.025)	(0.025)	(0.016)	(0.016)
Disclosed Compensation framing, *pooled*		0.240***		0.188***		0.085***
		(0.023)		(0.020)		(0.014)
p-values						
$\text{Disclosed}_{non-targeted} - \text{Non-Disclosed}_{non-targeted}$	[.047]		[.048]		[.057]	
$\text{Disclosed}_{targeted} - \text{Non-Disclosed}_{non-targeted}$	[.071]		[.138]		[.29]	
$\text{Disclosed}_{targeted} - \text{Disclosed}_{non-targeted}$	[.851]		[.618]		[.41]	
$\text{Disclosed}_{pooled} - \text{Non-Disclosed}_{non-targeted}$		[.013]		[.029]		[.051]
Control Mean	0.063	0.063	–	–	–	–
Covariates	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection
N	2660	2660	849	849	849	849

This table shows the rates of sending associated with assignment to treatment and assignment to each of the framing arms within treatment, relative to the control group which never received the campaign to share. The framing arms are mutually exclusive and together comprise the complete treatment group, therefore coefficients on the framing arms are not additive. Differences in rates of sharing between framing arms are reported in the second panel. P-values are reported in brackets below the associated difference in point estimates. The dependent variable in columns 1-3 is an indicator for whether the sender of the recipient sent any message to anyone in the friend group. A sender is recorded to have sent any message if the sender shared a screenshot with the study documenting having shared the message, or any of the sender's recipients said in the midline survey (the week after the campaign) that they received messages, or any of the sender's recipients said in the endline survey that they received a campaign message. The dependent variable in columns 4 and 5 is an indicator for whether there were any clicks to links that were included in the senders' content to the recipients. These clicks may have been by anyone. The dependent variable in columns 6 and 7 is an indicator for more than 1 click by different devices. In the appendix I restrict the variable to take 1 only for instances of more than 1 click from different devices and find a similar pattern of results. The last comparison in the second panel comes from running the same specification except that framings 1 and 2 are pooled together. That specification is reported in the appendix but not here to avoid encouraging over-interpretation of the comparison. Robust standard errors clustered at the sender level and reported in parentheses. Covariates are selected using the lasso double-selection procedure from a list of sender and recipient covariates following Belloni et al. 2014. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.4: Robustness: Sender follow-through excluding recipients with duplicate treatments

	(1) Received campaign (recipient-level)	(2) Received campaign (recipient-level)
Disclosed Compensation framing, *non-targeted*	0.245***	
	(0.031)	
Disclosed Compensation framing, *targeted*	0.227***	
	(0.030)	
Disclosed Compensation framing, *pooled*		0.237***
		(0.023)
Non-Disclosed Compensation framing, *non-targeted*	0.144***	0.144***
	(0.028)	(0.028)
p-values		
$\text{Disclosed}_{non-targeted} - \text{Non-Disclosed}_{non-targeted}$	[.008]	
$\text{Disclosed}_{targeted} - \text{Non-Disclosed}_{non-targeted}$	[.02]	
$\text{Disclosed}_{targeted} - \text{Disclosed}_{non-targeted}$	[.812]	
$\text{Disclosed}_{pooled} - \text{Non-Disclosed}_{non-targeted}$		[.004]
Control Mean	0.063	0.060
Covariates	Lasso Double Selection	Lasso Double Selection
N	2549	2549

4.5% of recipients were linked to multiple treated senders and so may have been exposed to the treatment more than once. This table tests the primary hypothesis that sending rates vary with the framing while excluding those individuals. This robustness check is not relevant for the click rate variable, because that outcome is directly tied to the unique sender. The dependent variable in all columns of this table is an indicator for whether the recipient received the campaign. The recipient is recorded to have received the campaign if the sender shared a screenshot with the study documenting having shared a message with that person, or any of the sender's recipients said in the midline survey that they received messages, or the recipient said in the endline survey that they received messages. Covariates are selected using lasso double-selection from a list of sender and recipient covariates following Belloni et al. 2014. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.5: Robustness: Sender follow-through analysis with logistic regression

	(1) Received campaign (recipient-level)	(2) Received campaign (recipient-level)	(3) Any clicks (0/1) (sender-level)	(4) Any clicks (0/1) (sender-level)
Disclosed Compensation framing, *non-targeted*	1.680***		0.713**	
	(0.209)		(0.289)	
Disclosed Compensation framing, *targeted*	1.641***		0.420	
	(0.221)		(0.303)	
Disclosed Compensation framing, *pooled*		1.833***		0.566**
		(0.206)		(0.264)
Non-Disclosed Compensation framing, *non-targeted*	1.290***	1.393***		
	(0.227)	(0.232)		
p-values				
$\text{Disclosed}_{non-targeted} - \text{Non-Disclosed}_{non-targeted}$	[.048]			
$\text{Disclosed}_{targeted} - \text{Non-Disclosed}_{non-targeted}$	[.075]			
$\text{Disclosed}_{targeted} - \text{Disclosed}_{non-targeted}$	[.851]		[.256]	
$\text{Disclosed}_{pooled} - \text{Non-Disclosed}_{non-targeted}$		[.012]		
Mean click rate in $\text{Non-Disclosed}_{non-targeted}$			.111	.111
Control Mean	0.063	0.063	–	–
Covariates	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection
N	2660	2660	642	643

This table presents the primary analysis using logistic regressions rather than linear regressions. The dependent variable in columns 1 and 2 is an indicator for whether the recipient received the campaign, and in columns 3 and 4 it is whether the sender's link was clicked on by anyone. In columns 3 and 4 the regression is run only on the treatment group and coefficients are relative to the non-disclosed non-targeted framing, because convergence does not occur when including the pure control group. Standard errors are robust in all specifications and clustered at the sender level in columns 1 and 2 where the analysis is at the recipient level. Covariates are selected using lasso double-selection from a list of sender and recipient covariates following Belloni et al. 2014. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.6: Robustness: Sender follow-through accounting for imbalance and variation in treatment intensity

	(1) Received campaign (recipient-level)	(2) Received campaign (recipient-level)	(3) Received campaign (recipient-level)	(4) Received campaign (recipient-level)	(5) Received campaign (recipient-level)	(6) Received campaign (recipient-level)
Disclosed Compensation framing, *non-targeted*	0.237***		0.228***		0.231***	
	(0.030)		(0.035)		(0.035)	
Disclosed Compensation framing, *targeted*	0.233***		0.178***		0.186***	
	(0.029)		(0.035)		(0.035)	
Disclosed Compensation framing, *pooled*		0.247***		0.212***		0.217***
		(0.023)		(0.027)		(0.027)
Non-Disclosed Compensation framing, *non-targeted*	0.174***	0.180***	0.138***	0.142***	0.157***	0.161***
	(0.028)	(0.028)	(0.030)	(0.030)	(0.031)	(0.031)
p-values						
$\text{Disclosed}_{non-targeted} - \text{Non-Disclosed}_{non-targeted}$	[.083]		[.031]		[.077]	
$\text{Disclosed}_{targeted} - \text{Non-Disclosed}_{non-targeted}$	[.097]		[.324]		[.48]	
$\text{Disclosed}_{targeted} - \text{Disclosed}_{non-targeted}$	[.916]		[.277]		[.311]	
$\text{Disclosed}_{pooled} - \text{Non-Disclosed}_{non-targeted}$		[.028]		[.044]		[.109]
Sample	All Observations	All Observations	Dropping entire affected week	Dropping entire affected week	Dropping entire affected week	Dropping entire affected week
Covariates	Imbalance Covariates And Lasso Double Selection	Imbalance Covariates And Lasso Double Selection	Lasso Double Selection	Lasso Double Selection	Imbalance Covariates And Lasso Double Selection	Imbalance Covariates And Lasso Double Selection
N	2660	2660	1764	1764	1764	1764

This table tests alternate ways to address imbalance and an implementation glitch that caused variation in the treatment intensity, by causing a random subset of senders in Framing 3 to not receive one of the three batches of campaign content. In column 1 the specification forces the inclusion of imbalanced covariates regardless of whether they were selected by lasso, and also controls for whether the sender experienced the implementation glitch. In column 2 the entire week affected by the implementation glitch is dropped, which leads to a large loss in power. In column 3 the affected week is dropped and imbalanced baseline covariates are forced to be included regardless of whether lasso selected them. Other covariates are selected using lasso double-selection from a list of sender and recipient covariates following Belloni et al. 2014. Robust standard errors are clustered at the sender level. The dependent variable in all columns of this table is an indicator for whether the recipient received the campaign. The results show that the primary results are robust to accounting for these concerns. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.7: Robustness: Sender follow-through measured at the sender level (extensive margin)

	(1) Sender shared (to anyone)	(2) Sender shared (to anyone)
Disclosed Compensation framing, *non-targeted*	0.331***	
	(0.041)	
Disclosed Compensation framing, *targeted*	0.384***	
	(0.042)	
Disclosed Compensation framing, *pooled*		0.380***
		(0.034)
Non-Disclosed Compensation framing, *non-targeted*	0.290***	0.302***
	(0.043)	(0.043)
p-values		
$\text{Disclosed}_{non-targeted} - \text{Non-Disclosed}_{non-targeted}$	[.411]	
$\text{Disclosed}_{targeted} - \text{Non-Disclosed}_{non-targeted}$	[.063]	
$\text{Disclosed}_{targeted} - \text{Disclosed}_{non-targeted}$	[.282]	
$\text{Disclosed}_{pooled} - \text{Non-Disclosed}_{non-targeted}$		[.073]
Control Mean	0.058	0.058
Covariates	Lasso Double Selection	Lasso Double Selection
N	849	849

This table replicates the primary result at the sender level rather than the recipient level. The dependent variable takes 1 if the sender was recorded to have shared the content with anyone, and 0 otherwise. Covariates are selected using lasso double-selection from a list of sender and recipient covariates following Belloni et al. 2014. Standard errors robust to heteroskedasticity. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.8: Heterogeneity in Disclosure Effect by Pre-Registered Sender Characteristics

	(1) Received campaign (recipient-level)	(2) Received campaign (recipient-level)	(3) Received campaign (recipient-level)	(4) Received campaign (recipient-level)	(5) Received campaign (recipient-level)
Disclosed Compensation framing, *pooled*	0.240***	0.239***	0.251***	0.224***	0.240***
	(0.023)	(0.023)	(0.024)	(0.028)	(0.022)
Non-Disclosed Compensation framing, *non-targeted*	0.166***	0.166***	0.163***	0.181***	0.168***
	(0.028)	(0.027)	(0.030)	(0.038)	(0.028)
Sender treatment efficacy beliefs					
X Compensation framing, *pooled*	0.045*				
	(0.024)				
X Non-Disclosed Compensation framing, *non-targeted*	0.017				
	(0.026)				
Mental health service efficacy beliefs (SD)	-0.015				
	(0.019)				
Sender own stigma views above median					
X Compensation framing, *pooled*		-0.035			
		(0.021)			
X Non-Disclosed Compensation framing, *non-targeted*		0.022			
		(0.025)			
Own stigma views (1st order) (SD)		0.014			
		(0.020)			
Sender altruism					
X Compensation framing, *pooled*			-0.002		
			(0.025)		
X Non-Disclosed Compensation framing, *non-targeted*			0.008		
			(0.027)		
Altruism (SD)			-0.003		
			(0.012)		
Sender female					
X Compensation framing, *pooled*				0.033	
				(0.043)	
X Non-Disclosed Compensation framing, *non-targeted*				-0.039	
				(0.054)	
Female				0.047	
				(0.068)	
Sender social desirability					
X Compensation framing, *pooled*					0.027
					(0.023)
Sender social desirability					
X Compensation framing, *pooled*					0.020
					(0.028)
Social desirability score (SD)					-0.006
					(0.015)
p-value: test of heterogeneous effect of disclosure	[.301]	[.05]	[.777]	[.23]	[.801]
Covariates	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection
N	2666	2660	2362	2668	2666

This table shows that the disclosure effect varies by only one of the five pre-specified dimension of sender heterogeneity – sender's own stigma toward mental health care seekers. The p-value at the bottom of the table tests whether the effect of disclosure (Disclosed – Non-disclosed) varies significantly by the given covariate. The pooled disclosed compensation framing group comprises the "disclosed compensation, non-targeted" and "disclosed compensation, targeted" groups, which were "An NGO is compensating me to share this *with all of my close friends /friends who I think can benefit from the information.* The non-disclosed compensation framing was always non-targeted, and was "I want to try to share this with all of my close friend." P-values are reported in brackets for the differences in point estimates. The dependent variable in columns 1 and 2 is an indicator for whether the recipient received a message from the sender. A recipient is recorded to have received a message if they report this in the midline or endline survey, or if their name shows as the message recipient in a screenshot shared by their sender. Standard errors clustered at the sender level and reported in parentheses. Covariates are selected using the lasso double-selection procedure from a list of sender and recipient covariates following Belloni et al. 2014. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.9: Attrition by Survey Round

	(1) Recipient Surveyed Baseline	(2) Recipient Surveyed Baseline	(3) Recipient Surveyed Baseline	(4) Recipient Surveyed Endline	(5) Recipient Surveyed Endline	(6) Recipient Surveyed Endline
Treatment (sender asked to share)	-0.023			-0.022		
	(0.027)			(0.027)		
Disclosed Compensation framing, *non-targeted*		-0.031			-0.021	
		(0.029)			(0.028)	
Disclosed Compensation framing, *targeted*		-0.037			-0.043	
		(0.030)			(0.030)	
Non-Disclosed Compensation framing, *non-targeted*		-0.036	-0.034		-0.029	-0.028
		(0.030)	(0.031)		(0.029)	(0.030)
Disclosed Compensation framing, *pooled*			-0.031			-0.031
			(0.027)			(0.026)
F-Statistic	.76	.734	.8220000000000001	.672	.734	.723
Control Mean	0.551	0.551	0.551	0.409	0.409	0.409
Covariates	No Covariates	No Covariates	No Covariates	No Covariates	No Covariates	No Covariates
N	2668	2668	2668	2668	2668	2668

This table shows that there was no difference in the probability of being treated for recipients reached for the baseline and endline surveys. Note that attrition is not relevant for senders since they are surveyed at the time of enrollment and not again after. Standard errors are clustered at the sender level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.10: Balance on Sender Covariates, Treatment versus Control

		(1) Total		(2) Control		(3) Treatment		(2)-(3) Pairwise t-test
Variable	N	Mean/(SE)	N	Mean/(SE)	N	Mean/(SE)	N	Mean difference
Used mental health services previously	848	0.173	207	0.198	641	0.165	848	0.033
		(0.013)		(0.028)		(0.015)		
Female	849	0.420	207	0.435	642	0.416	849	0.019
		(0.017)		(0.035)		(0.019)		
Age quartile	849	2.370	207	2.290	642	2.396	849	-0.106
		(0.037)		(0.075)		(0.042)		
Lives in refugee camp	849	0.153	207	0.130	642	0.160	849	-0.030
		(0.012)		(0.023)		(0.014)		
Education quartile	849	1.910	207	1.860	642	1.927	849	-0.067
		(0.042)		(0.082)		(0.048)		
Working	849	0.353	207	0.353	642	0.354	849	-0.001
		(0.016)		(0.033)		(0.019)		
Own stigma views (1st order) (SD)	846	-0.038	206	-0.190	640	0.011	846	-0.201**
		(0.034)		(0.065)		(0.040)		
2nd order stigma beliefs (SD)	826	-0.012	199	-0.035	627	-0.005	826	-0.031
		(0.034)		(0.071)		(0.039)		
Altruism (SD)	753	-0.023	184	0.002	569	-0.031	753	0.033
		(0.036)		(0.075)		(0.041)		
Social desirability score (SD)	848	-0.017	207	-0.025	641	-0.014	848	-0.012
		(0.034)		(0.063)		(0.040)		
PHQ-2 depression score (SD)	848	0.050	207	0.066	641	0.045	848	0.021
		(0.034)		(0.067)		(0.040)		
GAD-2 anxiety score (SD)	848	0.032	207	0.061	641	0.023	848	0.038
		(0.034)		(0.068)		(0.040)		
Social connectedness (SD)	848	0.041	207	0.102	641	0.021	848	0.081
		(0.034)		(0.065)		(0.040)		
Jordanian	849	0.110	207	0.101	642	0.112	849	-0.011
		(0.011)		(0.021)		(0.012)		
Mental health service efficacy beliefs (SD)	848	0.033	206	0.067	642	0.022	848	0.045
		(0.032)		(0.065)		(0.038)		
Friend group size	849	3.296	207	3.208	642	3.324	849	-0.116
		(0.059)		(0.118)		(0.068)		
F-test of joint significance (F-stat)								1.018
F-test, number of observations								723

Pair-wise regressions and F-test additionally control for governorate, survey week fixed effects, and randomization strata, which are the controls used in the main analysis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.11: Balance on Sender Covariates, Framing 1 versus Framing 3

	(1)		(2)		(3)		(2)-(3)	
		Total	Framing: Disclosed, nontargeted		Framing: Non-disclosed, nontargeted		Pairwise t-test	
Variable	N	Mean/(SE)	N	Mean/(SE)	N	Mean/(SE)	N	Mean difference
Used mental health services previously	434	0.150	215	0.149	219	0.151	434	-0.002
		(0.017)		(0.024)		(0.024)		
Female	435	0.418	216	0.440	219	0.397	435	0.043**
		(0.024)		(0.034)		(0.033)		
Age quartile	435	2.409	216	2.356	219	2.461	435	-0.105
		(0.051)		(0.072)		(0.071)		
Lives in refugee camp	435	0.170	216	0.185	219	0.155	435	0.030
		(0.018)		(0.026)		(0.025)		
Education quartile	435	1.920	216	1.875	219	1.963	435	-0.088
		(0.059)		(0.082)		(0.084)		
Working	435	0.336	216	0.343	219	0.329	435	0.014
		(0.023)		(0.032)		(0.032)		
Own stigma views (1st order) (SD)	434	0.012	215	0.076	219	-0.051	434	0.128
		(0.049)		(0.072)		(0.068)		
2nd order stigma beliefs (SD)	426	0.025	209	0.108	217	-0.054	426	0.162*
		(0.049)		(0.069)		(0.069)		
Altruism (SD)	381	-0.001	187	-0.021	194	0.019	381	-0.040
		(0.049)		(0.071)		(0.068)		
Social desirability score (SD)	434	-0.027	216	-0.019	218	-0.034	434	0.015
		(0.048)		(0.068)		(0.067)		
PHQ-2 depression score (SD)	434	0.040	216	0.031	218	0.049	434	-0.017
		(0.048)		(0.069)		(0.067)		
GAD-2 anxiety score (SD)	434	0.011	216	-0.031	218	0.053	434	-0.084
		(0.048)		(0.065)		(0.070)		
Social connectedness (SD)	435	0.024	216	0.148	219	-0.098	435	0.246***
		(0.048)		(0.069)		(0.067)		
Jordanian	435	0.117	216	0.167	219	0.068	435	0.098***
		(0.015)		(0.025)		(0.017)		
Mental health service efficacy beliefs (SD)	435	0.072	216	0.063	219	0.081	435	-0.018
		(0.043)		(0.063)		(0.058)		
Friend group size	435	3.340	216	3.361	219	3.320	435	0.041
		(0.079)		(0.113)		(0.112)		
F-test of joint significance (F-stat)								1.932**
F-test, number of observations								369

Pair-wise regressions and F-test additionally control for governorate, survey week fixed effects, and randomization strata, which are the controls used in the main analysis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.12: Balance on Sender Covariates, Framing 1 versus Framing 2

	(1)		(2) Framing:		(3) Framing:		(2)-(3)	
		Total	Disclosed, nontargeted		Disclosed, Targeted		Pairwise t-test	
Variable	N	Mean/(SE)	N	Mean/(SE)	N	Mean/(SE)	N	Mean difference
Used mental health services previously	426	0.174	207	0.198	219	0.151	426	0.047
		(0.018)		(0.028)		(0.024)		
Female	426	0.404	207	0.411	219	0.397	426	0.013
		(0.024)		(0.034)		(0.033)		
Age quartile	426	2.415	207	2.367	219	2.461	426	-0.094
		(0.052)		(0.075)		(0.071)		
Lives in refugee camp	426	0.148	207	0.140	219	0.155	426	-0.015
		(0.017)		(0.024)		(0.025)		
Education quartile	426	1.953	207	1.942	219	1.963	426	-0.021
		(0.060)		(0.085)		(0.084)		
Working	426	0.359	207	0.391	219	0.329	426	0.063*
		(0.023)		(0.034)		(0.032)		
Own stigma views (1st order) (SD)	425	-0.022	206	0.010	219	-0.051	425	0.061
		(0.048)		(0.068)		(0.068)		
2nd order stigma beliefs (SD)	418	-0.061	201	-0.068	217	-0.054	418	-0.014
		(0.048)		(0.066)		(0.069)		
Altruism (SD)	382	-0.036	188	-0.092	194	0.019	382	-0.111
		(0.051)		(0.075)		(0.068)		
Social desirability score (SD)	425	-0.011	207	0.013	218	-0.034	425	0.047
		(0.049)		(0.072)		(0.067)		
PHQ-2 depression score (SD)	425	0.052	207	0.056	218	0.049	425	0.007
		(0.049)		(0.071)		(0.067)		
GAD-2 anxiety score (SD)	425	0.051	207	0.049	218	0.053	425	-0.003
		(0.050)		(0.072)		(0.070)		
Social connectedness (SD)	425	-0.043	206	0.015	219	-0.098	425	0.113
		(0.049)		(0.071)		(0.067)		
Jordanian	426	0.085	207	0.101	219	0.068	426	0.033
		(0.013)		(0.021)		(0.017)		
Mental health service efficacy beliefs (SD)	426	0.001	207	-0.083	219	0.081	426	-0.164*
		(0.047)		(0.073)		(0.058)		
Friend group size	426	3.305	207	3.290	219	3.320	426	-0.030
		(0.084)		(0.127)		(0.112)		
F-test of joint significance (F-stat)								1.465
F-test, number of observations								370

Pair-wise regressions and F-test additionally control for governorate, survey week fixed effects, and randomization strata, which are the controls used in the main analysis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.13: Balance on Sender Covariates, Framing 2 versus Framing 3

	(1)		(2)		(3)		(2)-(3)	
		Total		Framing: Non-disclosed, nontargeted		Framing: Disclosed, Targeted		Pairwise t-test
Variable	N	Mean/(SE)	N	Mean/(SE)	N	Mean/(SE)	N	Mean difference
Used mental health services previously	422	0.173	215	0.149	207	0.198	422	-0.049
		(0.018)		(0.024)		(0.028)		
Female	423	0.426	216	0.440	207	0.411	423	0.029
		(0.024)		(0.034)		(0.034)		
Age quartile	423	2.362	216	2.356	207	2.367	423	-0.011
		(0.052)		(0.072)		(0.075)		
Lives in refugee camp	423	0.163	216	0.185	207	0.140	423	0.045
		(0.018)		(0.026)		(0.024)		
Education quartile	423	1.908	216	1.875	207	1.942	423	-0.067
		(0.059)		(0.082)		(0.085)		
Working	423	0.366	216	0.343	207	0.391	423	-0.049
		(0.023)		(0.032)		(0.034)		
Own stigma views (1st order) (SD)	421	0.044	215	0.076	206	0.010	421	0.066
		(0.050)		(0.072)		(0.068)		
2nd order stigma beliefs (SD)	410	0.022	209	0.108	201	-0.068	410	0.176**
		(0.048)		(0.069)		(0.066)		
Altruism (SD)	375	-0.057	187	-0.021	188	-0.092	375	0.071
		(0.052)		(0.071)		(0.075)		
Social desirability score (SD)	423	-0.003	216	-0.019	207	0.013	423	-0.033
		(0.049)		(0.068)		(0.072)		
PHQ-2 depression score (SD)	423	0.043	216	0.031	207	0.056	423	-0.024
		(0.049)		(0.069)		(0.071)		
GAD-2 anxiety score (SD)	423	0.008	216	-0.031	207	0.049	423	-0.080
		(0.048)		(0.065)		(0.072)		
Social connectedness (SD)	422	0.083	216	0.148	206	0.015	422	0.134
		(0.049)		(0.069)		(0.071)		
Jordanian	423	0.135	216	0.167	207	0.101	423	0.065**
		(0.017)		(0.025)		(0.021)		
Mental health service efficacy beliefs (SD)	423	-0.009	216	0.063	207	-0.083	423	0.146
		(0.048)		(0.063)		(0.073)		
Friend group size	423	3.326	216	3.361	207	3.290	423	0.071
		(0.085)		(0.113)		(0.127)		
F-test of joint significance (F-stat)								1.544*
F-test, number of observations								359

Pair-wise regressions and F-test additionally control for governorate, survey week fixed effects, and randomization strata, which are the controls used in the main analysis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.14: Baseline Recipient Attrition by Sender's Ranking of Recipient Need

	(1) Recipient Surveyed Baseline	(2) Recipient Surveyed Baseline
Ranked Recipient Need	-0.003	
	(0.007)	
Highest need recipient in friend group		0.011
		(0.021)
Control Mean	1	1
Covariates	Network Size	Network Size
N	2551	2551

This table shows that recipients who were reached for baseline were not ranked by senders and more or less in need than those recipients who were not reached at baseline. The regression restricts to friend groups of more than 1 person, and controls for the friend group size. Standard errors are clustered at the sender level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.15: Recipient Attrition for Mental Health Take-Up Outcomes

	(1)	(2)	(3)	(4)
	Recipient Surveyed Endline	Reached for survey helpline consent	Reached for survey used mental health services after 6 months	Reached for survey used mental health services after 6 months
Panel A: Pooled IV Estimates				
Sender shared (to anyone)	-0.030	0.041	0.006	0.012
	(0.075)	(0.094)	(0.066)	(0.074)
Control Mean	0.41	0.68	0.34	0.55
N	2668	1160	2668	1487
Sample Defined By:	All recipienthouseholds	Female recipient households	All recipienthouseholds	Male recipient households
Survey Respondents	Original recipient (male and female)	Only original female recipients	Females only: Original female recipient or any female in male recipient households	Women in household of male recipient
Panel B: ITT Estimates by Message Framing Arm				
Disclosed Compensation framing, *non-targeted*	-0.021	0.011	0.012	-0.008
	(0.028)	(0.037)	(0.027)	(0.032)
Disclosed Compensation framing, *targeted*	-0.043	0.030	-0.051**	-0.034
	(0.030)	(0.037)	(0.025)	(0.032)
Non-Disclosed Compensation framing, *non-targeted*	-0.029	-0.023	-0.042	-0.042
	(0.029)	(0.041)	(0.027)	(0.030)
Control Mean	0.41	0.68	0.34	0.55
N	2668	1160	2668	1487
Sample Defined By:	All recipient households	Female recipient households	All recipient households	Male recipient households
Survey Respondents	Original recipient (male and female)	Only original female recipients	Females only: Original female recipient or any female in male recipient households	Female in household of male recipient

This table tests for differential attrition across the sets of outcomes collected on recipients. Column 3 shows that there was differential attrition in the "disclosed compensation, targeted" message framing arm, relative to control, on the outcome of whether the respondent or anyone in their household had ever used mental health services, measured roughly 6 months after the main experiment. Note that only original female recipients (from the main experiment) were asked the helpline consent question used in the analysis (column 2) due to helpline programmatic priorities. For the outcome in column 3 *all* households of original recipients (male and female) were contacted for the 6 month follow-up questions, but the study interviewed only females in those households, again due to helpline programmatic priorities. Either the original female recipient or a (new) female respondent in the households of original male recipients was surveyed. Column 4 restricts the analysis to the sample of original male recipients and shows that there is no differential attrition for the outcome of using mental health services when restricting to female respondents in the households of original male recipients. Specifications cluster standard errors at the sender level and include no control covariates. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.16: Recipient Mental Health Take-Up by Restricted and Un-Restricted Samples

	(1) Ever used mental health services (after 6 months)	(2) Ever used mental health services (after 6 months)
Panel A: Pooled IV Estimates		
Sender shared (to anyone)	0.118	0.178
	(0.096)	(0.135)
Control Mean	0.230	0.190
Covariates	Lasso Double Selection	Lasso Double Selection
Sample	All respondents	Restricted, see notes
N	905	417
Panel B: ITT Estimates by Message Framing Arm		
Disclosed Compensation framing, *non-targeted*	0.072*	0.086
	(0.041)	(0.058)
Disclosed Compensation framing, *targeted*	0.062	0.100*
	(0.044)	(0.059)
Non-Disclosed Compensation framing, *non-targeted*	0.042	0.040
	(0.044)	(0.063)
Control Mean	0.23	0.19
Covariates	Lasso Double Selection	Lasso Double Selection
Sample	All Respondents	Restricted, see notes
N	905	417

This table shows that the marginally significant effect on mental health care take-up for recipients in the disclosed compensation arm remains qualitatively the same when considering the sample with or without attrition (noting that only the "disclosed compensation, targeted" arm exhibited differential attrition). Ever used mental health services is an indicator taking 1 if the respondent reports in the endline or 6 month follow-up that anyone in the household has ever used mental health services. Specification in column 1 includes all surveyed participants and column 2 is restricted to only female respondents in the household of an original male recipient – a subsample which did not display differential attrition (see Table D.15). Due to helpline programmatic priorities the survey measure was only collected with female respondents, who comprised both original female recipients, and, female respondents in the households of original male recipients. Covariates selected by lasso following Belloni et al. 2014. Standard errors are clustered at the sender level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.17: Types of Mental Health Care Taken-Up by Recipient

	(1) Called helpline (after 6 months)	(2) Used therapy in-person or remote (after 6 months)	(3) Used medication (after 6 months)	(4) Used other mental health care (after 6 months)
Panel A: Pooled IV Estimates				
Sender shared (to anyone)	-0.018	0.030	0.014	-0.012
	(0.044)	(0.075)	(0.069)	(0.070)
Control Mean	0.050	0.130	0.110	0.110
Covariates	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection
N	904	903	905	903
Panel B: ITT Estimates by Message Framing Arm				
Disclosed Compensation framing, *non-targeted*	-0.006	0.045	-0.010	0.026
	(0.018)	(0.034)	(0.027)	(0.030)
Disclosed Compensation framing, *targeted*	-0.006	0.013	0.006	-0.016
	(0.018)	(0.035)	(0.031)	(0.030)
Non-Disclosed Compensation framing, *non-targeted*	-0.008	0.028	0.029	-0.029
	(0.019)	(0.036)	(0.033)	(0.030)
Control Mean	0.05	0.13	0.11	0.11
Covariates	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection	Lasso Double Selection
N	904	903	905	903

This table show the types of mental health that the marginally significant effect on mental health care take-up for recipients in the disclosed compensation arm remains qualitatively the same when considering the sample with or without attrition (noting that only the "disclosed compensation, targeted" arm exhibited differential attrition). Ever used mental health services is an indicator taking 1 if the respondent reports in the endline or 6 month follow-up that anyone in the household has ever used mental health services. Specification in column 1 includes all surveyed participants and column 2 is restricted to only female respondents in the household of an original male recipient – a subsample which did not display differential attrition (see Table D.15). Due to helpline programmatic priorities the survey measure was only collected with female respondents, who comprised both original female recipients, and, female respondents in the households of original male recipients. Covariates selected by lasso following Belloni et al. 2014. Standard errors are clustered at the sender level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.18: Heterogeneity in Sender ability to target

	(1) Recipient depressed at baseline (0/1)	(2) Recipient depressed at baseline (0/1)	(3) Recipient depressed at baseline (0/1)	(4) Recipient depressed at baseline (0/1)	(5) Recipient depressed at baseline (0/1)	(6) Recipient depressed at baseline (0/1)
Highest need recipient in friend group	0.104***	0.106***	0.111***	0.106***	0.103***	0.106***
	(0.029)	(0.030)	(0.031)	(0.037)	(0.029)	(0.039)
Highest Need X Sender stigma 1st order	0.036					
	(0.029)					
Highest Need X Sender stigma 2nd order		-0.004				
		(0.029)				
Highest Need X Sender altruism			0.007			
			(0.031)			
Highest Need X Sender female				-0.002		
				(0.058)		
Highest Need X Sender social desirability					0.027	
					(0.029)	
Highest Need X Sender depressed						-0.001
						(0.058)
Control Mean	0.436	0.436	0.436	0.436	0.436	0.436
Covariates	No controls	No controls	No controls	No controls	No controls	No controls
N	1325	1308	1162	1330	1330	1329

This table shows that there is no significant heterogeneity by sender characteristics in senders' ability to identify which of their friends is in need. Observations are at the recipient level. The sample is restricted to instances when the sender has more than 1 friend and includes only the recipients that were reached for the baseline survey. The independent variable is a binary variable of the sender having indicated that the recipient would benefit the most from mental health information. The dependent variable is an indicator for whether the recipient's PHQ-9 score at baseline indicates that the recipient likely has moderate to severe depression (10 or higher). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.19: Recipient Impacts: Outcome Family 1

	Called Helpline	Benefit from MH services Relative to Stigma (SD)	Willingness to Share Own Story (SD)	Willing to accept call from helpline (after 6 months)
Panel A: Pooled IV Estimates				
Sender shared (to anyone)	-0.017	-0.001	0.080	-0.045
	(0.021)	(0.188)	(0.217)	(0.126)
FDR-adjusted q-value	1	1	1	1
Control Mean	0.016	0.003	1.012	0.536
Double selection	Yes	Yes	Yes	Yes
N	1041	906	1042	812

	Called Helpline	Benefit from MH services Relative to Stigma (SD)	Willingness to Share Own Story (SD)	Willing to accept call from helpline (after 6 months)
Panel B: ITT Estimates by Message Framing Arm				
Disclosed Compensation framing, *non-targeted*	-0.009	-0.017	0.020	0.036
	(0.009)	(0.091)	(0.104)	(0.050)
Disclosed Compensation framing, *targeted*	-0.004	-0.060	0.150	-0.078
	(0.010)	(0.089)	(0.109)	(0.049)
Non-Disclosed Compensation framing, *non-targeted*	-0.008	-0.095	-0.012	0.002
	(0.009)	(0.098)	(0.108)	(0.051)
FDR-adjusted q-value				
Disclosed_non-targeted	1	1	1	1
Disclosed Compenation_targeted	1	1	1	1
Non-Disclosed_non-targeted	1	1	1	1
Control Mean	0.02	0.00	1.01	0.54
Double selection	Yes	Yes	Yes	Yes
N	1041	906	1042	812

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.20: Recipient Impacts: Outcome Family 2

	PHQ-9 Score	GAD-2 Score (SD)	Friend Support Index (SD)	Social Connectedness Index (SD)	Any conversations about mental health	Labor Assistance Freq. (SD)
Panel A: Pooled IV Estimates						
Sender shared (to anyone)	-0.096	0.015	0.135	0.372**	0.159**	0.448**
	(0.160)	(0.155)	(0.176)	(0.186)	(0.063)	(0.202)
FDR-adjusted q-value	.494	.856	.494	.083	.07	.071
Control Mean	-0.001	-0.004	0.000	0.000	0.111	-0.006
Double selection	Yes	Yes	Yes	Yes	Yes	Yes
N	1042	1042	1040	1042	1038	1042

	PHQ-9 Score	GAD-2 Score (SD)	Friend Support Index (SD)	Social Connectedness Index (SD)	Any conversations about mental health	Labor Assistance Freq. (SD)
Panel B: ITT Estimates by Message Framing Arm						
Disclosed Compensation framing, *non-targeted*	-0.019	-0.061	0.040	0.107	0.050*	0.092
	(0.077)	(0.073)	(0.085)	(0.091)	(0.030)	(0.095)
Disclosed Compensation framing, *targeted*	-0.038	0.077	0.110	0.205**	0.070**	0.201**
	(0.077)	(0.075)	(0.080)	(0.092)	(0.034)	(0.102)
Non-Disclosed Compensation framing, *non-targeted*	-0.027	0.094	0.030	0.075	0.041	0.085
	(0.081)	(0.085)	(0.093)	(0.092)	(0.031)	(0.098)
FDR-adjusted q-value						
Disclosed_non-targeted	1	.798	1	.798	.54	.798
Disclosed Compenation_targeted	1	.798	.793	.423	.423	.423
Non-Disclosed_non-targeted	1	.798	1	.798	.793	.798
Control Mean	-0.00	-0.00	0.00	0.00	0.11	-0.01
Double selection	Yes	Yes	Yes	Yes	Yes	Yes
N	1042	1042	1040	1042	1038	1042

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.21: Recipient Impacts: Outcome Family 3

	Expected Benefit from MH Care (SD)	Concern: Not considered reliable by friends (SD)	Concern: Confidentiality Breached (SD)	Stigma Beliefs 2nd Order (SD)	Own Stigma Index (SD)
Panel A: Pooled IV Estimates					
Sender shared (to anyone)	0.066	0.208	-0.051	0.052	0.066
	(0.168)	(0.176)	(0.176)	(0.198)	(0.176)
FDR-adjusted q-value	1	1	1	1	1
Control Mean	0.001	0.006	0.008	-0.005	0.005
Double selection	Yes	Yes	Yes	Yes	Yes
N	1035	1042	1041	909	1027

	Expected Benefit from MH Care (SD)	Concern: Not considered reliable by friends (SD)	Concern: Confidentiality Breached (SD)	Stigma Beliefs 2nd Order (SD)	Own Stigma Index (SD)
Panel B: ITT Estimates by Message Framing Arm					
Disclosed Compensation framing, *non-targeted*	0.015	0.062	-0.074	0.020	0.057
	(0.082)	(0.083)	(0.081)	(0.090)	(0.088)
Disclosed Compensation framing, *targeted*	-0.097	0.058	-0.013	-0.022	-0.005
	(0.083)	(0.086)	(0.087)	(0.092)	(0.090)
Non-Disclosed Compensation framing, *non-targeted*	-0.024	0.144*	0.082	0.112	0.149
	(0.091)	(0.085)	(0.091)	(0.093)	(0.102)
FDR-adjusted q-value					
Disclosed_non-targeted	1	1	1	1	1
Disclosed Compenation_targeted	1	1	1	1	1
Non-Disclosed_non-targeted	1	1	1	1	1
Control Mean	0.00	0.01	0.01	-0.01	0.01
Double selection	Yes	Yes	Yes	Yes	Yes
N	1035	1042	1041	909	1027

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.22: Recipient Impacts: Outcome Family 4

	Ever used mental health services (after 6 months)	Used any MH care 30 days	Knowledge Index (SD)	Knows of the Helpline	Exercise Past 7 Days	Shared MH Information With others (0/1)
Panel A: Pooled IV Estimates						
Sender shared (to anyone)	0.118	-0.008	0.143	0.031	-0.051	0.070
	(0.096)	(0.038)	(0.179)	(0.019)	(0.177)	(0.055)
FDR-adjusted q-value	.799	.799	.799	.799	.799	.799
Control Mean	0.234	0.073	0.000	0.012	-0.002	0.098
Double selection	Yes	Yes	Yes	Yes	Yes	Yes
N	905	1037	1042	1034	1042	1042

	Ever used mental health services (after 6 months)	Used any MH care 30 days	Knowledge Index (SD)	Knows of the Helpline	Exercise Past 7 Days	Shared MH Information With others (0/1)
Panel B: ITT Estimates by Message Framing Arm						
Disclosed Compensation framing, *non-targeted*	0.072*	-0.011	0.053	0.006	0.038	0.028
	(0.041)	(0.018)	(0.084)	(0.010)	(0.085)	(0.027)
Disclosed Compensation framing, *targeted*	0.062	0.018	0.110	0.006	-0.056	0.023
	(0.044)	(0.022)	(0.083)	(0.012)	(0.087)	(0.029)
Non-Disclosed Compensation framing, *non-targeted*	0.042	-0.000	0.073	0.001	-0.028	0.034
	(0.044)	(0.020)	(0.089)	(0.010)	(0.086)	(0.028)
FDR-adjusted q-value						
Disclosed_non-targeted	1	1	1	1	1	1
Disclosed Compenation_targeted	1	1	1	1	1	1
Non-Disclosed_non-targeted	1	1	1	1	1	1
Control Mean	0.23	0.07	0.00	0.01	-0.00	0.10
Double selection	Yes	Yes	Yes	Yes	Yes	Yes
N	905	1037	1042	1034	1042	1042

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.23: Recipient Impacts: Outcome Family 5

	Employed (0/1)	Earnings Monthly (JD)	Borrowed (0/1)	Lent (0/1)
Panel A: Pooled IV Estimates				
Sender shared (to anyone)	-0.060	-2.728	0.183*	-0.025
	(0.072)	(15.421)	(0.101)	(0.048)
FDR-adjusted q-value	1	1	.392	1
Control Mean	0.363	57.649	0.594	0.086
Double selection	Yes	Yes	Yes	Yes
N	1040	1040	1041	1041
	Employed (0/1)	Earnings Monthly (JD)	Borrowed (0/1)	Lent (0/1)
Panel B: ITT Estimates by Message Framing Arm				
Disclosed Compensation framing, *non-targeted*	-0.027	-1.486	0.033	-0.005
	(0.034)	(7.234)	(0.044)	(0.022)
Disclosed Compensation framing, *targeted*	-0.025	-4.639	0.106**	-0.014
	(0.033)	(6.934)	(0.045)	(0.025)
Non-Disclosed Compensation framing, *non-targeted*	-0.001	7.436	0.114**	-0.009
	(0.034)	(8.603)	(0.044)	(0.023)
FDR-adjusted q-value				
Disclosed_non-targeted	1	1	1	1
Disclosed Compenation_targeted	1	1	.138	1
Non-Disclosed_non-targeted	1	1	.138	1
Control Mean	0.36	57.65	0.59	0.09
Double selection	Yes	Yes	Yes	Yes
N	1040	1040	1041	1041

* p < 0.10, ** p < 0.05, *** p < 0.01

Table D.24: Experiment 2: Robustness of Impact of Exogenous Messaging on Demand for Phone Counseling

	(1) Willing to accept call from helpline	(2) Willing to accept call from helpline	(3) Willing to accept call from helpline
Disclosed Compensation framing, *pooled*	0.034	0.038	-0.019
	(0.051)	(0.059)	(0.077)
Targeted framing, *pooled*	0.048	0.087	0.021
	(0.048)	(0.055)	(0.079)
Compensation X Targeted framing	-0.122*	-0.147*	0.068
	(0.070)	(0.080)	(0.105)
Recipient Ranked Most Distressed			
X Disclosed Compensation framing, *pooled*			0.096
			(0.103)
X Targeted framing, *pooled*			0.106
			(0.107)
X Disclosed X Targeted framing			-0.447***
			(0.146)
Recipient Ranked Most Distressed			0.053
			(0.080)
Reference category mean:			
Non-Disclosed compensation, *non-targeted*	0.694	0.694	0.694
Double selection	Yes	Yes	Yes
N	676	539	539

This table shows that restricting the analysis to only senders who ranked their friends' distress does not meaningfully impact the results on willingness to accept a call from the helpline. In column 1, missing values of the highest need indicator, due to some senders not providing this ranking, is imputed and an indicator is included for imputation, while columns 2 and 3 includes only observations from senders who ranked their friends' need. The dependent variable is the willingness of the new recipients to be contacted by the helpline to receive phone counseling. The measure of recipient being the most distressed is based on the sender's ranking, because this experimental design did not allow for a recipient baseline survey. Robust standard errors clustered at the original recipient level. Covariates are selected using lasso double-selection from a list of sender and recipient covariates following Belloni et al. 2014. * p < 0.10, ** p < 0.05, *** p < 0.01

Figure D.1: Recipients' Baseline Self-Assessed Distress Levels: Current and If Hypothetically Started Using Care

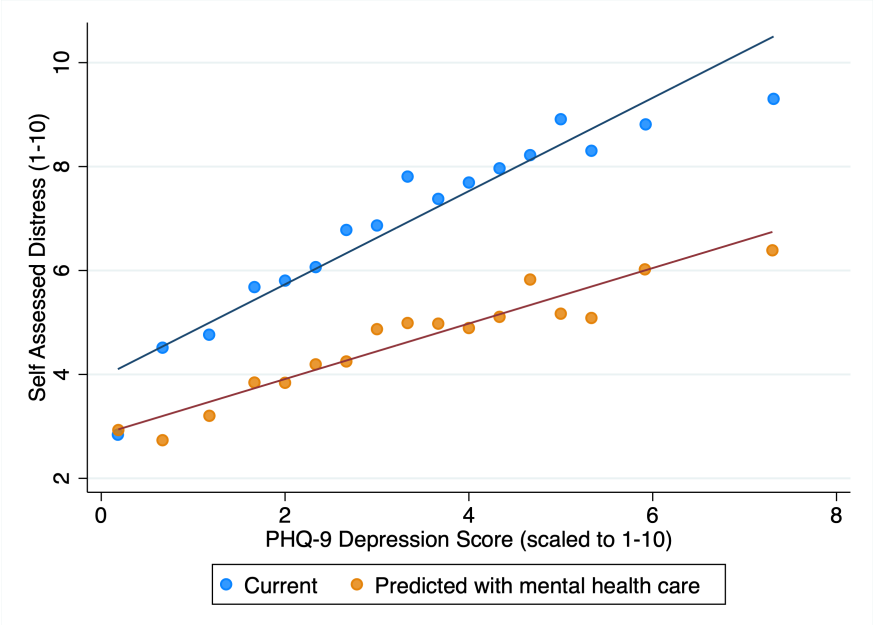


Figure D.2: Proportion of Representative Sample Agreeing: "If I were young and unmarried I would not marry someone who ever used mental health services."