Guy Talk: Catalyzing Peer Effects on IPV through Virtual Support Groups for Men *

Christopher Boyer, † Erica Field, ‡ Rachel Lehrer, § Andrew Morrison, ¶
Claudia Piras ‖

March 12, 2025

Abstract

We experimentally evaluate a novel approach to IPV prevention that harnesses social media to recruit and engage men in a virtual support group delivered by trained male facilitators via WhatsApp. The program succeeded in recruiting men at high risk of committing IPV through self-targeting alone: 52% of partners of men who enroll in the program in response to social media ads report experiencing IPV at baseline, more than four times the national average and nearly twice the rates observed in men recruited through targeted and untargeted invitations. Moreover, on average, participation in the program reduced the probability that female partners report sexual violence at endline by 20%. Treatment effects are concentrated among younger men (-36%), men who exhibit violence at baseline (-27%), and among those whose wives report that they do not drink alcohol (-40%). Program effects are also highly sensitive to group composition, which was randomly assigned. Segregating individuals based on baseline risk appears to magnify program impacts on high-risk individuals, and hence the program impact overall.

*We thank Daniel Hurtado, Gshan Irigoin & Juan Pablo Rossi for excellent research assistance. This study is possible thanks to the efforts of the Inter-American Development Bank (IDB) and the collaboration with the Airbel Impact Lab of the International Rescue Committee (IRC) for the adaptation of the program in Peru. Likewise, this initiative is also possible thanks to the support of the American people through the United States Agency for International Development (USAID). The content of this paper is the responsibility of the researchers and Innovations for Poverty Action (IPA). These do not necessarily reflect the views of the IDB, IRC, USAID, or the United States government.

†Case Western University. cbb100@case.edu
‡Duke University and NBER. field.erica@duke.edu
§Puddle Consultancy. rachel@puddledrip.com
¶Instituto de Estudios Peruanos. andrew.morrisong@gmail.com
‖Inter-American Development Bank. claudiapi@iadb.org

1 Introduction

As the primary perpetrators of violence, men have a vital role to play in ending violence against women. Over the last decade, there has been a surge of interest in engaging men in programming to reduce intimate partner violence (IPV).[1] However, evidence about what works to both enroll men in IPV programs and succeed in shifting their behavior is limited. These are two distinct problems, both confounded by the fact that IPV is a crime with potential legal and social consequences for men who reveal a propensity for violence.[2] First, how do we target men most in need of IPV prevention, and second, how do we structure these programs to best encourage behavior change so that they are not only well-targeted but effective? While directing programs to men at high risk for violence is necessary to produce change, doing so is difficult in practice and may discourage participation when there is potential for social backlash. In addition, in group-based interventions, the importance of targeting violent men will depend on whether mixing types compromises or activates the positive peer effects of the support-group model that are believed to be valuable for behavior change (Solomon, 2004; Steinmetz et al., 2016).

In this paper we evaluate a novel approach to IPV prevention that harnesses social media to recruit and engage men in a group-based violence reduction program. "Guy Talk" (in Spanish, "Hablemos entre Patas", henceforth HEP) is delivered as a virtual support group (VSG) by trained male facilitators exclusively via WhatsApp, an instant messaging application. The program operates on the theory that changes in norms and behavior are best achieved in a group setting, but also that men are more likely to engage with content that challenges prevailing norms around masculinity when offered outside the confines of their immediate social circles (Burri, Baujard, and Etter, 2006; Humphreys and Klaw, 2001).[3]

To provide evidence on the efficacy of this approach as well as the ideal way to structure such an intervention, we conducted a randomized-controlled trial of HEP in Peru. In 2022, a national

[1] Flood, 2020; Jewkes, Flood, and Lang, 2015, and Casey et al., 2018 discuss male-centered IPV programming.

[2] As shown in Alhabib, Nur, and Jones, 2010, IPV is severely under-reported by both men and women worldwide, which holds true in Peru (J. Agüero and Frisancho, 2017). Victims are discouraged from reporting due to the tacit acceptability of IPV (Gracia and Herrero, 2006), hope that the relationship will improve (Pugh, Li, and Sun, 2021), and shame (Tonsing and Barn, 2017), while perpetrators fear legal consequences (Williams, 1992) and the possibility of the victim leaving the household (García-Ramos, 2021).

[3] This is because social norms tend to be self-reinforcing among peers, such that anonymity might encourage participants to question prevailing norms or admit a desire for behavior change without fear of recrimination (Burri, Baujard, and Etter, 2006; Humphreys and Klaw, 2001).

sample of 2,710 men was recruited, primarily through social media, and a randomly chosen subset was assigned to participate in the intervention with a group of strangers. We interviewed both men and their female partners prior to the start of the program and again six months after the program concluded in order to assess the program's impact on female-reported IPV as well as male norms around violence.

Our first key finding is that a social media campaign advertising a relationship improvement program succeeded in recruiting men at high risk of committing IPV through self-targeting alone. That is, according to the reports of their partners, men who self-enrolled in response to social media ads had higher than average incidence of committing IPV: 50% of their partners report experiencing physical IPV at baseline and 21% report experiencing sexual IPV at baseline, which is almost six times the national average reported in household survey data and nearly twice the rates observed in men recruited through targeted and untargeted invitations.[4] Moreover, the program attracted a large proportion of men with strong norms of justification of IPV, which we show are powerful predictors of relationship violence.

These results provide new evidence that self-recruiting through social media advertising can be a highly cost-effective means of targeting men most in need of IPV behavior change interventions, at least with the appropriate advertisements. In the case of HEP, the social media ads explicitly avoided mention of domestic violence, appealing instead to men's self-interest in reducing household conflict by seeking a more harmonious and intimate relationship with their partner, thereby circumventing potential social image concerns of expressing interest in the program.

Second, we find that this highly scalable and cost-effective behavior change program reduced IPV: on average, participation in HEP reduced the probability that female partners report sexual violence at endline by 20%. Heterogeneity analysis reveals that the program was significantly more effective in reducing sexual IPV among younger men ages 20-31 (-35%), men whose partners report that they do not drink alcohol (-43%) and men who exhibit violence at baseline (-24%). Physical IPV falls significantly by 29% among men ages 32-38 and by 11% among men who are violent at baseline. The program effect on sexual violence is similar in magnitude among men who think IPV is sometimes justified and those who do not and the program did not change

[4] According to the data from the 2022 ENDES (peruvian Demographic and Health Survey - INEI, 2022), 8.64% of women report experiencing physical IPV in the past 12 months, and 2.16% report experiencing sexual IPV in the past 12 months.

attitudes towards violence, which suggests that the program did not operate primarily through norms change, but rather influenced men's propensity to regulate emotional responses. To date, few IPV programs have proven to be effective, and the vast majority have been targeted to women. While the importance of targeting men is frequently discussed in policy circles, enlisting men to participate in programs that relate to an increasingly stigmatized and criminalized behavior is a non-trivial challenge, and there is to date no evidence that this strategy works at scale.

Third, we find that program effects on violence were highly sensitive to group composition, which was randomly assigned. In particular, men assigned to groups whose members had a relatively high propensity for violence – measured either in terms of incidence of sexual IPV or the index of IPV justification at baseline – experience significantly larger decreases in sexual violence at endline. Moreover, sensitivity to group composition was observed only among men prone to violence. In other words, placing men at high risk of violence among like-minded individuals has a significantly larger impact on their behavior than placing them in a group of more progressively minded men, but not vice versa.

These results provide novel evidence that peer effects play an important role in mediating the delivery of curricular content in behavior change programs. While group delivery is the norm across a host of behavior change interventions based on a presumption of (net) positive spillovers across participants, there is to date little empirical evidence that the group component matters for outcomes.[5] Although our study does not compare group to individual delivery, the observed influence of group composition provides indirect evidence that peer effects play a role, and can either help or hurt. This mirrors the education literature, which emphasizes the potential for both positive and negative spillovers within classrooms. Our results indicate a strong role of peer effects in influencing outcomes not only in schools, but also in behavior change programs (Scott, 2004).

The patterns by group composition also deliver insight into the nature of peer effects on behavior change, with implications for the optimal composition of group interventions. In particular, the existing literature posits competing dynamics in group learning environments that make it unclear how program effects should move with group diversity (E. Watson, 2017, Burlingame, Strauss, and Joyce, 2013). On the one hand, heterogeneous groups might promote behavior change because

[5]Galizzi and Whitmarsh, 2019 and Fang et al., 2023 conclude that the vast amount of complementary factors in behavioral interventions require very extensive data to holistically measure behavioral spillovers for the group components of these interventions.

they expose men with traditional masculinity norms to more progressive-minded men, which could trigger a role model effect that magnifies program impact through positive spillovers (Rao, 2019, Beaman et al., 2009). In a parallel fashion, minimizing exposure to other men prone to violence might minimize negative spillovers from such individuals that would otherwise reinforce masculinity norms (Dinarte, 2024). On the other hand, positive spillovers might be greatest in segregated groups if program content better resonates when men experience it alongside participants with similar behavioral challenges, for instance by increasing exposure to experiential knowledge that encourages behavior change. Likewise, men may feel more comfortable opening up about their own struggles when surrounded by men who face similar challenges, encouraging more tailored feedback (Silvergleid and Mankowski, 2006). Or, they may quickly disengage from the program if they infer from group discussion that the program content is not intended for people like them (Scott, 2004). Consistent with the latter set of mechanisms, our findings indicate that segregating individuals based on baseline popensity magnifies program impacts on high-risk individuals – or those with the most to gain from the intervention -, and hence the program impact overall. This finding is novel in the behavior change literature in large part because support groups almost always target only individuals in need of behavioral modification.

Finally, by analyzing chat transcripts, we establish three specific ways in which group homogeneity activates positive and reduces negative peer effects. First, we record the incidence of positive group interaction, in which members either reveal personal relationship struggles or facilitators or group members offer problem-solving advice or challenge norms in a respectful manner. These dynamics are meant to capture the primary mechanisms through which peer support groups are thought to facilitate behavior change – personal identification and modeling of positive social behaviors, and the exchange of experiential knowledge (Solomon, 2004, Panahi, J. Watson, and Partridge, 2012). To capture two channels of negative peer effects found in other settings – intra-group conflict and dropout – we also code the incidence of participants arguing with or insulting one another, and record cases in which individuals exit the group before the end of the program.

Our analysis reveals that groups with higher concentrations of men who justify violence experience significantly more positive interactions. These results indicate that behavior change groups "self-tailor" curricular content through group discussion, and imply a significant benefit of sorting groups by baseline behavior. Homogeneous groups are also significantly less likely to experience

negative interactions, implying greater group harmony. Both results are in line with several studies from school settings that document benefits of tracking students by skill level so that curricula can be better tailored and teacher effort better directed, particularly for underperforming students (Duflo, Dupas, and Kremer, 2011; Cortes and Goodman, 2014; Girard et al., 2015).

In addition, we find novel evidence that group *diversity* generates negative peer effects on the men most in need. Men prone to violence are more likely to exit midway through the program when they are placed with a majority of non-violent men, a pattern that magnifies the negative impacts of group composition on violent men that remain in the program. These patterns contradict some previous studies that demonstrate positive spillovers from high-performing to low-performing peers (Rao, 2019, Beaman et al., 2009). One potential explanation is that negative peer effects may manifest more readily in a virtual environment (Banbury et al., 2018) such that diversity runs an especially high risk of derailing group cohesion in virtual classrooms or VSGs.[6] It may also be the case that group homogeneity is particularly beneficial for programs that are focused on changing stigmatized behaviors or polarized norms, in which direct confrontation more readily alienates those who do not share majority views.

While it would be tempting to conclude from these findings that male-centric IPV programs should *only* target high-risk individuals, our results underscore an underlying tension in maximizing the impact of programs designed to discourage stigmatized behaviors such as IPV. In such cases, it is arguably crucial to minimize explicit targeting of high-risk individuals at the recruitment stage in order to avoid stigma; however, doing so will generate a mix of participants with different program objectives whose inclusion appears to have a direct effect on the program experience of high-risk individuals. In such a program, it is arguably valuable to segregate participants *post-recruitment* according to baseline risk of violence in order to maximize program impacts.

[6]For instance, disengagement may happen more readily than it does in face-to-face programs. Likewise, subtleties of tone might be lost in chat-based communication such that small differences in viewpoints are easily magnified and the community more easily polarized.[7]

2 Data & Methods

2.1 Setting

Our study takes place in Peru, where gender-based violence remains a significant problem, despite long-term progress. In 2018, 38% of Peruvian women reported physical or sexual violence by an intimate partner during their lifetime, as compared to only 27% of women globally (Sardinha et al., 2022).[8] The problem was exacerbated by the Covid-19 pandemic, during which severe economic dislocations and the shuttering of services led to a surge in call volume to national victims hotlines (J. M. Agüero, 2021), and sizeable and sustained increases in IPV due to economic losses (J. M. Agüero et al., 2022).

The urgency of addressing this "shadow pandemic" prompted a renewed interest in developing male-focused IPV programs that directly engaged prospective perpetrators. Historically, IPV prevention efforts have focused on reforming prospective abusers primarily by lobbying for harsher punitive measures rather than developing programs that assist with behavior change, and controversy around the idea of directing therapeutic resources to abusers rather than victims has stalled progress on how to make these programs effective or implement them at scale (Pappas, 2023). Over the past couple of decades, a movement towards working with boys and men has increased the number of interventions directed at male perpetrators, yet there is still little evidence-based research on what works, and a large number of such programs operate within the criminal justice system. Several meta-analyses of batterer intervention programs show inconclusive evidence on the effectiveness of psychoeducational, motivational, and CBT interventions on recidivism (Feder and Wilson, 2005; Feder, Wilson, and Austin, 2008; Smedslund et al., 2011; Tarzia et al., 2017; Miller and Rollnick, 2002; Stephens-Lewis et al., 2021). However, existing evaluations suffer from inconsistent measurement of outcomes and poor follow-up (Gondolf, 2012).

Meanwhile, the inability to provide in-person services prompted a search for violence prevention interventions that could be conducted remotely and capitalize on the revolution in access to feature-rich mobile devices and the popularity of social media platforms in low and middle-income countries such as Peru. By 2023, there were nearly 25 million Facebook users in Peru comprising 93% of the

[8] At the same time, rates of the most extreme forms of violence against women were increasing: between 2017 and 2019, the rate of femicides in Peru increased by more than 10 percent (Quispe et al., 2023), sparking a wave of protests and renewed government commitments to reduce violence against women.

eligible population (those above the age of 13) (Lokmanoglu et al., 2023). Likewise, WhatsApp, a popular social media and messaging app owned by Facebook's parent company Meta, had more than 20 million users (59% of the population). To date, there is no rigorous evidence that remotely-delivered programs can reduce IPV.

2.2 The Hablemos Entre Patas program

Hablemos entre Patas (HEP) is an IPV prevention program that was designed to meet both of these criteria. The month-long program is delivered by trained male facilitators to groups of 50 men through the WhatsApp social media platform, and functions as a facilitated virtual support group (VSG). Over the course of 30 days, each facilitator shares daily messages including specific behavioral and skill-building "challenges" that men are asked to practice, often with their partner. The challenges are intended to help men identify and regulate emotions and help couples better resolve conflict. The facilitator engages the group in daily post-challenge follow-up discussion in which participants are encouraged to share their experiences with the group.

The program was designed with a number of key features intended to engage men at risk of violence and maximize cost-efficacy and scalability, many of which were learned through an extensive pilot phase that involved multiple rounds of focus group discussions. First, the group-based intervention builds on a wide body of evidence documenting the benefits of promoting behavior change through group interaction (Solomon, 2004; Steinmetz et al., 2016; E. Watson, 2017). In one particularly relevant experimental evaluation of a batterer intervention program delivered individually versus in a group, the group model had a significantly larger impact on recidivism, despite the fact that the individual program could be tailored to specific needs (Murphy et al., 2020). Group interventions not only harness potential positive peer effects on behavior change, but minimize delivery costs, and those delivered remotely are even better able to accommodate large groups than a typical support-group model. During piloting, we found that groups of roughly 50 men struck the right balance between having a sufficient number of participants to ensure a lively conversation while retaining enough intimacy to feel like both the facilitator and other men were accessible and the conversation could be followed without being overwhelming.

Second, the program aims to achieve violence prevention primarily by promoting healthy mas-

culinities and solid relationships, with no explicit discussion of IPV.[9] The intention of this indirect approach to violence prevention was to avoid alienating men prone to violence, and instead appeal to their self-interest in behavior change that improved their relationship quality and own mental health.[10] While there is growing consensus that male participation in IPV programs is key to realizing meaningful change in outcomes, recruitment is a first-order constraint to building successful male-centric programming. High-risk men have little incentive to volunteer for behavior change interventions, and possibly also fear backlash from revealing themselves to be prone to violence. As a result, we recruited men into the program under the guise of "relationship improvement" and targeted those who sought to reduce the degree of conflict with their partners.,

Similarly, self-targeted recruitment via social media advertising also avoided alienating men who might feel singled out by a direct invitation and reduced program costs at the recruitment stage. To recruit participants, we enlisted a social advertising firm to generate a series of advertising "posts" that appealed to the self-interests of men prone to violence without inducing stigma. The three most effective ads in terms of generating clicks are shown in Figure A.6. The main strategy was to capture the attention of men who were unhappy with the degree of conflict in their relationship with ad language describing HEP as offering a "second chance" to improve their relationship. For example, one ad describes the program as "30 days of challenges for a happy relationship", while another shows a remorseful man lamenting "another day of fighting with his girlfriend." A key feature of all ads was not only the absence of any mention of violence, but also the absence of blame assignment with respect to conflict, and the promise of personal reward in the form of conflict-free intimacy. The advertisements also enlisted traditionally masculine themes such as sports, with one popular ad showing a referee "giving a red card" to relationship conflict. Social media recruitment was extremely fast and highly cost-effective, requiring only $1.68 in marketing costs per registered user.

Third, the remote delivery of HEP provides a high degree of anonymity among program participants that could encourage participation by removing fear of social backlash. In particular, extensive piloting revealed that men interested in relationship improvement were drawn to the

[9] While the program contained elements of "gender transformative" programming, much of the content veered towards more "gender neutral" behavioral and therapeutic approaches to violence prevention. See Dworkin, Fleming, and Colvin, 2015 for a description of the different approaches.

[10] Casey et al., 2018 describe this approach as *recruitment through "hooks."*

support group model with the possibility of learning and sharing experiences with other men, but also preferred the confidentiality provided by participating with men outside of their direct social circles. Building groups among strangers also facilitates scale-up by not requiring a critical mass of participants in any one locale.

Fourth, the fact that the program was led by a trained facilitator unknown to the group rather than peer-led encouraged participation and reduced the burden of identifying effective facilitators who could enlist a sufficient number of acquaintances. That is, in contrast to the most common strategies for engaging men and boys in violence prevention programs which largely rely on recruitment through social networks and community role models (Casey et al., 2018), extensive piloting in our setting revealed that men not only did not want to participate with peers, but prospective facilitators were also reluctant to lead their peers in discussions of intimate relationships. While relying on trained facilitators reduces the scope for enlisting volunteer moderators, the fact that facilitators can reach a large number of men remotely (250 participants per month) increases the scalability of employing trained facilitators and likely improves the consistency of program delivery.

Finally, remote delivery minimizes participation costs by providing more flexibility in timing of engagement with material and no commuting hassle or coordination costs. Qualitative interviews with participants as well as data from program transcripts indicate that flexibility was important for the group-based model to thrive: almost 40% of chat activity occurred during late night hours not typically served by violence programs. This 'always-on' aspect of the VSG meant that chat was available when men were actually encountering issues in their relationship and when their thoughts were fresh.

2.3 Evaluation Design

We evaluated the impact of HEP on IPV through a randomized control trial (Boyer et al., 2022). Between January and April 2022 we recruited men nationally via: (1) an advertising campaign on social media, primarily Facebook (62% of final sample); (2) random digit dialing (RDD) using numbers from a verified, nationally representative, phone database (11% of final sample); and (3) nationally-dispersed promoters employed by the Ministry of Women (27% of final sample). During registration men were directed to an online web form where they could record their interest in

the program and provide their name and number.[11] Trained interviewers then called interested men and invited them to complete a short survey (hereafter, *men's baseline survey*) that assessed eligibility and collected basic socio-demographic data and information about their relationship, conflict resolution strategies, and views on gender and violence.[12]

Although our program targeted men, it was crucial that female partners were included in the study primarily so that we could assess relationship outcomes in an unbiased fashion and accurately measure IPV at endline as a primary outcome.[13] Hence, to be eligible to participate, men were required to provide a contact number for their (female) partner. We then contacted these women to verify that they were in a relationship with the interested man and approved of their partner participating in the program, and to ask them to complete a short survey (hereafter, *women's baseline survey*) that asked similar questions as the men's survey with the addition of questions about pre-existing IPV and their involvement in control and decision-making within the relationship.[14] It is important to note that our study population includes a narrower set of male participants than would be possible at scale, when there is no need to collect data from partners. We anticipate that this inclusion criteria generates a downward bias in our program estimate since violent men and their partners are likely to be the most reluctant to participate.

In total, we obtained contact numbers for 3,866 men who were in eligible relationships, completed the baseline survey, and agreed to have their partner contacted, either via the registration form or from lists obtained from the Ministry or the phone database provider. A total of 2,710 partners were successfully contacted and agreed to their partner participating, who comprise our primary study sample.[15] The final numbers recruited by source are shown in Table A.1. The majority (62%) of participants were recruited from social media advertising (SMA), 27% from the Ministry database (MIMP), and only 11% from random digit dialing (RDD).

Figure A.8 shows the geographic distribution of study participants. Overall, participants were

[11] In the case of Facebook registrants, this form was hosted directly by Facebook's online advertising platform. All others were directed to a website hosted by the study team.

[12] Men were not asked directly about perpetration of violence and throughout the program was branded more broadly as a relationship improvement program rather than a violence reduction program specifically.

[13] A wide body of evidence shows that men dramatically under-report IPV. See, for instance, Szinovacz and Egley, 1995 and Katz, Carino, and Hilton, 2002

[14] Although we could have sought the inclusion of female partners at endline only since only endline data were an absolute requirement, failing to screen on women's willingness to participate at baseline might have generated differential selection into the study if some treatment women became willing to participate as a result of treatment.

[15] In 131 cases the woman agreed to her partner's participation in the program but did not want to be interviewed at baseline.

widely dispersed geographically across 555 of 1,874 districts in Peru, reflecting one of the scale advantages of a remotely-delivered intervention. However, study districts also tended to be more urban (86.5% versus 78.6% nationally) and more populous (79.39 per square km) as compared to the nation as a whole (26.34 per square km), and participants had significantly greater levels of higher education than the national average (70% have some post-secondary education in our sample vs. 34.0% nationally), reflecting the fact that accessibility is not uniformly distributed.[16]

Our primary experimental manipulation was the assignment of recruited men to either receive HEP or to a pure control condition in which they received no programming.[17] To avoid lengthy delays between recruitment and implementation, randomization to the experimental condition occurred in batches of 500 formed as eligible couples completed the baseline surveys. Once a batch was formed, we further stratified participants based on whether their partner reported (at baseline) no IPV, physical IPV, or physical and sexual IPV. Within these strata, we randomized couples (1:1) to either treatment or control.[18]

Among the men in couples randomized to receive HEP, we then performed a second stage randomization to form WhatsApp groups of 50 men. This secondary randomization was done via a single complete random assignment of men within a particular batch and without regard to other baseline characteristics such as violence. This second stage of randomization induces exogenous variation in the composition of men across groups, which we leverage to evaluate how group composition mediates program impact.

Randomization took place in six batches from March 13 to May 1, 2022. In total, 1,355 men were randomized to receive the program across 27 WhatsApp groups and 1,355 men were randomized to the control group. Table 1 shows baseline characteristics across experimental conditions. A joint test confirms that measured characteristics were well balanced across conditions ($p = 0.858$), and none of the 25 baseline variables are individually unbalanced at conventional levels.

[16]Nation-wide and study district population figures were calculated using the 2017 population report by Peru's "Instituto Nacional de Estadística e Informática" (INEI), and the World Bank Open data portal. Schooling data was acquired from Peru's 2017 National Census.

[17]In accordance with standard ethical practice in violence research, we did provide all interviewed women with information about support services that were available to them and we set up referral pathways for women we encountered who were in distress. As other research suggests, this could affect subsequent violence and therefore may constitute a minimal intervention.

[18]Unlike previous research on violence prevention, here we were not concerned about the potential for spillover effects from treatment to control couples as participants were geographically quite dispersed and hence extremely unlikely to know one another at baseline.

Approximately six months after the program ended, we conducted separate follow-up surveys with men and their (baseline) partner via telephone. The follow-up surveys contained detailed questions on recent experiences with IPV, communication, conflict resolution, relationship quality and mental health as well as attitudes towards gender roles and justification of IPV. As in the baseline survey, only women were asked questions about their direct experience of violence. To minimize the potential for men to influence their partner's responses, we attempted to interview women first, which occurred in 90.77% of cases, and all survey participants were offered monetary incentives to complete the survey.

Of the 2,710 couples enrolled, we successfully interviewed 1,997 women (74%) and 1,747 men (65%) at endline, in line with rates of follow up achieved for telephone surveys in other rigorous evaluations (Gibson et al., 2017).[19] In Table A.2, we show that those who responded to the follow-up survey were better educated, younger, better able to self-regulate, more progressive, and less violent than those who attritted, which is likely to further bias downwards our estimates of program impact. As shown in Table A.5, attrition patterns were balanced across experimental conditions both marginally (HEP=73% vs. C=74%) and across observable characteristics (Joint F-test p-value = 0.8).

3 Take-up and engagement with HEP

We first report on two intermediate outcomes that are key areas of program success: recruitment patterns and level of participation of men in our study. If a program fails to recruit men prone to violence, it will have no potential to reduce IPV. Likewise, if the intervention model fails to engage participants in a meaningful way, it cannot generate behavior change. Because the VSG approach is novel in this policy space, and arguably risky in these two dimensions because both recruitment and engagement are approached in a relatively "blind" fashion, it is crucial to assess both intermediate outcomes in order to interpret channels of program success or failure.

[19] Because of the importance of IPV as a primary outcome, we prioritized re-interviewing female over male respondents in tracking efforts, which explains the difference in rates of attrition.

3.1 Social media advertising as a tool for recruiting

Our recruitment phase provided a unique opportunity to learn about the efficacy of using social media advertising campaigns to recruit men into a program designed to reduce IPV, particularly those who are violent or at risk of becoming violent, versus more traditional direct invitations. Mean differences across participants drawn from the three different strategies are shown in Table 2. In comparison with our other recruitment streams – random digit dialling of a national phone database and in-person recruitment via nationally-placed Ministry promoters – the social media strategy yielded both a higher number of participants per dollar of recruiting effort and significantly more men who were violent or prone to violence. Men recruited via social media had about twice the rates of physical violence (50% vs 25%; $p < 0.001$) and sexual violence (21% vs. 10%; $p < 0.001$) compared to those recruited via the other streams. They also have significantly higher levels of self-reported relationship conflict. Even more striking, self-targeted men have five times the rate of physical IPV and ten times the rate of sexual IPV than recent national estimates from a representative subpopulation.

Importantly, rates of violence among those recruited via direct approach methods were slightly below that of men in the general population. Hence, while the VSG model itself was likely important in encouraging participation through features such as convenience and anonymity, those features alone were insufficient to achieve any degree of program targeting towards men prone to violence when they were approached directly by recruiters. In other words, door-to-door recruiting of men into a face-to-face support group - the most common approach to such interventions - would presumably do even worse in terms of enrolling those most in need of help.

3.2 Participant engagement

A unique feature of a digitally-delivered intervention relative to traditional in-person programming was the availability of complete transcripts of the discussion that took place that allowes us to centrally monitor engagement and drop out. In particular, we were able to monitor (1) whether men joined the WhatsApp group, (2) how long men remained in the group, and (3) the number and nature of their responses to the facilitator and to one another as well as the time that interactions tended to occur throughout the day.

Although WhatsApp familiarity was incredibly high, we were concerned initially that some men who were enthusiastic about the program when they saw or heard it advertised might not actually join, particularly given that so many were recruited via social media. However, over 99% of participants successfully joined their assigned group. After joining, men were free to leave the WhatsApp group at any time. Of those who joined, 84% stayed for at least a week, 75% stayed for half the program, and 67% remained in their group for the entire 30 day intervention (full distribution: Figure A.13). In follow up interviews, men's most commonly cited reasons for leaving the group were lack of time, lack of cellular data, and other problems with their phone (including broken, lost, or stolen), comprising collectively about 60% of the reasons cited, although a vocal minority felt that either the content did not meet their expectations (and often mentioned more intensive psychotherapy) or that what was being shared was too personal or did not align with their views (overall about 15% reported the chats made them feel uncomfortable).

Engagement, as measured by the number of messages sent by men to the group, was strong but also highly heterogeneous across individuals and time. While the mean number of engagements throughout the length of the program was about 3, a vocal minority were the source of a majority of messages (coefficient of variation: 9.31). Similarly, messages were not uniformly distributed across topics. The program content was divided into four major thematic areas: "communication and emotional regulation", "health and sexuality", "finances" and "life at home". By far, the most popular topic in terms of male engagement was "communication and emotional regulation" (28.9 average daily interactions, Figure A.9). More specifically, content about relationship struggles and healthy communication strategies prompted the most discussion (32.8 average daily interactions, Figure A.10). Finally, messages also tended to be highly clustered outside of working hours in the morning and at night (particularly between the hours of 8 and 10 pm, Figure A.12). These are hours not typically serviced by violence prevention programming, demonstrating one of the advantages of a remote "always on" intervention.

In qualitative follow ups, we noted that the most active participants tended to be men searching to improve their relationships. Often these men were living through a difficult moment in their relationship, such as a trial separation. They tended to be in longer term relationships and had kids, and tended to view the program as a complement or alternative to couples counseling or therapy.

4 Impact of HEP on Relationship Outcomes

In this section we present the results from our experimental evaluation of HEP on relationship outcomes measured in endline surveys.

4.1 Estimation strategy

Prior to the start of endline we registered a pre-analysis plan (PAP) that specified how our outcomes would be constructed, our subgroups of interest, and our primary estimation and inference strategies (Boyer et al., 2022). Additional analyses or deviations from the PAP made after registration of the initial protocol are identified throughout as "non pre-registered" analyses.

We estimate the impact of the WhatsApp intervention on violence and relationship outcomes via least squares regression using the following specification

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2'(\mathbf{X}_i - \overline{\mathbf{X}}_i) + \beta_3'\{Z_i \times (\mathbf{X}_i - \overline{\mathbf{X}}_i)\} + \varepsilon_i \quad (1)$$

where Y_i is the outcome reported by individual i (either man or woman), Z_i is an indicator of assignment to the program, and $(\mathbf{X}_i - \overline{\mathbf{X}}_i)$ is a vector of mean-centered covariates that includes indicators for the randomization strata and pre-treatment covariates from the baseline survey selected via the double post-selection lasso (Belloni, Chernozhukov, and Hansen, 2014). To increase precision and reduce the possibility for misspecification, we interact the mean-centered covariates and assignment indicator as suggested in Lin, 2013. Under random assignment, the coefficient β_1 estimates the intention to treat effect of HEP assuming no interference between units [cite Rubin 1980]. However, because the intervention is delivered in a group-setting via WhatsApp in which interaction is encouraged, interference may occur within groups due to peer effects. In this case, β_1 estimates the average effect of assignment to HEP marginalized over all possible peer groups as discussed further in Appendix 6.2.

As mentioned previously, the second stage of randomization – assigning men to a particular WhatsApp group – induced exogenous variation in the composition of peer groups. Therefore, we further estimate the moderating effect of group composition on the impact of HEP via another

least squares regression using the following specification:

$$Y_i = \alpha_0 + \alpha_1 Z_i + \alpha_2 (Z_i \times G_i) + \alpha_3' (\mathbf{X}_i - \overline{\mathbf{X}}_i) + u_i \tag{2}$$

where G_i is a measure of an individual's exposure to group composition formed by taking the average characteristic at the group level at baseline minus the individual's response, i.e. $G_i = \frac{1}{n-1} \sum_{j=1}^{n-1} G_{j,-i}$, and Y_i and Z_i are defined as previously. The vector $(\mathbf{X}_i - \overline{\mathbf{X}}_i)$ here includes batch level fixed effects which, given that each batch was synonymous with a particular facilitator, account for any facilitator specific effects. We code G_i such that α_2 is the effect of ranging group composition from the minimum to the maximum observed value. Where possible we use indices of conceptually related items to increase efficiency and to help reduce the number of hypothesis tests and avoid multiplicity of testing. For all primary outcomes that are indices we determine statistical significance based on the index, but include graphical estimates of effects on individual items to help explain movement or lack of movement on the overall measure.

In primary intention to treat analyses, we calculate standard errors using a heteroskedasticity-robust (HC2) estimator that is consistent with random assignment. We also calculate non-parametric randomization-based p-values via 10,000 permuted assignments. These p-values form the basis for our inferences about statistical significance. Following Zhao (2021), we use a studentized test statistic with HC2 standard errors, which is well powered under both the sharp and weak null. Under interference due to peer effects, these p-values remain valid under the joint null of no effect of the intervention or peers. However, HC2 standard errors may no longer be optimal, especially for superpopulation inference. Therefore, in analyses of group composition and in a robustness check for our main result we also calculate cluster-robust standard errors (CR2).

4.2 Effects on IPV

Our primary outcome of interest – as specified in our pre-analysis plan – was intimate partner violence in the 6 months preceding the endline survey as reported by the female partners of the men in our study. We measured violence using a version of the standardized scale published by the World Health Organization and adapted for Peru. The scale asks about the frequency with which women have experienced six acts of violence of which two can be categorized as constituting

sexual violence (henceforth "sexual IPV") and four as other forms of physical violence (henceforth "physical IPV"). Because the determinants of and mechanisms for reducing physical and sexual violence often differ, we report program effects on indices of these items separately in addition to effects on any form of violence (henceforth "any IPV").

In intention to treat (ITT) analyses in Table 3, we find that the probability of any IPV declined by 1.3 percentage points among women whose partners were randomized to HEP ($T = 0.252$ vs. $C = 0.265$), although results were also consistent with no effect of HEP ($p = 0.548$). Similarly, we see little evidence that HEP shifted women's reports of physical acts of violence ($\widehat{\beta}_1 = -0.006$, $p = 0.727$). The estimate is nearly identical with and without the lasso-selected control variables. By contrast, we do find that HEP reduced the probability of sexual violence by 2.3 percentage points ($p = 0.096$) both with and without flexible covariate adjustment. Although imprecise, the estimated effect on sexual violence is large in magnitude, reflecting a relative reduction of nearly 20% compared to the control group ($T = 0.091$ vs. $C = 0.114$).

The relative size of the estimated effect on sexual violence compared to other forms is not wholly unsurprising given the amount of time devoted to discussions of sexual pleasure, consent, and intimacy communication in the HEP program. Comparing effects of treatment on the two sub-items in the sexual violence index, we see that the effect is driven by movement on the proportion of women reporting that their partner forced intercourse, which falls by 27% and is statistically significant at conventional levels when control variables are included ($p < 0.05$; Table 4).

The mean program effect obscures a large amount of variation in estimated program impact across groups, and across facilitators. Figure A.4 shows group-specific estimates of treatment impact on violence, grouped by facilitator. The bold markers at the bottom of each grouping show the facilitator-level estimate of program impact. In general, facilitators 2, 4, and 6 performed notably better than the other three, although no one facilitator achieved uniform success across groups. F2 had one group and F4 had two groups out of five that failed to move at all on violence measures from baseline to endline, despite being led through the program curriculum by a competent mediator. Facilitator 3 was by far the least successful, with no single group showing evidence of reductions in violence. Dropping this facilitator from the estimates gives a sense of how large program effects might be if only the highest skilled mediators are retained. Estimated program impacts among all facilitators but 3 yields an estimate of program impact on sexual violence of -25%.

As specified in our PAP, we examined treatment effects among the following subgroups: men whose wives reported versus did not report any instance of violence at baseline; men with secondary education or less versus those with technical tertiary education versus those with some university education; men in three terciles of age groupings; and men whose wives reported that they did not drink alcohol at baseline versus those reported to be occasional drinkers versus habitual drinkers (Table 2). The subgroup where we see arguably the strongest pattern of heterogeneity is baseline violence (columns 1 and 2): among couples in which the female partner reported either physical or sexual IPV at baseline, rates of physical IPV fall by 11% ($p = 0.058$) and rates of sexual violence fall by 26% ($p = 0.074$). This is reasonable given that rates of ongoing violence are far higher than entry into violence: in our sample only 8% of women in the control group in relationships without violence at baseline reported new violence six months later, while roughly 55% of women in violent relationships at baseline reported continuing violence at endline.

Table 2 also reveals that reductions in sexual violence were heavily concentrated among younger men (column 4), men with technical education (column 6), and especially among men who do not drink alcohol (column 8). Reductions in sexual violence were 35% among younger men ($p = 0.040$) and 43% among men who do not drink ($p = 0.034$). Young men likely respond more because of higher rates of baseline IPV and lower rates of program drop-out, which may reflect their level of comfort engaging with a chat-based messaging app. The fact that men who drink are less responsive to the program suggests that drinking reduces men's ability to take control over emotive violence, which was a topic emphasized by the program. Moreover, physical IPV falls significantly by 29% among men ages 32-38.

In addition to whether physical or sexual IPV was reported by a study participant's female partner, we examined effects on less normatively severe forms of IPV, including controlling behaviors, psychological violence, and cyber violence (Table A.6). Estimated program effects suggest little impact on these margins. Figure 1 shows program effects on individual index components of all IPV domains. Of these, forced intercourse is the only IPV measure for which significant program impacts are detected, although we also observe suggestive declines in the incidence of particpants reviewing their partner's phone messages ($p = 0.11$), and the point estimates on all sexual and controlling behaviors are negative. Overall, 11 out of 15 individual items are negative in value.

As specified in our pre-analysis plan, we also examined effects on outcomes that reflect the

frequency and severity of violence reported by women. First, as reported in Appendix Tables A.12 -A.13, we looked at effects on continuous indices that capture the intensity of violence reported by women. These indices take into account the frequency that each item is reported (rather than just whether it was reported) coded as "once", "a few times", or "many times" in the last six months. To form the index, we take the arithmetic mean across items and scale the result between 0 and 1. In all cases, we find point estimates indicative of small improvements due to HEP on intensive margins of violence, but none are significant at conventional decision thresholds.

4.3 Effects on other outcomes

Our analysis also investigates the program impact on a range of secondary outcomes, including emotional IPV, measures of communication within the couple, emotional regulation, decision-making power, perceived control over sex, attitudes towards IPV, alcohol use, and satisfaction and well-being. As shown in Appendix Tables A.11-A.23, estimated effects of HEP on pre-registered secondary outcomes are generally signed in the hypothesized direction but are not statistically significant at pre-specified or conventional levels.

Most notably, there is no significant effect of the program on men's attitudes towards IPV, measured as the number of circumstances in which they believe that IPV is justified at endline (Table 7). This pattern suggests that the program effect on violence is more likely to operate through providing concrete strategies and advice to men who already wish to reform their behavior, rather than convincing them to adopt a different set of fundamental beliefs about women's right to protection from violence. An important caveat is that rates of attrition in the men's survey are significantly higher than they are in the woman's survey (34%), although the attrition is balanced by treatment arm. However, potential selective attrition by baseline norms confounds our ability to fully evaluate the program's impacts on male norms.

4.4 Role of group composition in mediating program impacts

Our intervention aims to reduce IPV by exposing participants to a curriculum of content designed to persuade them of the benefits of non-violence and give them tools to assist with behavior change. However, we also designed the intervention as a group-based model since it is often argued that behavior change is best achieved in a group setting, for a number of reasons. E. Watson 2017

provides a comprehensive discussion of potential mechanisms, and emphasizes the roles of modeling behavior change and exchange of knowledge. First, groups enhance learning by providing a platform for the exchange of experiential knowledge, which can expand participants' knowledge base and thereby increase the set of strategies available to achieve program goals. Second, groups give participants the opportunity to observe the positive behaviors of others in their reference group, which can trigger norms-related behavior change through personal identification or aspirational social comparisons. A related channel through which groups can magnify program impacts is by leading participants to update their second-order beliefs about what members of their reference group believe about a certain norm. For instance, in the HEP intervention, knowing that 50 other men also wanted to join the program and are learning the same information could change your normative expectations about the beliefs of your reference group.

Both channels (learning and norms) are likely to depend on group composition in comparable manners. In particular, the degree to which group engagement reinforces program content as well as the degree to which beliefs about your reference group are influenced by a group-based intervention will depend on the how much group members have in common with respect to the focal norm or behavior. That is, if an individual is surrounded by individuals who do not share the same behavioral challenges or are outside of their social reference group, the group-level component will do little to enhance the curriculum's effect on behavior change.[20]

To examine this prediction, we leverage variation in group composition induced by the second stage of randomization in which men are assigned to a particular WhatsApp group in order to test whether program effects are sensitive to the composition of the group.This second stage of randomization allows us to test whether putting men with others who share the same propensity for violence influences their response to the program.

We measure an individual's propensity for violence using two baseline variables that were pre-registered as potential sources of heterogeneity in program response: sexual violence reported by female partners and an index of men's views about when violence is justified (hereafter justification index). Our rationale for selecting these variables was their strong relationship with physical

[20]In the medical literature, the two group characteristics most frequently investigated as mediators of group-based therapeutic outcomes are "cohesion" - group members' level of connectedness - and "alliance" - the degree to which members are aligned with the program leader's goals (Alldredge et al., 2021). Presumably, homogeneity would support cohesion among men at risk of violence, while the presence of low-risk men could help high-risk men align with the facilitator's objectives through positive modeling.

and sexual violence. As shown in Table A.10, both correlate strongly with violence at baseline and at endline. In terms of violence norms, the strongest predictors are positive response to survey questions asking whether a man believes that jealousy is love, and whether he believes that sometimes women are to blame for sexual harassment. However, as pre-specified, we combine answers to all five violence norms questions into an index of IPV justification by summing the number of statements where an individual agreed violence was justified. The resulting index is also a strongly significant predictor of IPV at baseline and endline (Columns 2-3, Table A.10; $p < 0.01$).

Overall, the program managed to attract men with a wide range of IPV norms. In our sample, 47% of participants believe that IPV is never justified, while 53% believe there are circumstances in which IPV is justified. To quantify an individual's exposure to group members with different propensities for violence, we calculate the mean justification index of the other men in their group (excluding their own), as well as the fraction of group members that commit sexual violence at baseline, for comparison. Figure 2 shows the distribution of group compositions with respect to these outcomes induced by the second stage of randomization.

The regression estimates in Table 6 reveal that the program impact on sexual violence is strongly sensitive to the IPV norms and behavior of other group members. In particular, the program impact falls strongly with the number of other group members prone to violent behavior at baseline, measured either in terms of IPV norms (column 1) or behavior (column 4). Moreover, the pattern is only observed among men who are themselves prone to violence. Columns 2-3 and 5-6 divide the sample of participants into those who say that IPV is never justified (henceforth non-violent men) versus those that believe there are circumstances in which IPV is justified (henceforth violent men). While the program response of non-violent men is insensitive to group composition (columns 2 and 5), men prone to violence are much more likely to exhibit behavioral change as a results of HEP if they are placed into groups with other violent men (columns 3 and 6), using either measure of group composition.

Figure 3 shows how the estimated effect of HEP on violent men grows with the fraction of other high-risk men in a participant's group. As shown in the figure, and implied by the coefficient estimates in column 3 of Table 6, reductions in sexual violence among violent men are concentrated in groups where more than 50% justified violence at baseline, ranging from about a 2 percentage point reduction in sexual violence when group composition is 50% to a 10 percentage point reduction

when group composition is above 70%. The corresponding estimate for sexual violence among group members implies that positive effects are only observed when at least half of group members have committed violence at baseline.

The sensitivity of program effects to group composition imply that peer effects are present even in a virtual support group, and operate only on those most in need of help. In the following section we unpack the specific group dynamics at play in order to better understand the nature of peer effects on violent men.

4.5 Effect of composition on group chat dynamics

Because the program was conducted remotely via WhatsApp, we have access to digital transcripts of all messages sent between participants in the group chats. Hence, to shed light on the channels through which individuals with relatively conservative IPV norms benefit from experiencing the intervention alongside others with similar norms, we scrutinize chat transcripts and record the incidence of three distinct types of group interaction that are potentially influenced by group composition: (1) the degree to which participants share specific behavioral challenges they are facing with the group ; (2) the extent of helpful feedback they receive from other group members or the facilitator; and (3) the degree of group conflict.

The full program corpus contained 19,557 messages across 27 groups. To categorize messages, we engaged a team of five "coders" - all native-speaker research assistants employed by our local partner organization - to systematically review all messages and categorize them using a pre-specified rubric (Appendix X). The exercise was intended to minimize oversight and flag instances where our categorization was unclear.[21] Coders were instructed to identify and flag three types of messages: incidents in which a group member opened up to the group about a problem they were having with their partner ("share challenges with group"); incidents in which another group member or the facilitator offered concrete problem-solving advice or challenged the belief of a participant who had shared a problem or frustration, in a friendly manner ("helpful feedback"); and incidents in which a group member insulted or argued with another group member in a non-friendly or

[21] Since the transcripts contained 20K messages, we assume that the primary source of measurement error is randomly overlooking an instance of key communication type. Each coder received identical written and verbal instructions, and the order of group-day transcripts was randomized to be different for all coders to avoid common order effects.

accusatory manner ("group conflict").[22]

Several concrete examples of chat activity in the transcripts provide qualitative evidence of these dynamics. For instance, there are many instances in which it is evident that participants in homogeneous group feel a greater sense of trust that enables them to open up to the group with both problems and well-meaning advice. In one example, a participant shared: *"I was very jealous and did not trust my wife ... we have 1 son, and for them I want to change ... I want advice that you could give me - thank you so much".* Another group member replied *"Look, I was separated due to mistrust, but now I accept that it was my fault because I was unfaithful to her."* A different group member chimed in: *"I think there is no other way than separating if your doubts are harming the other person, and if you can't trust, you should fix yourself."* Added a fourth, *"I had the same experience as you, ... , I had to ask my family for forgiveness and blame myself for everything that was going on in the house so that they could trust me, and change the way I expressed myself to talk in a more calm manner."*

We also see clear cases in which the shared experiences of men who have also had conflict in their relationships lead to genuinely helpful feedback. In one group that was discussing the topic of expenses, a group member expressed frustration over fighting with his partner about money, and another member shared the following anecdote in response: *"When we first got married I didn't accept that I had to consult my wife before buying things. But with the passage of time I started to understand that what I earned was not just mine, but shared, and that I had to propose to my partner what I wanted to spend money on. From then on we'd make a list (from 1 to 5) of .. priorities and made a budget for expenses."* Another group member chimed in, *"When it comes to [savings goals] we both talk about it, and then later consult my son so that we can all get behind a single positive idea."*

The facilitator's interaction with group members also appears to benefit from the more open exchange. For instance, in one relatively homogeneous group, a member came forward asking, *"What would you say about a situation in which all investing in the home, health and savings comes only from the man and everything earned by the women goes for personal spending?"* To which the facilitator responded, *"Umm, friend, this is a complicated subject. In my opinion I think*

[22]In constructing the variables, messages were assigned values consistent with the majority of coders, and messages in which only 1 or 2 coders identified an instance of positive or negative communication were scrutinized for interpretation and reconciled by the research team.

everyone who lives in the household should contribute to household expenses in a coordinated and proportional manner considering the priorities of everyone who lives in it. That is what I do with my partner and from what you seem to tell me, your household division doesn't seem to be equal."

Finally, we also see extensive evidence of chat interactions in heterogeneous groups that leave group members with long-established beliefs feeling alienated from the program curriculum. For example, in one relatively non-violent group, one frustrated participant with more traditional norms stated to the group, *"It seems to me that most of the group members are lying about their [relationships] so that they don't appear too macho, which is not helping anyone!"* In a different group, after members began discussing the equitable division of housework, a more traditional member remarked: *"What??? Now most men are like this? Everyone in the group is a good husband and helps their partners?? Is there anyone like me that doesn't help at home?"*. In contrast, in more homogeneous groups, it is common to observe instances in which group members debate gender norms in a more respectful and productive manner. For instance, in one such group, when one member remarked that *"I don't do woman stuff,"* another group member replied, *"I think that doing [housework] makes me more independent... for instance if I travel or she gets sick."*

We also record the incidence of individuals exiting the chat group before the end of the program ("remain in chat"), along with the number of messages posted across the four broad categories of program topics outlined in the curriculum. For the latter, we are particularly interested in the amount of chat activity occurring on the topics of healthy communication and emotional regulation ("Messages: healthy communication"), as opposed to the program topics that are less central to reducing violent conflict ("Messages: other"). See Figure A.10 for an overview of program content by domain of topics. Not only does program content in the domain of *Healthy Communication and Emotional Regulation* touch on the key triggers of arguments between couples (e.g. "jealousy", "avoid impulsive reactions", and the fifth most discussed topic, "how your partner feels"), but program content in this category receives the most attention from group members, with the top 7 daily messages in terms of chat activity falling into this category (see Figure A.10). This is in part because it was the first topic covered by the program; hence, it is also the case that the atmosphere of the group chat is most likely to get established during this round of chat activity.

The regression analysis of group composition on chat activity is presented in Tables 6- 10. Table 8 shows that men prone to violence are increasingly likely to remain in the program as the fraction

of group members like them rises. In groups with the lowest average group-level justification scores, only 57% of these men stay for the duration of the program, while nearly 80% remain in the groups with the highest group-level scores. The non-parametric relationship between group composition and drop-out of men who justify violence is shown in Figure 4, which reveals a pattern of sharp increases in dropout when the majority of the group does not justify IPV, and increases in retention when at least 2/3 of the group does.

We also observe significant shifts in the nature of conversation happening in the group. While the overall number of group messages is insensitive to group composition, the fraction of messages on topics related to healthy communication and emotional regulation rises significantly as the number of men prone to violence in the group increases, and there is a corresponding decrease with respect to chat activity on topics in other domains. In other words, in groups that contain more men prone to violence, the program results in more group conversation around topics pertinent to emotional regulation and conflict management.

Likewise, in Table 9 we see evidence from an analysis of individual message content that the *nature* of group interaction shifts with group composition. In particular, as the fraction of men prone to violence increases, so does the fraction of chat interactions that involve a group member sharing a behavioral challenge, as well as the fraction of messages involving a different member or the facilitator offering concrete advice for overcoming relationship challenges.

In contrast, the overall rate of group conflict does not rise significantly with the fraction of violence-prone types. On the other hand, as shown in Figure 4, the rate of group conflict *is* lower for relatively homogeneous groups at both ends of the spectrum. In particular, the degree of conflict rises significantly as groups approach a balanced composition of types, at just around the 50-50 mark. As a result of this pattern, we create a measure of group homogeneity that is independent of type by calculating the distance between the group-level proportion who justify any form of violence at baseline and 50%, and regress group outcomes on homogeneity. These estimates are presented in Table 10. Here, just as in Figure 4, we observe that the rate of group arguments is significantly higher for less homogeneous groups, although the point estimate is only significant in estimates with controls.

5 Conclusion

Our study provides novel evidence that a relationship improvement program directed to men and delivered through virtual support groups can be an effective tool in the fight against intimate partner violence. Moreover, support group composition has a large impact on the efficacy of the program for high-risk individuals. Placing men prone to violence in groups with a higher fraction of like-minded individuals greatly enhanced the program effect on sexual violence. In data coded from chat transcripts, we observe that homogeneity among violence-prone men leads to greater willingness to share relationship challenges, which in turn leads to more helpful feedback being provided by group members, which in turn leads to lower group conflict and lower drop-out of violence-prone individuals.

While the result might extrapolate to an array of settings, we attribute some of the benefits of homogeneity in this setting to the fact that our recruitment strategy went out of the way to pitch the intervention as a relationship improvement program so as not to alienate men prone to violent behavior. As such, the program attracted a significant number of men at low risk of committing violence, unlike most batterer intervention programs worldwide that target only convicted perpetrators. Indeed, baseline survey data and also qualitative interviews with participants reveal that two different types of men were drawn to the program – those with less acceptance of IPV norms seeking greater intimacy with their spouse, and those with long-established masculinity norms seeking to minimize domestic conflict. While IPV programs are likely to have a bigger impact if they target the latter, recruiting men at high risk of violence is likely to require some level of discretion so as to not alienate high-risk individuals, which will inevitably draw in a nontrivial number of low-risk men.

Taken together, our results suggest that the ideal program structure is to recruit men via a discrete approach, and then segregate them according to baseline IPV norms in order to maximize program impact. Although some men who do not exhibit IPV at baseline nor strong norms of IPV justification are still susceptible to entering into violence, and a program such as HEP can reduce this incidence, those most at risk are best served by being placed alongside individuals who are struggling with the same behavior challenges. An important caveat is that we are unable to draw strong conclusions on the impact of a very high degree of segregation since our program included

no cases of purely homogeneous groups.

More generally, our analysis of group interaction also sheds light on the channels through which virtual support groups promote behavior change around IPV. Notably, groups in our study did not change behavior by persuading members to adopt different ethical norms (Table 7). Instead, they appear to reduce violence by helping group members formulate concrete strategies to manage behavior change that is in their self-interested objective of reducing relationship conflict.

From a policy perspective, the program was a highly cost-effective means of promoting behavior change, aided by both recruitment through social media and delivery through mobile messaging. Our estimates of cost per participant encompass both recruitment and implementation costs and factor in yields from recruitment efforts. Our estimate of the cost of recruiting via social media is $1.68 per registered participant, which encompasses the cost of posting ads on Facebook and Instagram, as well as the cost of designing these ads with a firm, including the cost of ad design and social media advertising strategy. Our measure of implementation costs includes the human resources needed to implement the program, which is primarily the cost of training and paying for facilitators. We also included other HR costs needed to design and run the program safely, including a psychologist and IPV expert who were on call throughout the program to address any risky or difficult situations that emerged from the chats. We estimate these costs to be $5.46 per beneficiary. It is worth noting that both costs would presumably fall with scale, making this a particularly cost-effective means of reducing incidence of sexual IPV.

References

Agüero, Jorge and Verónica Frisancho (2017). "Misreporting in sensitive health behaviors and its impact on treatment effects: an application to intimate partner violence". In: *Inter-American Development Bank*.

Agüero, Jorge M (2021). "COVID-19 and the rise of intimate partner violence". In: *World Development* 137, p. 105217.

Agüero, Jorge M et al. (2022). "Is Remote Sensing Data Useful for Studying the Association between Pandemic-Related Changes in Economic Activity and Intimate Partner Violence?" In: *AEA Papers and Proceedings*. Vol. 112. American Economic Association, pp. 277–281.

Alhabib, Samia, Ula Nur, and Roger Jones (2010). "Domestic violence against women: Systematic review of prevalence studies". In: *Journal of Family Violence* 25, pp. 369–382.

Alldredge, Cameron T et al. (2021). "Alliance in group therapy: A meta-analysis." In: *Group Dynamics: Theory, Research, and Practice* 25.1, p. 13.

Babić Rosario, Ana, Cristel Antonia Russell, and Doreen Ellen Shanahan (2022). "Paradoxes of social support in virtual support communities: A mixed-method inquiry of the social dynamics in health and wellness Facebook groups". In: *Journal of Interactive Marketing* 57.1, pp. 54–89.

Banbury, Annie et al. (2018). "Telehealth interventions delivering home-based support group videoconferencing: systematic review". In: *Journal of Medical Internet Research* 20.2, e25.

Beaman, Lori et al. (2009). "Powerful Women: Does Exposure Reduce Bias?*". In: *The Quarterly Journal of Economics* 124.4, pp. 1497–1540. DOI: 10.1162/qjec.2009.124.4.1497.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2014). "Inference on treatment effects after selection among high-dimensional controls". In: *Review of Economic Studies* 81.2, pp. 608–650.

Boyer, Christopher et al. (2022). *'Hablemos Entre Patas': A Randomized-Controlled Trial of a WhatsApp Intervention to Reduce Intimate Partner Violence*. DOI: 10.1257/rct.10043-1.1. URL: https://www.socialscienceregistry.org/trials/10043. preprint.

Burlingame, Gary M, Bernhard Strauss, and A Joyce (2013). "Change mechanisms and effectiveness of small group treatments". In: *Bergin and Garfield's handbook of psychotherapy and behavior change* 6, pp. 640–689.

Burri, Mafalda, Vincent Baujard, and Jean-François Etter (2006). "A qualitative analysis of an internet discussion forum for recent ex-smokers". In: *Nicotine & Tobacco Research* 8.Suppl_1, S13–S19.

Casey, Erin et al. (2018). "Gender transformative approaches to engaging men in gender-based violence prevention: A review and conceptual model". In: *Trauma, Violence, & Abuse* 19.2, pp. 231–246.

Cortes, Kalena E and Joshua S Goodman (2014). "Ability-tracking, instructional time, and better pedagogy: The effect of double-dose algebra on student achievement". In: *American Economic Review* 104.5, pp. 400–405.

Dinarte, Lelys (2024). "Peer Effects on Violence: Experimental Evidence from El Salvador". In: *Social Science Research Network (SSRN)*.

Duflo, Esther, Pascaline Dupas, and Michael Kremer (2011). "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya". In: *American Economic Review* 101.5, pp. 1739–1774.

Dworkin, Shari L, Paul J Fleming, and Christopher J Colvin (2015). "The promises and limitations of gender-transformative health programming with men: critical reflections from the field". In: *Culture, Health & Sexuality* 17.sup2, pp. 128–143.

Fang, Ximeng et al. (2023). "Complementarities in behavioral interventions: Evidence from a field experiment on resource conservation". In: *Journal of Public Economics* 228, p. 105028.

Feder, Lynette and David B Wilson (2005). "A meta-analytic review of court-mandated batterer intervention programs: Can courts affect abusers' behavior?" In: *Journal of Experimental Criminology* 1, pp. 239–262.

Feder, Lynette, David B Wilson, and Sabrina Austin (2008). "Court-mandated interventions for individuals convicted of domestic violence". In: *Campbell Systematic Reviews* 4.1, pp. 1–46.

Flood, Michael (2020). "Engaging Men and Boys in Violence Prevention". In: *Men, Masculinities and Intimate Partner Violence*.

Galizzi, Matteo M and Lorraine Whitmarsh (2019). "How to measure behavioral spillovers: a methodological review and checklist". In: *Frontiers in Psychology* 10, p. 342.

García-Ramos, Aixa (2021). "Divorce laws and intimate partner violence: Evidence from Mexico". In: *Journal of Development Economics* 150, p. 102623.

Gibson, Dustin G et al. (2017). "Mobile phone surveys for collecting population-level estimates in low-and middle-income countries: a literature review". In: *Journal of Medical Internet Research* 19.5, e139.

Gondolf, Edward W (2012). *The future of batterer programs: Reassessing evidence-based practice*.

Gracia, Enrique and Juan Herrero (2006). "Acceptability of domestic violence against women in the European Union: A multilevel analysis". In: *Journal of Epidemiology & Community Health* 60.2, pp. 123–129.

Humphreys, Keith and Elena Klaw (2001). "Can targeting nondependent problem drinkers and providing internet-based services expand access to assistance for alcohol problems? A study of the moderation management self-help/mutual aid organization." In: *Journal of Studies on Alcohol* 62.4, pp. 528–532.

INEI, Peru (2022). "Perú Encuesta Demográfica y de Salud Familiar - ENDES 2022". In: *Instituto Nacional de Estadística e Informática*.

Jewkes, Rachel, Michael Flood, and James Lang (2015). "From Work with Men and Boys to Changes of Social Norms and Reduction of Inequities in Gender Relations: A Conceptual Shift in Prevention of Violence against Women and Girls". In: *The Lancet* 385.9977, pp. 1580–1589. DOI: 10.1016/S0140-6736(14)61683-4.

Katz, Jennifer, Andrew Carino, and Angela Hilton (2002). "Perceived verbal conflict behaviors associated with physical aggression and sexual coercion in dating relationships: A gender-sensitive analysis". In: *Violence and Victims* 17.1, p. 93.

Lin, Winston (2013). "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique". In: *Annals of Applied Statistics*.

Lokmanoglu, Ayse D et al. (2023). "Social media sentiment about COVID-19 vaccination predicts vaccine acceptance among Peruvian social media users the next day". In: *Vaccines* 11.4, p. 817.

Miller, William R and Stephen Rollnick (2002). *Motivational interviewing: Preparing people for change*.

Murphy, Christopher M et al. (2020). "Individual versus group cognitive-behavioral therapy for partner-violent men: A preliminary randomized trial". In: *Journal of Interpersonal Violence* 35.15-16, pp. 2846–2868.

Niemz, Katie, Mark Griffiths, and Phil Banyard (2005). "Prevalence of pathological Internet use among university students and correlations with self-esteem, the General Health Questionnaire (GHQ), and disinhibition". In: *Cyberpsychology & Behavior* 8.6, pp. 562–570.

Panahi, Sirous, Jason Watson, and Helen Partridge (2012). "Social media and tacit knowledge sharing: Developing a conceptual model". In: *World Academy of Science, Engineering and Technology* 64, pp. 1095–1102.

Pappas, S (2023). "Preventing intimate partner violence by focusing on abusers". In: *Monitor on Psychology* 54.3, p. 62.

Pugh, Brandie, Luye Li, and Ivan Y Sun (2021). "Perceptions of why women stay in physically abusive relationships: A comparative study of Chinese and US college students". In: *Journal of Interpersonal Violence* 36.7-8, pp. 3778–3813.

Quispe, Antonio M et al. (2023). "Femicides and victim's age-associated factors in Peru". In: *Hispanic Health Care International* 21.3, pp. 166–173.

Rao, Gautam (2019). "Familiarity does not breed contempt: Generosity, discrimination, and diversity in Delhi schools". In: *American Economic Review* 109.3, pp. 774–809.

Sardinha, Lynnmarie et al. (2022). "Global, regional, and national prevalence estimates of physical or sexual, or both, intimate partner violence against women in 2018". In: *The Lancet* 399.10327, pp. 803–813.

Scott, Katreena L (2004). "Stage of change as a predictor of attrition among men in a batterer treatment program". In: *Journal of Family Violence* 19, pp. 37–47.

Silvergleid, Courtenay S and Eric S Mankowski (2006). "How batterer intervention programs work: Participant and facilitator accounts of processes of change". In: *Journal of Interpersonal Violence* 21.1, pp. 139–159.

Smedslund, Geir et al. (2011). "Cognitive behavioural therapy for men who physically abuse their female partner". In: *Campbell Systematic Reviews* 7.1, pp. 1–25.

Solomon, Phyllis (2004). "Peer support/peer provided services underlying processes, benefits, and critical ingredients." In: *Psychiatric Rehabilitation Journal* 27.4, p. 392.

Steinmetz, Holger et al. (2016). "How effective are behavior change interventions based on the theory of planned behavior?" In: *Zeitschrift für Psychologie*.

Stephens-Lewis, Danielle et al. (2021). "Interventions to reduce intimate partner violence perpetration by men who use substances: a systematic review and meta-analysis of efficacy". In: *Trauma, Violence, & Abuse* 22.5, pp. 1262–1278.

Szinovacz, Maximiliane E and Lance C Egley (1995). "Comparing one-partner and couple data on sensitive marital behaviors: The case of marital violence". In: *Journal of Marriage and the Family*, pp. 995–1010.

Tarzia, Laura et al. (2017). "Interventions in health settings for male perpetrators or victims of intimate partner violence". In: *Trauma, Violence, & Abuse*, pp. 1524838017744772–1524838017744772.

Tonsing, Jenny and Ravinder Barn (2017). "Intimate partner violence in South Asian communities: Exploring the notion of 'shame'to promote understandings of migrant women's experiences". In: *International Social Work* 60.3, pp. 628–639.

Watson, Emma (2017). "The mechanisms underpinning peer support: a literature review". In: *Journal of Mental Health*.

Williams, Kirk R (1992). "Social sources of marital violence and deterrence: Testing an integrated theory of assaults between partners". In: *Journal of Marriage and the Family*, pp. 620–629.

Wright, Kevin (2002). "Social support within an on-line cancer community: An assessment of emotional support, perceptions of advantages and disadvantages, and motives for using the community from a communication perspective". In: *Journal of Applied Communication Research* 30.3, pp. 195–209.

Table 1: Baseline characteristics, all randomized couples

	HEP		Control		
	N	Mean/SE	N	Mean/SE	Difference
Woman's age	1273	32.687	1272	32.699	−0.022
		(0.230)		(0.236)	(0.328)
Man's age	1354	35.419	1354	35.162	0.248
		(0.237)		(0.230)	(0.330)
Woman has some post-secondary education	1273	0.697	1272	0.698	−0.001
		(0.013)		(0.013)	(0.018)
Man has some post-secondary education	1273	0.756	1272	0.762	−0.006
		(0.012)		(0.012)	(0.017)
Years together (m)	1354	8.772	1354	9.021	−0.254
		(0.194)		(0.203)	(0.279)
Household size	1273	4.292	1272	4.291	0.000
		(0.054)		(0.052)	(0.075)
Woman is employed	1273	0.919	1272	0.932	−0.012
		(0.008)		(0.007)	(0.010)
Man is employed	1354	0.979	1354	0.983	−0.004
		(0.004)		(0.004)	(0.005)
Man uses WhatsApp daily	1351	0.955	1353	0.945	0.009
		(0.006)		(0.006)	(0.008)
Man was recruited via social media	1354	0.613	1354	0.616	−0.001
		(0.013)		(0.013)	(0.017)
Man drinks alcohol (w)	1273	0.696	1272	0.689	0.008
		(0.013)		(0.013)	(0.018)
Decision-making power (w)	1257	7.700	1265	7.695	0.001
		(0.058)		(0.056)	(0.080)
Control index (w)	1130	0.612	1157	0.609	0.001
		(0.007)		(0.007)	(0.009)
Communication index (w)	1199	0.675	1229	0.668	0.006
		(0.006)		(0.006)	(0.007)
Perceived control over sex (w)	1203	0.733	1224	0.727	0.004
		(0.007)		(0.007)	(0.009)
Men's ability to self-regulate (w)	1219	0.480	1226	0.480	−0.001
		(0.008)		(0.008)	(0.010)
Conflict index (m)	1143	0.332	1137	0.336	−0.002
		(0.005)		(0.005)	(0.007)
Any justification of violence (w)	1273	0.208	1272	0.217	−0.008
		(0.011)		(0.012)	(0.016)
Justification of violence index (w)	1273	0.235	1272	0.232	0.003
		(0.014)		(0.013)	(0.019)
Any justification of violence (m)	1354	0.540	1354	0.540	0.001
		(0.014)		(0.014)	(0.019)
Justification of violence index (m)	1354	0.971	1354	0.963	0.011
		(0.031)		(0.031)	(0.043)
Any psychological violence (w)	1273	0.385	1272	0.370	0.019
		(0.014)		(0.014)	(0.017)
Any physical violence (w)	1273	0.396	1272	0.405	0.000
		(0.014)		(0.014)	(0.000)
Any sexual violence (w)	1273	0.153	1272	0.179	−0.024
		(0.010)		(0.011)	(0.014)
Man's father was violent	1282	0.486	1286	0.510	−0.022
		(0.014)		(0.014)	(0.020)
Joint F-statistic (df1=24, df2=1722)					0.992
P-value					0.473

Note: This balance table shows the mean differences between the groups randomized for HEP and controls for various demographic characteristics. None of the differences between both randomized groups is statistically significant.

Table 2: Baseline violence by recruitment source

Source:	Social Media		Ministry		RDD		DHS 2022	
	N	Mean/SE	N	Mean/SE	N	Mean/SE	N	Mean/SE
Men aged 18-31	1665	0.426	742	0.261	300	0.319	20718	0.325
		(0.000)		(0.000)		(0.000)		(0.000)
Men aged 32-38	1665	0.321	742	0.315	300	0.289	20718	0.295
		(0.000)		(0.000)		(0.000)		(0.000)
Men aged 39-60	1665	0.253	742	0.423	300	0.392	20718	0.379
		(0.000)		(0.000)		(0.000)		(0.000)
Completed Secondary or Less (men)	1665	0.261	742	0.175	300	0.398	24698	0.268
		(0.000)		(0.000)		(0.000)		(0.000)
Some technical education (men)	1665	0.345	742	0.285	300	0.302	24698	0.190
		(0.000)		(0.000)		(0.000)		(0.000)
Some university education (men)	1665	0.394	742	0.540	300	0.389	24698	0.542
		(0.000)		(0.000)		(0.000)		(0.000)
Never drinks alcohol (men)	1665	0.301	742	0.396	300	0.419	21321	0.377
		(0.000)		(0.000)		(0.000)		(0.000)
Drinks alcohol occasionally (men)	1665	0.304	742	0.347	300	0.349	21321	0.574
		(0.000)		(0.000)		(0.000)		(0.000)
Drinks alcohol habitually (men)	1665	0.395	742	0.257	300	0.233	21321	0.049
		(0.000)		(0.000)		(0.000)		(0.000)
Any justification of violence (men)	1665	0.570	742	0.467	300	0.555	-	-
		(0.012)		(0.018)		(0.029)		-
Justification of violence index (men)	1665	1.031	742	0.793	300	1.047	-	-
		(0.028)		(0.039)		(0.070)		-
Conflict index (men)	1386	0.377	640	0.261	253	0.282	-	-
		(0.004)		(0.006)		(0.010)		-
Any psychological violence (wom)	1543	0.464	710	0.249	286	0.237	14198	0.134
		(0.013)		(0.016)		(0.025)		(0.003)
Any physical violence (wom)	1543	0.497	710	0.250	286	0.254	14198	0.086
		(0.013)		(0.016)		(0.026)		(0.002)
Any sexual violence (wom)	1543	0.214	710	0.097	286	0.084	14198	0.019
		(0.010)		(0.011)		(0.016)		(0.001)

Note: This table shows sample characteristics at baseline by source of recruitment. Men were recruited via: (1) social media advertising campaign (mostly Facebook), (2) nationally-dispersed promoters employed by the Ministry of Women, or (3) random digit dialing (RDD) using numbers from a verified, nationally representative, phone database. The estimates for psychological, physical, and sexual violence from our baseline survey are based on self-reported violence in the past 6 months. (4) displays results from a representative urban subpopulation from the 2022 Peruvian ENDES survey, the national Demographic and Health Survey. Information in this table came from different modules with different eligibility criteria. The results of this national survey report violence in the past 12 months and only interview women. All responses related to men's demographic characteristics from ENDES were reported by their partners. Current partners reported information related to alcohol consumption of men. "Drinks alcohol occasionally" is defined in our sample and the ENDES as responses of their partner drinking "sometimes". "Drinks alcohol habitually" is defined in our sample as responses indicating their partner drinking a specific amount of times a week, and in ENDES as the responses indicating "high frequency" of alcohol consumption. Columns show mean and robust standard errors (HC2) by source.

Table 3: Effects on violence outcomes

	Any Physical		Any Sexual	
	(1)	(2)	(3)	(4)
HEP	−0.007	−0.006	−0.023*	−0.022*
	(0.018)	(0.017)	(0.014)	(0.013)
RI p-value	0.689	0.710	0.093	0.096
Control Mean	0.231	0.231	0.114	0.114
Covariates	No	Yes	No	Yes
Observations	1989	1989	1990	1990

Notes: First column for each outcome is a design-based least squares estimator. Second column includes baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. All estimations include fixed effects for randomization strata and batch. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses. Randomization-based p-values are calculated from 10,000 permutted assignments using a studentized test statistic with HC2 standard errors.
* p < 0.1, ** p < 0.05, *** p < 0.01

Table 4: Effects on specific items related to sexual violence

	Forced sex		Forced other acts	
	(1)	(2)	(3)	(4)
HEP	−0.020*	−0.024**	−0.010	−0.012
	(0.012)	(0.012)	(0.011)	(0.012)
RI p-value	0.095	0.0474	0.397	0.313
Control Mean	0.090	0.090	0.073	0.073
Covariates	No	Yes	No	Yes
Observations	1912	1912	1991	1991

Note: First column for each outcome is a design-based least squares estimator. Second column includes baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. All estimations include fixed effects for randomization strata and batch. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses. Randomization-based p-values are calculated from 10,000 permutted assignments using a studentized test statistic with HC2 standard errors.
* p < 0.1, ** p < 0.05, *** p < 0.01

Table 5: Heterogeneity by baseline violence, age, education, and alcohol consumption

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Phys.	Sex.	Phys.	Sex.	Phys.	Sex.	Phys.	Sex.
HEP	0.026	0.003	0.008	-0.044**	-0.037	-0.015	-0.008	-0.042**
	(0.016)	(0.010)	(0.029)	(0.021)	(0.038)	(0.031)	(0.027)	(0.020)
HEP × (Baseline IPV)	-0.072*	-0.053*						
	(0.038)	(0.030)						
Baseline IPV	0.334*	0.181						
	(0.196)	(0.166)						
HEP ×(32≤ Man's Age ≤ 38)			-0.072*	0.044				
			(0.042)	(0.032)				
HEP ×(39≤ Man's Age ≤ 60)			0.031	0.034				
			(0.040)	(0.031)				
32 ≤ Man's Age ≤ 38			0.004	-0.019				
			(0.030)	(0.023)				
39 ≤ Man's Age ≤ 60			-0.051*	-0.016				
			(0.027)	(0.023)				
HEP × (Man - some technical education)					0.031	-0.017		
					(0.049)	(0.040)		
HEP × (Man - some university education)					0.051	0.002		
					(0.044)	(0.035)		
Man - some technical education					0.027	0.008		
					(0.034)	(0.029)		
Man - some university education					-0.048	-0.059**		
					(0.031)	(0.025)		
HEP × (Man occasionally drinks)							0.006	0.019
							(0.038)	(0.029)
HEP × (Man habitually drinks)							-0.002	0.046
							(0.043)	(0.034)
Man occasionally drinks							-0.017	-0.006
							(0.027)	(0.022)
Man habitually drinks							0.098***	0.025
							(0.030)	(0.024)
Control Mean for Excluded Group	0.070	0.031	0.246	0.127	0.281	0.150	0.175	0.097
Control SD	0.256	0.173	0.431	0.334	0.451	0.358	0.381	0.296
Observations	1989	1990	1989	1990	1989	1990	1989	1990
R-sqr	0.191	0.085	0.191	0.083	0.192	0.091	0.200	0.088

All columns display regression results of physical and sexual violence on different heterogeneity measures interacted with participation in the HEP program. Columns 1 and 2 evaluate a binary for any physical or sexual violence reported at baseline. Columns 3 and 4 consider the top 2 terciles of men's age reported at baseline relative to the first tercile. Columns 5 and 6 evaluate different levels of educational attainment by men relative to those who at most, finished high school. Columns 7 and 8 consider different levels of alcohol consumption by men relative to those who never drink. Current partners reported information related to alcohol consumption of men. "Drinks alcohol occasionally" is defined as responses indicating partner drinking "sometimes". "Drinks alcohol habitually" is defined as responses that indicate their partner drinking a specific amount of times a week. HC 2 Standard errors displayed in parentheses.

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

Table 6: Effects on sexual violence by group composition

	Any Sexual					
	(1)	(2)	(3)	(4)	(5)	(6)
HEP	0.043	0.006	0.085*	0.037	0.007	0.079
	(0.026)	(0.029)	(0.041)	(0.034)	(0.049)	(0.052)
HEP × just. index (group)	−0.131**	−0.039	−0.222***			
	(0.044)	(0.054)	(0.068)			
HEP × % sexual violence (group)				−0.093*	−0.031	−0.172**
				(0.052)	(0.080)	(0.076)
Sample	Full	No just.	Any just.	Full	No just.	Any just.
RI p-value	0.024	0.462	0.007	0.065	0.866	0.022
Control Mean	0.114	0.075	0.149	0.114	0.078	0.146
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1990	940	1050	1904	900	1004

Note: Columns 1, 2, and 3 are least squares regressions of endline sexual violence on participation in the HEP program interacted with the group-level justification index. The justification index is an aggregate of 5 violence norms questions from our baseline survey which measures men's views about when violence is justified. Columns 4, 5, and 6 are least squares regressions of endline sexual violence on the percentage of group members identified by their spouses as perpetrators of sexual violence at baseline. Results from columns 1 and 4 were calculated on our complete sample ("Full"). Results from columns 2 and 5 were calculated on the subsample of women whose partners never justified any violence at baseline ("No just."). Results from columns 3 and 6 were calculated on the subsample of women whose partners justified violence for 1 or more of our 5 violence norms questions ("Any just."). All standard errors are robust (CR2) and clustered by WhatsApp group and all estimations include fixed effects for randomization strata and batch. Randomization-based p-values for the composition effect are calculated from 10,000 permutted assignments of the program and group cross-randomizations using a studentized test statistic with CR2 standard errors. All estimations include fixed effects for randomization strata and batch.
* p < 0.1, ** p < 0.05, *** p < 0.01

Table 7: Effects on justification of violence by group composition

	Justification Index		
	(1)	(2)	(3)
HEP	−0.124	−0.004	−0.137
	(0.096)	(0.095)	(0.153)
HEP × just. index (group)	0.156	−0.107	0.202
	(0.179)	(0.179)	(0.275)
Sample	Full	No just.	Any just.
RI p-value	0.704	0.331	0.536
Control Mean	0.780	0.451	1.075
Observations	1997	942	1055

Note: This table shows least squares regressions of individual- level justification index on participation in the HEP program interacted with the group-level justification index. The justifi- cation index is an aggregate of 5 violence norms questions from our baseline survey which measures men's views about when violence is justified. Results from column 1 were calculated on our complete sample ("Full"). Results from column 2 were calculated on the subsample of women whose partners never justified any violence at baseline ("No just."). All standard er- rors are robust (CR2) and clustered by WhatsApp group and all estimations include fixed effects for randomization strata and batch. Randomization-based p-values for the composition effect are calculated from 10,000 permutted assignments of the program and group cross-randomizations using a studentized test statistic with CR2 standard errors. All estimations include fixed effects for randomization strata and batch.

* p < 0.1, ** p < 0.05, *** p < 0.01

Table 8: Potential mechanisms for effect of group composition: remaining in chat.

	Remain in Chat			Days in Chat		
	(1)	(2)	(3)	(4)	(5)	(6)
just. index (group)	0.056	−0.137	0.221**	1.398	−2.219	4.515**
	(0.097)	(0.139)	(0.096)	(2.315)	(3.628)	(2.150)
Sample	HEP	No just.	Any just.	HEP	No just.	Any just.
RI p-value	0.590	0.373	0.046	0.563	0.582	0.072
Control Mean	0.635	0.717	0.566	22.23	23.84	20.86
Observations	1355	623	732	1355	623	732

Note: Columns 1-3 show least squares regressions of the binary indicator for an individual remaining in the group chat during the whole duration of the HEP intervention on the group-level justification index. Columns 4-6 show least squares regressions of the number of days an individual remained in the group chat during the HEP intervention on the group-level justification index. The justification index is an ag- gregate of 5 violence norms questions from our base- line survey which measures men's views about when violence is justified. Results from column 1 were cal- culated on our complete sample of treated individuals ("HEP"). Results from column 2 were calculated on the subsample of women whose partners never justi- fied any violence at baseline ("No just."). All stan- dard errors are robust (CR2) and clustered by What- sApp group and all estimations include fixed effects for randomization strata and batch. Randomization- based p-values for the composition effect are calcu- lated from 10,000 permutted assignments of the group cross-randomizations using a studentized test statistic with CR2 standard errors. All estimations include fixed effects for randomization strata and batch.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Potential mechanisms for effect of group composition

	Share problems (%)		Helpful feedback (%)		Group conflict (%)	
	(1)	(2)	(3)	(4)	(5)	(6)
just. index (group)	0.119***	0.080*	0.022	0.023*	−0.001	−0.002
	(0.042)	(0.042)	(0.021)	(0.013)	(0.007)	(0.008)
RI p-value	0.003	0.088	0.126	0.053	0.891	0.802
Fixed Effects	No	Yes	No	Yes	No	Yes
Control Mean	0.15	0.15	0.11	0.11	0.01	0.01
Observations	1355	1355	1355	1355	1355	1355

Note: This table displays various results of regressions of various types of messages that serve as potential mechanisms for the effects of the HEP program on the group-level justification index. The justification index is an aggregate of 5 violence norms questions from our baseline survey which measures men's views about when violence is justified. Columns 1 and 2 display results for the fraction of group interactions in which a participant shares the personal problems they are having with their spouses on the group chat. Columns 3 and 4 show results for the fraction of group interactions in which a participant or the facilitator offers concrete advice for overcoming relationship challenges. Columns 5 and 6 display results for the fraction of group interactions in which participants generate conflict within the group chat by insulting others or questioning the authority of the facilitator. All results in this table were calculated on the subsample of men who participated in the HEP program. All standard errors are robust (HC2).
* p < 0.1, ** p < 0.05, *** p < 0.01

Table 10: Effect of group homogeneity

	Share problems (%)		Helpful feedback (%)		Group conflict (%)	
	(1)	(2)	(3)	(4)	(5)	(6)
homogeneity (group)	0.054	0.087	−0.071	−0.024	−0.057	−0.039**
	(0.153)	(0.125)	(0.055)	(0.040)	(0.035)	(0.018)
RI p-value	0.767	0.547	0.241	0.558	0.160	0.042
Fixed Effects	No	Yes	No	Yes	No	Yes
Control Mean	0.13	0.13	0.03	0.03	0.01	0.01
Observations	1355	1355	1355	1355	1355	1355

Note: This table displays various results of regressions of various types of messages that serve as potential mechanisms for the effects of the HEP program on group-level homogeneity. The measure of homogeneity is generated by calculating the distance between the group-level proportion who justify any form of violence at baseline and 50%. Columns 1 and 2 display results for the fraction of group interactions in which a participant shares the personal problems they are having with their spouses on the group chat. Columns 3 and 4 show results for the fraction of group interactions in which a participant or the facilitator offers concrete advice for overcoming relationship challenges. Columns 5 and 6 display results for the fraction of group interactions in which participants generate conflict within the group chat by insulting others or questioning the authority of the facilitator. All results in this table were calculated on the subsample of men who participated in the HEP program. All standard errors are robust (HC2).
* p < 0.1, ** p < 0.05, *** p < 0.01

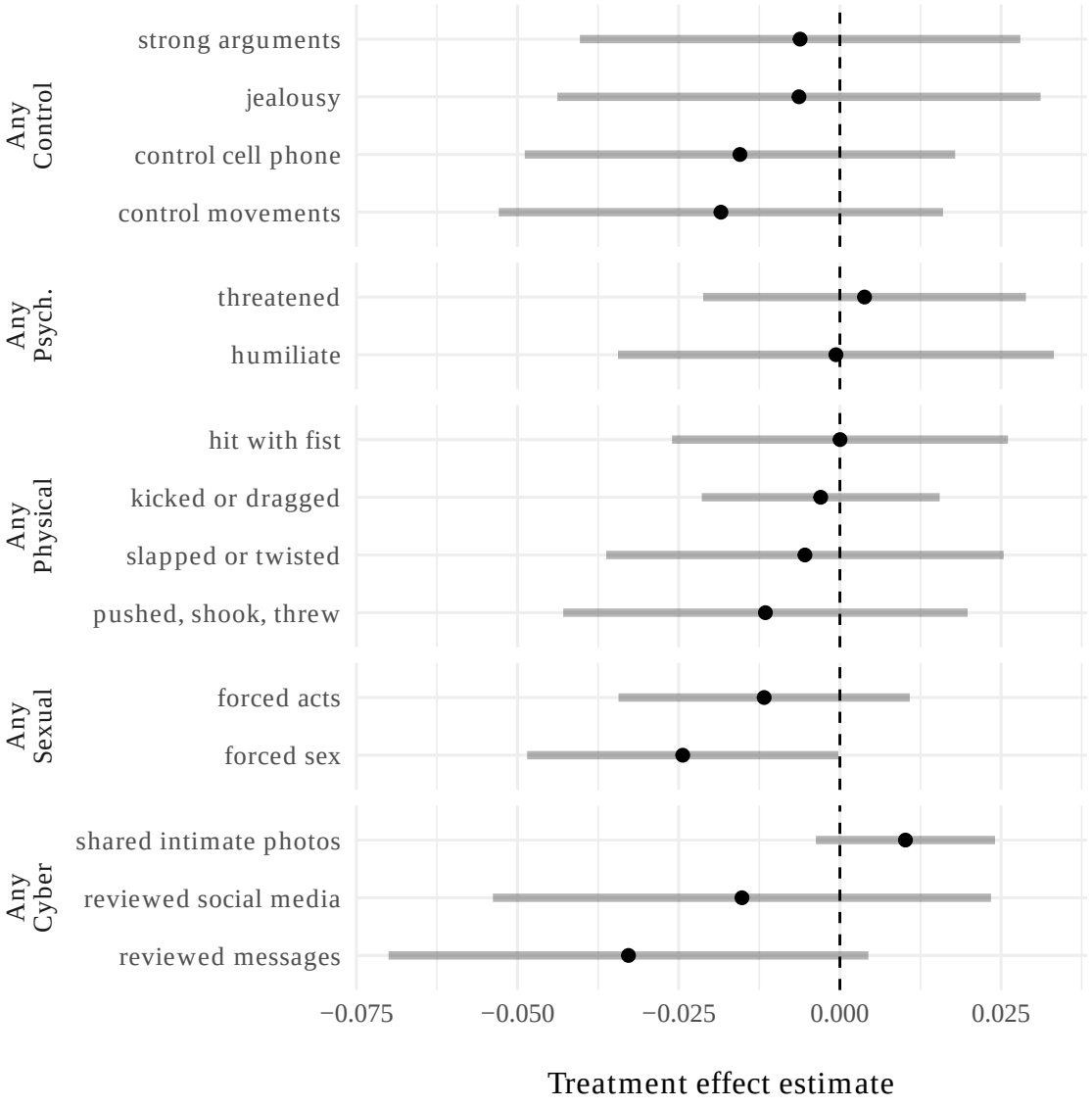

Figure 1: Estimated effect of HEP on individual violence items.

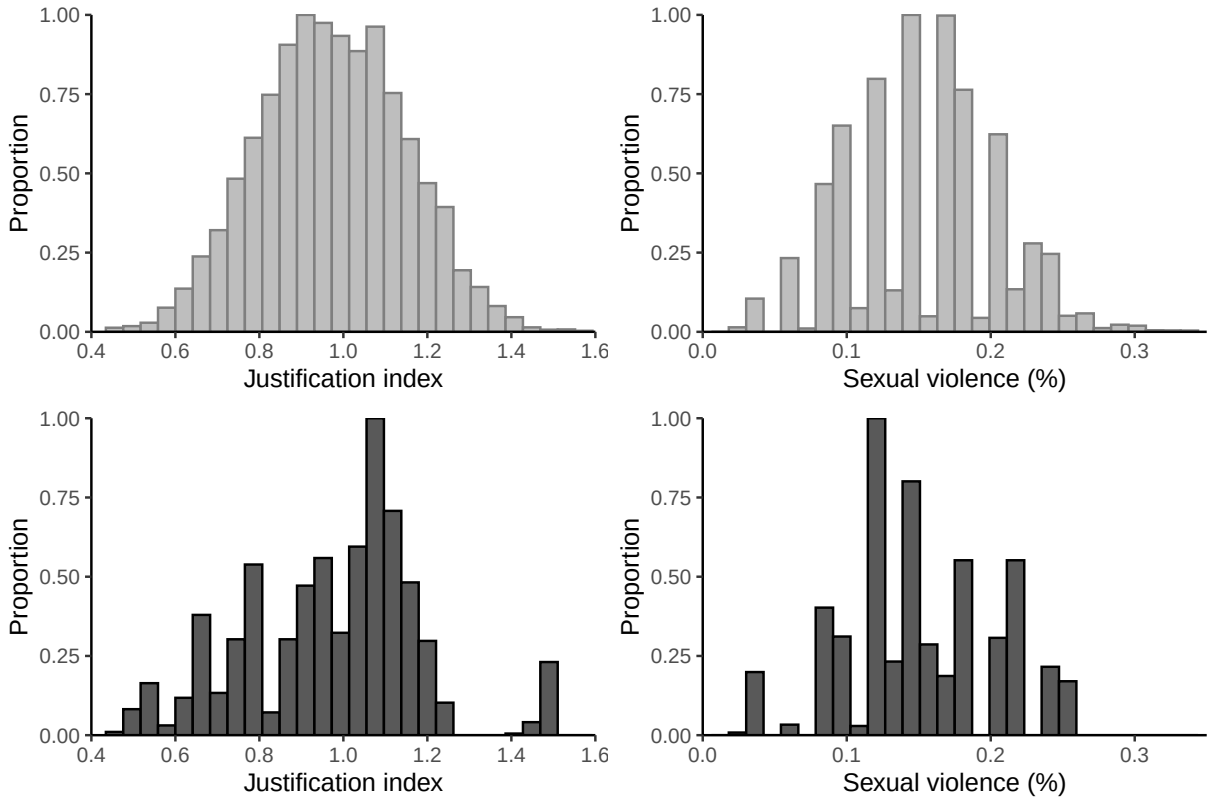

Figure 2: Distribution of group composition variables. Top panel shows simulated distributions under repeated randomizations in grey. Bottom panel shows observed distributions in black.

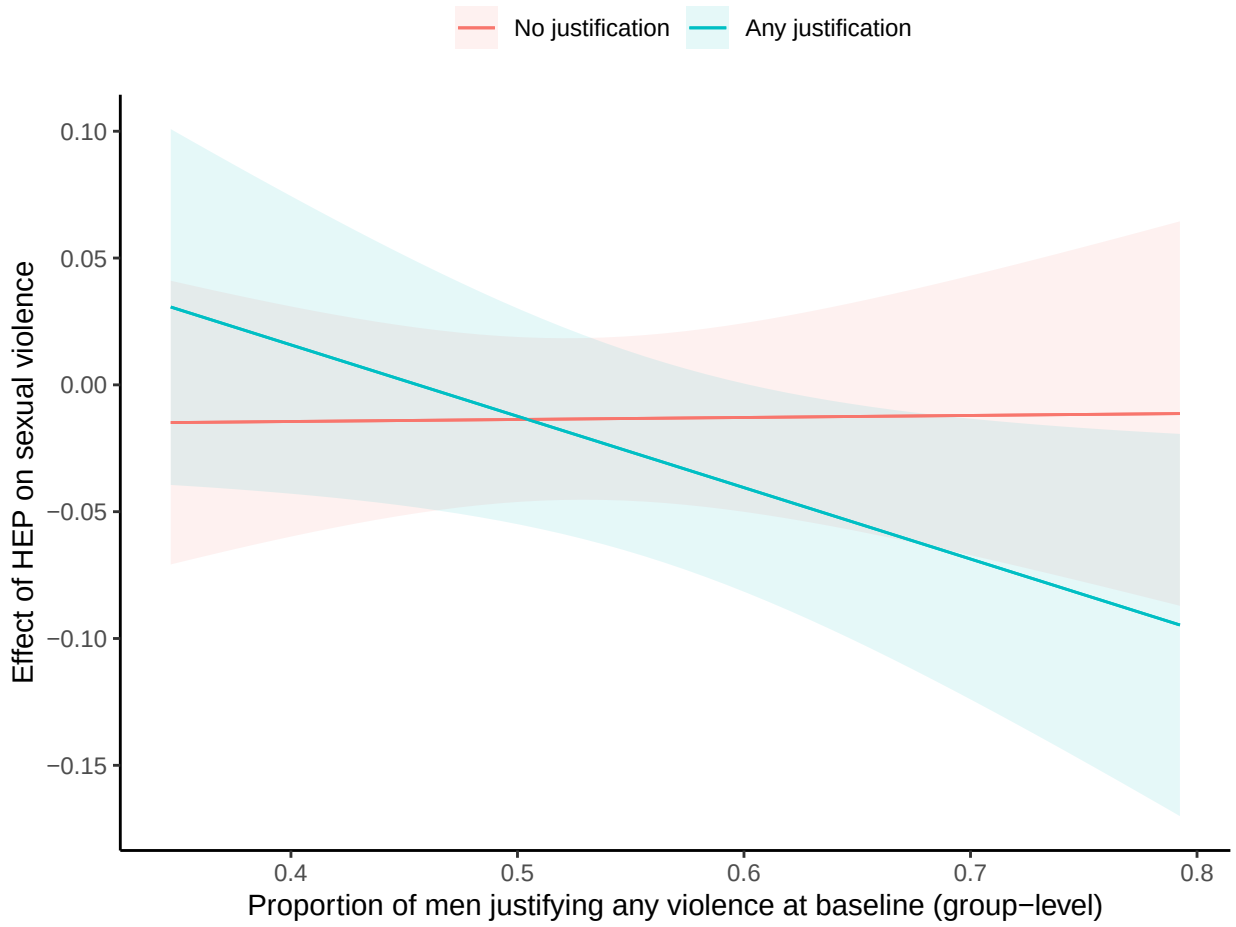

Figure 3: Effect of HEP on sexual violence grows with increasing concentration of men justifying violence, particularly among violent men.

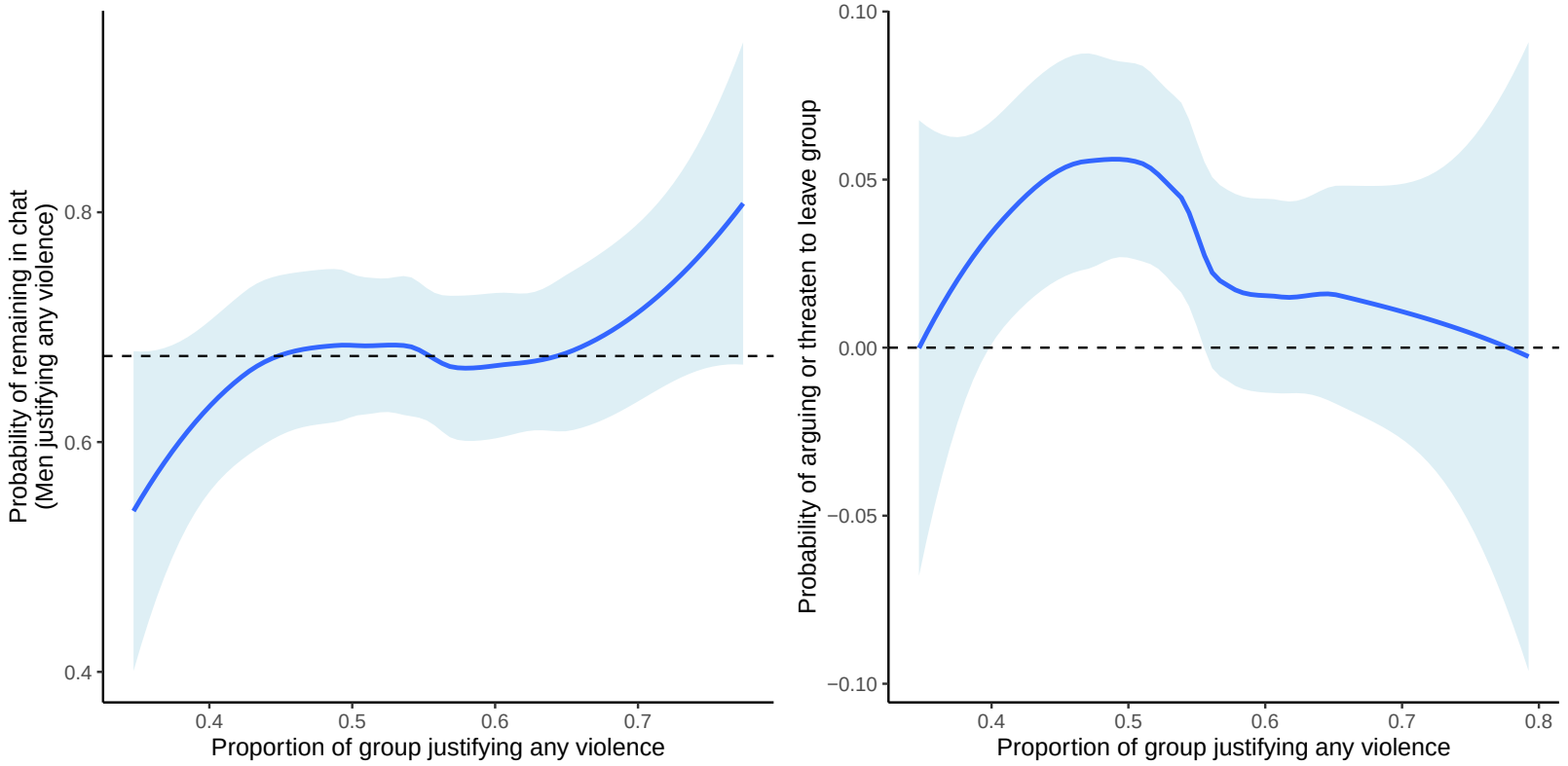

Figure 4: Potential mechanisms for the effects of group composition. Plots show LOESS fits of group-level (left) probability of remaining in chat among men justifying any violence at baseline and (right) probability of participants arguing or threatening to leave as a function the proportion of the group justifying any form of violence at baseline.

6 Appendix

6.1 Recruitment

Table A.1: Summary of recruitment of study participants by source

	Men's Number Obtained		Men Recruited		Study Participants	
Source	N	%	N	%	N	%
SMA	7,353	22.7	2,414	62.4	1,666	61.5
MIMP	2,327	7.2	960	24.8	743	27.4
RDD	22,696	70.1	481	12.4	296	10.9
Other	11	0.0	11	0.3	5	0.2
Total	**32,387**	**-**	**3,866**	**-**	**2,710**	**-**

Table A.2: Baseline characteristics of responders versus non-responders at follow up

	Responder		Non-responder		
	N	Mean/SE	N	Mean/SE	Difference
Woman's age	1908	32.65	637	32.81	−0.33
		(0.190)		(0.328)	(0.378)
Man's age	1746	35.11	494	35.74	−0.74*
		(0.207)		(0.387)	(0.441)
Woman has some post-secondary education	1908	0.72	636	0.64	0.07***
		(0.010)		(0.019)	(0.022)
Man has some post-secondary education	1746	0.78	494	0.73	0.04**
		(0.001)		(0.020)	(0.022)
Years together (reported by man)	1746	8.97	494	8.78	0.10
		(0.174)		(0.334)	(0.377)
Household size	1908	4.30	637	4.27	0.02
		(0.044)		(0.071)	(0.084)
Woman is employed	1908	0.93	637	0.93	0.00
		(0.006)		(0.010)	(0.012)
Man is employed	1746	0.98	494	0.98	0.01
		(0.003)		(0.007)	(0.008)
Man uses WhatsApp daily	1743	0.95	494	0.95	−0.01
		(0.005)		(0.010)	(0.011)
Man was recruited via social media	1746	0.59	494	0.62	−0.02
		(0.012)		(0.022)	(0.024)
Man drinks alcohol (reported by woman)	1908	0.68	637	0.72	−0.02
		(0.011)		(0.018)	(0.021)
Decision-making power (reported by woman)	1892	7.72	630	7.62	0.06
		(0.046)		(0.085)	(0.097)
Control index (reported by woman)	1719	0.62	568	0.59	0.02
		(0.005)		(0.009)	(0.011)
Communication index (reported by woman)	1824	0.68	604	0.65	0.01
		(0.005)		(0.008)	(0.009)
Perceived control over sex (reported by woman)	1829	0.74	598	0.71	0.02
		(0.006)		(0.010)	(0.011)
Men's ability to self-regulate (reported by woman)	1836	0.49	609	0.43	0.03***
		(0.007)		(0.011)	(0.011)
Conflict index (reported by man)	1473	0.33	421	0.34	0.00
		(0.004)		(0.008)	(0.009)
Justification of violence (reported by woman)	1846	0.59	613	0.58	0.02***
		(0.003)		(0.005)	(0.005)
Justification of violence (reported by man)	1746	0.93	494	1.02	−0.06
		(0.026)		(0.052)	(0.058)
Any psychological violence (reported by woman)	1908	0.36	637	0.42	−0.02
		(0.011)		(0.020)	(0.020)
Any physical violence (reported by woman)	1908	0.38	637	0.47	−0.09***
		(0.011)		(0.020)	(0.023)
Any sexual violence (reported by woman)	1908	0.16	637	0.18	−0.02
		(0.008)		(0.015)	(0.017)
Man's father was violent	1656	0.49	467	0.50	0.01
		(0.012)		(0.023)	(0.026)
Joint F-statistic (df1=14, df2=2131)					3.051
P-value					0.000

6.2 Peer Effects

Suppose our recruited sample is a fixed population of n individuals to be randomly assigned a treatment $Z_i \in \{0, 1\}$ for $i = 1, \ldots, n$ with $Z = 1$ denoting assignment to WhatsApp intervention and $Z = 0$ to pure control. We represent the vector of assignments for all individuals by $\mathbf{Z} = (Z_1, \ldots, Z_n)$. Let $Y_i(\mathbf{z})$ denote the potential outcome of individual i under an intervention that sets $\mathbf{Z} = \mathbf{z}$. Assume, first, no interference between individuals such that $Y_i(z_i, \mathbf{z}_{-i}) = Y_i(z_i, \mathbf{z'}_{-i}) = Y_i(z_i)$ for all $\mathbf{z}$ and $\mathbf{z'}$. We define the average treatment effect estimand in this case as:

$$ATE := \frac{1}{n} \sum_{i=1}^{n} \{Y_i(1) - Y_i(0)\}.$$

Under simple random assignment it is identified by the regression estimator for β_1 in the specification in equation (1) and it has been shown that HC2 standard errors converge to the same asymptotic limit as the variance of $\widehat{\tau}$ under random assignment.

Now suppose, those in the treated group are further assigned, randomly, to a group $A_i \in \{1, \ldots, G\}$ in which the treatment will be delivered (i.e. WhatsApp chat) with vector of assignments represented by $\mathbf{A} = (A_1, \ldots, A_n)$. Because the treatment involves interactions within groups, we assume that an individual's outcome under treatment may be affected by the peers in the group that they are assigned to, i.e. there is interference between units, however it only occurs within groups that are randomly formed. In this case, define $Y_i(\mathbf{z}, \mathbf{a})$ as the potential outcome under treatment assignment $\mathbf{Z} = \mathbf{z}$ and subsequent group assigment $\mathbf{A} = \mathbf{a}$. Since an individual is only affected by peers in same group $Y_i(\mathbf{z}, \mathbf{a}) = Y_i(z_i, \mathcal{A}(a_i))$ where $\mathcal{A}(a_i)$ represents the units in group a_i. We can then define the average marginal peer effect as:

$$AMPE := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|\mathcal{A}_i|} \sum_{\omega \in \mathcal{A}_i} \{Y_i(1, \omega) - Y_i(0, \omega)\},$$

where $\overline{\mathcal{A}(a_i)}$ is the set of possible peers for individual i. In words this represents the direct effect of treatment when assigned an "average" peer group (or the average over all possible peer groups). It is still identified by the regression estimator for β_1 in the specification in equation (1). However, as shown by Fu and Samii a better estimator of the standard error is the CR2 estimator.

Table A.3: Sensitivity to peer effects, sexual violence.

	Any Sexual			
	(1)	(2)	(3)	(4)
HEP	−0.023*	−0.022*	−0.023*	−0.022*
	(0.014)	(0.013)	(0.013)	(0.013)
RI p-value	0.093	0.096	0.075	0.086
Control Mean	0.114	0.114	0.114	0.114
Covariates	No	Yes	No	Yes
Fixed Effects	Yes	Yes	Yes	Yes
Standard Errors	HC2	HC2	CR2	CR2
Estimand	ATE	ATE	AMPE	AMPE
Observations	1990	1990	1990	1990

Notes: First and third columns are design-based least squares estimator that includes fixed effects for randomization strata and batch. Second and fourth columns includes baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. First and second columns target the average treatment effect (ATE) under no peer effects and use Heteroscedasticity-consistent robust standard errors (HC2). The third and fourth columns allow for peer effects in which case the estimand is the direct effect marginalized over peers (AMPE) and use cluster-robust standard errors (CR2) adjusting for clustering within treatment groups. Randomization-based p-values are calculated from 10,000 permutted assignments using a studentized test statistic using either HC2 or CR2 standard errors based on specification.
* p < 0.1, ** p < 0.05, *** p < 0.01

Table A.4: Sensitivity to peer effects, physical violence.

	Any Physical			
	(1)	(2)	(3)	(4)
HEP	−0.007	−0.006	−0.007	−0.006
	(0.018)	(0.017)	(0.015)	(0.015)
RI p-value	0.675	0.711	0.689	0.710
Control Mean	0.231	0.231	0.231	0.231
Covariates	No	Yes	No	Yes
Fixed Effects	Yes	Yes	Yes	Yes
Standard Errors	HC2	HC2	CR2	CR2
Estimand	ATE	ATE	AMPE	AMPE
Observations	1990	1990	1990	1990

Notes: First and third columns are design-based least squares estimator that includes fixed effects for randomization strata and batch. Second and fourth columns includes baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. First and second columns target the average treatment effect (ATE) under no peer effects and use Heteroscedasticity-consistent robust standard errors (HC2). The third and fourth columns allow for peer effects in which case the estimand is the direct effect marginalized over peers (AMPE) and use cluster-robust standard errors (CR2) adjusting for clustering within treatment groups. Randomization-based p-values are calculated from 10,000 permutted assignments using a studentized test statistic using either HC2 or CR2 standard errors based on specification.
* p < 0.1, ** p < 0.05, *** p < 0.01

6.3 Group Composition

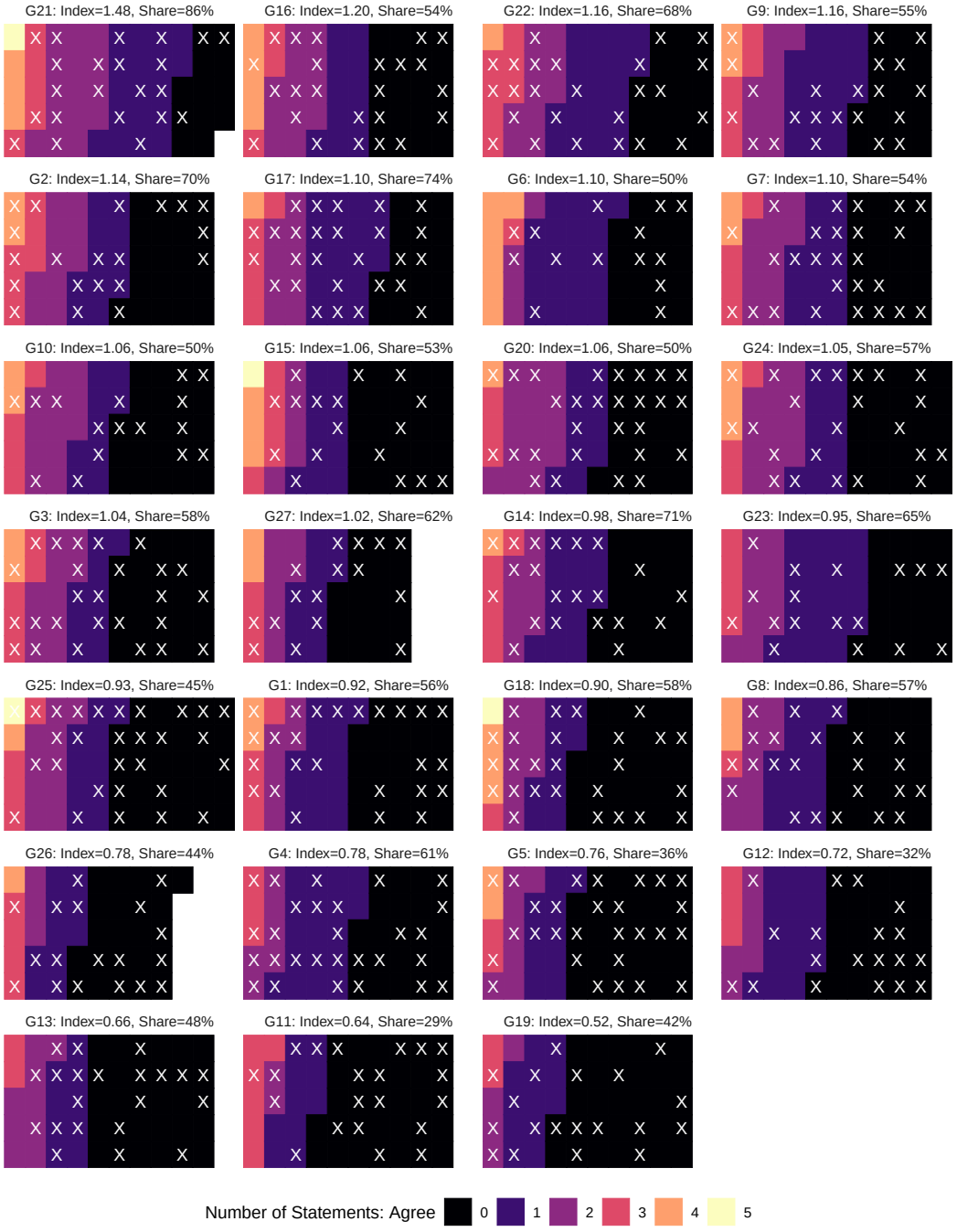

Figure A.1: Variation in group composition and who engages across groups. Subplots represent a WhatsApp group with one square for each individual colored according to the number of statements where they agreed a man would be justified in using violence at baseline. Individuals who sent more than one message to group during the 30-day program are denoted by a white X. Titles state mean index value and share of messages from men ever justifying violence.

6.4 Attrition

Table A.5: Attrition by Gender

		Women's Response Rate				Men's Response Rate			
	N	T	C	T - C	P-value	T	C	T - C	P-value
Panel A. Overall									
	2710	0.728	0.746	-0.018	0.276	0.650	0.639	0.011	0.547
Panel B. Within baseline violence strata									
No violence	1391	0.763	0.796	-0.034	0.129	0.695	0.697	-0.001	0.960
Sexual only	123	0.803	0.823	-0.019	0.785	0.803	0.661	0.142	0.074
Physical or sexual	1019	0.698	0.713	-0.016	0.580	0.607	0.618	-0.011	0.714
No baseline	177	0.573	0.489	0.084	0.262	0.438	0.295	0.143	0.048
Joint F-test					*0.508*				*0.077*

Notes: This table evaluates the attrition rates of both men and women for the endline survey. The columns labeled "T-C" represent the mean difference in response rates between the treatment and control groups across various subsamples.

6.5 Index Components

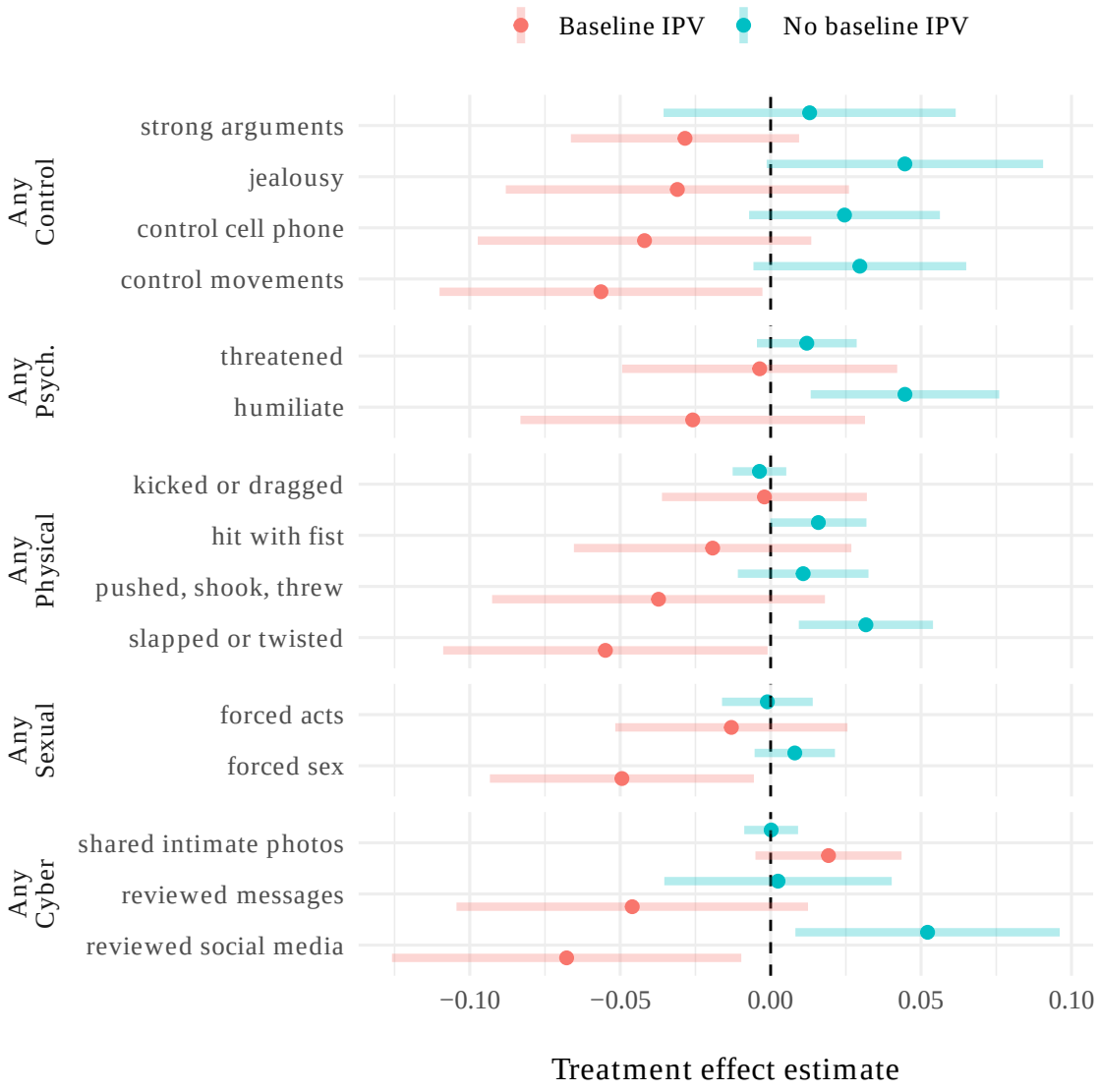

Figure A.2: Estimated effect of HEP on individual violence items among couples reporting any baseline violence versus no violence.

6.6 Other primary outcomes

Table A.6: ITT estimates of effects of HEP on other binary violence outcomes

	Any Control		Any Psych.		Any Cyber	
	(1)	(2)	(3)	(4)	(5)	(6)
treatment	−0.011	−0.011	0.011	0.005	−0.010	−0.020
	(0.017)	(0.016)	(0.019)	(0.017)	(0.022)	(0.020)
	[0.541]	[0.490]	[0.573]	[0.760]	[0.639]	[0.309]
Covariates	No	Yes	No	Yes	No	Yes
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Control Mean	0.801	0.801	0.262	0.262	0.411	0.411
Control SD	0.399	0.399	0.440	0.440	0.492	0.492
Observations	1993	1993	1988	1988	1956	1956
R^2	0.074	0.231	0.164	0.301	0.093	0.211

Notes: First column for each outcome is design-based least squares estimator that includes fixed effects for randomization strata and batch. Second column adjusts for baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses and p-values in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

6.7 Effect heterogeneity

Table A.7: ITT estimates of effects of HEP on primary violence outcomes by woman's education

	Any IPV		Any Physical		Any Sexual	
	(1)	(2)	(3)	(4)	(5)	(6)
treatment	−0.020	−0.029	−0.043	−0.043	0.003	0.001
	(0.032)	(0.032)	(0.031)	(0.030)	(0.025)	(0.025)
	[0.538]	[0.353]	[0.169]	[0.155]	[0.899]	[0.957]
woman's education ≥ median	−0.061**	−0.042	−0.058**	−0.059**	−0.026	0.006
	(0.027)	(0.027)	(0.026)	(0.025)	(0.021)	(0.022)
	[0.023]	[0.122]	[0.028]	[0.020]	[0.216]	[0.794]
treatment × woman's education ≥ median	0.025	0.043	0.057	0.071**	−0.034	−0.030
	(0.039)	(0.040)	(0.038)	(0.036)	(0.030)	(0.032)
	[0.530]	[0.280]	[0.131]	[0.050]	[0.255]	[0.342]
Covariates	No	Yes	No	Yes	No	Yes
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Control Mean	0.263	0.263	0.232	0.232	0.114	0.114
Control SD	0.441	0.441	0.422	0.422	0.318	0.318
Observations	1898	1898	1901	1901	1904	1904
R^2	0.216	0.292	0.203	0.274	0.094	0.156

Notes: First column for each outcome is design-based least squares estimator that includes fixed effects for randomization strata and batch. Second column adjusts for baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses and p-values in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.8: ITT estimates of effects of HEP on primary violence outcomes by woman's age

	Any IPV		Any Physical		Any Sexual	
	(1)	(2)	(3)	(4)	(5)	(6)
treatment	−0.010	−0.023	0.000	0.000	−0.028	−0.031
	(0.026)	(0.026)	(0.026)	(0.025)	(0.019)	(0.019)
	[0.711]	[0.361]	[0.997]	[0.996]	[0.140]	[0.100]
woman's age ≥ median	−0.029	−0.035	−0.034	−0.041*	−0.011	−0.009
	(0.025)	(0.024)	(0.024)	(0.023)	(0.019)	(0.019)
	[0.250]	[0.146]	[0.162]	[0.075]	[0.570]	[0.639]
treatment × woman's age ≥ median	0.013	0.043	−0.010	0.006	0.020	0.027
	(0.036)	(0.035)	(0.035)	(0.034)	(0.027)	(0.026)
	[0.725]	[0.219]	[0.767]	[0.854]	[0.469]	[0.302]
Covariates	No	Yes	No	Yes	No	Yes
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Control Mean	0.263	0.263	0.232	0.232	0.114	0.114
Control SD	0.441	0.441	0.422	0.422	0.318	0.318
Observations	1898	1898	1901	1901	1904	1904
R^2	0.214	0.292	0.203	0.274	0.089	0.156

Notes: First column for each outcome is design-based least squares estimator that includes fixed effects for randomization strata and batch. Second column adjusts for baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses and p-values in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.9: ITT estimates of effects of HEP on primary violence outcomes by recruitment source

	Any IPV		Any Physical		Any Sexual	
	(1)	(2)	(3)	(4)	(5)	(6)
treatment	−0.019	−0.026	−0.027	−0.026	−0.022	−0.028
	(0.027)	(0.028)	(0.026)	(0.026)	(0.018)	(0.019)
	[0.472]	[0.359]	[0.290]	[0.322]	[0.227]	[0.142]
recruited from social media	0.077***	0.002	0.053**	0.000	0.051***	0.021
	(0.025)	(0.027)	(0.024)	(0.025)	(0.019)	(0.018)
	[0.002]	[0.929]	[0.026]	[0.996]	[0.007]	[0.242]
treatment × recruited from social media	0.016	0.022	0.034	0.033	−0.002	0.009
	(0.037)	(0.039)	(0.035)	(0.036)	(0.026)	(0.025)
	[0.655]	[0.575]	[0.333]	[0.359]	[0.951]	[0.730]
Covariates	No	Yes	No	Yes	No	Yes
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Control Mean	0.265	0.265	0.232	0.232	0.114	0.114
Control SD	0.442	0.442	0.422	0.422	0.318	0.318
Observations	1986	1986	1989	1989	1990	1990
R^2	0.212	0.280	0.199	0.260	0.091	0.151

Notes: First column for each outcome is design-based least squares estimator that includes fixed effects for randomization strata and batch. Second column adjusts for baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses and p-values in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.10: Presence of violence and conservative beliefs among men about when violence is justified predict sexual violence at baseline and endline

	Any sexual: baseline		Any sexual: endline	
	(1)	(2)	(3)	(4)
(Intercept)	0.130***	0.129***	0.072***	0.051***
	(0.009)	(0.009)	(0.008)	(0.005)
punish disrespect	0.007			
	(0.017)			
jealousy is love	0.079***			
	(0.021)			
asking to be harassed	0.055**			
	(0.024)			
punish infidelity	0.004			
	(0.019)			
always willing for sex	0.015			
	(0.025)			
justification index		0.029***	0.034***	
		(0.007)	(0.007)	
Any sexual violence at baseline				0.335***
				(0.028)
Observations	2710	2710	1990	1904
R^2	0.013	0.008	0.015	0.162

Note: The first column is a least squares regression of sexual violence at baseline on justification items. The second column shows the same result applied to the index. The third shows persistance of the association regressing endline violence on baseline justification index. The fourth shows the corresponding association between sexual violence at baseline and sexual violence at endline. All standard errors are robust (HC2).
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.11: ITT estimates of effects of HEP on other primary outcomes

	Control & DM Index		Consent Index		Comm. Index	
	(1)	(2)	(3)	(4)	(5)	(6)
treatment	−0.001	−0.002	0.007	0.005	0.009	0.003
	(0.006)	(0.006)	(0.010)	(0.010)	(0.008)	(0.007)
	[0.861]	[0.751]	[0.473]	[0.609]	[0.234]	[0.646]
Covariates	No	Yes	No	Yes	No	Yes
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Control Mean	0.802	0.802	0.773	0.773	0.702	0.702
Control SD	0.140	0.140	0.237	0.237	0.170	0.170
Observations	1817	1817	1878	1878	1551	1551
R^2	0.031	0.166	0.069	0.279	0.188	0.416

Notes: First column for each outcome is design-based least squares estimator that includes fixed effects for randomization strata and batch. Second column adjusts for baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses and p-values in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

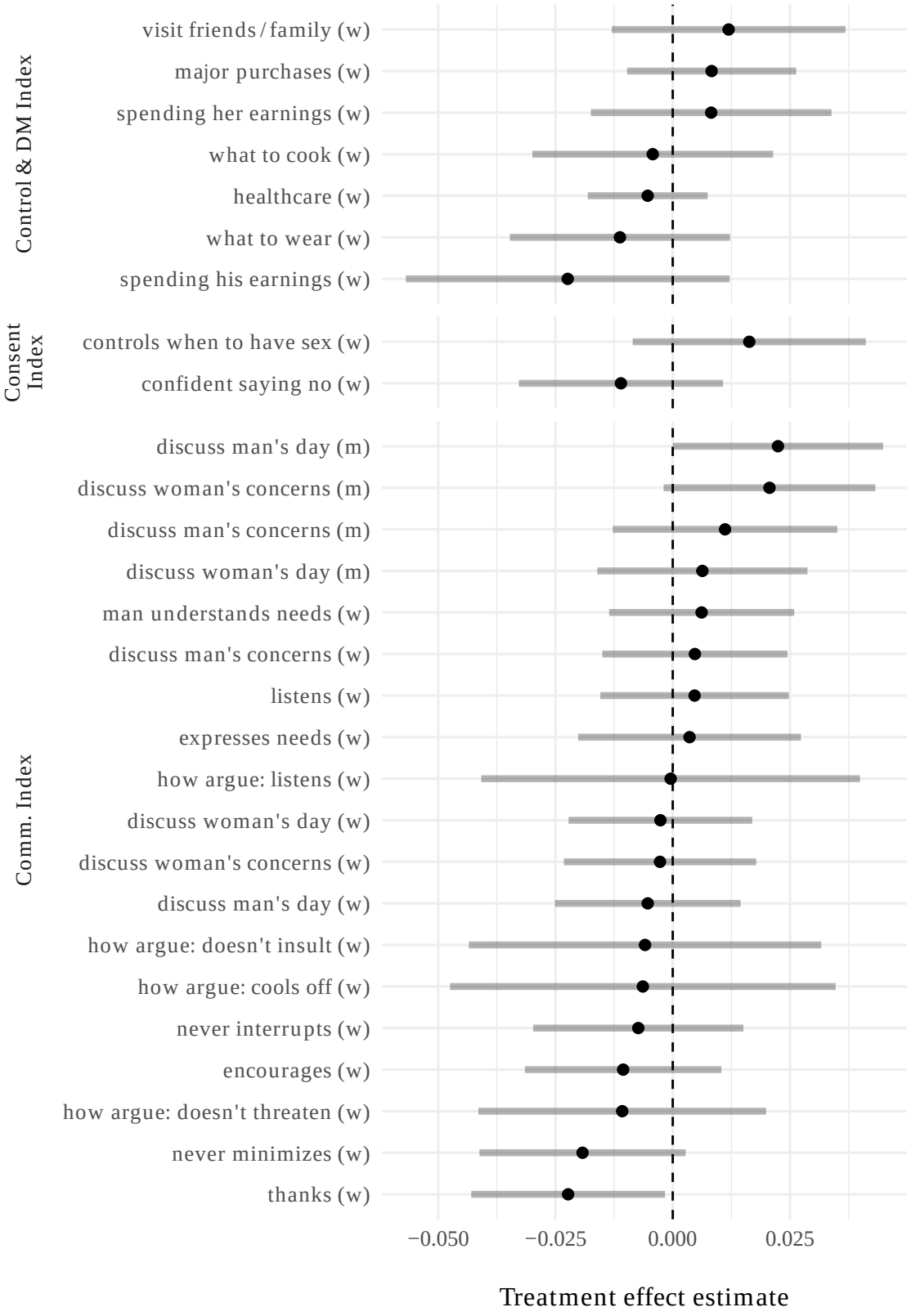

Figure A.3: Estimated effect of HEP on components of other primary outcome indices.

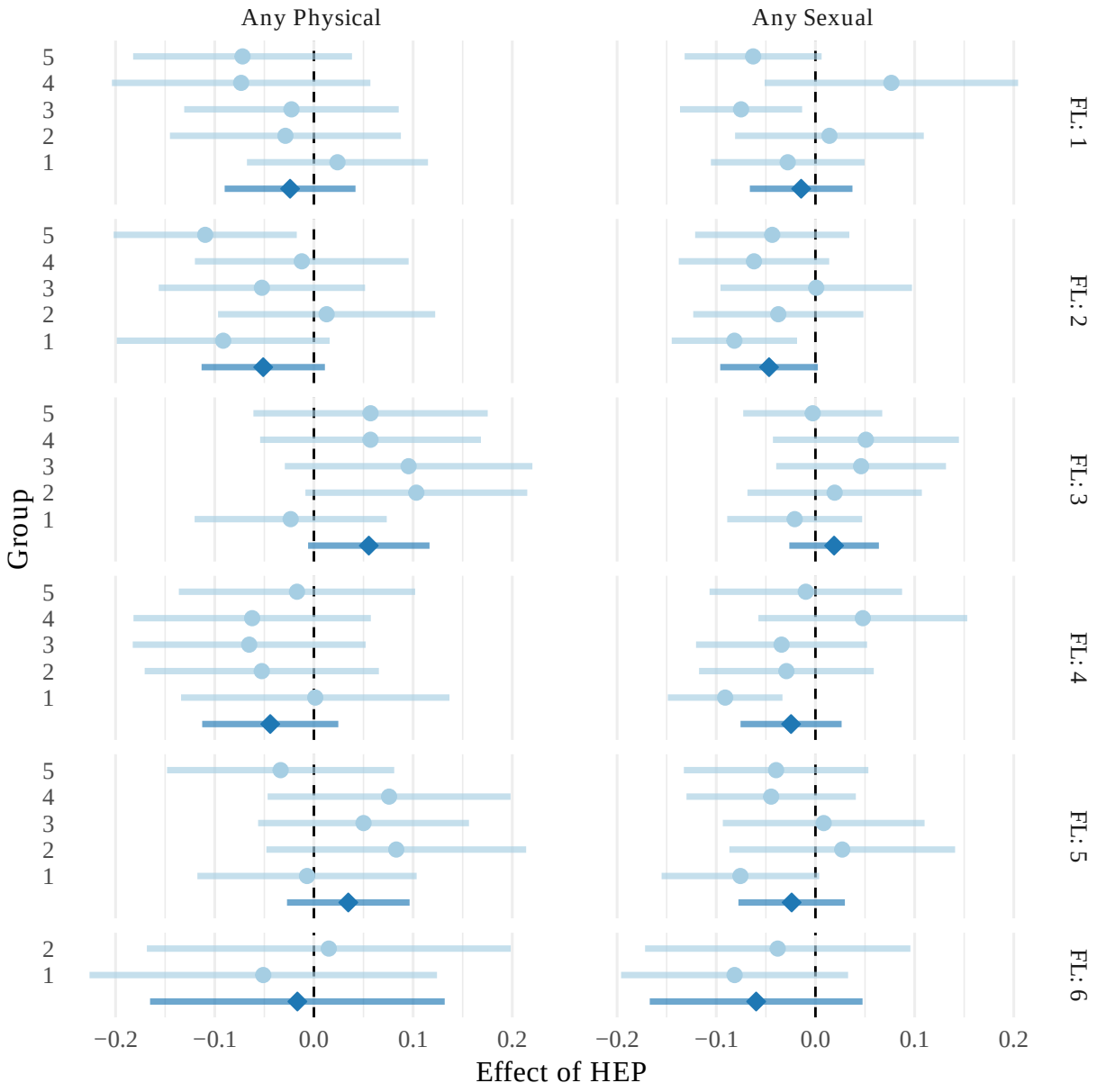

Figure A.4: Estimated effect of HEP on physical and sexual violence by group and facilitator (FL)

6.8 Effect of HEP on indices of severity of violence

Table A.12: ITT estimates of effects of HEP on continuous violence indices

	IPV Index		Physical Index		Sexual Index	
	(1)	(2)	(3)	(4)	(5)	(6)
treatment	−0.007	−0.016	0.004	−0.004	−0.008	−0.007
	(0.025)	(0.024)	(0.019)	(0.019)	(0.010)	(0.010)
	[0.779]	[0.521]	[0.846]	[0.825]	[0.427]	[0.490]
Covariates	No	Yes	No	Yes	No	Yes
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Control Mean	0.241	0.241	0.176	0.176	0.065	0.065
Control SD	0.564	0.564	0.434	0.434	0.212	0.212
Observations	1916	1916	1915	1915	1914	1914
R^2	0.145	0.287	0.147	0.261	0.067	0.146

Notes: First column for each outcome is design-based least squares estimator that includes fixed effects for randomization strata and batch. Second column adjusts for baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses and p-values in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.13: ITT estimates of effects of HEP on other continuous violence indices

	Control Index		Psych. Index		Cyber Index	
	(1)	(2)	(3)	(4)	(5)	(6)
treatment	−0.006	−0.014	0.003	−0.003	−0.012	−0.019
	(0.030)	(0.026)	(0.013)	(0.012)	(0.021)	(0.020)
	[0.854]	[0.605]	[0.793]	[0.832]	[0.576]	[0.353]
Covariates	No	Yes	No	Yes	No	Yes
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Control Mean	0.722	0.722	0.146	0.146	0.318	0.318
Control SD	0.709	0.709	0.298	0.298	0.483	0.483
Observations	1916	1916	1914	1914	1913	1913
R^2	0.146	0.405	0.140	0.339	0.094	0.235

Notes: First column for each outcome is design-based least squares estimator that includes fixed effects for randomization strata and batch. Second column adjusts for baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses and p-values in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

6.9 Effect of HEP on Secondary Outcomes

6.9.1 Alternative control and decision-making measures

Table A.14: ITT estimates of effects of HEP on control ladder

	Decision-making power (W)	
	(1)	(2)
treatment	0.004	0.002
	(0.008)	(0.008)
	[0.650]	[0.801]
Covariates	No	Yes
Fixed Effects	Yes	Yes
Control Mean	0.818	0.818
Control SD	0.191	0.191
Observations	1918	1918
R^2	0.033	0.135

Notes: First column for each outcome is design-based least squares estimator that includes fixed effects for randomization strata and batch. Second column adjusts for baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses and p-values in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

6.9.2 Alternative communication and conflict resolution measures

Table A.15: ITT estimates of effects of HEP on alternative communcation indices

	Comm. Index (W)		Emo. Reg. (W)	
	(1)	(2)	(3)	(4)
treatment	0.000	−0.009	−0.008	−0.006
	(0.008)	(0.007)	(0.012)	(0.012)
	[0.995]	[0.223]	[0.535]	[0.627]
Covariates	No	Yes	No	Yes
Fixed Effects	Yes	Yes	Yes	Yes
Control Mean	0.683	0.683	0.585	0.585
Control SD	0.187	0.187	0.278	0.278
Observations	1862	1862	1890	1890
R^2	0.173	0.434	0.120	0.223

Notes: First column for each outcome is design-based least squares estimator that includes fixed effects for randomization strata and batch. Second column adjusts for baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses and p-values in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.16: ITT estimates of effects of HEP on communication questions only

	Comm. (W)		Comm. (M)	
	(3)	(4)	(1)	(2)
treatment	0.003	−0.008	0.014	0.015
	(0.008)	(0.007)	(0.010)	(0.010)
	[0.727]	[0.258]	[0.140]	[0.119]
Covariates	No	Yes	No	Yes
Fixed Effects	Yes	Yes	Yes	Yes
Control Mean	0.719	0.719	0.751	0.751
Control SD	0.184	0.184	0.210	0.210
Observations	1887	1887	1663	1663
R^2	0.155	0.455	0.073	0.154

Notes: First column for each outcome is design-based least squares estimator that includes fixed effects for randomization strata and batch. Second column adjusts for baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses and p-values in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

6.9.3 Alternative consent measures

Table A.17: ITT estimates of effects of HEP on arguments about sex

	Arg. Infidelity (M)		Arg. Sex (M)	
	(1)	(2)	(3)	(4)
treatment	−0.009	−0.011	−0.006	−0.008
	(0.014)	(0.012)	(0.014)	(0.013)
	[0.501]	[0.394]	[0.674]	[0.562]
Covariates	No	Yes	No	Yes
Fixed Effects	Yes	Yes	Yes	Yes
Control Mean	0.225	0.225	0.271	0.271
Control SD	0.281	0.281	0.286	0.286
Observations	1662	1662	1653	1653
R^2	0.052	0.244	0.036	0.163

Notes: First column for each outcome is design-based least squares estimator that includes fixed effects for randomization strata and batch. Second column adjusts for baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses and p-values in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.18: ITT estimates of effects of HEP on attitudes about sex

	Provocative (M)		Always willing (M)	
	(1)	(2)	(3)	(4)
treatment	−0.005	−0.001	0.002	0.007
	(0.008)	(0.007)	(0.008)	(0.007)
	[0.494]	[0.854]	[0.775]	[0.293]
Covariates	No	Yes	No	Yes
Fixed Effects	Yes	Yes	Yes	Yes
Control Mean	0.579	0.579	0.562	0.562
Control SD	0.162	0.162	0.157	0.157
Observations	1656	1656	1658	1658
R^2	0.026	0.206	0.016	0.251

Notes: First column for each outcome is design-based least squares estimator that includes fixed effects for randomization strata and batch. Second column adjusts for baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses and p-values in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

6.9.4 Relationship quality

Table A.19: ITT estimates of effects of HEP on self-reported relationship satisfaction among men and women

	Satisfaction (M)		Satisfaction (W)	
	(1)	(2)	(3)	(4)
treatment	0.008	0.009	0.004	−0.004
	(0.009)	(0.009)	(0.009)	(0.008)
	[0.373]	[0.292]	[0.685]	[0.671]
Covariates	No	Yes	No	Yes
Fixed Effects	Yes	Yes	Yes	Yes
Control Mean	0.683	0.683	0.641	0.641
Control SD	0.191	0.191	0.208	0.208
Observations	1661	1661	1900	1900
R^2	0.050	0.207	0.116	0.335

Notes: First column for each outcome is design-based least squares estimator that includes fixed effects for randomization strata and batch. Second column adjusts for baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses and p-values in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.20: ITT estimates of effects of HEP on perception of partner's satisfaction with relationship among men and women

	Partner's Satisfaction (M)		Partner's Satisfaction (W)	
	(1)	(2)	(3)	(4)
treatment	0.012	0.012	0.003	−0.003
	(0.010)	(0.009)	(0.009)	(0.008)
	[0.206]	[0.194]	[0.754]	[0.736]
Covariates	No	Yes	No	Yes
Fixed Effects	Yes	Yes	Yes	Yes
Control Mean	0.641	0.641	0.657	0.657
Control SD	0.203	0.203	0.195	0.195
Observations	1616	1616	1873	1873
R^2	0.077	0.205	0.102	0.266

Notes: First column for each outcome is design-based least squares estimator that includes fixed effects for randomization strata and batch. Second column adjusts for baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses and p-values in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.21: ITT estimates of effects of HEP on conflict outcomes

	Broke up (W)		Arguments (M)	
	(1)	(2)	(3)	(4)
treatment	0.012	0.012	0.008	0.010NA
	(0.013)	(0.013)	(0.009)	
	[0.333]	[0.324]	[0.373]	
Covariates	No	Yes	No	Yes
Fixed Effects	Yes	Yes	Yes	Yes
Control Mean	0.083	0.083	0.336	0.336
Control SD	0.276	0.276	0.173	0.173
Observations	1997	1997	1426	1426
R^2	0.028	0.130	0.078	0.359

Notes: First column for each outcome is design-based least squares estimator that includes fixed effects for randomization strata and batch. Second column adjusts for baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses and p-values in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

6.9.5 Attitudes towards gender roles and violence

Table A.22: ITT estimates of effects of HEP on attitudes

	VAW attitudes (M)		Gender attitudes (W)	
	(1)	(2)	(3)	(4)
treatment	0.003	0.009*	0.002	−0.001
	(0.006)	(0.005)	(0.005)	(0.004)
	[0.670]	[0.079]	[0.630]	[0.847]
Covariates	No	Yes	No	Yes
Fixed Effects	Yes	Yes	Yes	Yes
Control Mean	0.519	0.519	0.563	0.563
Control SD	0.121	0.121	0.109	0.109
Observations	1606	1606	1811	1811
R^2	0.028	0.329	0.014	0.409

Notes: First column for each outcome is design-based least squares estimator that includes fixed effects for randomization strata and batch. Second column adjusts for baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses and p-values in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

6.9.6 Mental health

Table A.23: ITT estimates of effects of HEP on alcohol consumption and mental health

	Depression (M)		Man's alcohol use (W)	
	(1)	(2)	(3)	(4)
treatment	−0.007	−0.001	0.005	−0.004
	(0.008)	(0.007)	(0.044)	(0.036)
	[0.365]	[0.830]	[0.909]	[0.914]
Covariates	No	Yes	No	Yes
Fixed Effects	Yes	Yes	Yes	Yes
Control Mean	0.406	0.406	2.150	2.150
Control SD	0.155	0.155	0.990	0.990
Observations	1657	1657	1971	1971
R^2	0.038	0.297	0.058	0.395

Notes: First column for each outcome is design-based least squares estimator that includes fixed effects for randomization strata and batch. Second column adjusts for baseline covariates selected using double-post-selection lasso. Covariates are mean-centered and interacted with treatment. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses and p-values in square brackets.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.24: Effects on physical violence by group composition.

	Any Physical					
	(1)	(2)	(3)	(4)	(5)	(6)
HEP	−0.012	−0.039	0.021	0.030	0.028	0.145
	(0.036)	(0.052)	(0.052)	(0.070)	(0.097)	(0.123)
HEP × just. index (group)	0.013	0.050	−0.037			
	(0.067)	(0.099)	(0.092)			
HEP × % physical violence (group)				−0.047	−0.034	−0.210
				(0.092)	(0.128)	(0.158)
Sample	Full	No IPV	Any IPV	Full	No just.	Any just.
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Control Mean	0.232	0.187	0.271	0.232	0.181	0.278
Observations	1989	938	1051	1901	897	1004

Note: Columns 1, 2, and 3 are least squares regressions of endline physical violence on participation in the HEP program interacted with the group-level justification index. The justification index is an aggregate of 5 violence norms questions from our baseline survey which measures men's views about when violence is justified. Columns 4, 5, and 6 are least squares regressions of endline physical violence on the percentage of group members identified by their spouses as perpetrators of physical violence at baseline. Results from columns 1 and 4 were calculated on our complete sample ("Full"). Results from columns 2 and 5 were calculated on the subsample of women whose partners never justified any violence at baseline ("No just."). Results from columns 3 and 6 were calculated on the subsample of women whose partners justified violence for 1 or more of our 5 violence norms questions ("Any just."). All standard errors are robust (CR2) and clustered by WhatsApp group and all estimations include fixed effects for randomization strata and batch. Randomization-based p-values for the composition effect are calculated from 10,000 permutted assignments of the program and group cross-randomizations using a studentized test statistic with CR2 standard errors. All estimations include fixed effects for randomization strata and batch.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

6.10 HEP Model

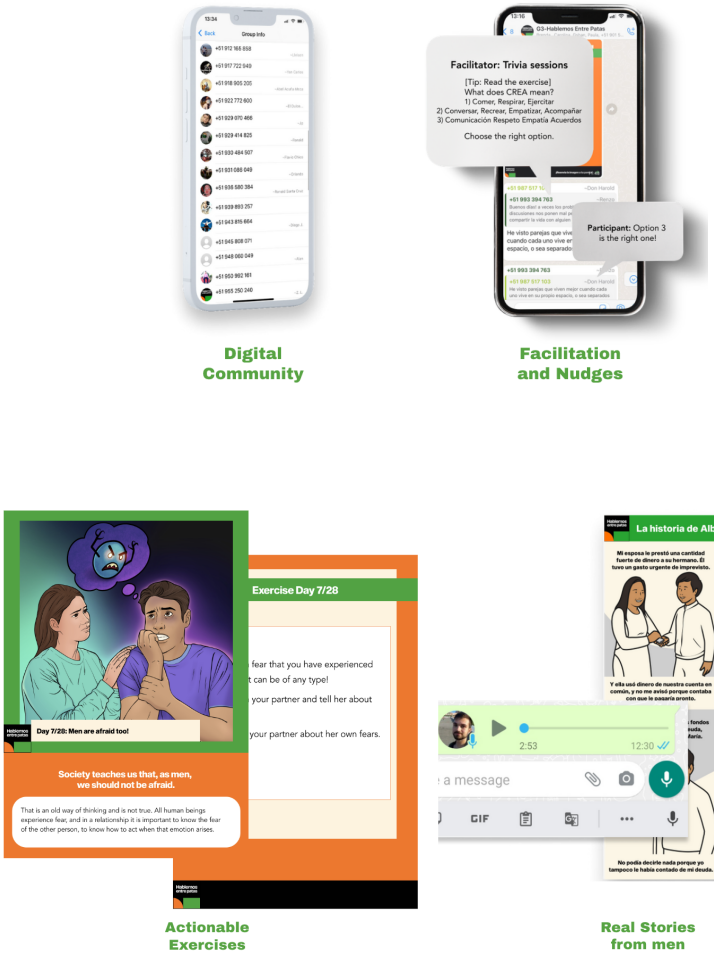

Figure A.5: Virtual support group model

Figure A.6: Example Facebook Ads

Figure A.7: HEP Website Frontpage

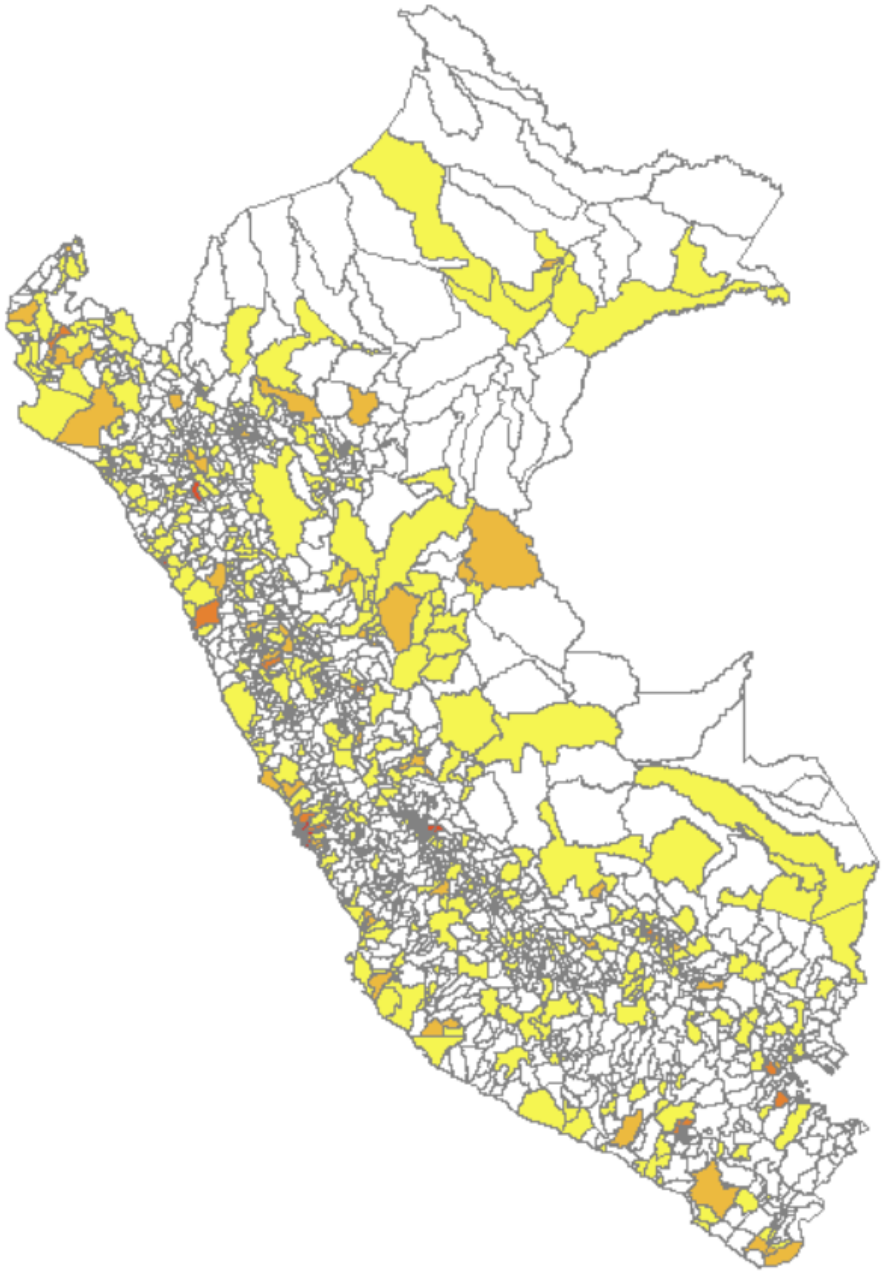

Figure A.8: Geographical distribution of the study sample

Note: Our total sample covers 555 districts, which represents 29.6% of total districts in Peru.

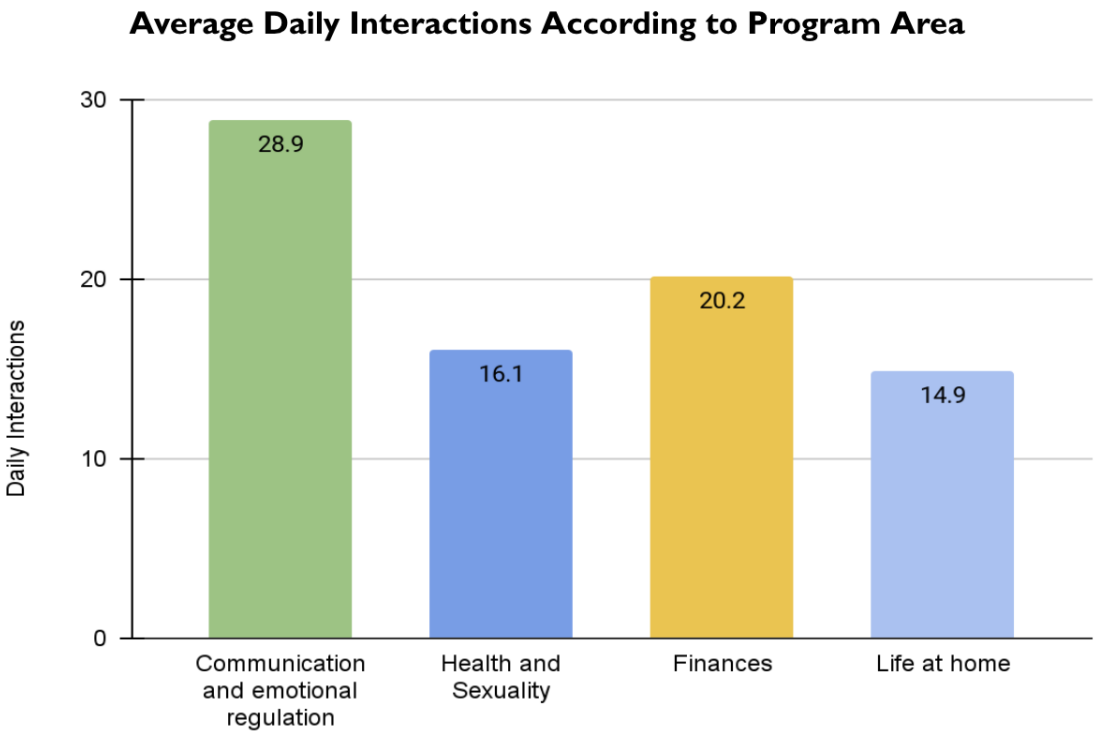

Figure A.9: Daily Interactions by Program Area

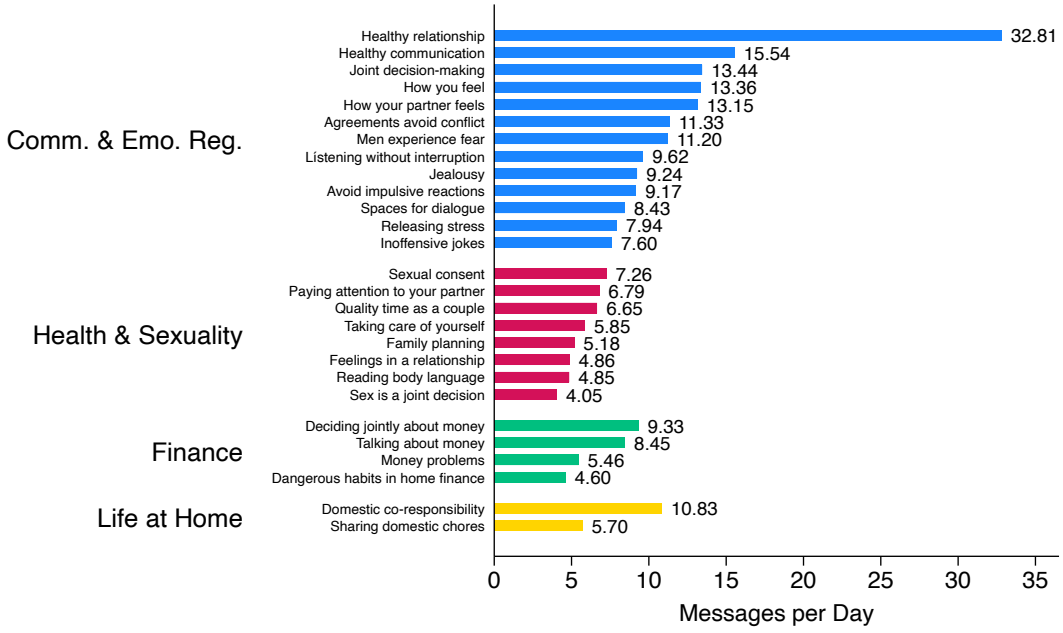

Figure A.10: Messages by Topic

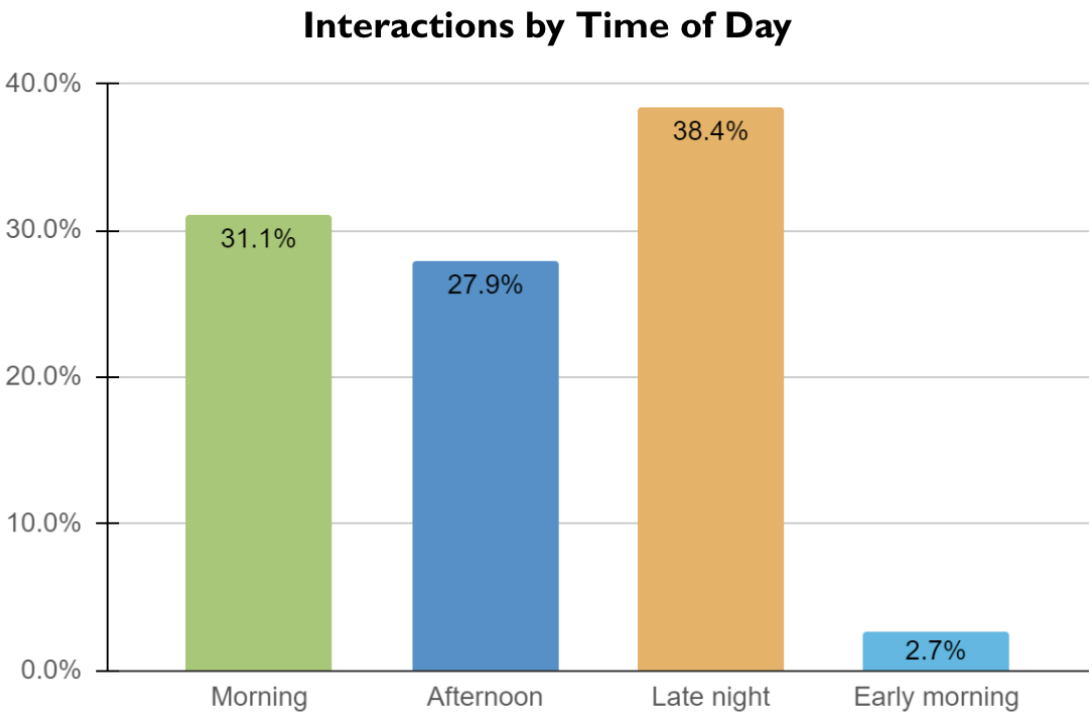

Figure A.11: Daily Interactions by Time

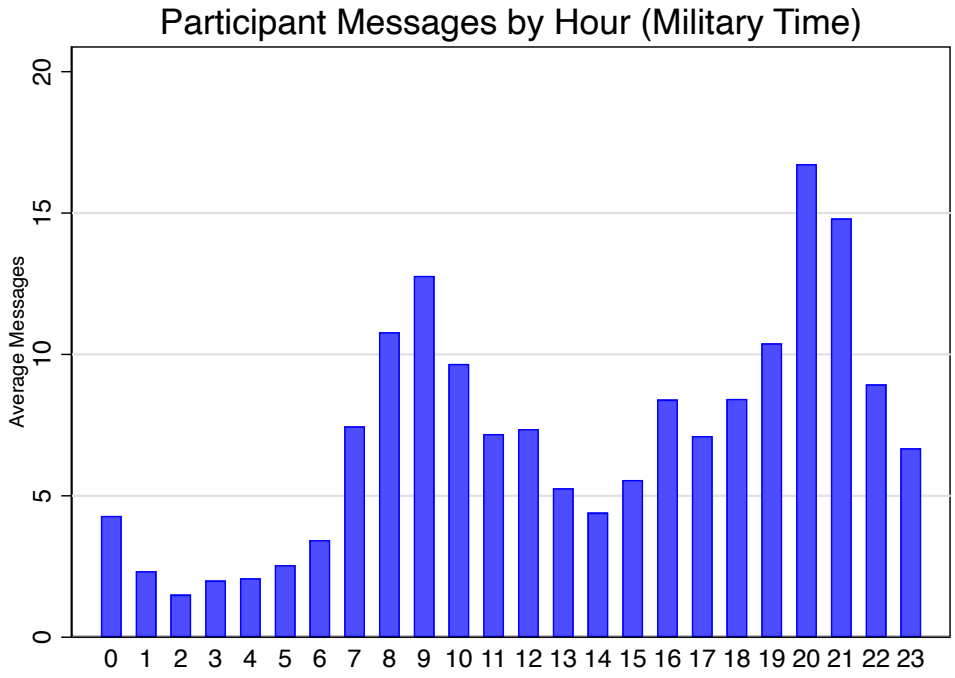

Figure A.12: Participant messages by hour

7 Compliance and per protocol effects

7.1 Exposure to program

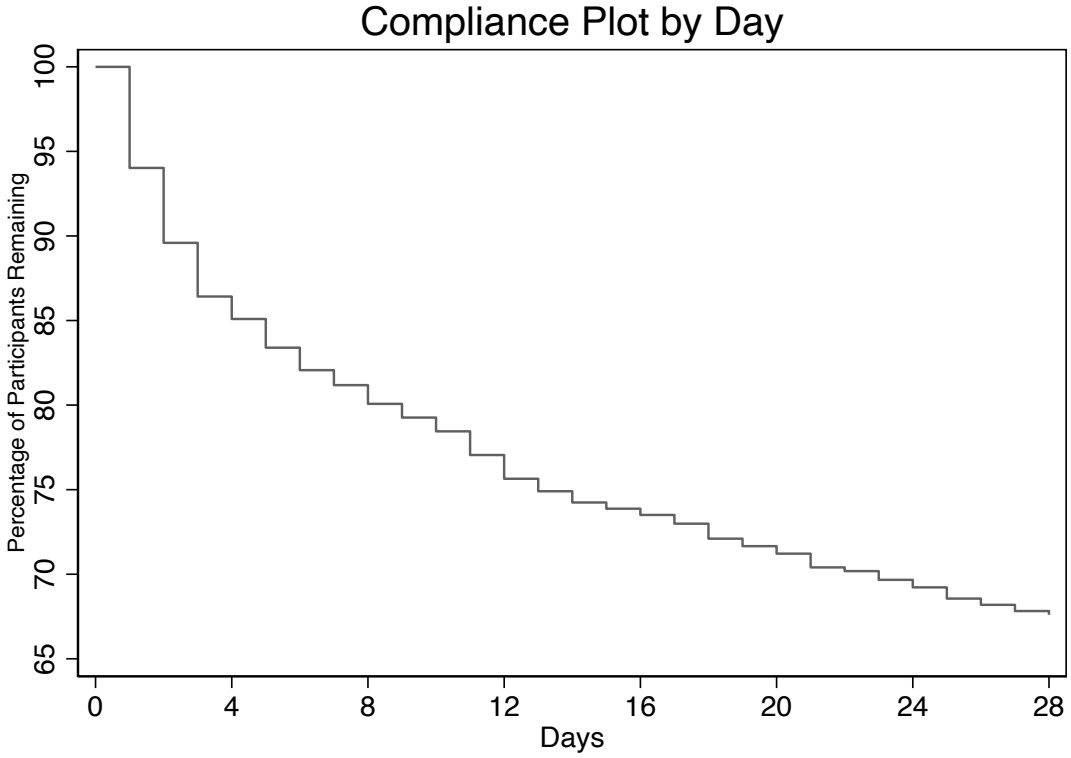

Figure A.13: Percentage of participants remaining in the chat by day

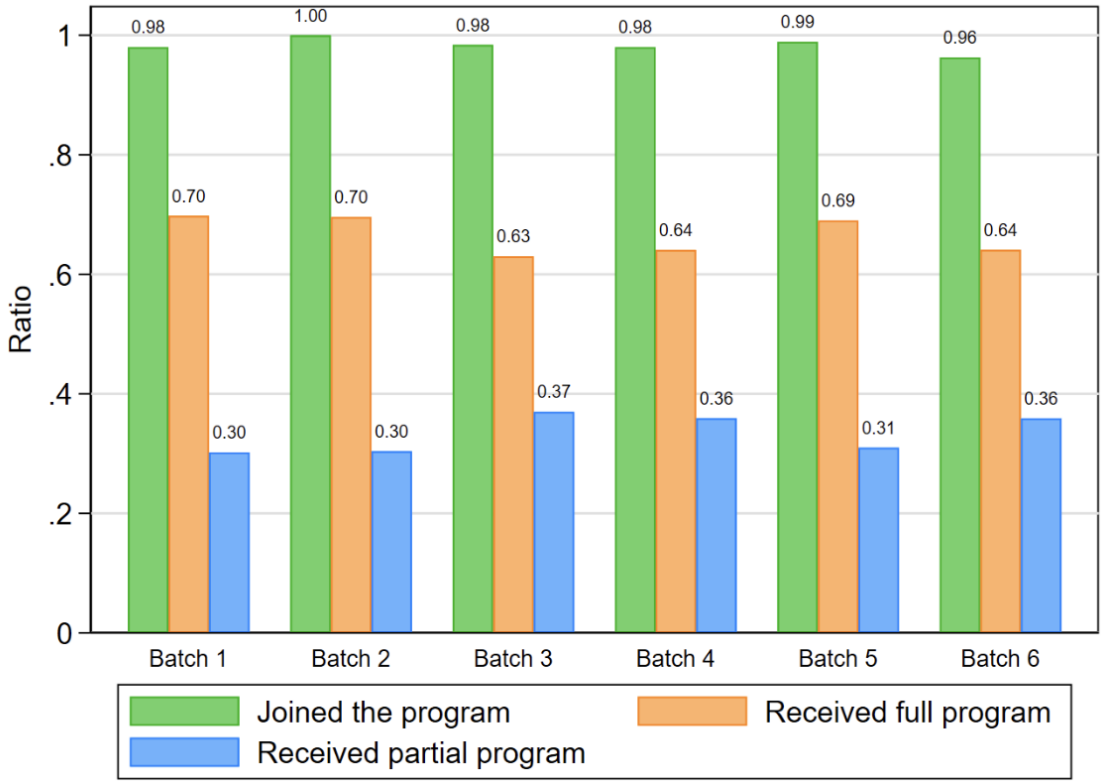

Figure A.14: Compliance and Program Completion

7.2 Experimenter demand

Table A.25: ITT estimates of effects of HEP on donating to charity (placebo)

	Donate to charity (W)		Donate to charity (M)	
	(1)	(2)	(3)	(4)
treatment	0.004	0.008	0.028	0.031
	(0.020)	(0.020)	(0.023)	(0.023)
	[0.857]	[0.697]	[0.240]	[0.181]
Covariates	No	Yes	No	Yes
Fixed Effects	Yes	Yes	Yes	Yes
Control Mean	1.260	1.260	1.272	1.272
Control SD	0.446	0.446	0.466	0.466
Observations	1969	1969	1642	1642
R^2	0.011	0.046	0.007	0.050

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: First column for each outcome is design-based estimator that includes fixed effects for randomization strata and batch. Second column is regression specification that adjusts for baseline covariates using mean-centered, interacted approach of Lin estimator. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses and p-values in square brackets.

Table A.26: ITT estimates of effects of HEP on Marlowe Social Desirability Index

	Social Desirability (W)		Social Desirability (M)	
	(1)	(2)	(3)	(4)
treatment	−0.001	0.000	−0.001	−0.002
	(0.007)	(0.007)	(0.008)	(0.008)
	[0.917]	[0.961]	[0.944]	[0.795]
Covariates	No	Yes	No	Yes
Fixed Effects	Yes	Yes	Yes	Yes
Control Mean	0.719	0.719	0.737	0.737
Control SD	0.162	0.162	0.170	0.170
Observations	1842	1842	1614	1614
R^2	0.047	0.074	0.018	0.127

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: First column for each outcome is design-based estimator that includes fixed effects for randomization strata and batch. Second column is regression specification that adjusts for baseline covariates using mean-centered, interacted approach of Lin estimator. Heteroscedasticity-consistent robust standard errors (HC2) for all specifications are shown in parentheses and p-values in square brackets.