Pre-Analysis Plan (PAP) for:

A Pair-Matched Randomized Evaluation of Faith-Based Couples Counselling in Uganda

THIS PRE-ANALYSIS PLAN WAS FIRST REGISTERED ON MARCH 11, 2019

FIRST DAY OF MIDLINE DATA COLLECTION PLANNED FOR MARCH 12, 2019

Jeannie Annan
Christopher Boyer
Jasper Cooper
Lori Heise
Betsy Levy Paluck

Table of Contents

1. Executive summary for reviewers

We include here a high-level overview of the study and analysis plan that we hope will help reviewers in quickly assessing the overall consistency of the final paper with the planned analyses, while pointing them to areas where greater detail is provided.

1. Intervention and randomization

The Becoming One program is a skills-based curriculum designed for faith leaders to use with couples in their congregation seeking counseling prior to or following a wedding. The curriculum uses a biblical framework to communicate about practical skills for improving communication, emotional regulation, shared control over financial resources, financial planning, and sexual consent and pleasure. Faith leaders lead the participatory sessions with groups of couples, asking both members of the couples to attend together. The 12 total sessions are intended to be delivered once a week with each session lasting approximately 90 min. At the heart of the Becoming One intervention is the idea that individuals and couples will change their behavior when they are already self-motivated to do so, and when the source of advice or counsel is a person of social, religious, or political significance for them.

This study measures the effect of the Becoming One (B1) program through a pair-matched randomized control trial. There are three waves of measurement: a pre-randomization baseline, a midline survey approximately five weeks after the "intervention" group has completed the couples counselling program, and an endline survey approximately one year after the "intervention" group has completed the program, but before the "comparison" group has gone through couples' counselling.

The random assignment works by forming couples into pairs within congregations (groups organized by the faith leaders) by matching on their baseline levels of physical violence, and then randomizing one couple in each pair to the intervention group, and one couple to the comparison group. The baseline levels of physical violence include an average of six binary indicators (standardized to have a mean of 0 and standard deviation of 1) for whether the woman in the couple reports having been pushed, slapped, arm-twisted / hair-pulled, punched, kicked, or choked over the preceding 12 months.

2. Main analyses

The main analyses of the experiment take place at the couple-level. We will construct non-parametric *p*-values calculated using randomization inference. Using simulation, we will report the alphas (type 1 error risks) one would need to apply at the test-level in order to reject at least one test in the family of main analyses only 5% and 10% of the time, under the global null of no effect on any outcome for any unit.

Our main estimator is a covariate-adjusted estimator below, which, in addition to conditioning on an indicator for the treatment assignment and block-level fixed effects, adjusts for predictive covariates using a mean-centering and interaction approach described in Lin (2013). For each outcome, we use a cross-validated Lasso estimator to select only those covariates that are predictive of the outcome, irrespective of whether those covariates are imbalanced. We justify this approach through a simulation study.

We expect to encounter two-sided non-compliance in this study. For our main analyses, our estimand is an intent-to-treat (ITT) effect.

To reduce the number of tests we are running, our main analyses focus on four outcomes that combine many theoretically-related measures:

1. Intimate Partner Violence. A binary indicator that is 1 if the female partner has experienced any of ten types of physical or sexual violence over the five months preceding the midline survey, and 0 otherwise. We expect that the ITT effect is negative and so will conduct a one-sided (lower-tailed) test.

2. Control and Decision-Making. An index of several groups of items, each of which is coded to vary between 0 and 1. The groups include decision-making about household finances and purchasing decisions, financial control, and social control. When at 0, this implies the female partner experiences all forms of financial and social control, and both partners perceive her to be excluded from all decision-making. When it is at 1, this implies the female partner does not report experiencing any forms of financial or social control, and is involved in all aspects of decision-making. We expect that the ITT effect is positive, and will conduct an upper-tailed test of the null hypothesis.

3. Sexual Consent and Autonomy. An index of items, mostly answered by the female partner but one answered by the male partner, about whether the male partner takes a non-coercive strategy when his partner refuses sex, and how the female partner feels about her control over decisions about and initiation of sex. The responses are coded to vary between 0 and 1. When the index is at 0, this implies that the female partner reports sexual coercion across all of the domains, and that the male partner reports coercive behavior. When it is at 1, this implies the absence of coercion.

4. Communication and Conflict Management Techniques. An index of items addressing a broad range of domains of communication, answered by both female and male partners. The domains include communication about sex, about daily events and companionship, and techniques for de-escalation when arguing about difficult topics. When the index is at 0, this implies that the partners agree that they have had no supportive or open communication or techniques for peaceful or de-escalating communication across many domains. When it is at 1, this implies the partners agree that they communicate with supportive or open communication or techniques for peaceful or de-escalating communication across all the domains.

We will base our main confirmatory inferences on these four outcomes, which also constitute the family of tests for which we will construct multiple comparison-adjusted alphas as above.

In order to explain why these outcomes move, we will report and describe treatment effects graphically by plotting point estimates on the underlying items. We also have a range of secondary analyses that target either variables that we expect mediate the effect of B1 on these four outcomes, or other outcomes that are not of primary interest. All of these will be reported as analyzed, possibly in an appendix. We intend to conduct analyses while blinded to the treatment status, in a verifiable procedure described at the end of the PAP. We will note departures from the analysis plan, labelling tests as pre-registered and / or conducted while blind to the true assignment, or not.

3. Overview of PAP sections

The rest of the PAP reads as follows: we provide a project background and more extensive intervention description, followed by a description of the study design and randomization procedure. We provide the details of our estimation strategy, followed by the description of and justification for our 4 primary outcomes, calculated as indices. We also describe our secondary outcomes, some of which we hypothesize to be possible mechanisms of the program's impact, and others as possible outcomes, of B1. We discuss extensions of and subgroups added to our main analyses, and finally describe our robustness checks and strategies for challenges to internal validity.

2. Project Background

Intimate partner violence (IPV) is thought to touch the lives of 30% of women globally and over 44% of married women living in rural Uganda. Efforts to reduce rates of abuse often seek to effect radical normative change through intervention by outside actors. Despite concerns that outside actors often lack legitimacy with local communities, there is comparatively little research on the capacity to reduce violence from within by working with local authority figures.

In 2016, the International Rescue Committee (IRC) Research and Development team began prototyping intervention ideas for the Becoming One (B1) intervention in Liberia. The design team started off with extensive field research to better understand perceptions of different factors relating to intimate partner violence (IPV) as well as existing strategies that communities employ to address IPV. This work identified pastors and other faith leaders as particularly influential sources of authority over a couple's behavior.

Based on feedback from pastors and the World Vision team, IRC created an initial set of couples counselling prototypes, which could be used in pre-marital counselling by faith leaders. After many iterations based on feedback from both faith leaders and couples, the team conducted a small 8 week pilot in Gulu, Uganda with 40 faith leaders and approximately 900 couples in September and October of 2017. Observations and interviews suggested that framing relationship skills as based in "biblical principles," seemed to increase the acceptance and practice of new skills. Additionally, providing faith leaders with new counseling methods through instructional videos was an efficient way for many of the faith leaders to digest and then replicate novel approaches.

The purpose of this research protocol is to provide a rigorous assessment of the potential of the Becoming One program to improve couples' communication and financial transparency, sexual agency for women and other markers of relationship quality.

Through our evaluation of the effectiveness of Becoming One, we also seek to contribute to understanding the causes of intimate partner violence and intrahousehold conflict more generally.

4. Theory of change

At the heart of the Becoming One intervention is the idea that individuals and couples will change their behavior when 1) they are already self-motivated to do so, and 2) when the source of advice or counsel is a person of social, religious, or political significance for them.

Put in the negative, Becoming One does *not* attempt to 1) create a motivation to change one's relationship within couples who have not already expressed some amount of interest in change, or 2) influence couples or individuals using a source of authority external to their pre-existing social networks or community.

What theories justify this approach? Behavioral science theories suggest that behavior change is more likely when interventions seek to allow pre-existing motivations to be expressed, as opposed to when they seek to create a new motivation from scratch (e.g., Prentice & Miller, 2012). Interventions might help pre-existing motivations to be expressed by making it *easier* to achieve goals (through reminders, forms of social or logistical assistance), or by *clarifying* ways in which people can achieve goals (through simplified instructions, using culturally relevant explanations or symbols). Becoming One is based on these theoretical principles because it recruits couples who are already interested in religious couples counseling as a first step to getting married (individuals and couples have a *pre-existing motivation* to examine their relationship), and because Becoming One's instruction is delivered in the language of religious text (which *clarifies* ways for couples to develop skills in their relationship using a familiar and revered system of belief) and in couples counseling classes with their peers (which makes it *easier* for couples to change with peer examples and support).

A second set of social and behavioral theories that justify the Becoming One approach posit that patterns of behavior for individuals and communities are influenced to a relatively large degree by individuals with cultural, religious, social, or political significance or authority (e.g., Rogers & Catano, 1965; Watts & Dodds, 2007). These theories identify "key" influencers as important for activating individual changes in behavior, and under some circumstances, triggering critical mass changes in behavior within communities (e.g., Centola, 2018; Paluck, Shepherd & Aronow, 2016). Becoming One is based on this broad theoretical assertion, because it operates through the influence of faith leaders, who have high status within the Ugandan communities under study, and who are regularly consulted by community members on social and religious matters, including conflicts within households.

We use these basic theories to justify the design of Becoming One, and to predict that participation in the couples counseling will change individual behavior and interactive couple behavior in **four primary ways.** Becoming One should:

Change couple dynamics through the Becoming One curriculum instruction and example of faith leaders, specifically:

1) Shift couple dynamics that represent *zero-sum power struggles* over finances, movement and general decision-making, resulting in a more equal balance of power in the couple,
2) Increase the use of skills within couples that *do not relate to zero-sum power struggles*, involving communication and conflict management techniques.

Reduce violence, through the Becoming One curriculum instruction and example of faith leaders, specifically:

3) reduce intimate partner violence (physical and sexual),
4) increase sexual autonomy and affirmative consent to sex within couples.

On a global level, we predict that these changes will occur because the Becoming One curriculum is delivered by a religious local influencer, the messages encouraging these behaviors will be framed and justified by religious texts, and they will be delivered within the context of peer groups. For this reason, we will collect data (outlined below) to measure the fidelity of program delivery (via program monitors) and to measure treatment participants' reactions to religious ideas, and perceptions of the peers in their Becoming One group (via a midline and baseline survey). In the survey, we plan to measure effects on a number of potential mediators of the outcomes of Becoming one, described below, in addition to these global predictors of change. They include individuals' gender attitudes and perceived norms of the community, men's levels of emotional regulation, greater gender-equitable interpretation of religious scripts, increased financial transparency in the couple, couples' convergence in trust and communication, and individuals' definitions of violence.

We also predict a number of secondary outcomes from this program related to the primary outcomes, detailed below. These include the degree of convergence in a couple regarding their beliefs about and perceptions of the relationship, a shifted distribution of domestic labor, lowered levels of emotional violence and of incidence of arguments, greater communication about sexual practices, and greater levels of time spent together; on an individual level, shifted perceptions of community norms regarding intimate partner violence, lowered levels of individual depression, and higher levels of confidence in expressing opinions publicly and in perceived social support. As a secondary outcome, we also plan to test separately the effect of Becoming One on different kinds of intimate partner violence, including emotional, physical, and sexual violence.

5. Actors and roles

The Principal Investigator (PI) team is composed of Jeannie Annan, Christopher Boyer, Jasper Cooper, Lori Heise, and Betsy Levy Paluck. The PI team works in close cooperation with Innovations for Poverty Action Uganda, who implements the baseline, midline, and endline survey measurements and monitoring of program implementation. IRC and World Vision are the primary actors in charge of developing and implementing the Becoming One program.

6. Ethics

In general, we expect the possible risks to study participants and staff to be minimal and where possible we have taken steps to minimize them. The research protocol, survey instruments, and consent forms for the baseline and midline were reviewed and approved by the IRB at IPA (protocol #14916), by the IRB in Uganda (MUREC protocol #REC REF 0508-2018), and by the Ugandan National Council for Science and Technology (protocol SS 4782). We describe below some of the steps taken to protect research participants' rights, as well as the ethical considerations involved in the measurement and randomization.

i. Protection of subjects' rights and minimization of risk

One possible risk to participants is if personally identifiable information were to be received by the wrong individuals. Given that (i) data will be electronically collected and very few individuals in the enumeration areas have access to technology to read digital data, (ii) all data will be password protected and encrypted and removed from phones on a daily basis, this is highly unlikely.

We also recognize that the survey includes several sensitive topics, including emotional, physical, and sexual abuse, that carry with them the following risks:

- The risk that the respondent is re-traumatized
- The risk of retaliation if disclosure of violence is made public
- The risk that the enumerator is traumatized and/or harassessed
- A legal risk if local laws regarding reporting of these incidents is not strictly followed

We took special precautions to ensure that the risk for both the respondent and the enumerator is minimal. Accordingly, as suggested in the IPA Guidelines for Safe and Ethical Conduct of Violence Research, we plan to undertake the following risk mitigation strategies in the midline that were also taken in the baseline:

1. Ensure participant safety: The interview will be done in a private area (allowing only children below 2 years of age to be in the same area as the respondent). Enumerators will be trained to change questions to non-sensitive subjects if the survey is interrupted and/or they notice someone else is listening. Moreover, no one else in the household or community will be informed that that the research includes questions on violence. Only women will be asked questions about victimization at the hands of a partner, so that male partners will not be aware that their spouses may have divulged this information (see W ONLY flag in survey document).
2. Minimize participant distress: due to the sensitive subject matter, it is possible that the interview can provoke a powerful emotional response among participants. The enumerators will be trained to be sensitive to respondent's experiences and recognize signs of distress and take appropriate steps to support the respondent and/or to terminate the interview. Moreover, we will gender-match enumerator and respondent.
3. Provide referrals for care and support: The research team has an ethical obligation to provide information to participants regardless of whether they report experiencing violence. All respondents will receive the contact information for appropriate services in their area. Moreover, enumerators will be trained to help understand their role in relation to respondents who report experiencing violence.
4. Providing referrals for emergency situations: The enumerators will be provided with safety training by a trained psychologist to respond to participants who express thoughts of suicidal or homicidal ideation or when the respondent appears to be in imminent danger. They will be trained to escalate the issue to Research Associate or Field Manager, who will contact legal aid providers, shelter homes for survivors, health services, local police officers, community officers and probation officers in the study districts. The different referral organizations will be alerted before the start of data collection and will be prepared to support respondents undergoing extreme

trauma. The trained psychologist will continue on the project as a consultant and will provide support in making safety plans for high-risk respondents.

5. Protect field staff: The research team will prepare a plan in advance to quickly extricate field teams from volatile situations. Additionally, during the enumerator training, opportunity will be given for enumerators to come to terms with their own experiences with abuse. Debriefs focusing on emotional well-being of the field staff will be held regularly.

ii. Mandatory reporting obligations and referral processes

In addition, to minimize legal risks we obtained a recent opinion about mandatory reporting requirements in Uganda. They referenced the Domestic Violence Act of 2010, the Children Act Cap 59, the Children (Amendment) Act of 2016, and the Prevention of Prohibition of Torture Act of 2012. Their conclusion is restated below:

> As discussed above the mandatory legal obligation to report violence is optional for adults however for children there is mandatory legal obligation report any infringement to their rights by persons such as medical practitioner; social worker; teacher; local councilors.

Given that only adults above the age of 18 are eligible to participate in our study, we believe that there is no legal obligation to report instances of violence discovered in the process of this survey to the authorities. Furthermore, it is critical that the decision about whether violence is reported to the authorities is a decision made by the woman herself, rather than one that is made for her.

To ensure the entire field team is informed of their responsibilities and trained well to execute the mitigation strategies above, we conducted a training, led by the Research Associate and Field Manager, and based upon the aforementioned IPA Guidelines as well as the Ethical and Safety Recommendations for Research on Domestic Violence Against Women from the World Health Organization. The training included a basic introduction to domestic violence issues as well as an overall orientation to the concepts of gender, and gender discrimination and inequality. Opportunity was provided to enumerators to acknowledge privately their own experiences with abuse. During the training, it was made clear that the subject of violence can always be openly discussed, and enumerators can withdraw from the project without prejudice. Lastly, enumerators are also led to understand their role in relation to respondents who report experiencing violence: i.e., they should be open to assisting the respondent if asked, but they should not tell her/him what to do or to take on the personal burden of trying to “save her/him.” Enumerators were advised not to take on a role as counsellor and any counselling activity that may be offered in the context of the study should be entirely separate from the data collection. They will have referral cards that they will provide to all participants with information about where they can get confidential professional psychosocial support services. As noted above, they have also been instructed on what to do in extreme circumstances if they feel that woman may be in imminent danger or suicidal.

The contacts for support services that will be offered to both men and women who are part of the study will be:

- The community development officer - At the Sub County level, the assistant community development officers are the frontline staff who offer guidance and counseling to the victims. They also do referrals to the nearest health centers for medical examinations and treatment as and when required or to the Police if the case is more of a legal issue.
- MIFUMI hotline/Communication for Development Foundation Uganda hotline– these are toll-free lines where counseling, advice and referral services are given to callers experiencing Domestic Violence. Through the helpline, survivors receive psychosocial counseling, GBV related information, referrals- survivors of who call in are referred and directed to the appropriate institutions, agencies or organizations. The helpline enables callers experiencing Domestic Violence access free convenient and confidential advice on GBV services
- Contacts of local health service providers

Overall, we believe that these precautions should ensure that the risks to the participants are minimal, despite the sensitive subject matter of our interviews. Indeed, there is even some evidence to suggest that respondents respond well to surveys on violence and may even derive some benefit.

Although we hope that the findings of the study will contribute to evidence-based policy formulation on the impact of counselling on the described key outcomes, there will be no immediate benefits to the respondents. However, each respondent will be given a bar of soap as a token of appreciation for each survey. Members of the PI team have used this compensation in similar areas of Uganda in previous studies and found it to be appreciated and appropriate.

iii. Ethics of randomization

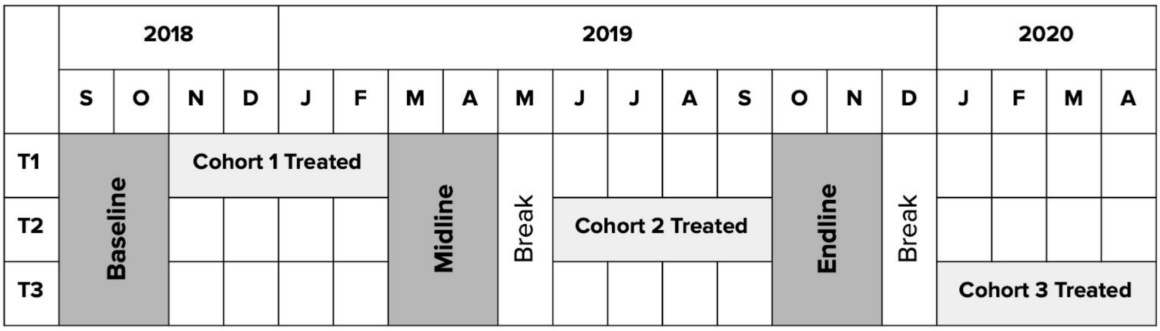
The purpose of this study is to understand whether and how the B1 program affects couple dynamics, but ultimately we operate on the presumption that the program will be beneficial to participants. Unlike other contexts in which the good being randomized is scarce, faith leaders in this context could eventually put all couples through the program. As such, we do not see it as ethical to have a "pure" control that would be kept from ever going through B1.

To address this equity concern we use a type of wait-list control design, in which every research participant is eventually entered into the program, but in successive phases, such that we are able to estimate the effects of the program on initial cohorts before subsequent cohorts have gone through it. The design is not a stepped-wedge in the sense that we do not exploit the wedge structure through the analysis of more than two cohorts.

There is some concern that even keeping couples from being treated for longer is unethical due to the possibility that – if the program reduces arguments, for example – research participants could enter the program earlier and thereby avoid having arguments the intervention would prevent. However, the advantage of this structure is that we are also able to assess any adverse effects of the program. If we do find compelling evidence of such effects, we will be able to halt the implementation of the program among subsequent cohorts. Thus, the phased structure of the rollout appears ethically defensible to us as a way of ensuring equitable distribution of a potentially beneficial resource, while still being able to assess any unforeseen risks it could pose.

7. Timeline

The first version of this document was produced and registered following baseline survey measurement and randomization, but prior to the collection of any post-randomization outcome data, which will happen through the midline survey.



	2018				2019												2020			
	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M	A
T1	Baseline		Cohort 1 Treated				Midline		Break					Endline		Break				
T2										Cohort 2 Treated										
T3																	Cohort 3 Treated			

3. Intervention description

8. Training of faith leaders

The faith leaders recruited to deliver the Becoming One program completed one or two video-based trainings that prepared them to lead couples in sessions designed to develop practical skills for reducing IPV. During the training they received a faith leader guide, couples guides, an attendance and planning log, marketing materials, “I love my wife/husband” bracelets, and a tablet with copies of the videos and additional instructional materials in a branded bag to help them jump start and administer the program. They were also connected with other faith leaders who would be leading Becoming One sessions and were invited to join a WhatsApp group where they could share insights and learn from one another as they implemented the program.

9. Becoming One syllabus

The Becoming One program includes a skills-based curriculum which is meant to provide couples with practical skills for improving communication, emotional regulation, shared control over financial resources, financial planning, and sexual consent and pleasure. Sessions are meant to be participatory with the couples often asked to complete interactive activities together. Each couple is also given a guide book with additional home practice activities that they are asked to complete and report back on during subsequent sessions. The content of the program is infused with gender-equitable religious and biblical materials to inspire and reinforce concepts while also providing a relatable social frame though which to integrate them. The design materials also showcase aspirational couple identities through vivid photos of Ugandan couples.

10. Number of sessions

The Becoming One program is comprised of 12 sessions delivered once a week with each session lasting approximately 90 min. The groups were responsible for selecting a time and place that was convenient for them to meet. The sessions are organized into three thematic modules: communication, finance, and sex. These are preceded by an introductory session and the program concludes with a final ceremony in which the couples are recognized. The exact sequence is described below:

1. Introduction
2. Communication - Session 1
3. Communication - Session 2
4. Communication - Session 3
5. Communication - Session 4
6. Finance - Session 1
7. Finance - Session 2
8. Finance - Session 3
9. Sex - Session 1
10. Sex - Session 2
11. Sex - Session 3

12. Final ceremony

In practice sessions were sometimes rescheduled due to conflicting responsibilities of faith leaders, community event, or lack of availability of meeting location, but faith leaders were strongly encouraged to complete all 12 sessions. All groups also stopped meeting for two weeks during the holidays.

4. Design and Randomization

This study measures the effect of the B1 program through a pair-matched randomized control trial.

There are three waves of measurement: a pre-randomization baseline, a midline survey approximately five weeks after the "intervention" group has completed the couples counselling program, and an endline survey approximately one year after the "intervention" group has completed the program, but before the "comparison" group has gone through couples' counselling.

Every couple in the study is invited to participate in B1. Intervention and comparison groups are formed from the three "cohorts" that go through the program. Specifically, those randomly assigned to the first cohort of B1 go through the program before the midline and endline surveys, whereas those assigned to the third cohort of B1 go through the program after the midline and endline. The design is not a typical "stepped-wedge," insofar as the outcomes of the second cohort, who go through B1 some time after the midline but before the endline, are not directly measured as part of the main experimental comparison. Rather, the creation of a second cohort makes feasible the one-year delay between the baseline survey and the eventual entry of the third cohort into B1. Details of this design are discussed below.

11. Recruitment of faith leaders

Having designed the intervention components (B1 training and curriculum), the first step of the evaluation is to recruit faith leaders who will deliver B1. Project partners World Vision and IRC worked to identify 145 faith leaders from different congregations who would be able to deliver the B1 programming.

In mid 2018, partners at World Vision and the IRC identified 145 faith leaders in the districts of Kagadi, Kakumiro and Kamwenge in Western Uganda who were interested in running the Becoming One program and met program criteria (see below). The faith leaders were invited to a training in which they were given an overview of the Becoming One program. During this training, they were given instructions on how to recruit couples for the program from within their community. Each Faith leader was instructed to identify 15 couples in his or her community whom they think would be suitable for Becoming One. Specifically, they were instructed by IRC and World Vision to invite fifteen couples from their congregation who have been together for at least one year, are in monogamous relationships and in the age bracket from 18-65. They were not instructed to target couples currently experiencing marital problems, or to engage in any research related recruitment.

12. Sample selection

The Faith Leader-led recruitment phase ran from July through August of 2018 and resulted in an initial list of 2,561 couples. A follow up survey exercise, conducted by IPA, confirmed the eligibility of the recruited couples and formally assessed their interest in participating in both the Becoming One program and the research study. Couples were also asked whether both members of the couple intended to reside in the area for at least 1 year.

Due to resource constraints we were only able to interview 1,960 couples in 140 Faith Leader groups at baseline. To select these, we used a two-stage sampling procedure:

- First, to select the Faith Leaders, we started by dropping 3 Faith Leaders who did not meet the recruitment targets, leaving 142, then we randomly selected 140 Faith Leaders whose recruits would form the basis for our baseline sample.
- Next, within each selected Faith Leader group, 14 couples were randomly chosen to participate in the baseline.

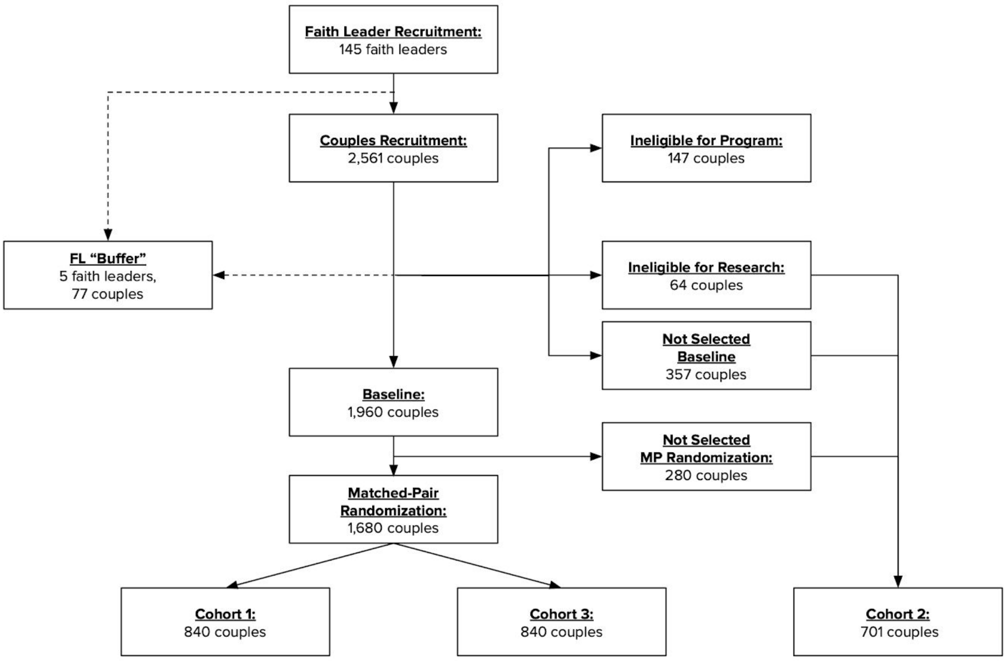
In the 5 Faith Leader groups that were not selected to participate in the baseline during the first stage of sampling, couples were still randomized into 3 cohorts and these groups became a “buffer” in the event that the other Faith Leader groups selected into the study were unable to initiate the program. The couples in the 140 Faith Leader groups who were not selected to complete a baseline survey during the second stage of sampling were randomly assigned a replacement number and held as a pool to draw from in the event that we were unable to interview those originally selected. During the baseline, 147 couples were deemed ineligible to participate in Becoming One and 64 were excluded from the research study (but were still eligible to participate in the program).

After the baseline survey was concluded, we used the following sampling procedure to determine the final set of 1,680 couples that would be included in the blocked experiment:

1. Given the heterogeneity in the size of the initial recruitment class by Faith Leader, the optimal number to sample (8, 10, 12, or 14) from each Faith Leader was determined such that class sizes were relatively stable across cohorts 1, 2, and 3.
2. Within each Faith Leader group, the optimal number was sampled from among the couples who successfully completed the baseline.

All remaining couples (i.e. those who were eligible to participate in the program but were not in the baseline sample or those that were in the baseline sample but not selected for the blocked experiment) were assigned to cohort 2.

The sample selection described in this section procedure is documented in the figure below:



13. Randomization

The random assignment works by forming couples into pairs within congregations by matching on their baseline levels of violence, and then randomizing one couple in each pair to the intervention group (first cohort), and one couple to the comparison group (third cohort).

Baseline violence was measured as follows: first, we clean the data and impute any missing items; second, we code six binary indicators for whether the woman in the couple reports having been pushed, slapped, arm-twisted / hair-pulled, punched, kicked, or choked over the preceding 12 months; third, we take the within-subject means of these binary indicators, and standardize it to have mean 0 and standard deviation of 1.

We form two-couple blocks within congregations using the `blockTools` algorithm in R. Some faith leaders were able to gather a bigger group than others, and some respondents recruited by faith leaders did not want to do the survey. However, the sample selection procedure described above ensures that each congregation has eight (n = 1), ten (n = 30), twelve (n = 77), or fourteen (n = 32) couples to randomize among (the numbers in parentheses indicate the frequencies of the different congregation sizes in the study). Thus, we are assured of an even number of couples within faith leaders, and so able to do pair-matching within congregations in every congregation.

We create blocks and randomize within them using the following code:

```
# Create blocks
-----------------------------------------------------------
block_on <- c("prop_physical_z")
bls <- blockTools::block(
  data = blwc,
  n.tr = 2,
  id.vars = "cup_id",
  groups = "fl_id",
  block.vars = block_on)
blwc$blocks <- createBlockIDs(obj = bls,data = blwc,id.var = "cup_id")

# Randomize
----------------------------------------------------------------
set.seed(5862007)
blwc$Z <- with(blwc, randomizr::block_ra(blocks = blocks, prob = .5))
```

14. "FL Buffer" couples

As mentioned above, prior to the baseline we dropped three faith leaders from the baseline survey because they had not recruited at least 14 couples each, and then randomly dropped two more in order to have a sample of 140 faith leaders from the original 145. In total, this resulted in 77 couples being excluded from the baseline.

At the randomization stage, we still randomized 54 of these couples into cohorts 1 and 3. The randomization was of course different from the main randomization, in that couples were assigned with 50/50 probabilities within *faith leader* blocks, and not within matched pair blocks (as we did not have baseline data upon which to match them).

We plan to sample and interview these 54 couples at midline. We will analyze their responses in two separate analyses. First, if we encounter a significant loss of power due to congregation-level attrition (one congregation in which all couples refuse to participate), we will report robustness results including the 54 FL buffer couples to make up for this power loss. Randomization inference will take account of the different assignment mechanisms, but the main specification will remain the same, just with FL block indicators for the buffer couples, rather than couple-pair block indicators. Second, we will compare the rates of violence reported by control respondents in the two congregations randomly dropped from the baseline to those of the control respondents in the main sample. This constitutes a test of the idea that being asked a baseline survey may change responses at midline.

15. Midline Survey Sampling Strategy

During midline we intend to monitor rates of attrition (see definition of expected types in section 9 below) by treatment status on a daily basis for signs that the attrition is above expected levels and/or is possibly differential across treatment and comparison (cohort 1 vs. cohort 3). Where possible we will seek to track and interview all respondents who completed a baseline survey and were randomly assigned to cohort 1 or 3. In the case of possible differential attrition, instead of tracking all possible cases of attrition with equal weight we may instead attempt to divert resources to a more concerted resampling method to correct for observed unbalance on baseline factors that are also believed to be prognostic of the outcome. In this case, we will detail our resampling methods and any associated changes to the estimation strategy in an addendum to the PAP.

5. Estimation

16. Main specification

Unless indicated in the tables below, most of the analyses will take place at the **couple-level.** For these analyses, we will construct a wide couple-level dataset by appending "_w" (for women) and "_m" to each variable, and putting observations for men and women in the same couple on the same row.

Some analyses will be conducted at the **individual-level**. For these analyses, we will construct a long dataset in which every row corresponds to the responses to an individual survey. Individuals are assigned to treatment through couple clusters, so standard errors will be clustered at the couple-level in any individual-level analyses.

In both cases, we will construct *p*-values using randomization inference, in a procedure described in more detail below.

Our main estimator is the "covariate-adjusted specification" below, which adjusts for predictive covariates using a mean-centering and interaction approach described in Lin (2013). This specification will serve as the basis for inference about the treatment effects. The procedure for selecting covariates and the justification for using this procedure are provided below. Any familywise corrections (described below) will be conducted on the covariate-adjusted estimator.

For transparency, we will also report (in main table or appendix) effects from a minimal, design-based estimator that only accounts for the randomization and does not adjust for covariates. The two estimators can be written as follows:

1. Covariate-adjusted specification: $y_{ij} = \gamma_j + \tau z_i + \beta^\top \bar{\mathbf{x}}_i + \lambda^\top \bar{\mathbf{x}}_i z_i + \epsilon_i$
2. Design-based specification: $y_{ij} = \gamma_j + \tau z_i + \epsilon_i$

Where gamma is a block-level fixed effect, tau is the average treatment effect, *z* is an indicator of assignment to cohort 1, ***x*** bar is a vector of mean-centered covariates, and epsilon an error term. The index i indicates individuals and couples in individual- and couple-level analyses, respectively. The index j indicates matched pair blocks.

Letting Y stand for a generic outcome, Z for treatment, and X_c for mean-centered covariates, and subgroup for the subset among whom the analysis is being conducted (when estimating among the whole sample we use subset = NULL), we will use the following code to estimate treatment effects with specifications 1 and 2 (using the estimatr package for R).

Covariate-adjusted specification, Couple-Level:

```
estimatr::lm_robust(formula = Y ~ Z + X_c + Z:X_c,
                    fixed_effects = ~ blocks,
                    se_type = "HC2",
                    subset = subgroup,
                    data = wide_data)
```

Design-based specification, Couple-Level:

```
estimatr::lm_robust(formula = Y ~ Z,
                    fixed_effects = ~ blocks,
                    se_type = "HC2",
                    subset = subgroup,
                    data = wide_data)
```

Covariate-adjusted specification, Individual-Level:

```
estimatr::lm_robust(formula = Y ~ Z + X_c + Z:X_c,
                    fixed_effects = ~ blocks,
                    se_type = "CR2",
                    clusters = cup_id,
                    subset = subgroup,
                    data = long_data)
```

Design-based specification, Individual-Level:

```
estimatr::lm_robust(formula = Y ~ Z,
                    fixed_effects = ~ blocks,
                    se_type = "CR2",
                    clusters = cup_id,
                    subset = subgroup,
                    data = long_data)
```

17. Covariate selection procedures

We will only condition on covariates that are predictive of the outcome, irrespective of whether those covariates are imbalanced. We made this decision in light of a simulation study, in which we assumed the following data-generating process:

$$\begin{aligned} Z &\sim Binom(.5) \\ x_1 &\sim N(0,1) \\ x_2 &\sim N(0,1) \\ Y_0 &\sim N(\rho x_1, \sqrt{1-\rho^2}) \\ Y_1 &= Y_0 + \tau \\ Y &= Y_1 Z + Y_0(1-Z) \end{aligned}$$

With rho = .7 and tau = .20, so that x_1 is correlated with Y but x_2 is not.

We then compare four strategies for estimating the effect of Z on Y:

1. Never condition on x_i
2. Condition on x_i iff x_i ~ Z is significant (x_i imbalanced), otherwise do not condition on x_i

3. Condition on x_i iff Y ~ x_i is significant (x_i predictive), otherwise do not condition on x_i
4. Condition on x_i iff x_i is imbalanced *and* predictive, otherwise do not condition on x_i

Ex ante, all approaches are unbiased but approach 3 minimizes variance. If, ex post, x_1 is imbalanced, approach 1 is biased due to imbalance, but approaches 2-4 all fare well. In situations where, ex post, x_2 is imbalanced, approach 3 provides the best estimates. By conditioning on a non-prognostic covariate, approaches 2 and 4 burn a degree of freedom and potentially generate collinearity between Z and x_2.

To choose covariates that are predictive of each outcome, we will perform lasso on baseline data that has item-level missingness removed through multiple chained equations, as described below.

The lasso procedure that we plan to use features a generalized linear model with lasso penalization, and is implemented in the `glmnet` package for R. The loss function requires selecting a regularization parameter, lambda, that determines the severity of the penalty for including extra covariates. Since this regularization parameter cannot be optimally chosen in advance, we will select it using 10-fold cross-validation. Specifically, for each outcome, we will choose the lambda that minimizes the 10-fold cross-validation error averaged over 10 runs (since the folds are chosen at random). Only the covariates retained by the lasso will be included in the specification.

We will perform this lasso variable selection method using the entire list of available covariates.

18. Estimands

Given that both couples and faith leaders were initially recruited purposefully rather than as a random sample from a population, for our main analyses we restrict inferences to estimands at the sample- and sub-sample levels.

We expect to encounter two-sided non-compliance in this study (defined below). For our main analyses, we will therefore estimate intent-to-treat effects. That is, $E[Y_i(Z_i = 1) - Y_i(Z_i = 0)]$, the effect of assignment to treatment, Z, on outcome Y. As a robustness check, we will also seek to estimate the average treatment effect, $E[Y_i(D_i = 1) - Y_i(D_i = 0)]$, using the inverse compliance-weighted IV estimator described below. D here stands for "delivered," since the treatment was actually delivered in such cases.

19. Statistical inference (p-value and standard error calculation)

In all couple-level analyses, we will calculate standard errors using a heteroskedasticity-robust (HC2) estimator as coded in `estimatr` for R. In all individual-level analyses, we will calculate cluster-robust standard errors (CR2).

Decisions about the significance of effect sizes will rely primarily on non-parametric *p*-values calculated using randomization inference (one exception is the multiple comparisons correction procedure described below, which relies on parametric p-values for speed of computation).

The randomization inference procedure works by re-estimating effects thousands of times under repeated simulations of the assignment. This provides a representative sample of effects from the distribution of effects that would obtain under the sharp null hypothesis of a constant 0 effect. The proportion of estimates under this null hypothesis that are at least as large in positive, negative, or absolute value as the observed effect gives us the *p*-value: the probability of observing an effect as large as the one we observe if indeed there is no effect for any unit in the sample.

Typically, the procedure will involve drawing from the null distribution of effects as follows

```
null_dist <- replicate(
  n = sims,
  expr = {
    wide_data$Z_sim <-
      with(wide_data, randomizr::block_ra(blocks = blocks, prob = .5))
    lm_robust(formula = Y ~ Z_sim,
              fixed_effects = ~ blocks,
              se_type = "HC2",
              subset = subgroup,
              data = wide_data) %>%
      tidy() %>% filter(term == "Z_sim") %>% select(estimate) %>%
      unlist()
  })
```

And comparing this to the observed estimate, which we calculate as follows:

```
obs <- lm_robust(formula = Y ~ Z,
                 fixed_effects = ~ blocks,
                 se_type = "HC2",
                 subset = subgroup,
                 data = wide_data) %>%
          tidy() %>% filter(term == "Z") %>% select(estimate) %>%
          unlist()
```

One-tailed hypotheses

A test of a one-tailed hypothesis focuses on the probability mass in one tail of the null distribution only. A negative one-tailed or lower-tailed hypothesis test involves calculating a *p*-value as follows:

```
lower_tailed_pval <- mean(obs >= null_dist)
upper_tailed_pval <- mean(obs <= null_dist)
```

Two-tailed

For two-tailed hypotheses we take the absolute values of the observed and null test statistics, in order to assess the proportion of effects under the null at least as large in absolute value.

```
two_tailed_pval <- mean(abs(obs) <= abs(null_dist))
```

Subgroup effects

When assessing heterogeneous effects, we will base inferences on a null distribution of differences in effects, computed among each subgroup on a given run:

```
null_dist_AB <- null_dist_A - null_dist_B
obs_AB <- obs_A - obs_B
A_bigger_than_B_pval <- mean(abs(obs_AB) <= abs(null_dist_AB))
```

6. Main Analyses

We analyze four primary outcomes: intimate partner violence; power inequality (control over finances, movement and decision-making); sexual consent and autonomy; and communication and conflict management. To reduce the risk of false positives, we construct indices. Having conducted power analyses using Z-score indexing, factor score indexing, and simple arithmetic mean indexing, we determined that arithmetic mean indices were better-powered if the treatment only affects some subset of the index, and also corresponded to a clearer estimand. All component measures are coded in a similar substantive direction and vary between 0 and 1: thus, for an index where positive outcomes are positively valued, 0 signifies the worst possible outcome across all outcomes and 1 the best.

Effects on all primary outcomes will be assessed using the main "covariate-adjusted" specification, as above. By way of explaining any observed movement, we will also report the effects on sub-indices graphically. All hypotheses pertain to the effect of the treatment. As indicated below, we plan one-tailed tests for primary outcomes.

1. Intimate partner violence (sexual and physical)

We estimate effects on intimate partner violence using one main outcome. In the appendix, we will report both the effects on the individual components that make up these composite measures, as well as effects on a multinomial outcome designed by Heise.

Outcome name:	any_violence
Coding:	Binary indicator that any one of the following is true: physical_push_5mo_w > 0; physical_slap_5mo_w > 0; physical_twist_5mo_w > 0; physical_punch_5mo_w > 0; physical_kick_5mo_w > 0; physical_choke_5mo_w > 0; physical_weapon_5mo_w > 0; sexual_forced_intercourse_5mo_w > 0; sexual_forced_other_acts_5mo_w > 0; sexual_threaten_5mo_w > 0.
Interpretation	Proportion of female partners who report having experienced any physical or sexual violence at the hands of male partner over preceding five to six months.
Hypothesis:	Negative (one-tailed)
Analysis level:	Couple

Ability to detect unanticipated adverse effects on violence

We do not expect to see an increase in violence caused by the B1 program. Hence, we primarily want to confirm the hypothesis of a treatment-induced reduction, and so conduct a one-tailed test. However, it is important to monitor for any unanticipated adverse effects of the program, in particular because project partners are planning to scale the program if it is found to be effective in some domains, and because cohorts 2 and 3 will go through the program following the midline. If we see that confidence intervals on treatment effects are bounded far from zero in a positive direction, we will interpret this as concerning but exploratory evidence of an adverse effect.

2. Power inequality (2 indices: control and decision-making, and sexual consent and autonomy)

To construct measures of power inequality in a couple, we collected all of the questions that measure an outcome that is plausibly "zero sum": for example, if a woman is given a greater role in financial decision-making, that diminishes her husband's ability to unilaterally decide how finances should be spent. We conducted an exploratory factor analysis of thirty-five such outcomes. This helped to identify two distinct families of sub-indices: first, those relating to financial control, social control, and decision-making; second, those related to practices around sexual consent and control of the body.

Index items are subscripted _i to indicate their recoding in order to point in the same substantive direction (in contrast to the original variables as coded in the survey, which do not always point numerically in the same direction).

Control and decision-making index

Outcome name:	control_index

Coding:	Mean of the following variables, coded between 0 and 1:
	fin_control_w_i = 1 if "Self/Own Choice", .5 if "Give Part to Husband/Partner", 0 if "Give All To Husband/Partner";
	fin_control_work_w_i = 1 if no, 0 if yes;
	fin_control_take_money_w_i = 1 if no, 0 if yes;
	fin_control_keep_money_w_i = 1 if no, 0 if yes;
	control_friends_w_i = 1 if no, 0 if yes;
	control_family_w_i = 1 if no, 0 if yes;
	control_whereabouts_w_i = 1 if no, 0 if yes;
	control_mobile_w_i = 1 if no, 0 if yes;
	dm_earnings_resp_w_i = 1 if female partner makes or is involved in decision-making; 0 otherwise;
	dm_earnings_partners_w_i = 1 if female partner makes or is involved in decision-making; 0 otherwise;
	dm_large_purchase_w_i = 1 if female partner makes or is involved in decision-making; 0 otherwise;
	dm_health_w_i = 1 if female partner makes or is involved in decision-making; 0 otherwise;
	dm_visit_family_w_i = 1 if female partner makes or is involved in decision-making; 0 otherwise;
	dm_windfall_resp_w_i = 1 if female partner makes or is involved in decision-making; 0 otherwise;

	dm_windfall_partner_w_i = 1 if female partner makes or is involved in decision-making; 0 otherwise; dm_earnings_resp_m_i = 1 if female partner makes or is involved in decision-making; 0 otherwise; dm_earnings_partners_m_i = 1 if female partner makes or is involved in decision-making; 0 otherwise; dm_large_purchase_m_i = 1 if female partner makes or is involved in decision-making; 0 otherwise; dm_health_m_i = 1 if female partner makes or is involved in decision-making; 0 otherwise; dm_visit_family_m_i = 1 if female partner makes or is involved in decision-making; 0 otherwise; dm_windfall_resp_m_i = 1 if female partner makes or is involved in decision-making; 0 otherwise; dm_windfall_partner_m_i = 1 if female partner makes or is involved in decision-making; 0 otherwise. Income_r_sep_amt_m = 1 if no, 0 if yes; control_general_w_i = 1 if I have control over all such decisions, .66 if I have control over most such decisions, .33 if I have control over some of these decisions, 0 if I don't control any of these decisions.
Interpretation	When the index is at 0, this implies the female experiences all forms of financial and social control, and both partners perceive her to be excluded from all decision-making. When it is at 1, this implies the female partner does not report experiencing any forms of financial or social control, and is involved in all aspects of decision-making.
Hypothesis:	Positive (one-tailed)
Analysis level:	Couple

Sexual consent and autonomy index

Outcome name:	consent_index
Coding:	Mean of the following variables, coded between 0 and 1: no_to_sex_strategies_m_i = 1 if male partner indicates he will take a non-coercive path when refused sex by his partner (don't have sex or talk about it); 0 otherwise (persuade her, she always wants to have sex, have sex anyway); control_sex_w_i = 1 if female partner feels she has control over all decisions about when to have sex with husband; .66 if "most such" decisions; .33 over "some of these decisions"; 0 if female partner expresses not having any control; initiate_sex_w_i = 1 if female partner indicates always, .66 if sometimes; .33 if rarely; 0 if never; no_to_sex_w_i = 1 if female partner feels very confident saying no to

	sex, .5 if somewhat confident, 0 if not at all confident; no_to_sex_threatens_w_i = 1 if female partner indicates no, 0 if yes; ever_pressured_w_i = 1 if female partner never feels pressured into sex, .66 if not often, .33 if sometimes, 0 if often.
Interpretation	When the index is at 0, this implies that the female partner reports sexual coercion across all of the domains, and that the male partner reports coercive behavior. When it is at 1, this implies the absence of coercion.
Hypothesis:	Positive (one-tailed)
Analysis level:	Couple

3. Skills and practices (communication and conflict management)

While the primary outcomes grouped under "power inequality" above pertain to zero sum dynamics, we theorize that other outcomes do not have this same feature. If a behavior adopted by one member of the couple did not detract from the other member’s ability to participate in that behavior or to obtain power in the relationship more broadly, we counted it as a skill or practice that was not “zero sum” in its effects on power in a relationship. Illustrative examples include: doing fun things together, feeling lonely with a partner (reverse coded), communicating about sex, discussing issues from the day, thanking the partner, and responding calmly to the partner. In each of these cases, if a female partner enacted one of these behaviors, it is not clear that a male partner would have to do less of this behavior.

Communication and conflict management index

Outcome name:	comm_index
Coding:	Mean of the following variables, coded between 0 and 1: disc_day_m_i = average of disc_day_p_m_i and disc_day_r_m_i, where the first concerns the male partner's perception of how often the couple discusses things in the female partner's day, and the second concerns the male partner's perception of frequency of discussing things in their own day: both sub-items coded 1 for many times, .66 for a few times, .33 for once, 0 for never; disc_worry_m_i = average of disc_worry_p_m_i and disc_worry_r_m_i, where the first concerns the male partner's perception of how often the couple discusses the female partner's worries or feelings, and the second concerns the male partner's perception of frequency of discussing his own worries or feelings: both sub-items coded 1 for many times, .66 for a few times, .33 for once, 0 for never; disc_sex_m_i = 1 if male partner reports discussing sexual relationship many times, .66 if a few, .33 if once, 0 if never; disc_day_w_i = average of disc_day_p_w_i and disc_day_r_w_i,

	where the first concerns the female partner's perception of how often the couple discusses things in the male partner's day, and the second concerns the female partner's perception of frequency of discussing things in their own day: both sub-items coded 1 for many times, .66 for a few times, .33 for once, 0 for never; disc_worry_w_i = average of disc_worry_p_w_i and disc_worry_r_w_i, where the first concerns the female partner's perception of how often the couple discusses the male partner's worries or feelings, and the second concerns the female partner's perception of frequency of discussing her own worries or feelings: both sub-items coded 1 for many times, .66 for a few times, .33 for once, 0 for never; disc_sex_w_i = 1 if female partner reports discussing sexual relationship many times, .66 if a few, .33 if once, 0 if never; comm_1_m_i = 1 if male partner reports female partner never interrupts, .66 if rarely, .33 if sometimes, 0 if always; comm_2_m_i = 1 if male partner reports female partner always listens, .66 if sometimes, .33 if rarely, 0 if never; comm_3_m_i = 1 if male partner reports female partner always comforts when he is having problems, .66 if sometimes, .33 if rarely, 0 if never; comm_4_m_1 = 1 if male partner reports female partner always thanks him for things he does, .66 if sometimes, .33 if rarely, 0 if never; comm_5_m_i = 1 if male partner reports female partner never says things that make him feel small, .66 if rarely, .33 if sometimes, 0 if always; comm_1_w_i = 1 if female partner reports male partner never interrupts, .66 if rarely, .33 if sometimes, 0 if always; comm_2_w_i = 1 if female partner reports male partner always listens, .66 if sometimes, .33 if rarely, 0 if never; comm_3_w_i = 1 if female partner reports male partner always comforts when she is having problems, .66 if sometimes, .33 if rarely, 0 if never; comm_4_w_i = 1 if female partner reports male partner always thanks her for things she does, .66 if sometimes, .33 if rarely, 0 if never; comm_5_w_i = 1 if female partner reports male partner never says things that make her feel small, .66 if rarely, .33 if sometimes, 0 if always; how_arg_calm_m_i = 1 if either there is no category of arguments the respondent reports having had or they had at least one kind of argument and respondent's partner often expressed feelings in calm and respectful way, 0 otherwise; how_arg_listened_m_i = 1 if either there is no category of arguments the respondent reports having had or they had at least one kind of argument and respondent's partner often tried to see respondent's side of things, 0 otherwise; how_arg_yell_m_i = 1 if either there is no category of arguments the respondent reports having had or they had at least one kind of argument and respondent's partner never yelled, insulted or swore, 0

	otherwise; how_arg_threat_m_i = 1 if either there is no category of arguments the respondent reports having had or they had at least one kind of argument and respondent's partner never threatened in some way, 0 otherwise; how_arg_left_m_i = 1 if either there is no category of arguments the respondent reports having had or they had at least one kind of argument and respondent's partner often left so that she could calm down when argument was heated, 0 otherwise; how_arg_calm_w_i = 1 if either there is no category of arguments the respondent reports having had or they had at least one kind of argument and respondent's partner often expressed feelings in calm and respectful way, 0 otherwise; how_arg_listened_w_i = 1 if either there is no category of arguments the respondent reports having had or they had at least one kind of argument and respondent's partner often tried to see respondent's side of things, 0 otherwise; how_arg_yell_w_i = 1 if either there is no category of arguments the respondent reports having had or they had at least one kind of argument and respondent's partner never yelled, insulted or swore, 0 otherwise; how_arg_threat_w_i = 1 if either there is no category of arguments the respondent reports having had or they had at least one kind of argument and respondent's partner never threatened in some way, 0 otherwise; how_arg_left_w_i = 1 if either there is no category of arguments the respondent reports having had or they had at least one kind of argument and respondent's partner often left so that he could calm down when argument was heated, 0 otherwise; Conflict_2_m_i = 1 if the partner disagrees strongly that they've held back feelings to avoid a conflict, .66 if they disagree, .33 if they agree, and 0 if they agree strongly. Conflict_3_m_i = 1 if the partner strongly agrees they have good strategies for resolving disagreements, .66 if they agree, .33 if they disagree, and 0 if they disagree strongly. Conflict_2_w_i = 1 if the partner disagrees strongly that they've held back feelings to avoid a conflict, .66 if they disagree, .33 if they agree, and 0 if they agree strongly. Conflict_3_w_i = 1 if the partner strongly agrees they have good strategies for resolving disagreements, .66 if they agree, .33 if they disagree, and 0 if they disagree strongly.
Interpretation	When the index is at 0, this implies that the partners agree that they have had no supportive or open communication or techniques for peaceful or de-escalating communication across many domains.. When it is at 1, this implies the partners agree that they communicate with supportive or open communication or techniques for peaceful or de-escalating communication across all the domains.
Hypothesis:	Positive (one-tailed)

Analysis level:	Couple

7. Secondary Analyses

We group secondary outcomes into two categories: those that we hypothesize are possible *mediators* between Becoming One and any effects on our primary outcomes, and those that we consider to be *additional outcomes* that participating in Becoming One may influence, but are not of primary concern.

We hypothesize that the following variables are potential mediators: Gender attitudes and norms, men's emotional regulation, religious interpretation, financial transparency, convergence in couples' trust and communication, and individuals' definitions of violence

The list of secondary outcomes is:
The degree of convergence in a couple regarding their beliefs about and perceptions of the relationship, a shifted distribution of domestic labor and of perceptions of community norms regarding intimate partner violence; within the couple, lowered levels of emotional violence and of incidence of arguments, greater communication about sexual practices, higher levels of time spent together; on an individual level, shifts in perceived community norms, lowered levels of individual depression and higher levels of confidence in expressing opinions publicly and in perceived social support. As a secondary outcome, we also plan to test separately the effect of Becoming One on different kinds of intimate partner violence, including emotional, physical, and sexual violence.

4. Potential Mediators

Gender attitudes and norms

Outcome name:	equitable_gender_att
Coding:	Note: at midline couples are randomized to receive 4 attitude questions from the list: att_1; att_3; att_7; att_10; att_11; norm_1; norm_2; norm_4; and then, of those four, two questions are randomly changed to their alternate version: att_1_a; att_3_a; att_7_a; att_10_a; att_11_a; norm_1_a; norm_2_a; norm_4_a; which cover the same construct but from the opposite substantive direction. Note also: all of these outcomes pertain to individual attitudes, and not perceptions of norms. The "norm" labelling reflects an earlier phrasing of the question that was abandoned in favor of a focus on individual attitudes. To construct the index we take the mean of indicators for the 4 attitude questions to which the individual was randomly assigned, ,

	coded towards the “gender equitable” position as described below:
	att_1_i = 1 if Disagree or Strongly disagree and 0 if Agree or Strongly Agree;
	att_3_i = 1 if Disagree or Strongly disagree and 0 if Agree or Strongly Agree;
	att_7_i = 1 if Disagree or Strongly disagree and 0 if Agree or Strongly Agree;
	att_10_i = 1 if Disagree or Strongly disagree and 0 if Agree or Strongly Agree;
	att_11_i = 1 if Disagree or Strongly disagree and 0 if Agree or Strongly Agree;
	norm_1_i = 1 if Disagree or Strongly disagree and 0 if Agree or Strongly Agree;
	norm_2_i = 1 if Disagree or Strongly disagree and 0 if Agree or Strongly Agree;
	norm_4_i = 1 if Disagree or Strongly disagree and 0 if Agree or Strongly Agree;
	att_1_a_i = 1 if Agree or Strongly agree and 0 if Disagree or Strongly disagree;
	att_3_a_i = 1 if Agree or Strongly agree and 0 if Disagree or Strongly disagree;
	att_7_a_i = 1 if Agree or Strongly agree and 0 if Disagree or Strongly disagree;
	att_10_a_i = 1 if Agree or Strongly agree and 0 if Disagree or Strongly disagree;
	att_11_a_i = 1 if Agree or Strongly agree and 0 if Disagree or Strongly disagree;
	norm_1_a_i = 1 if Agree or Strongly agree and 0 if Disagree or Strongly disagree;
	norm_2_a_i = 1 if Agree or Strongly agree and 0 if Disagree or Strongly disagree;
	norm_4_a_i = 1 if Agree or Strongly agree and 0 if Disagree or Strongly disagree;

Interpretation	Proportion of equitable attitudes towards the role of women in the household subscribed to by the respondent..
Hypothesis:	Positive (one-tailed)
Analysis level:	Individual

Outcome name:	anti_violence_att
Coding:	Note: at midline couples are first randomized to receive 2 attitude questions from the list: att_5; att_6; att_8; att_9 and then, of those two questions, 1 was randomly changed to their alternate version: att_5_a; att_6_a; att_8_a; att_9_a

	which cover the same construct but from the opposite substantive direction. Additionally they were randomized to receive 3 justification attitudes from the list:
	att_abuse_unfaithful; att_abuse_goes_out; att_abuse_argues; att_abuse_neglects; att_abuse_burns_food; att_abuse_no_sex;
	
	To construct the index we take the mean of indicators for the 5 attitude questions to which the individual was randomized as well as an indicator for their responses to att_abuse questions coded towards the "gender equitable" position as described below:
	att_5_i = 1 if Disagree or Strongly disagree and 0 if Agree or Strongly Agree;
	att_6_i = 1 if Disagree or Strongly disagree and 0 if Agree or Strongly Agree;
	att_8_i = 1 if Disagree or Strongly disagree and 0 if Agree or Strongly Agree;
	att_9_i = 1 if Disagree or Strongly disagree and 0 if Agree or Strongly Agree;
	att_abuse_disobeys_i = 1 if no, 0 if yes;
	att_abuse_disobeys_no_i = 1 if no, 0 if yes;
	att_abuse_disobeys_yes_i = 1 if no, 0 if yes;
	att_abuse_unfaithful_i = 1 if not, 0 if yes;
	att_abuse_goes_out_i = 1 if no, 0 if yes;
	att_abuse_argues_i = 1 if no, 0 if yes;
	att_abuse_neglects_i = 1 if no, 0 if yes;
	att_abuse_burns_food_i = 1 if no, 0 if yes;
	att_abuse_no_sex_i = 1 if no, 0 if yes;
	att_5_a_i = 1 if Agree or Strongly agree and 0 if Disagree or Strongly disagree;
	att_6_a_i = 1 if Agree or Strongly agree and 0 if Disagree or Strongly disagree;
	att_8_a_i = 1 if Agree or Strongly agree and 0 if Disagree or Strongly disagree;
	att_9_a_i = 1 if Agree or Strongly agree and 0 if Disagree or Strongly disagree;

Interpretation	Proportion of anti violence attitudes subscribed to by the respondent.
Hypothesis:	Positive (one-tailed)
Analysis level:	Individual

Men's emotional regulation

Outcome name:	emotional_regulation
Coding:	The following are summed for each respondent, emo_reg_2_m

	emo_reg_3_m emo_reg_6_m emo_reg_10_m emo_reg_11_m emo_reg_14_m coded 0 for “Not at all”, 1 for “Several days”, 2 for “More than half the days”, and 3 for “Nearly every day” and divided by the maximum possible score (3 x 6 = 18).
Interpretation	When the score is at 0, respondent engages in no constructive emotional regulation practices, when at 1 they are able to emotionally regulate at highest possible level across measured indicators.
Hypothesis:	Positive (one-tailed)
Analysis level:	Individual (Men only)

Outcome name:	num_calm_methods
Coding:	Number of calming methods (from calm_method variable) mentioned by the respondent fitting into the categories: 1. Use “I feel…” instead of “you…” 2. Speak at one time 3. Go outside and be quiet 4. Go to sleep 5. Keep personal space 6. Think before you speak 7. Go to church 8. Ask ‘what is the goal?’ 9. Drink a glass of water 10. Breathe deeply
Interpretation	Number of calming methods spontaneously mentioned by the as being used by the individual.
Hypothesis:	Positive (one-tailed)
Analysis level:	Individual

Scriptural Interpretation

Outcome name:	scripture
Coding:	Binary indicator indicating "gender equitable" interpretation of scripture for one of: scripture_adam_eve == 1; scripture_creation == 2; scripture_wife_body == 2; scripture_obey == 2.
Interpretation	Probability that respondent subscribes to a more gender equitable interpretation of scripture.
Hypothesis:	Positive (one-tailed)
Analysis level:	Individual

Violence definitions

Outcome name:	includes_sexual_violence
Coding:	Binary indicator that definition includes sexual violence (i.e. violence_def_binary == 2 or violence_def_binary == 3)
Interpretation	Probability that respondent thinks of acts of sexual violence as "violence."
Hypothesis:	Positive (one-tailed)
Analysis level:	Individual

Financial Planning (Note: see section on convergence in beliefs and attitudes for two additional measures of planning and transparency that are predicted to be mediators, income_hiding_meta_belief *and* income_share_divergence*)*

Outcome name:	financial_planning
Coding:	1 if financial_plan_w = yes and financial_plan_m = yes, 0 otherwise.
Interpretation	Couple has engaged in financial planning
Hypothesis:	Positive (one-tailed)
Analysis level:	Couple

NOTE: for three additional mediators regarding convergence in trust and communication, please again see section on convergence in beliefs and attitudes: trust_meta_belief, discuss_worry_meta_belief, discuss_day_meta_belief

5. Additional outcomes

Perceived community norms

Outcome name:	anti_violence_norm
Coding:	Mean of the following: perc_att_disobey_i = 0 if “no one believes this”, 0.33 if “some…”, 0.66 if “most…”, 1 if “everyone..”; perc_att_abuse_i = 0 if “no one believes this”, 0.33 if “some…”, 0.66 if “most…”, 1 if “everyone..”;
Interpretation	Degree to which individual perceives their community to subscribe to “anti-violence” attitudes.
Hypothesis:	Positive (one-tailed)
Analysis level:	Individual

Outcome name:	perc_norm_intervention
Coding:	Mean of two variables about respondent perceptions of community's approach towards intervention: - recode comm_intervene to be 1 if “they would intervene…” 0 if “they would mind their own business…” - recode should_intervene to be 1 if “more like the first place”, 0 if “more like the second place”
Interpretation	Strength of respondent's belief belief that their community would intervene to stop a man from beating his wife.
Hypothesis:	Two-tailed
Analysis level:	Individual

Outcome name:	domestic_labor_share
Coding:	sum of activities_hours_2, activities_hours_3, activities_hours_4, activities_hours_5, and activities_hours_6 divided by total activities hours (activities_hours_1 through activities_hours_6)
Interpretation	Proportion of hours spent on domestic activities
Hypothesis:	No overall hypothesis: positive for men (one-tailed), negative for women (one-tailed),

Analysis level:	Individual

Incidence of Arguments

Outcome name:	arg_intensity
Coding:	Mean of the following: (1) Sum of arg_resp_w, arg_prov, arg_money_w, arg_drink_w, arg_infidelity_w, arg_sex_w, arg_children_w, arg_anything_else_w, divided by maximum possible score (24). (2) Sum of arg_resp_m, arg_prov, arg_money_m, arg_drink_m, arg_infidelity_m, arg_sex_m, arg_children_m, arg_anything_else_m, divided by maximum possible score (24).
Interpretation	Intensity/frequency of arguments about responsibilities, providing for family, money, alcohol use, infidelity, sex, the children or anything else. At 0 the couple had lowest possible frequency of arguments, and at 1 they had the highest.
Hypothesis:	Negative (one-tailed)
Analysis level:	Couple

Sexual communication

Outcome name:	sex_satisfying_i
Coding:	1 if sex_satisfying = Very satisfying, .66 if Satisfying, .33 if Unsatisfying, 0 if Very unsatisfying.
Interpretation	Amount of self-reported satisfaction R derives from sexual relationship with partner.
Hypothesis:	Positive (one-tailed)
Analysis level:	Individual

Depression Scale

Outcome name:	depression_score
Coding:	sum of phq_1, phq_2, phq_3, phq_4, phq_5, phq_6, phq_7, phq_8, phq_9 with responses coded 0 for “Not at all”, 1 for “Several days”, 2

	for “More than half the days”, and 3 for “Nearly every day” divided by the maximum possible score (27).
Interpretation	Depression score [0-1] indicating severity of depressive symptoms expressed by respondent. When the score is 0 they express never experiencing any depressive symptoms over last 2 weeks, when the score is 1 they express the maximum level of severity across all items over the last two weeks.
Hypothesis:	Negative (one-tailed)
Analysis level:	Individual

Time together

Outcome name:	quality_time
Coding:	Mean of the following variables, coded between 0 and 1: pair_13_i = 1 when respondent reports always finding time to do pleasurable things together, .66 when sometimes, .33 when rarely, 0 when never; pair_14_i = 1 when respondent reports never feeling lonely when together with partner, .66 when rarely, .33 when sometimes, 0 when always; pair_15_i = 0 when respondent reports partner always seems disinterested in sex, 1 otherwise.
Interpretation	When the index is at 0, this implies that the partners agree that they never enjoyed spending time together. When it is at 1, this implies they always enjoyed time together.
Hypothesis:	Positive (one-tailed)
Analysis level:	Individual

Voice

Outcome name:	voice

Coding:	Mean of the following binary indicators:
	hh_voice_4weeks_i = 1 if yes, 0 if no;
	hh_voice_i = 1 if “somewhat confident” or “very confident” 0 if “not at all”
	comm_voice_differ_i = 1 if “somewhat confident” or “very confident”, 0 if “not at all”;

Interpretation	Degree of confidence in expressing thoughts and opinions publicly.

Hypothesis:	Positive (one-tailed)
Analysis level:	Individual

Outcome name:	social_support
Coding:	Mean of the following: support_1 = 1 if strongly agree or agree, =0 if strongly disagree or disagree; support_6 = 1 if strongly agree or agree, =0 if strongly disagree or disagree; fl_seek_support= 1 if strongly agree or agree, =0 if strongly disagree or disagree
Interpretation	A score of 1 means that the individual can find emotional and financial support from someone in their community as well as personal support from their faith leader, a score of 0 means that an individual has none of these types of support, and scores in the middle indicate they receive some of this type of support
Hypothesis:	Positive (one-tailed)
Analysis level:	Individual

Secondary Violence Outcomes:

Outcome name:	severity_emotional_violence
Coding:	Sum of emotional_humiliate_freq_i, emotional_threaten_freq_i, emotional_insult_freq_i coded as 0 if “never or not in last 5 months”, 1 if “once”, 2 if “a few times”, 3 if “many times” and then divided by the maximum possible score (9).
Interpretation	Intensity of emotional violence experienced by respondent at the hands of partner.
Hypothesis:	Negative (one-tailed)
Analysis level:	Individual

Outcome name:	severity_physical_violence
Coding:	Sum of the following: physical_push_5mo_freq_w_i, physical_slap_5mo_freq_w_i, physical_twist_5mo_freq_w_i, physical_punch_5mo_freq_w_i,

	physical_kick_5mo_freq_w_i, physical_choke_5mo_freq_w_i, physical_weapon_5mo_freq_w_i coded as “0” Not in last 5 months “1” Once in last 5 months “2” A few times (2 - 5) in last 5 months and “3” Many times (5+) in last 5 months and then divided by the maximum possible score.
Interpretation	Intensity of physical violence experienced by female partners at the hands of male partners.
Hypothesis:	Negative (one-tailed)
Analysis level:	Couple-level

Outcome name:	severity_sexual_violence
Coding:	Sum of the following: sex_forced_5mo_freq_w_i, sex_forced_other_5mo_freq_w_i, sexual_threaten_5mo_freq_w_i coded as “0” Not in last 5 months “1” Once in last 5 months “2” A few times (2 - 5) in last 5 months and “3” Many times (5+) in last 5 months and then divided by the maximum possible score.
Interpretation	Intensity of physical violence experienced by female partners at the hands of male partners.
Hypothesis:	Negative (one-tailed)
Analysis level:	Couple-level

Ancillary violence outcomes

Outcome name:	fear_partner_bin
Coding:	Binary indicator that female partner expresses fear of male partner at least “Sometimes”: 1 if fear_partner_w > 1; equal to 0 otherwise.
Interpretation	Female partner expresses fear of male partner at least some of the time.
Hypothesis:	Negative (one-tailed)
Analysis level:	Couple-level

Outcome name:	w_any_hit

Coding:	Binary indicator: 1 if hit_partner_w > 1; 0 otherwise.
Interpretation	Proportion of female partners that report ever hitting their male without first being hit themselves in the past 5 months.
Hypothesis:	Negative (one-tailed)
Analysis level:	Couple-level

Convergence in beliefs and perceptions

B1 seeks to promote greater communication and empathy between couples. One possible outcome of this attempt is that partners begin to converge in their understanding of their shared situation.

We specify effects on two such outcomes. First, meta-beliefs. Certain outcomes pertain to objects in the world about which respondents have beliefs, but they also express "meta-beliefs" about their partner's beliefs pertaining to those same objects. For example, we ask respondents whether they trust their partners and whether they believe their partners trust them. Here, the object is the respondent's trust of their partner: we have the respondent's belief about it, and the respondent's beliefs about their partner's beliefs. These outcomes are defined at the **individual**-level.

Second, perception divergence. We have questions that simply ask respondents about beliefs about some object. For example, how close they perceive their relationship to their partner as being. With these outcomes, it is not a case of understanding whether respondents' beliefs about their partners' beliefs match their partners' true beliefs, but more simply whether partners' beliefs do or do not align. These outcomes are defined at the **couple**-level.

We denote outcomes that correspond to the respondent with _r and those that correspond to their partner with _p

*Convergence in trust and communication (*trust_meta_belief, discuss_worry_meta_belief, discuss_day_meta_belief)

NOTE: We hypothesize that the convergence in trust and communication outcomes trust_meta_belief, discuss_worry_meta_belief, discuss_day_meta_belief *will be mediators of our main outcomes, as stated above. Because they are calculated differently, they are listed in this section on couple convergence*

Outcome name:	trust_meta_belief
Coding:	(P_trusts_R_r - R_trusts_P_p)^2

Interpretation	Meta belief. When 0, the respondent is correct in their assessment of whether their partner trusts them, 1 otherwise -- there is some divergence in the respondent's understanding of whether their partner trusts them.
Hypothesis:	Negative (one-tailed)
Analysis level:	Individual-level

Outcome name:	discuss_worry_meta_belief
Coding:	(disc_worry_R_r - disc_worry_P_p)^2
Interpretation	Meta belief. When 0, the respondent is correct in their assessment of whether their partner discusses their worries or feelings, 1 otherwise -- there is some divergence in the respondent's understanding of whether their partner discusses their worries or feelings.
Hypothesis:	Negative (one-tailed)
Analysis level:	Individual-level

Outcome name:	discuss_day_meta_belief
Coding:	(disc_day_R_r - disc_day_P_p)^2
Interpretation	Meta belief. When 0, the respondent is correct in their assessment of whether their partner discusses their day, 1 otherwise -- there is some divergence in the respondent's understanding of whether their partner discusses their day.
Hypothesis:	Negative (one-tailed)
Analysis level:	Individual-level

Financial transparency

NOTE: We hypothesize that the financial transparency outcomes income_hiding_meta_belief *and* income_share_divergence *will be mediators of our main outcomes, as stated above. Because they are calculated differently, they are listed in this section on couple convergence*

Outcome name:	income_hiding_meta_belief
Coding:	(income_p_sep_r - income_r_sep_p)^2

Interpretation	Meta belief. When 0, the respondent is correct in their assessment of whether their partner hides money from them, 1 otherwise -- there is some divergence in the respondent's understanding of whether their partner hides money from them.
Hypothesis:	Negative (one-tailed)
Analysis level:	Individual-level

Outcome name:	income_share_divergence
Coding:	(income_share_w_i - income_share_m_i)^2, where: income_share_w_i = 0 if less than partner, .5 if same, 1 if more than. income_share_m_i = 1 if less than partner, .5 if same, 0 if more than.
Interpretation	When 0, partners converge in their beliefs about who makes more, otherwise they diverge.
Hypothesis:	Negative (one-tailed)
Analysis level:	Couple-level

Other convergence measures:

Outcome name:	relation_quality_divergence
Coding:	Now I want you to think about your relationship. Imagine a ladder with steps numbered from zero at the bottom to ten at the top. Suppose we say that the top of the ladder represents the best possible relationship for you and the bottom of the ladder represents the worst possible relationship for you. On which step of the ladder do you feel you personally stand today? [Show ladder, ranging from 0 - 10] (vignette_3_m - vignette_3_m)^2
Interpretation	When 0, partners converge in their beliefs about the quality of their relationship, otherwise they diverge.
Hypothesis:	Negative (one-tailed)
Analysis level:	Couple-level

Outcome name:	overlap_divergence

Coding:	(overlap_1_m - overlap_1_w)^2
Interpretation	When 0, partners converge in their beliefs about how close they are, otherwise they diverge.
Hypothesis:	Negative (one-tailed)
Analysis level:	Couple-level

Outcome name:	violence_divergence
Coding:	(violence_def_binary_m - violence_def_binary_w)^2
Interpretation	When 0, partners converge in their definitions of what counts as violence against a woman, otherwise they diverge.
Hypothesis:	Negative (one-tailed)
Analysis level:	Couple-level

Our primary outcomes will be reported in the main paper that is written up as a result of the B1 study. We detail here additional evidence that we will report in online appendices, and possibly in any papers if it sheds light on understanding the main results on primary outcomes.

8. Extensions and Subgroup Analyses

We speculate that the effects of the treatment on couples and individuals will be moderated by various characteristics measured at baseline or as part of the monitoring of implementation.

6. Gender

We list here outcomes that we believe may move in different ways in response to treatment depending on gender. We will test these hypotheses by subsetting analyses to men and women, and will test for differences in effects by including an interaction between the treatment and being a woman in the specification. The outcomes, all of which are coded above, and associated hypotheses are as follows:

- domestic_labor_share: positive (one-tailed) for men, negative (one-tailed) for women
- equitable_gender_att: positive (one-tailed) for both men and women, stronger for women (one-tailed hypothesis on interaction)
- anti_violence_att: positive (one-tailed) for both men and women, stronger for women (one-tailed hypothesis on interaction)
- control_general: positive (one-tailed) for women, (two-tailed) for men (we expect failure to reject null for men)
- dm_ladder: positive (one-tailed) for women, negative (one-tailed) for men
- control_ladder: positive (one-tailed) for women, negative (one-tailed) for men

7. Alcohol Use

We will test for interactions between the treatment indicator and an indicator for whether the female partner reported the male partner drinks alcohol at baseline. We will conduct this as a two-tailed test: effects may be stronger among such couples, simply because baseline rates of violence are higher in couples where the man has a drinking problem; conversely, the effect may be smaller among such couples, because drinking may lead to lower program engagement.

8. Variation in implementation

We have many measures of how the B1 program was implemented, all of which correspond to different unobservable dimensions of the implementation quality. In order to study how implementation quality produces variation in effect sizes, we therefore construct a confirmatory factor analysis model, that specifies which measured variables load onto different factors. Specifically, we imagine that implementation quality is the combination of four underlying factors for which we have several measurements:

1. Competence of the faith leader: is the faith leader better able to convey and teach B1 because they are educated, charismatic, and literate?

2. Gender progressivism of the faith leader: is the faith leader more likely to convey gender-equitable attitudes and practices because they already held such beliefs prior to going through B1?
3. Attendance: did people attend the B1 sessions enthusiastically and often?
4. Program fidelity: were the sessions of the appropriate length, did they convey the content they were supposed to, in a style consistent with the intent?

Data on these factors comes from four main sources listed below. We will construct a confirmatory factor model using the `lavaan` package for R and use it to generate scores for each of these latent dimensions of implementation quality. We will then sum them to obtain an overall score, and partition the sample into the bottom 25%, middle 50%, and top 25% on the aggregate and constitutive scores. For our main heterogeneous effects analysis, we will report our main specification, subsetted to each implementation quality stratum of the overall score. We will derive RI p-values for the differences in effects from a model that interacts treatment with implementation quality indicators. This will be our "main" heterogeneous effects model (using primary outcomes and overall score). We will report results by constitutive scores graphically in order to explain main effects.

We derive these measures from four main data sources:

1. FLBL: a baseline survey conducted with faith leaders during their first training
2. BL: the baseline survey with couples in the study
3. SC: randomized spotchecks (audit surveys) that took place at least once during each FL's B1 sessions
4. Photos: photos of sessions taken by faith leaders and sent to research team for coding attendance.

The table below maps latent variable to measured variables. With 21 measured variables and four latent variables ("Competence of faith leader," "Gender progressivism of faith leader," "Attendance," and "Program delivery"), we should have enough degrees of freedom to estimate scores for each of the 145 faith leaders in the sample. However, the model has not yet been constructed. We will describe and justify any departures in the eventual model from the one described below.

Latent variable	**Measured variable name**	**Description**	**Source**
Competence of faith leader	education	FL educational attainment as integer	FLBL
	age	FL age as integer	FLBL
	literacy	FL can read / write	FLBL
	years_village	Years FL in village (indicator of familiarity with community)	FLBL
	years_congregation	Years FL in congregation	FLBL
	main_language_same	Main language FL speaks at home is same as that spoken by congregation	FLBL & BL
	fl_experience	Years working as faith leader	FLBL

	fl_seminary	Received formal training in a seminary	FLBL
	N_sampled	Number of people recruited for baseline survey (indicator of charisma / network)	BL
Gender progressivism of faith leader	gender_equity	Mean of questions on attitudes towards gender equality: granchildren_boys_q3_5 boys_school_q7_1 able_marry_q7_2 final_say_q7_3 child_rearing_q7_4 achiever_q7_1 kneel_q7_3 woman_earns_q7_3 women_leaders_q7_5	FLBL
	anti_vaw	Mean of questions on VAW attitudes coded to be indicators of anti-violence attitudes: disobeys gossip_q8_1b unfaithful_q8_1c neglects_q8_1d no_housework_q8_1e no_sex_q8_1f	FLBL
	pro_intervention	Mean of questions on pro-intervention attitudes: action_q8_1a_4 action_q8_1b_4 action_q8_1b_2_4 action_q8_1c_4 action_q8_1d_4 action_q8_1d_2_4	FLBL
	biblical_progressive	mean of scripture_obey and scripture_adam_eve	SC
Attendance		Mean of binary attendance variables indicating if each person in cohort 1 came to session: came_female came_male	SC
		Average number of attendees: - from photos sent by FLs of couples after the sessions - from attendance log filled out by FLs (c1_cupf_log and c1_cupm_log)	Photos
Program delivery	fl_session_fidelity	Mean of following variables coded to 0-1 scale: fl_session_fidelity: Is the FL doing the appropriate sessions according to the schedule? (yes/no) fl_follow_sess: How well does the FL follow to the order of activities in the session? (3scale) fl_follow_script: How well does the FL follow the script - key messages, stories, questions, answers? (3scale)	SC
	session_length	Length of the spotchecked session	SC
	fl_norm_b1	Binary indicators for whether FL conveys gender-equitable "norm" during B1: norm_sex	SC

		norm_equal norm_gifts norm_whore norm_submit norm_money	
	fl_performance	Mean of binary indicators: fl_prepared = Is the FL prepared for his or her session? fl_participate = Does the FL encourage participation from both genders? fl_confident = Is the FL a confident facilitator?	SC
	rating	Enumerator's subjective rating of how well is the FL able to deliver the message of the sesion	SC

9. COH

Some of the faith leaders completed another World Vision training, called Channels of Hope (COH), which encouraged them to incorporate more gender equitable themes into their sermons/messages; however it did not provide them with the means to lead sessions with couples. As a subgroup, we want to explore whether the effects on couples are stronger among faith leaders who participated in this training compared to those who did not. Our hypothesis is that the effects will be larger in these groups because the faith leader has some prior experience with similar material. This will pertain only to the main analyses. Because of endogeneity in the original selection into COH (i.e. it is likely that those FLs who participated in original COH training were different than those who did not across characteristics that may also influence their ability to deliver Becoming One) we cannot make any causal inferences about whether COH itself is responsible for any observed differences in effect size.

10. "Quasi-Solomon" survey experiment

As part of our baseline, we conducted a survey experiment to understand whether being asked questions about violence at baseline would influence responses at midline. Specifically, a random sample of 30% of the women in the baseline were not asked any questions on violence. We will regress our primary violence outcome on an indicator for whether the woman in the couple was asked a question about violence in the baseline, using the couple-level specifications 1 and 2 above. We will make our inference on the basis of a two-tailed p-value.

Broader than the question of whether the introduction of the violence questions specifically at baseline may influence responses at midline is the question of whether participating in the baseline at all may influence responses at midline. Therefore, we will compare the midline violence outcomes for the control couples (i.e. those in cohort 3) in the main sample to those of control couples who belong to the congregations of the faith leaders that were randomly assigned to be excluded from the baseline (see "Buffer Faith Leaders" above). We will regress our primary violence outcome on an indicator for whether the woman in the couple participated in the baseline survey, but limited to the sample of couples in cohort 3. We will make our inference on the basis of a two-tailed p-value constructed by generating the sharp null distribution generated by the random sampling mechanism for the baseline.

9. Robustness and Threats to Internal Validity

11. Alternative measures of violence

We will consider the effects of Becoming One on several alternative measures of violence as robustness checks of the results on the proportion experiencing any sexual or physical violence and the frequency measures specified in the primary and secondary analyses.

First, we will consider a multinomial measure of violence, developed by one of the authors (Heise), which classifies violence into the following types:

1. None - experienced no instances of violence in last 5 months,
2. Moderate - experienced only 1 form of "moderate" violence in the last 5 months,
3. Severe - experienced any "severe" form of violence or more than 1 moderate form in the last 5 months.

Where a "moderate" form is a slap or push, and a "severe" form is any of the other acts in the scale, excluding sexual violence.

Second, we will consider a measure of the "breadth" of violence experienced, defined as the proportion of the total number of acts experienced by the female partner over the preceding 5 months, including sexual violence. We will report the results of these checks in an online supplement / appendix.

12. Alternative measures of control and decision-making

We will also consider, as a robustness check for the results of the control and decision-making index, the effects on a simpler general measure formed by asking respondents to conceptualize a ladder with 10 rungs and use it to rate their level of control and decision-making on a 10 point scale. The definitions for these outcomes are provided below.

Outcome name:	control_ladder
Coding:	Integer values, 0-10 from variable vignette_1, divided by 10. A 0 represents the lowest rung on a ladder where stand "people who feel they have no free choice and no control over their lives," and on the step ten, "are people who feel they have completely free choice and total control over their lives."
Interpretation	At 0 the respondent expresses feeling no free choice and control over their lives, at 1 they express feeling fully in control.
Hypothesis:	Positive (one-tailed)
Analysis level:	Individual

Outcome name:	dm_ladder
Coding:	Integer values from vignette_2, 0-10, divided by 10. 0 represents the lowest rung on a ladder where stand "people who feel they have no decision making power," and on the step ten, "are people who feel they are able to make all decisions they wish."
Interpretation	At 0 the respondent expresses feeling they have no decision making power and at 1 they express feeling able to make all the decisions they wish.
Hypothesis:	Positive (one-tailed)
Analysis level:	Individual

13. Imputation for item-level missingness

In main analyses, non-response to outcome questions and covariates will be dealt with through imputation methods. We will impute missing items at the couple level by merging baseline and midline data at the couple-level, and conducting multivariate imputation via chained equations (MICE) as implemented in the `mice` package for `R`. Imputations will be performed using code such as the following:

```
mlw <- mlw %>% arrange(cup_id)
tmp <- mice(data = mlw[rows_to_impute, columns_to_impute],
            m = 1, seed = 7819135)
tmp <- complete(tmp)
mlw[rows_to_impute, columns_to_impute] <- tmp
```

Here, `mlw` stands for "midline, in wide format," meaning the couple-level midline dataset. The mice package is row- and column-sensitive, so we sort by couple ids before imputations. `rows_to_impute` and `columns_to_impute` are logical vectors that indicate which units and variables should be imputed. Importantly, we include all variables relevant to the analysis in the columns to be imputed **except** the treatment indicator, Z. We thereby avoid inducing any spurious associations between the treatment and the outcomes through the imputation. The rows to impute are those couples that were actually in the B1 randomization (more couples than were in the randomization were included in the baseline).

As a robustness check, we will include in the appendix a version of the results that deals with missingness through listwise deletion (dropping observations that have at least one missing value for the variables in the analysis).

i. Questions on sex

Our survey provides respondents an explicit opportunity to opt out of the questions on sexual activity. It is possible that there are simply reticent "types" in the population, in which case we can treat the non-responders to these items as belonging to a principal stratum unaffected by treatment, and limit our analysis to always-responders. However, it is also possible the B1 program will cause certain people to respond to these questions and not others. This is a special case of item-level missingness in which a large number of variables will be missing, and we do not deem imputation appropriate for the analysis of such outcomes.

In the case of these questions, will adopt the same approach as that used to deal with unit-level missingness as described below.

14. Approach to unit-level missingness (attrition)

There are four main ways we anticipate individuals will go missing from the midline data collection efforts:

1. Refusals.
2. Break-up.
3. Moving for other reasons.
4. Death.

We stipulate a strong belief that B1 will have no causal effect on death. Thus, for cases that attrit due to death, we will assume the missingness is unrelated to the treatment assignment and simply condition analyses to those alive.

Respondents in the baseline may attrit from the midline for reasons 1-3. Note here that 1 includes the case of refusing to answer the questions on sexual activity.

We conduct two tests.

First, we will perform a two-tailed unequal-variances t-test of the hypothesis that treatment does not affect the attrition rate. Per the Green lab SOP (https://github.com/acoppock/Green-Lab-SOP), we will implement the test as a permutation test that compares the observed t–statistic with its empirical distribution under thousands of repeated random reassignments of treatment.

Second, using a linear regression of an attrition indicator on treatment, baseline covariates, and treatment-covariate interactions, we will perform a heteroskedasticity-robust F-test of the hypothesis that all the interaction coefficients are zero.

The first test establishes whether missingness is related to the treatment, while the second test establishes whether missingness is related to baseline covariates.

If both tests result in a rejection of the null at the .05 level, we will report in an appendix estimates of covariate-adjusted Lee trimming bounds as well as extreme value (Manski-type) bounds. We will also report the analysis that was specified in the PAP using unit-wise deletion with inverse propensity weights that account for differential missingness of certain covariate profiles. Specifically, we will estimate the probability of attriting conditional on baseline covariates and treatment status using a logistic regression model on the full baseline sample of the form:

logit Pr(attrit = 1 | X, Z) = beta_0 + beta_1^T X + beta_2 Z + beta_3^T Z X

We select which baseline covariates to include via same LASSO procedure we use to determine which covariates to adjust for in covariate-adjusted estimator described above.

Finally, separate from the question of bias due to attrition is that of a loss of statistical power. If we experience a large amount of attrition -- say, at the congregation-level -- we will incorporate into our robustness analyses an analysis of our findings that includes the 54 couples from congregations in which we did not conduct a baseline survey.

15. Strategy for non-compliance

We define four mutually-exclusive and exhaustive types of individuals. We think of these types as unaffected by treatment: they are "principal strata."

Let $Z_i \in \{0,1\}$ denote an indicator for assignment to treatment and $D_i \in \{0,1\}$ an indicator for whether the respondent has attended any session of B1 prior to surveying. The principal strata are as follows.

Compliers	Attend at least one session during the first cohort when assigned to treatment, attend no sessions in first or second cohort when assigned to control.	$D_i = \begin{cases} 0 \text{ if } Z_i = 0 \\ 1 \text{ if } Z_i = 1 \end{cases}$
Always-Takers	Attend at least one session during the first or second cohort whether assigned to control or to treatment.	$D_i = \begin{cases} 1 \text{ if } Z_i = 0 \\ 1 \text{ if } Z_i = 1 \end{cases}$
Never-Takers	Never attend at least one session, irrespective of the cohort to which they were assigned.	$D_i = \begin{cases} 0 \text{ if } Z_i = 0 \\ 0 \text{ if } Z_i = 1 \end{cases}$
Defiers	Attend at least one session during the first or second cohort when assigned to control, and no sessions when assigned to treatment.	$D_i = \begin{cases} 1 \text{ if } Z_i = 0 \\ 0 \text{ if } Z_i = 1 \end{cases}$

From these individual-level strata we can construct a 4^4 couple-level typology. We are interested in two estimands of special interest: first, the effect of the treatment among compliers; second, the effect of the treatment on never-takers whose partners are compliers.

Denoting a stratum indicator $S_i \in \{\mathcal{C}, \mathcal{A}, \mathcal{N}, \mathcal{D}\}$ for compliers, always-takers, never-takers, and defiers, respectively, we can define these two "complier" and "spillover" estimands for partners i and j as follows:

1. Complier average treatment effect: $E[Y_i(Z_i = 1) - Y_i(Z_i = 0) \mid S_i = \mathcal{C}]$
2. Spillover to non-compliers: $E[Y_i(Z_i = 1) - Y_i(Z_i = 0) \mid S_i = \mathcal{N}, S_j = \mathcal{C}]$

We take two approaches to estimating these estimands, each relying on separate assumptions.

Approach 1: Assume no defiers and use instrumental variables.

In the first approach, we assume that there are no defiers in our sample: $\sum_{i=1}^{n} \mathbb{I}(S_i = \mathcal{D}) = 0$. In that case, the assignment variable Z exerts a monotonic effect on D. We can use the two-stage least squares estimator here as a consistent estimator of equation 1.

Approach 2: Compliance modeling

In the second approach, we again assume that there are no defiers, and use machine-learning techniques to build a predictive model for always-takers (using the control group data) and never-takers (using the treatment group data). Specifically, we will use Bayesian Additive Regression Tree (BART) classification methods, as implemented in the `bartMachine` package for R, with all available baseline data, in order to train a model capable of classifying control units as always-takers and treatment units as never-takers with a low prediction error rate. We will train the models using 10-fold cross validation, and then obtain predictions of always-takers in the treatment and never-takers in the control. Those who are predicted to be neither will be thought of as compliers. We will then estimate the estimands described above by subsetting to the relevant groups.

16. Multiple comparisons corrections

We use the term "testwise alpha" to refer to the probability that a given test rejects the null of no effect for all units (sharp null) when the sharp null is in fact true, and the term "familywise alpha" to refer to the probability that at least one test in a family of tests rejects the sharp null when the "global sharp null" is true.

The goal of our multiple comparison correction is to ensure that the testwise alpha is set to a level that ensures a familywise alpha of .05 and .10. We will not adjust p-values, but will report in the paper the testwise alphas below which our p-values would have to fall in order to reach the targeted familywise error rates.

To calculate the testwise alpha for the main outcomes analysis, we will use code such as the following:

```
library(DeclareDesign)
design <-
  declare_population(data = blwc) +
  declare_assignment(blocks = blocks, prob = .5) +
  declare_estimator(any_violence ~ Z,
                    se_type = "HC2",
                    fixed_effects = ~ blocks,
                    model = lm_robust,
                    label = "any_violence") +
  declare_estimator(control_index ~ Z,
                    se_type = "HC2",
                    fixed_effects = ~ blocks,
                    model = lm_robust,
                    label = "control_index") +
  declare_estimator(comm_index ~ Z,
                    se_type = "HC2",
                    fixed_effects = ~ blocks,
                    model = lm_robust,
                    label = "comm_index") +
  declare_estimator(consent_index ~ Z,
                    se_type = "HC2",
                    fixed_effects = ~ blocks,
                    model = lm_robust,
                    label = "consent_index")

simulations <- simulate_design(design,sims = 500)
```

This code essentially re-randomizes and re-estimates the treatment many hundreds of time -- and thus provides the exact sampling distribution of the estimates under the global sharp null of no effect of the treatment on any of the outcomes. Thus, any rejection is a false positive. It also provides the asymptotically-derived p-values on every simulation (we do not compute RI p-values on each run due to the enormous computational requirements).

Having simulated the distribution of families of p-values we would have obtained under the global sharp null, we are able to assess the familywise alpha that obtains given the application of a testwise alpha of *x_j*, where j indexes the number of testwise alphas considered. We use the following code to generate a familywise alpha given a stipulated testwise alpha.

```
get_alpha_per_family <- function(alpha_per_test){
  simulations %>%
    mutate(test_rejection = p.value <= alpha_per_test) %>%
    group_by(sim_ID) %>%
    summarize(family_rejection = any(test_rejection)) %>%
    ungroup() %>%
    with(., mean(family_rejection))
}
```

Now we can apply this function across the range of testwise alphas we might use.

```
alpha_per_tests <- seq(.001,.10,.001)
alpha_per_familys <- sapply(alpha_per_tests, get_alpha_per_family)
```

We have obtained a vector of familywise alphas that correspond to the x_j testwise alphas we considered. We set our familywise alpha target to.05 initially, then later could set it to .10. We calculate how far the proposed testwise alphas got us from our goal of .05, and then choose the largest testwise alpha that minimizes this difference.

```
target_alpha <- .05
dist_from_goal <- (alpha_per_familys - target_alpha)^2

# Choose biggest alpha which gets you closest to target alpha
new_alpha_per_test <- max(alpha_per_tests[which.min(dist_from_goal)])

new_alpha_per_test
```

We have now obtained the testwise alpha we would need to get a familywise alpha of .05 given: the members in our family of tests; the global null is true; and our tests may be correlated.

We will report these two alphas alongside a discussion of our main results, including the primary outcomes in the family of tests.

17. Balance tests

We will report balance on all available covariates unaffected by treatment in the baseline and midline, using specification 1 (Design-based specification) above, in which *Y* is the covariate. P-values will be calculated using randomization inference.

10. Using PAP to Guide Analysis

While the authors firmly believe that pre-specification of analytical strategies, whenever possible, promotes transparency and helps to reduce researcher degrees of freedom, it is unavoidable that there are exigencies in implementation or analysis that the researchers had not considered in advance. Similarly, it is often impractical that every data cleaning and analytic decision, no matter how minute, is captured. In this section we discuss additional strategies that we intend to employ during the cleaning and analysis phases to reduce the possibility of biasing our results.

18. Independent, blind analysis

Once post-treatment data have been collected, we will conduct an independent, blind analysis,[1] by this we mean that, independently from one another, two of the PIs will code and run the analysis as described in this pre-analysis plan using a “dummy” treatment assignment, i.e. a different random assignment vector that is completely unrelated to the actual treatment as delivered. We believe the independence of the coding procedure reduces the possibility of errors and/or bugs in the analysis code and ensures that two or more independent and reasonable approaches lead to the same results, while the blinded assignment ensures that any coding or analytic choices at this stage are independent of any knowledge of their influence over results on primary and secondary outcomes. This allows us to respond to new challenges in coding and to identify analytic approaches that will maximize our ability to answer the questions of interest without the concomitant risk that these decisions will inadvertently introduce bias.

The specific procedure that we intend to carry out for this independent, blind analysis is described below. To make the process transparent to other researchers and reviewers we will use Github to track each stage of the analysis.

1. Prior to receiving any post-treatment data from our institutional partner, IPA, the identifiers will be scrambled such that we, the PIs, will be unable to link them to the original assignment vector. IPA will retain the true identifiers throughout the blind analysis process and will reveal them once a final candidate analysis has emerged. (commit tag **start blinding**)

[1] Although relatively new in the context of the social sciences, blind analysis is common within a subset of the physical and biomedical sciences. Further discussion of the technique can be found in:

1. MacCoun, R., & Perlmutter, S. (2015). Blind analysis: hide results to seek the truth. *Nature News*, *526*(7572), 187.
2. MacCoun, R. J., & Perlmutter, S. (2017). Blind analysis as a correction for confirmatory bias in physics and in psychology. *Psychological science under scrutiny: Recent challenges and proposed solutions*, 297-322.
3. Nuzzo, R. (2015). How scientists fool themselves–and how they can stop. *Nature News*, *526*(7572), 182.
4. Leek, J., McShane, B. B., Gelman, A., Colquhoun, D., Nuijten, M. B., & Goodman, S. N. (2017). Five ways to fix statistics.

2. Once post-treatment data is received by the researchers and added into a folder of the repository encrypted through Boxcryptor, we will tag the commit such that changes from this stage are easy to track (commit tag **post-treatment received**)
3. Two of the PIs will create separate branches from the master branch of the analysis repository and begin an independent analysis with the blind assignment (commit tag **start independent analysis**)
    a. We will track any divergences from the PAP that take place during this process.
4. Once each author is satisfied with their code, they will begin to compare and resolve and discrepancies between the two (commit tag **end independent analysis**)
5. Eventually a candidate analysis with the blinded assignment will emerge that both authors are satisfied with (commit tag **final blind analysis**)
6. At this point the authors will merge in the true identifiers and produce results using the analysis constructed while using the dummy assignment (commit tag **unblinded analysis**)
7. Any other changes that happen past this stage will be considered “post-blind”

For clarity, we will use the following conventions when reporting results obtained during this procedure in any publications:

If the analysis is conducted...	**If the analysis procedure used is….**	
	Pre-registered	Not pre-registered
Blind	Pre-registered and blind	Exploratory and blind
Post-Blind	Pre-registered, post-blind	Exploratory and post-blind

Code appendix

This code illustrates the benefits of only adjusting treatment effect estimation for prognostic covariates. Vary rho to vary degree of correlation between Y and X.

```
rho <- 0

custom_estimator <- function(data,if_imbal = FALSE, if_pred = FALSE,and =
FALSE,thresh = .1){
  terms <- "Z"
  p_Z_x2 <- p_Z_x1 <- p_x2_Y <- p_x1_Y <- NA
  if(if_imbal){
    p_Z_x1 <- lm_robust(x1 ~ Z, data)$p.value["Z"]
    p_Z_x2 <- lm_robust(x2 ~ Z, data)$p.value["Z"]
    if(p_Z_x1 <= thresh){
      terms <- c(terms,"x1")
    }
    if(p_Z_x2 < .1){
      terms <- c(terms,"x2")
    }
  }
  if(if_pred){
    p_x1_Y <- lm_robust(Y ~ x1, data)$p.value["x1"]
    p_x2_Y <- lm_robust(Y ~ x2, data)$p.value["x2"]
    if(p_x1_Y <= thresh){
      terms <- c(terms,"x1")
    }
    if(p_x2_Y < .1){
      terms <- c(terms,"x2")
    }
  }
  if(and){
    terms <- "Z"
    if(p_x1_Y <= thresh & p_Z_x1 <= thresh){
      terms <- c(terms,"x1")
    }
    if(p_x2_Y <= thresh & p_Z_x2 <= thresh){
      terms <- c(terms,"x2")
    }
  }
  terms <- unique(terms)
  tidy(lm_robust(reformulate(termlabels = terms, response = "Y"),data =
data)) %>%
    filter(term == "Z") %>%
    cbind(., x1_imbal = p_Z_x1 <= thresh, x2_imbal = p_Z_x2 <= thresh)
}

design <-
  declare_population(
    N = 300,
    x1 = rnorm(N),
    x2 = rnorm(N),
    Y0 = rnorm(N, x1 * rho, sqrt(1 - rho^2))
  ) +
  declare_potential_outcomes(
    Y_Z_0 = Y0,
    Y_Z_1 = Y0 + .2
  ) +
  declare_estimand(ate = mean(Y_Z_1 - Y_Z_0)) +
  declare_assignment(prob = .5) +
  declare_reveal(Y, Z) +
  declare_estimator(
    handler = tidy_estimator(custom_estimator),
    if_imbal = FALSE,
    if_pred = FALSE,
    estimand = "ate",
    label = "No x"
  ) +
  declare_estimator(
    handler = tidy_estimator(custom_estimator),
    if_imbal = TRUE,
    if_pred = FALSE,
    estimand = "ate",
    label = "x if imbalanced"
  ) +
  declare_estimator(
    handler = tidy_estimator(custom_estimator),
    if_imbal = FALSE,
    if_pred = TRUE,
    estimand = "ate",
    label = "x if predictive"
  ) +
  declare_estimator(
    handler = tidy_estimator(custom_estimator),
    if_imbal = TRUE,
    if_pred = TRUE,
    and = TRUE,
    estimand = "ate",
    label = "x if predictive and imbalanced"
  )
```

```
design


diagnosis <- diagnose_design(design,sims = 5000)

diagnosis
```

This code illustrates power gains from indexing.

```
library(DeclareDesign)
design <- declare_population(
  N = 200,
  Y1_Z_0 = rbinom(N, 1, .2),
  Y1_Z_1 = rbinom(N, 1, .25),
  Y2_Z_0 = rbinom(N, 1, .3),
  Y2_Z_1 = rbinom(N, 1, .35),
  Y3_Z_0 = rbinom(N, 1, .5),
  Y3_Z_1 = rbinom(N, 1, .55),
  Y4_Z_0 = rbinom(N, 1, .5),
  Y4_Z_1 = rbinom(N, 1, .55)
) +
  declare_assignment(prob = .5, assignment_variable = "Z") +
  declare_reveal(
    outcome_variables = c("Y1","Y2","Y3","Y4"),
    assignment_variable = "Z"
  ) +
  declare_step(Y_index = rowMeans(cbind(Y1,Y2,Y3,Y4)),
               handler = fabricate) +
  declare_estimator(Y1 ~ Z, label = "one outcome") +
  declare_estimator(Y_index ~ Z, label = "index")

diagnose_design(design)
```