Analytical Methods for the IMPower and Sources of Strength cluster randomized controlled trial

Evan Rosenman, Clea Sarnquist, Rina Friedberg, Mary Nyamongo, Gabriel Oguda, Dorothy Otieno, Michael Baiocchi

Stanford University

August 8, 2018

Contents

1 Background

Sexual assault is a serious public health issue among adolescents. Worldwide, more than 10% of girls experience forced intercourse at some point in their lifetimes (UNICEF et al., 2014). Prior work has shown victimization rates as high as 31% among African girls, and perpetration rates as high as 40% among African boys (Shamu et al., 2015).

Research on sexual assault has typically focused on intimate partner violence among adults or older adolescents (Livingston et al., 2007) and typically in high-income countries (Koss et al., 2007; Ellsberg et al., 2015). The literature offers a more limited understanding of sexual assault prevalence and prevention among children at an earlier stage of development. Vulnerable subgroups, such as economically and socially disadvantaged children living in informal settlements, are further understudied.

Previous work has established the effectiveness of sexual assault prevention programs in early adolescents in these settings (Jewkes et al., 2014). In particular, a randomized trial indicates that such programs can reduce the rate of rape in the prior twelve months by 3.7% (95% confidence interval: [0.4%, 8.0%]) from a baseline of 7.3% (Baiocchi et al., 2017). Such interventions – referred to as "empowerment" programs – place emphasis on (i) awareness of autonomy and right to refuse, (ii) understanding of common patterns of perpetration, and (iii) skill development to appropriately intervene and mitigate perceived threats.

This paper concerns data from a stratified sample of student participants in a cluster-randomized trial underway in five informal settlements around Nairobi, Kenya. The clusters in this intervention are the schools attended by the boys and girls. The tested intervention is a combination of a previously-tested girls' intervention, IMPower, and a newly revised boys' intervention, Source of Strength.

Analysis of baseline data includes establishing point estimates and confidence intervals for outcome measures and covariates at baseline. Additional work involves an exploratory analysis of modifiable causal pathways (a.k.a. "mediating variables") for rape, and assessing predictors of each school's need for the intervention Methodological challenges arise around several key areas in this work, including estimating uncertainty, data adjudication, and modeling risk factors. This paper details the methodologies used to address each of these challenges.

2 Related Literature

The design and analysis of surveys to gauge experiences of sexual assault is an active topic of research (Cook et al., 2011; Davis et al., 2014). The most widely used tool is the Sexual Experiences Survey, which measures unwanted sexual experiences from age 14 onward (Koss et al., 2007). In this and other surveys (Brener et al., 1999), the potential for inconsistent responses is a well-known difficulty. For example, Fisher and Cullen (2000) found that when surveying college women about their experience with sexual assault, nearly 20% of incidents initially classified as rape were later classified as "undetermined" because the respondents provided inconsistent responses to follow-up questions. They hypothesize that "some 'real' rape victims were not counted because they did not understand questions in the incident report or wearied at having to answer a second round of questions about a potentially painful event in their lives" (Fisher and Cullen, 2000).

Due to these inconsistencies, researchers have suggested balancing the need for information collection against the possibility that follow-up questions may suppress disclosure (Koss et al., 2007). Current adjudication procedures frequently involve categorizing respondents according to their most severe reported assault experience (Davis et al., 2014). Researchers also make use of Cronbach's Alpha (Cronbach and Meehl, 1955), a measure of the internal consistency of survey data, to validate that the responses are reliable (Alpha scores greater than 0.7 are often considered acceptable). However, explicit best practices for conflicting responses were not found in a review of the relevant literature.

The cluster-randomized setting of this trial also presents several statistical challenges that are absent from studying treatments randomized to individuals. For a full treatment, see Gail et al. (1996). Chief among these challenges is homophily, the tendency of individual units within a cluster to be more similar to one another (in terms of pre-intervention covariates) than to units in other clusters. Because of homophily, randomization can be challenging, as "cluster randomization does not guarantee balance of potential person-level confounders between conditions" (Vuchinich et al., 2012).

A second set of issues in cluster-randomized trials concerns multiple sources of variability. It is tempting in this setting to compute summary statistics for each cluster and then to treat each of these statistics as independent draws from a population. But if clusters are very different sizes or if individuals are more variable within some clusters as compared to others, then this approach may lead to incorrect inference. Gail et al. (1996) discuss a number of situations in which standard randomization tests will fail to achieve the nominal significance level.

A potential solution is found via bootstrapping, originally proposed by Efron (1992) as a way to compute the uncertainty of a sample estimate of a target parameter. In a bootstrap procedure, "pseudo-datasets" are constructed by resampling with replacement from the original datasets. Estimates of the target parameter are then computed on each of the pseudo-datasets, and the spread of these estimates allows us to compute confidence intervals for the original sample estimate. In the cluster setting, researchers make use of a variant – the clustered bootstrap – which permits heteroskedasticity and within-cluster error correlation as long as there is a reasonably large number of clusters (Cameron et al., 2008).

Research in this area also frequently makes use of structural equation modeling (SEM) to identify pathways that lead to sexual assault (Shamu et al., 2015; Russell et al., 2014). SEM, also known as causal modeling is a collection of statistical techniques which allow researchers to tease out linear causal relationships between a set of independent variables and dependent variables. Using SEM, a researcher can posit a causal relationship between variables (represented by a directed graph), estimate coefficients associated with such a model of the data, and obtain goodness-of-fit measures to evaluate the plausibility of the relationships (Ullman and Bentler, 2012). SEM is widely used in the social sciences, and various guidelines exist for its practical utility (Anderson and Gerbing, 1988). In the context of sexual violence, a typical analysis might use SEM to identify the ways that gender attitudes mediate the relationship between socioeconomic status and rape frequency.

While SEM is quite flexible, it also imposes parametric assumptions that can be indefensible and biasing in practice (Ullman and Bentler, 2012). Alternatively, one can use a matching-based approach to the principal stratification technique proposed by Frangakis and Rubin (2002) in order to estimate the effect of intermediate variables. In this approach, a matching algorithm (Lu et al., 2001) is used to create pairs of individuals such that they are as similar as possible on all variables except a candidate modifiable variable of interest. The algorithm is nonbipartite, meaning that girls can be used as matches more than once. Computing differences on an outcome measure across the matched pairs, we attempt to isolate the relationship between the outcome and our candidate variable. While such an approach can emphatically not establish causality in the absence of a randomized controlled trial (nor can SEM), by isolating variation from baseline covariates and other candidate causal pathways, it may more reliably generate hypotheses for causal pathways to sexual assault.

3 Study Design

3.1 Partners and Materials

These data come from a two-arm, parallel cluster-randomized trial (CRT). The treatment is a combination of a previously-tested girls' intervention, IMPower, and a newly revised boys' intervention, Source of Strength. Clusters are defined as the schools which serve students from the informal settlements. Participants are adolescent girls and boys in class 6, generally between the ages of 10-14 at baseline.

Data collection for the study began in January 2016 and will continue through December 2018. The baseline data period collected data from January 2016 through October 2016. The study is being conducted in six informal settlements around Nairobi: Dandora, Huruma, Kariobangi, Kibera, Korogocho, and Mukuru. Due to proximity, and a small number of schools in Kariobangi, all subsequent analysis will classify schools in Kariobangi as being part of Dandora. Two of the five participating organizations are: Kenya-based, non-governmental organizations Ujamaa Africa and the African Institute for Health and Development (AIHD). Ujamaa implemented the intervention and AIHD is the Nairobi-based research partner for the evaluation component of the study. The San Francisco based NGO No Means No Worldwide (NMNW) developed the intervention. Stanford University is the external evaluator of the intervention.

The primary outcome for the CRT is the change in incidence of self-reported rape among girls from baseline, compared to a life skills standard of care intervention. Secondary outcomes include determining the impact of the interventions on self-efficacy, self-esteem, gender attitudes and beliefs, and experiences of physical and emotional violence. The data reported here were collected before any trainings.

Ethical approval for the study was given by the KEMRI and Stanford IRBs. School administrators determined the suitability of the program for their students. For the evaluation component (i.e., participation in surveys) learners were randomly sampled within school using a predetermined number of colored marbles to indicate selection into the study. The girls were then informed of their rights and potential risks and invited to participate; those that assented into the study then had their parents and caregivers given the opportunity to decline student participation. The trial is registered at ClinicalTrials.gov #NCT02771132. This research was supported by the South African Medical Research Council through the What Works to Prevent Violence Innovation Grant (#52069) as a result of the support obtained from the Secretary of State for International Development at the Department for International Development. This research was also conducted with Government support under and awarded by DoD, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a. We are grateful for support from the Marjorie Lozoff Fund, Michelle R. Clayman Institute for Gender Research, Stanford University. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of the Secretary of State for International Development at the Department for International Development or the South African Medical Research Council, or any other funding agencies.

3.2 Measures

There are few validated scales available for this group (10-14 year old girls living in informal settlements) regarding sexual abuse, and variables on the pathway to sexual assault. Therefore, the study draws from a variety of sources in creating the surveys. In order to measure the primary outcome, change in incidence of self-reported sexual assault among girls from baseline, the study creates an index based on several variables. Key surveys that informed these questions included the Kenya VACS (Kenya, 2012) and surveys from the Stepping Stones project in South Africa (Jewkes et al., 2014).

Survey items include experiences of intimate partner violence, or IPV (for example, "In the past 12 months, how many times has your current or a previous boyfriend physically forced you to have sex when you did not want to?" and "In the past 12 months, how many times has your current or a previous boyfriend used threats or intimidation to get you to have sex when you did not want to?"). They also include sexual violence by non-partners ("In the past 12 months, how many times has a man who is NOT your boyfriend forced or persuaded you to have sex against your will?" and "In the past 12 months how many times was there an occasion when you agreed to have sex with one man or boy and one or more others who you had not agreed to have sex with forced you to have sex with them as well?").

Validated, widely used scales are more readily available for some of the secondary outcomes. Specifically, we used the "Self-Efficacy Questionnaire for Children" (Muris, 2001) for self-efficacy, and the "Rosenburg Self-Esteem Scale" (Rosenberg, 2015) for self-esteem. For experiences of social and emotional violence, we relied on the VACS and Stepping Stones surveys. For gender norms and relations, some questions were removed from existing surveys and modified for this population during piloting, and others were created based on the field team's knowledge of this population, then modified during piloting.

4 Methods

4.1 Adjudication

The baseline survey contains several questions pertaining to rape, covering a range of possible assailants and situations (for example: perpetrator, methods of coercion, threat types, physical assault during encounter, use of alcohol or drugs during assault). The survey also asks the girls how many times they have been assaulted, at the end of that line of questioning. Several logical inconsistencies are present in student-provided answers. For example, some girls answered they had been sexually assaulted in at least one of the aforementioned scenarios, and then reported 0 or did not answer for how many times they had been assaulted in their lives. This particular pattern is present in about 10% of the surveys where a girl reported an assault. These inconsistencies in reporting require consideration.

It is important to note that this data adjudication process is being put into place at this point in the research project, using the pre-intervention survey data. As data from the next phases of the survey are collected, this same procedure – implemented using the same code we developed for the baseline survey cleaning – will be replicated. What follows below is our work on justifying and validating our adjudication process.

It appears that some respondents interpreted the questions as "asked and answered," meaning that if they answered affirmatively on an earlier question then they removed that experience from consideration when answering future questions that might be overlapping. In the data, it was observed that more than half (151 respondents out of the 253 who provided seemingly contradictory responses) had a yes followed by a series of no responses (that is, consistent with "asked and answered"). The remaining answer patterns appear to show some confusion in how the respondents counted the number of assaults. Given this line of reasoning, we considered to adjudicate any inconsistent answers into an indication of a rape.

Formalizing the above: we classified a respondent as having been raped if a girl gave a nonzero answer to any of thirteen questions indicating frequency of forced sex. To estimate the number of times a girl has been raped, we took the maximum between her answer to "How many times in your life have you been forced to have sex against your will?" and the sum across her answers to the previous questions.

To empirically validate our adjudication procedure, we analyzed the covariate profiles of girls with conflicting survey responses. We built a logistic regression model predicting the likelihood of sexual assault, using as predictors all covariates other than reports of forced sexual activity. We refer to this as the "Adjudication Model." We fit our model using only fully-consistent reports from girls who were (i) raped (n = 194) or (ii) not raped (n = 3,679), excluding the 253 reports with some internal inconsistency. Using this fitted model the 253 reports with some internal inconsistency were then predicted as "out of sample" observations.

The logistic regression model was fit to the data using the `glm()` function in R. In all, 63 covariates were provided as candidates for the model (56 were numeric variables, and seven were categorical). A forward stepwise variable selection procedure, using AIC as the goodness of fit criterion, was used for variable selection. Fourteen variables were selected by the model, including many indicators of an abusive relationship ("Insulted by boyfriend," "Threatened with a weapon by boyfriend") as well as indicators of domestic strife and drug use. Many variables were not included by the stepwise procedure, including socioeconomic indicators ("household size," "parents' education"), the majority of self-efficacy measures, and all measures of gender normative thinking.

To validate the accuracy of the Adjudication Model, we used a ten-fold cross-validation. The data was split into ten mutually exclusive equal-sized sub-samples (or "folds") via random permutation. We then looped over each fold, removing it from the dataset and fitting the model to the remaining 90% of the data. The model was used to predict the probability of being raped for each girl in the held-out fold, and the resultant predictions were recorded. This process was coded using built-in R functions, rather than any cross-validation library.

The predictions for each girl were then compared against the true reports to assess the model's accuracy. This is summarized in the ROC curve below, generated with the `roc()` function from the `pROC` library (Robin et al., 2011). The curve is a common measure of performance in binary classification problems. For any classifier, the ROC plots the false positive rate (i.e. the percent of the girls who are classified as raped who were not actually raped) that must be tolerated in order to achieve a desired true positive rate (i.e. the percent of all girls who were raped that are identified as such by the classifier). If the ROC line is arced high above the 45-degree line, this indicates good performance, summarized by an AUC ("area under curve") near 1.0. If the line is essentially on top of the 45-degree line, this indicates performance no better than a random guess, with an AUC near 0.5.

In our case, the line is inflated relative to the 45-degree, and the AUC is 0.78, indicating moderately good performance. The model is imperfect: predicting an outcome like rape is challenging. The model identifies broad risk factors that make it more likely for a girl to be victimized, but cannot fully determine all of the factors that may result in a girl being victimized.

Now we apply Adjudication Model to the 253 girls with conflicting reports. Averaging across all 253 girls, we get an average prediction of 19.4% probability of sexual assault. Among the girls in the dataset on which the model was built – which contained only reports without any internal inconsistencies – 5.0% had been raped, so the average is four times higher than the average from the dataset. We also applied the model to the data to which it was fit, finding that the model predicted an average 22.1% probability of rape among the non-contradictory records from girls who were raped; and 4.1% probability among the records of girls who were not raped.

Restating the above: the average prediction for the population of girls who gave contradictory rape

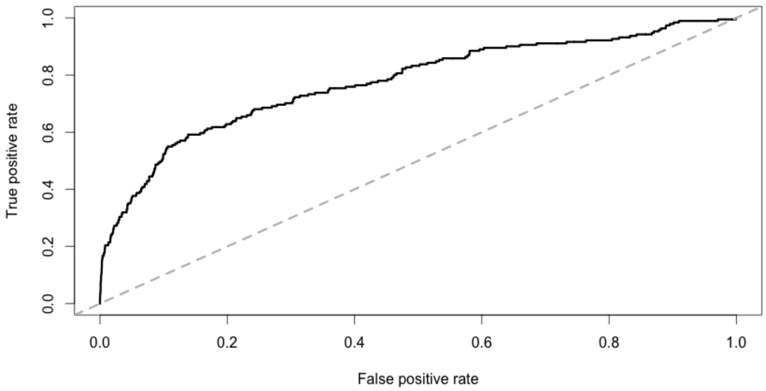

Figure 1: ROC curve for the Adjudication Model. The area under the curve (AUC) is 0.78.

Table 1: Comparison of average predicted rape probability by subgroup, using the adjudication model

Group	**Number of Girls**	**Average Predicted Probability of Rape**
Girls reporting rape, with consistent responses	194	22.1%
Girls reporting rape, with inconsistent responses	253	19.4%
Girls not reporting rape	3,679	4.1%

reports is far closer to the average prediction for the girls with non-contradictory rape reports than to the girls who were not raped. This lends further credence to our adjudication choices.

4.2 Bootstrap standard errors

With the adjudication process in place, point estimates of baseline rape prevalence can estimated using the survey data. As these measurements occurred prior to the intervention, such estimates combine both treatment levels from the study. Though difficult to assert due to the dynamism of the environment, we believe the study achieved inclusion of nearly all primary schools in the five settlements. Within each school, girls were sampled using a simple random sampling procedure for all girls in Class 6 who were in attendance on the day of study initiation at the school.

Standard confidence intervals, for example the familiar Wald interval, fail to account for the structure of this study's sampling procedure, and are thus unlikely to produce accurate measurements of sampling error. In the example at hand, we expect significant homophily within schools; that is, we expect two girls in the same school to be more similar to each other, as compared to two girls from two different schools. Hence, treating each girl as an independent observation will lead to a nave variance estimate that we should be hesitant to trust. In the example at hand, we might expect Wald standard errors to be inappropriately small.

There is a straightforward and intuitive statistical fix in the clustered bootstrap, which is appropriate in our setting because our number of schools exceeds the threshold of thirty described by Cameron et al. (2008). The bootstrap standard error estimate, as opposed to formulaic estimates, is a Monte-Carlo estimate which means it is derived from a series of random samples. In the example at hand, we sample individuals within schools with replacement. The number of students from a given school stays the same, but within in a school it may be that a student is replicated a few times or not at all. This sample-resample procedure can be used to generate many pseudo-datasets that should have covariate patterns similar to the one we observed in our actual data set, because we imposed the important clustered structure. Observing the proportion of girls who have been raped in each of these bootstrapped, pseudo-datasets yields a sampling distribution for the observed proportion. In a formula-driven estimate of sampling variance, such as a Wald confidence interval, this sampling distribution is assumed to be generated from a Normal or Binomial distribution; here, we generate the sampling distribution using information from our data, and use this to estimate standard errors.

4.3 Assessing candidate causal pathways

The dataset being analyzed is cross-sectional – i.e., the data are collected at one point in time rather than at several time points ("longitudinally"). It is challenging to use such a dataset to obtain causal insights without researcher-directed randomization. Thus we do not attempt an analysis that establishes causal links between proposed causal pathways and the outcome rape. Instead, we attempt to isolate variation in the outcome (rape) as much as possible so as to inform the relative contribution of candidate causal pathways. At the end of the CRT study, there will be three timepoints measured for participants, and similar causal pathway analyses will be performed.

Based on theoretical models of prevention, as well as findings in prior literature, we identified four candidate modifiable variables which may prevent rape: former experience of IPV, generalized self-efficacy, prior alcohol use, and belief in gender norms. These are girl-specific experiences and beliefs, not measures of the entire group. Self-efficacy was estimated using the girl's average response across three questions related to her ability to keep herself calm, rely on social networks, and deal with unpleasant thoughts. Higher scores represented greater self-efficacy. Gender-normative thinking was estimated by the girl's average response across twelve questions about the relationships between women and men, with higher scores representing greater gender-normative thinking. Alcohol use was estimated using a girl's self-reported number of times consuming alcohol. Intimate partner violence was estimated as the average of a girl's response across three questions about whether she had been slapped, insulted, or intimidated by a boyfriend.

There are three types of variables in this analysis: (i) covariates – i.e., socioeconomic indicators, parental violence, whether parents are alive or dead, and intimate partner status; (ii) candidate causal pathways – i.e., prior IPV, alcohol and drug use, generalized self-efficacy, and belief in gender norms; and (iii) outcome – i.e., whether a girl has been raped. Because there are four candidate modifiable variables under consideration, we perform four separate analyses. For each candidate modifiable variable of interest, girls were matched one-to-one based on a pool of covariates. These included the three other pathway variables; a set of variables reflecting a girl's experience of parental violence, whether her parents were alive or dead, her socioeconomic status, and whether or not she had a boyfriend; and missing value indicators for each of the aforementioned variables.

For the matching variables, missing values were imputed with a large value outside of the usual range of the variable (e.g. 100). This encourages matches between girls who were both missing the value. For the pathway variable, girls were dropped from the analysis if they were missing a value for the pathway variable; this resulted in a loss of no more than 11% but no less than 6% of the girls for each pathway explored.

For example, to isolate the effect of prior IPV on probability of rape, we attempted to match two girls who were nearly identical on all covariates as well as identical on alcohol and drug use, generalized self-efficacy, and belief in gender norms. At the same time, we also tried to make sure the two matched girls were as different as possible in their prior experiences of IPV. Thus the two sets we create in this analysis are nearly identical in all ways, expect for having quite different experiences in terms of prior IPV. We then compare the probability of experiencing rape in the prior twelve months.

We repeat the analysis described in the previous paragraph three more times – for alcohol and drug use, for generalized self-efficacy, and for belief in gender norms.

The non-bipartite matching algorithm was implemented using the `nbpMatching` package in R (Beck et al., 2016). Adequate balance was assessed using standardized mean difference, with values with magnitude less than 0.10 deemed acceptable balance. We were able to force separation of standardized mean differences for prior IPV, generalized self-efficacy, alcohol use, and gender norms of 0.48, 2.97, 0.67, and 2.41 respectively.

4.4 Assessing and proposing predictors of school-level prevalence of rape

Two analyses are performed to provide insight on which kinds of schools interventionalists may want to approach first with the treatment program. That is, we try to find the best predictors of school-level rape prevalence. The first analysis uses easily accessible school-level covariates (e.g., number of students, test scores, building materials) to predict the school-level prevalence of rape. The second analysis starts with individual girls' responses, aggregates these responses to school-level variables, and then predicts. The second analysis acknowledges that the questions asked of individual-level girls are much more private and difficult to obtain from school administrator before building trust within the school. But, the analysis of these aggregated responses – and their predictive power – may be useful for developing new questions interventionalists can ask of the administrators to assess the school's prevalence of rape before entering the school.

To investigate the associations between school-level covariates and school-level rape prevalence, univariate linear regressions were computed, in which the sexual assault frequency was the dependent variable and each covariate the independent variable in its own model. Fourteen regressions were fit in total, using the standard `lm()` function in R. For numeric variables, the reported p-value is that of the t-test for the coefficient. For categorical variables, the reported p-value is that of the F-test for all levels of the variable. Coefficient signs were also reported for all numeric variables. All regressions were weighted by the number of girls interviewed at the school, down-weighting the influence of extreme values realized due to small sample sizes for small schools. The results were not corrected for multiple comparisons, though the majority of covariates were nonetheless not significant at the uncorrected $p = 0.10$ level.

An identical approach was utilized to understand the marginal relationship between aggregated girl-level covariates and rape frequency. Covariates observed at the individual level (such as whether each girl has a boyfriend) were aggregated up to school-level averages (such as "percentage of girls who have a boyfriend"). Univariate regressions were then fit using school-level rape frequency as the dependent variable and each of these aggregated covariates as the independent variables.

Lastly, to understand the relative predictive ability of these variables conditional on the presence of the other variables, we fit a LASSO model (Tibshirani, 1996) to the union of the school-level and aggregated individual-level covariates. The model was fit using the `cv.glmnet()` function in the `glmnet` package in R (Friedman et al., 2010). Lambda, the regularization penalty, was selected using the default behavior, "lambda.1se," which chooses the largest value of lambda such that the cross-validated error is within a single standard error of the minimum. This enforces relatively aggressive regularization. Such a procedure typically generates parsimonious models with relatively few nonzero coefficients, and can be viewed as an alternative to a forward or backward stepwise variable selection procedure. Hence, the variables selected for model inclusion tend to be those most predictive when accounting for the effect of the other variables. Using such a model allows us to get a sense of which variables remain predictive even in the presence of other predictors.

5 Results

Below, we discuss results obtained at the baseline of the trial, with a particular focus on statistical methodology. For a more in-depth discussion of the baseline results from a public health perspective, see Baiocchi et al. (2018) (**SELF-CITATION**).

5.1 Adjudication

Our adjudication procedure is consequential insofar as it alters the baseline estimate of rape prevalence. As discussed in 4.1, we found empirical support for our adjudication procedure by fitting a predictive model of rape to all nonconflicting reports. We validated the model via cross-validation, and applied it the model to the girls who gave conflicting reports of rape, finding that these girls' covariate profiles were much similar to those of rape victims with nonconflicting reports than to non-victims.

Thus, all girls who reported rape were designated as rape victims, even if their survey responses were self-contradictory. This procedure yields a baseline estimate of rape prevalence of 10.83%. Had only nonconflicting reports been counted, the estimate would have been less than half that: 4.70%.

5.2 Bootstrap standard errors

The effect of using bootstrap standard errors can be seen in our uncertainty estimates. Ninety-five percent confidence intervals for a selected set of variables can be found in Table 2.

In most cases, the bootstrap confidence intervals are similar in width the Wald Intervals, though they are somewhat wider in the case of measuring the proportion of girls who have had a recent boyfriend. The bootstrap intervals are also not necessarily symmetric about our point estimate, while the Wald intervals are symmetric by definition. This asymmetry may help us better quantify the direction of possible uncertainty.

Table 2: Point estimates and confidence intervals for selected variables, using both Wald and bootstrap procedures

Quantity	**Point Estimate**	**Wald 95% CI**	**Bootstrap 95% CI**
Proportion of girls who have had a boyfriend in last six months	21.34%	(20.08%, 22.61%)	(19.75%, 23.72%)
Proportion of girls who have been forced to have sex	10.83%	(9.79%, 11.68%)	(9.81%, 11.68%)
Proportion of girls who have consumed alcohol	11.67%	(10.69%, 12.65%)	(10.74%, 12.62%)
Average gender norms score	2.48%	(2.46%, 2.50%)	(2.47%, 2.49%)
Proportion of girls who have tried drugs	0.71%	(0.45%, 0.96%)	(0.46%, 0.97%)
Proportion of girls who have had intercourse	7.36%	(6.56%, 8.15%)	(6.59%, 8.16%)

5.3 Assessing candidate causal pathways

We interrogated candidate modifiable causal variables, making use of the nonbipartite algorithm discussed previously. These analyses reduce differences in all variables except the candidate modifiable variable of interest, attempting to isolate the relationship between the outcome and this target variable.

We explored four potential causal pathways: the effect of a girl's self-efficacy, gender-normative thinking, alcohol use, and experience of intimate partner violence. For each candidate pathway, we matched a control girl to every treated girl. Once the matches were computed, balance on the match variables and the pathway variable were assessed. The balance table for the analysis in which IPV is the candidate causal pathway variable can be found in Table 3. The remaining tables can be found in the appendix.

Covariate balance was quite good in all four analyses; all covariates had absolute standardized mean differences of less than 0.10. To compute the relationship between the candidate modifiable variable of interest and the outcome, we computed the percent of girls who were raped among those in each pair who had the lower value of the pathway variable. We then found the analogous quantity for the girls with the higher value of the pathway variable, and computed the difference.

The results are summarized in Figure 2 below. There is not a natural quantity on which to compare the four candidate modifiable variables. It might seem that one could standardize the variable to shift in one standard deviation, but that may be an unrealizable goal for an intervention to obtain. We were able to force separation of standardized mean differences for prior IPV, generalized self-efficacy, alcohol use, and gender norms of 0.48, 2.97, 0.67, and 2.41 respectively. These values can be read off of the balance tables in the appendix. Future interventionalists would need to assess how realistic it is for a given intervention to shift the candidate modifiable variable. As we can see below, intimate partner violence and alcohol use were associated with a substantially higher proportion of girls being victimized. Self-efficacy was associated with a substantially lower proportion. The association between gender-normative and rape frequency was positive but smaller in magnitude. It may be the case that shifting gender norms within an individual, without shifting gender-norms for the surrounding community, does not have a profound effect on outcomes.

It should be repeated that these are exploratory analyses and no claim of establishing causality is being made here. It is entirely possible that the direction of causality is in the reverse of how it is discussed above.

Table 3: Covariate balance, near-far matching algorithm, for intimate partner violence candidate modifiable variable exploration.

Covariate	**Higher value (n = 1919)**	**Lower value (n = 1919)**	**Absolute Standardized Difference**
Intimate partner violence	0.103	0.001	0.475
Parental violence: missing indicator	0.005	0.004	0.016
Parental violence	0.178	0.170	0.020
Socioeconomic status: missing indicator	0.001	0.002	0.014
Socioeconomic status	2.853	2.855	0.002
Mother dead: missing indicator	0.006	0.006	0.007
Mother dead	0.034	0.034	0.000
Father dead: missing indicator	0.045	0.045	0.000
Father dead	0.095	0.095	0.000
Has boyfriend: missing indicator	0.018	0.018	0.000
Has boyfriend	0.208	0.207	0.004
Times consumed alcohol: missing indicator	0.008	0.008	0.000
Times consumed alcohol	0.100	0.100	0.000
Self efficacy: missing indicator	0.016	0.016	0.000
Self efficacy	3.767	3.774	0.007
Gender normative thinking: missing indicator	0.053	0.052	0.005
Gender normative thinking	2.345	2.348	0.006

5.4 Assessing and proposing predictors of school-level prevalence of rape

One trend immediately evident in the data is the high degree of variation in rape rates across the schools. At eight of the schools, none of the girls reported being raped, while at four of the schools, more than 30% of the interviewees reported being raped. The overall distribution is shown in the left panel in Figure 3. The mean of this distribution (weighted by the number of girls interviewed at each school) is 10.8%.

Certainly some of this variation in estimated school-prevalence is attributable to the low number of girls interviewed at certain schools. With few interviewees, it is substantially more likely to get an extremely high or extremely low proportion of girls who report having been raped. If we exclude all schools with fewer than 20 interviewees, the most extreme values disappear from the distribution (as can be seen in the right panel in Figure 3). But there remains substantial heterogeneity between schools.

This heterogeneity is significant from a public health perspective. If the intervention is found to be effective at reducing the incidence of sexual assault, then schools with higher rates of sexual assault should be prioritized first for deployment of the intervention. But, as was mentioned earlier, there remains a technical challenge in obtaining estimated sexual assault rates or even reliable predictors at a given school, as this information is frequently quite sensitive to obtain and requires investments in trust and relationship-building at a given institution. In the interest of more effective triaging of resources, we explore the relationship between rape frequency and several easier-to-obtain administrative variables. While most variable definitions should be intuitive, a full list of the variables and how they were computed can be found in the appendix. To understand the variation in rate of rape from school-to-school, we analyzed the associations between each of our covariates and the sexual assault frequency at each school, using the techniques described in our methods section. These associations were computed via univariate linear regressions, where the prior rape frequency is the dependent variable and each covariate the independent variable in its own model (so 14 regressions were fit in all). Regressions were computed in R, using the standard `lm()` function. For numeric variables, the reported p-value is that of the t-test for the coefficient. For categorical variables, the reported p-value is that of the F-test for all levels of the variable. Coefficient signs were also reported for all numeric variables. All regressions were weighted by the number of girls interviewed at the school, down-weighting the influence of extreme values realized due to low sample sizes. Results are given in Table 4:

No correction was made for multiple comparisons, but most variables were nonetheless not significant at

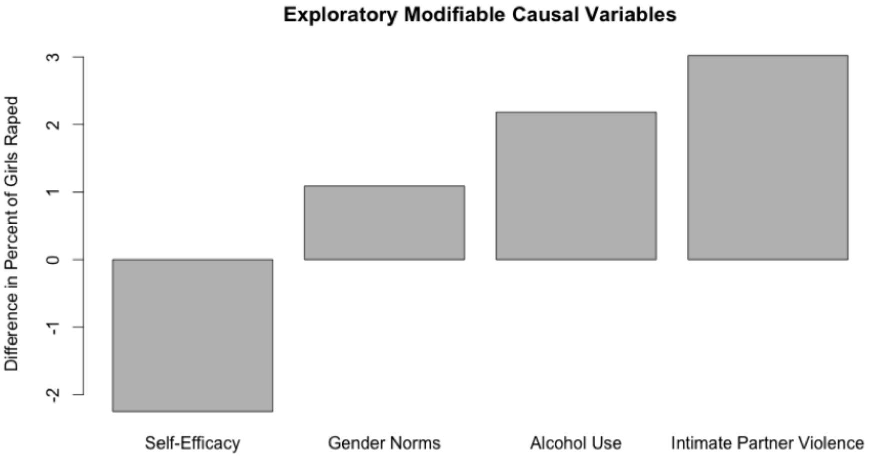

Figure 2: Difference in percent of girls raped among girls paired according to bipartite matching algorithm. Distance within pairs is induced only on the candidate variable. Girls are similar in other respects.

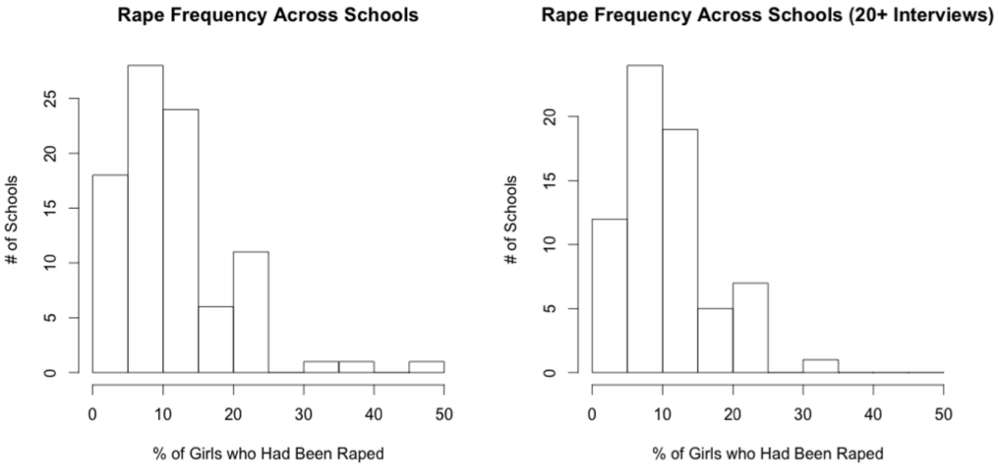

Figure 3: Percentage of girls who have been raped across schools in the study. The right panel filters to schools with more interviews, which provide less variant estimates of the school-level rape frequency.

Table 4: Assessment of school-level GBV predictors

Covariate	**p-Value**	**Sign**
Relative Dropout Rate	0.03	−
Area	0.06	NA
Roof	0.16	NA
Floor	0.17	NA
Sponsor	0.36	NA
Total Classrooms	0.40	−
Mean Test Score	0.41	−
Students Per Classroom	0.42	+
Girls' Toilet	0.44	−
Gender Ratio	0.46	−
Boys' Toilets per Boy	0.61	+
Students per Teacher	0.62	−
Total Teachers	0.65	+
Boys' Toilets	0.74	−
Girls Toilets Per Girl	0.80	−
Total Students	0.90	−

the $p = 0.10$ level.

The relative dropout rate was significant at the $p = 0.03$ level, indicating a possible association. Relative dropout rate tracks the change in gender-ratios within class, across classes 5-8. If there are relatively more boys in the school in class 8 than there were in in class 5 then we say that the "relative dropout rate" is higher for girls. This variable is intuitively appealing, as lower retention of girls relative to boys may reflect underlying issues in a particular community – which may, in turn, affect the prevalence of sexual assault. But several alternative explanations are also possible, including a "reversed" causal pathway whereby high rates of sexual assault discourage girls from attending school, leading to higher dropout rates and lower retention. This is not an extraordinarily strong predictor, and will require replication in future settings before it should be considered for regular use in triaging intervention into schools.

The area where a school was located was found to be highly significant, with the neighborhood of Kogorocho yielding much higher average rates of sexual assault. No other variables showed strong indications of significance.

We had hypothesized that some information that was harder to collect, only obtained from surveys completed by the girls, and then aggregated to the school-level, would be predictive of the school-level rate of rape. This indeed was borne out by the results. Aggregating our individual-level covariates to the school level, we found that many of these were much more predictive than the school-level covariates, with 36 of 87 aggregated covariates significant at the 0.05 level (though many of these variables are redundant). The full set of univariate regression p-values can be found in the appendix. Lastly, to develop a parsimonious model a LASSO was fit to explore the effect of each of these variables conditional on the others. Using the procedure described in the Methods section, the model in Table 5. Note that all variables have been standardized, such that their coefficients should be directly comparable.

The resulting model is quite parsimonious, including just seven variables, each of which is an aggregation of an individual-level covariate, rather than a school-level covariate. This further supports the idea that these variables are much stronger predictors of rape frequency. The proportion of girls who have had sex or been bribed for sex is strongly associated with rape frequency at a given school (observe that some girls also refused to answer the question, explaining why the coefficients for "have not had sex" and "have had sex" are different in magnitude). There is also a weakly negative relationship between the proportion of girls who have been trained in self-defense and no-means-no and the girls who have been victimized. One might anticipate that the causal relationship would point in the opposite direction, but it is possible that girls in more dangerous areas are also more likely to be taught self-defense skills and also more likely to be victimized.

It is possible that proxies for some of these variables may be almost as easy to collect as the administrative

Table 5: LASSO model coefficients for union of school-level and aggregated individual-level predictors

Covariate	**Coefficient**
(Intercept)	−0.045
Proportion of girls who have not been taught self-defense	−0.046
Proportion of girls who have not been taught a no-means-no skill	−0.044
Proportion of girls who have a boyfriend	−0.005
Proportion of girls who have been insulted by their boyfriends	0.053
Proportion of girls who have not had sex	−0.119
Proportion of girls who have had sex	0.156
Proportion of girls who have not been bribed for sex	−0.148

variables used earlier. For example, administrators may be able to estimate the proportion of girls who have a boyfriend. Likely more difficult to obtain from school administrators, the prevalence of sexual debut in the school looks to be quite predictive. This analysis may allow future interventionalists to target schools even beyond the administrative variables discussed earlier.

6 Discussion

As the study of sexual assault prevention develops, the literature must advance statistical practices commensurate with the complexity and challenges of the data. It is crucially important to provide unbiased estimates of baseline rape prevalence in a given area, and valid confidence intervals around these estimates. In turn, academics can assist interventionalists by helping (i) to direct energies to the most in need ("triaging"), and (ii) developing evidence about what aspects of interventions are most beneficial for reducing rates of sexual assault ("causal pathways")

This paper aims to document analytical practices both for the goal of transparency and improved statistical rigor. With regards to data adjudication, we have proposed both a particular approach and a novel validation procedure. Documenting our analysis at this time will be useful, as similar issues with data inconsistency will likely manifest at the midline and endline of this study. Moreover, as there is evidence that conflicting results are a common issue in studying sexual assault in a variety of contexts, our validation procedure could be adapted by other researchers in this area.

Our approach to uncertainty quantification is not novel, but offers both strong theoretical justification and relative simplicity. The approach we have suggested for identifying modifiable casual variables is a variant on existing algorithms in the literature. This approach reduces some of the burdensome parametric assumptions present in structural equation modeling, and may allow for more targeted hypothesis generation. It is also appealingly straightforward and can be implemented easily with existing statistical software.

Lastly, we take an empirical approach to developing school-level measurements to better triage intervention. We think it is of particular importance to separate variables based on ease of accessibility to researchers, and to assess the relative predictive value of easily obtainable variables versus those requiring live interviews. If aggregated summary statistics of questions currently asked in live interviews are highly predictive of the outcome of interest then it may be worthwhile to develop new questions that can be asked of administrators.

References

Anderson, J. C. and Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological bulletin*, 103(3):411.

Baiocchi, M., Nyamongo, M., Oguda, G., Akinyi, D., Friedberg, R., Rosenman, E., Omondo, B., and Sarnquist, C. (2018). Prevalence and risk factors for sexual assault among class 6 students in unplanned settlements of nairobi, kenya: Methods from the impower & sources of strength cluster randomized controlled trial.

Baiocchi, M., Omondi, B., Langat, N., Boothroyd, D. B., Sinclair, J., Pavia, L., Mulinge, M., Githua, O., Golden, N. H., and Sarnquist, C. (2017). A behavior-based intervention that prevents sexual assault: The results of a matched-pairs, cluster-randomized study in nairobi, kenya. *Prevention science*, 18(7):818–827.

Beck, C., Lu, B., and Greevy, R. (2016). *nbpMatching: Functions for Optimal Non-Bipartite Matching*. R package version 1.5.1.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brener, N. D., McMahon, P. M., Warren, C. W., and Douglas, K. A. (1999). Forced sexual intercourse and associated health-risk behaviors among female college students in the united states. *Journal of consulting and clinical psychology*, 67(2):252.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427.

Cook, S. L., Gidycz, C. A., Koss, M. P., and Murphy, M. (2011). Emerging issues in the measurement of rape victimization. *Violence against women*, 17(2):201–218.

Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, 52(4):281.

Davis, K. C., Gilmore, A. K., Stappenbeck, C. A., Balsan, M. J., George, W. H., and Norris, J. (2014). How to score the sexual experiences survey? a comparison of nine methods. *Psychology of violence*, 4(4):445.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer.

Ellsberg, M., Arango, D. J., Morton, M., Gennari, F., Kiplesund, S., Contreras, M., and Watts, C. (2015). Prevention of violence against women and girls: what does the evidence say? *The Lancet*, 385(9977):1555–1566.

Fisher, B. S. and Cullen, F. T. (2000). Measuring the sexual victimization of women: Evolution, current controversies, and future research. *Criminal justice*, 4:317–390.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1):21–29.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Gail, M. H., Mark, S. D., Caroll, R. J., Green, S. B., and Pee, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine*, 15(11):1069–1092.

Jewkes, R., Gibbs, A., Jama-Shai, N., Willan, S., Misselhorn, A., Mushinga, M., Washington, L., Mbatha, N., and Skiweyiya, Y. (2014). Stepping stones and creating futures intervention: shortened interrupted time series evaluation of a behavioural and structural health promotion and violence prevention intervention for young people in informal settlements in durban, south africa. *BMC public health*, 14(1):1325.

Kenya, V. (2012). Violence against children in kenya: Findings from a 2010 national survey.

Koss, M. P., Abbey, A., Campbell, R., Cook, S., Norris, J., Testa, M., Ullman, S., West, C., and White, J. (2007). Revising the ses: A collaborative process to improve assessment of sexual aggression and victimization. *Psychology of Women Quarterly*, 31(4):357–370.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Livingston, J. A., Hequembourg, A., Testa, M., and VanZile-Tamsen, C. (2007). Unique aspects of adolescent sexual victimization experiences. *Psychology of Women Quarterly*, 31(4):331–343.

Lu, B., Zanutto, E., Hornik, R., and Rosenbaum, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association*, 96(456):1245–1253.

Muris, P. (2001). A brief questionnaire for measuring self-efficacy in youths. *Journal of Psychopathology and Behavioral Assessment*, 23(3):145–149.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Mller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77.

Rosenberg, M. (2015). *Society and the adolescent self-image*. Princeton university press.

Russell, M., Cupp, P. K., Jewkes, R. K., Gevers, A., Mathews, C., LeFleur-Bellerose, C., and Small, J. (2014). Intimate partner violence among adolescents in cape town, south africa. *Prevention Science*, 15(3):283–295.

Shamu, S., Gevers, A., Mahlangu, B. P., Jama Shai, P. N., Chirwa, E. D., and Jewkes, R. K. (2015). Prevalence and risk factors for intimate partner violence among grade 8 learners in urban south africa: baseline analysis from the skhokho supporting success cluster randomised controlled trial. *International health*, 8(1):18–26.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Ullman, J. B. and Bentler, P. M. (2012). Structural equation modeling. *Handbook of Psychology, Second Edition*, 2.

UNICEF, O. et al. (2014). Hidden in plain sight: a statistical analysis of violence against children. 2014. *Nova York*.

Vuchinich, S., Flay, B. R., Aber, L., and Bickman, L. (2012). Person mobility in the design and analysis of cluster-randomized cohort prevention trials. *Prevention Science*, 13(3):300–313.

7 Appendix

7.1 Variable Definitions

The cluster-level covariates are summarized into two types of covariates. The first group of covariates was collected pre-randomization. There are 10 of these variables, six of which are numeric and four of which are categorical. The numeric variables are number of total classrooms, number of girls' toilets, number of total students, number of boys' toilets, the school's prior year mean standardized test score, and number of total teachers. The categorical variables are summarized below.

- **Area**: the neighborhood in which the school is located, taking on five distinct values (Dandora, Huruma, Kibera, Korogocho, Mukuru)
- **Roof**: the material from which the roof is constructed, taking on seven distinct values (Ironsheets, Bricks, Tiles, Tiles & Asbestos, Asbestos, Concrete, Tiles and Concrete)

- **Floor**: the material from which the floor is constructed, taking on six distinct values (Cemented, Tiles, Concrete, Mud, Non-Cement, Earthen)
- **Sponsor**: the sponsor of the school, taking on four distinct values (Private, Government, Church/mission, Others)

Additionally we have six covariates that are composites of continuous covariates. We considered these variables because the give more precise statements of the relative resources available in the schools.

- **Relative dropout rate**: gender ratio (number of boys divided by number of girls) in grade 5 at a school, divided by gender ratio in grade 8. This value will be lower if the ratio increases through the grades, indicating that girls drop out at a higher frequency than boys.
- **Students per teacher**: total number of students at a school, divided by total number of teachers at that school
- **Boys' toilets per boy**: total number of boys' toilets at a school, divided by total number of boys at that school
- **Girls' toilets per girl**: total number of girls' toilets at a school, divided by total number of girls at that school
- **Gender ratio**: total number of boys at a school divided by total number of girls at the school
- **Students per classroom**: total number of students at a school, divided by total number of classrooms at that school

Summary statistics for each of the numeric covariates are provided below.

Variable	**Mean**	**Std. Deviation**	**Range**
Mean Test Score	263.45	35.48	(189.55, 378.58)
Girls Toilets	9.16	4.8	(1, 20)
Boys Toilets	10.96	5.8	(1, 25)
Total Students	560.86	364.35	(117, 1772)
Total Teachers	15.75	9.37	(4, 66)
Total Classrooms	10.22	4.36	(4, 21)
Students Per Classroom	52.89	19.98	(17.62, 126.57)
Students Per Teacher	44.22	31.51	(4.88, 188)
Gender Ratio	0.96	0.14	(0.62, 1.52)
Relative Dropout Rate	1.14	0.44	(0.48, 3.25)
Girls' Toilets Per Girl	0.05	0.03	(0.01, 0.13)
Boys' Toilets Per Boy	0.04	0.02	(0.01, 0.16)

7.2 Balance Tables for Assessing Candidate Causal Pathways

Table 6: Covariate balance, near-far matching algorithm, for self efficacy candidate modifiable variable exploration

Covariate	**Higher value (n = 1911)**	**Lower value (n = 1911)**	**Absolute Standardized Difference**
Self efficacy	4.580	2.938	2.974
Parental violence: missing indicator	0.005	0.004	0.016
Parental violence	0.168	0.190	0.054
Socioeconomic status: missing indicator	0.003	0.003	0.010
Socioeconomic status	2.846	2.848	0.002
Mother dead: missing indicator	0.009	0.007	0.018
Mother dead	0.042	0.042	0.000
Father dead: missing indicator	0.047	0.051	0.019
Father dead	0.093	0.093	0.000
Has boyfriend: missing indicator	0.020	0.021	0.004
Has boyfriend	0.211	0.213	0.005
Times consumed alcohol: missing indicator	0.013	0.015	0.022
Times consumed alcohol	0.107	0.113	0.015
Intimate partner violence: missing indicator	0.016	0.018	0.012
Intimate partner violence	0.045	0.052	0.032
Gender normative thinking: missing indicator	0.057	0.058	0.005
Gender normative thinking	2.342	2.340	0.003

8 Temp: Collect Citations

(Liaw and Wiener, 2002) (Breiman, 2001)

Table 7: Covariate balance, near-far matching algorithm, for alcohol use candidate modifiable variable exploration

Covariate	**Higher value (n = 1923)**	**Lower value (n = 1923)**	**Absolute Standardized Difference**
Times consumed alcohol	0.242	0.000	0.669
Parental violence: missing indicator	0.003	0.002	0.011
Parental violence	0.173	0.169	0.008
Socioeconomic status: missing indicator	0.001	0.001	0.000
Socioeconomic status	2.867	2.865	0.001
Mother dead: missing indicator	0.005	0.005	0.007
Mother dead	0.037	0.037	0.000
Father dead: missing indicator	0.043	0.042	0.005
Father dead	0.083	0.083	0.000
Has boyfriend: missing indicator	0.016	0.016	0.000
Has boyfriend	0.205	0.203	0.004
Self efficacy: missing indicator	0.019	0.017	0.012
Self efficacy	3.738	3.734	0.004
Intimate partner violence: missing indicator	0.011	0.009	0.016
Intimate partner violence	0.047	0.044	0.018
Gender normative thinking: missing indicator	0.057	0.056	0.005
Gender normative thinking	2.341	2.343	0.004

Table 8: Covariate balance, near-far matching algorithm, for gender normative thinking candidate modifiable variable exploration

Covariate	**Higher value (n = 1836)**	**Lower value (n = 1836)**	**Absolute Standardized Difference**
Gender normative thinking	2.749	2.207	2.411
Parental violence: missing indicator	0.003	0.004	0.009
Parental violence	0.172	0.170	0.003
Socioeconomic status: missing indicator	0.001	0.001	0.000
Socioeconomic status	2.853	2.868	0.010
Mother dead: missing indicator	0.008	0.008	0.000
Mother dead	0.034	0.034	0.000
Father dead: missing indicator	0.045	0.045	0.000
Father dead	0.087	0.086	0.002
Has boyfriend: missing indicator	0.018	0.019	0.008
Has boyfriend	0.203	0.202	0.002
Times consumed alcohol: missing indicator	0.011	0.011	0.000
Times consumed alcohol	0.102	0.103	0.002
Self efficacy: missing indicator	0.016	0.017	0.004
Self efficacy	3.771	3.783	0.013
Intimate partner violence: missing indicator	0.014	0.014	0.000
Intimate partner violence	0.040	0.045	0.025