Lynette M Smith
Nicholas A Hein
Danstan S Bagenda

Cash transfers and HIV/HSV-2 prevalence

A replication of a cluster randomized trial in Malawi

October 2017

Replication Paper 12

Health

International Initiative for Impact Evaluation

About 3ie

The International Initiative for Impact Evaluation (3ie) is an international grant-making NGO promoting evidence-informed development policies and programs. We are the global leader in funding, producing and synthesizing high-quality evidence of what works, for whom, how, why and at what cost. We believe that using better and policy-relevant evidence helps to make development more effective and improve people's lives.

3ie Replication Paper Series

The 3ie Replication Paper Series is designed to be a publication and dissemination outlet for internal replication studies of development impact evaluations. Internal replication studies are those that reanalyze the data from an original paper in order to validate the results. The series seeks to publish replication studies with findings that reinforce or challenge the results of an original paper. To be eligible for submission, a replication study needs to be of a paper in 3ie's online Impact Evaluation Repository and needs to include a pure replication. 3ie invites formal replies from the original authors. These are published on the 3ie website together with the replication study.

The 3ie Replication Program also includes grant-making windows to fund replication studies of papers identified on our candidate studies list. Requests for proposals are issued one to two times a year. The candidate studies list includes published studies that are considered influential, innovative or counterintuitive. The list is periodically updated based on 3ie staff input and outside suggestions. The aim of the 3ie Replication Program is to improve the quality of evidence from development impact evaluations for use in policymaking and program design.

About this report

The publication of this report was supported by the Bill & Melinda Gates Foundation. The content of this report is solely the responsibility of the authors and does not necessarily represent the official views of the Bill & Melinda Gates Foundation or 3ie, its donors or its Board of Commissioners. Please direct any comments or queries to the corresponding author, Lynette M Smith at lmsmith@unmc.edu.

Suggested citation: Smith, LM, Hein, NA and Bagenda, DS, 2017. *Cash transfers and HIV/HSV-2 prevalence: a replication of a cluster randomized trial in Malawi*. 3ie Replication Paper 12. Washington, DC: International Initiative for Impact Evaluation (3ie).

3ie Replication Paper Series executive editor: Sara Pacqué-Margolis
Managing editor: Benjamin D Wood
Production manager: Brigid Monaghan
Copy editor: Yvette Charboneau
Proof reader: Jaime L Jarvis
Cover design: Akarsh Gupta

Cash transfers and HIV/HSV-2 prevalence: a replication of a cluster randomized trial in Malawi

Lynette M Smith
University of Nebraska Medical Center

Nicholas A Hein
University of Nebraska Medical Center

Danstan S Bagenda
University of Nebraska Medical Center

3ie Replication Paper 12

October 2017

Acknowledgements

This work was funded under the 3ie Replication Program. We are grateful to Berk Özler and Sarah Baird for preparing and providing the data and code from their original paper and for answering our questions. Thanks to 3ie reviewers Benjamin Wood and Eric Djimeu, as well as several anonymous reviewers, for their helpful comments.

Summary

The study selected for replication was *Effect of a cash transfer programme for schooling on prevalence of HIV and herpes simplex type 2 in Malawi: a cluster randomised trial* by Sarah Baird and others published in *The Lancet* in 2012. The original study was a two-year cluster randomized intervention in Malawi. Enumeration areas were randomized to either the intervention group (cash transfers) or control. In the intervention group, baseline schoolgirls were further required to attend school to receive payment or given no school attendance requirements. The study included 1,289 Malawian girls, 13–22 years old, who had never been married, were enrolled in school at baseline and had biological testing for HIV and herpes simplex virus type 2 (HSV-2). Among the schoolgirls at baseline, the intervention group was found to be more likely to choose younger partners and report less frequent sexual activity, though there was no effect on the frequency of unprotected sex. Also, HIV prevalence was 64 percent lower and HSV-2 was 76 percent lower in the cash transfer group compared to control, regardless of whether school attendance was required. The study also included a second cohort of 419 Malawian girls who were not enrolled in school at baseline (baseline dropouts). All the baseline dropout girls in the intervention group were required to attend school to receive payment. In the baseline dropout cohort, no significant difference was detected between the intervention and control groups in terms of HIV and HSV-2 prevalence.

The first objective of this replication research is to conduct a pure replication of the study; that is, establish whether the published findings can be reproduced using the study's own data and methods. Other than a few minor discrepancies, the original study was replicated. In a measurement and estimation analysis, the robustness of the results was examined with alternative methods; it was found that the intervention effect on HIV prevalence was somewhat sensitive to model choice. In the baseline schoolgirls cohort, the point estimate of the odds ratio for HIV increased by slightly more than 50 percent from 0.36 to 0.54 and the 95 percent confidence interval widened from 0.14–0.91 to 0.19–1.54 when using a generalized linear mixed model. The HSV-2 odds ratio point estimate was also sensitive, changing by 40 percent, but was still statistically significant. In permutation analysis and in a generalized linear mixed model, the intervention effect on HIV prevalence was no longer significant, but the results for HSV-2 prevalence were retained. A theory of change analysis showed no effect of intervention on a composite HIV awareness variable. In an analysis of the causal pathway, there were several variables that were partial and one variable that was a full mediator of intervention on outcome.

Contents

List of figures and tables

List of appendix tables

Abbreviations and acronyms

ALR	Alternating logistic regression
CI	Confidence interval
CCT	Conditional cash transfer
EA	Enumeration area
GLMM	Generalized linear mixed model
HSV-2	Herpes simplex virus type 2
OR	Odds ratio
PBR	Push button replication
PCA	Principal component analysis
SD	Standard deviation
UCT	Unconditional cash transfer

1. Introduction

The study selected for replication is *Effect of a cash transfer programme for schooling on prevalence of HIV and herpes simplex type 2 in Malawi: a cluster randomised trial*, by Sarah Baird, Richard Garfein, Craig McIntosh and Berk Özler, published in *The Lancet* in 2012. This study uses a fairly new approach to address structural drivers of HIV/AIDS described as physical, social, cultural, organizational, community, economic, legal or policy aspects of the environment that influence the risks and vulnerability environment and thus act as barriers to, or facilitators of, HIV prevention and treatment behavior (Blankenship et al. 2000, Sumartojo et al. 2000). In the current study, monthly cash transfer (to influence an economic structural driver), which was not accompanied by a program or training directly related to HIV prevention, was associated with decreases in the prevalence of both HIV and herpes simplex virus 2 (HSV-2), as well as decreases in high-risk sexual behavior of the baseline schoolgirls receiving transfers for 18 months.

The study was conducted in the Zomba district of southern Malawi. The Zomba district is made up of 550 enumeration areas (EAs) and tends to have high poverty rates, high HIV prevalence and low school enrollment (Baird et al. 2012a). Never married girls who were aged 13−22 years old were eligible for the study. These girls were separated into two cohorts, baseline schoolgirls who currently attend school and baseline dropouts who were not enrolled in school at baseline. EAs were randomized to intervention (cash transfer) or control (no cash transfer) and the randomization was stratified by the geographic stratum (urban, near rural, far rural). For the baseline schoolgirls cohort, the EAs randomly assigned to the intervention group were further randomized to a conditional cash transfer (CCT) group, where the girl was required to attend school to receive payment, or an unconditional cash transfer (UCT) group, where no attendance was required. All participants in the baseline dropout cohort intervention group were assigned to CCT. Intervention group participants either randomly received cash monthly (US$1−US$5 monthly and US$4–US$10 for their parents) or received nothing (to test for spillover effects of the intervention). All participants provided written informed consent, and further consent was obtained from guardians of girls younger than 18 years old. After 18 months, HIV prevalence was 64 percent lower (adjusted odds ratio [OR]=0.36) and HSV-2 was 76 percent lower (adjusted OR=0.24) among young women in the intervention group compared to the control group, regardless of whether school attendance was required (CCT and UCT groups combined).

The impact of this study lies both in the study population considered and the absence of intentional HIV prevention training during the intervention. Young women between the ages of 15 and 24 represent 30 percent of new HIV infections in southern Africa (Dellar et al. 2015), and while antiretroviral treatments have shown promise to prevent HIV acquisition, none of the trials to date have been tested in adolescents (Abdool Karim and Dellar 2014). Other studies have improved economic empowerment of young women through microfinance loans (Erulkar et al. 2006; Pronyk et al. 2008; Dunbar et al. 2010) or subsidies to pay for school uniforms or other education costs (Duflo et al. 2006), but typically have only measured sexual behavior post-intervention or measured other sexually transmitted infections as a proxy for sexually risky behavior, rather than HIV prevalence directly.

The current study measures prevalence of HIV and HSV-2 directly in participants at the end of the study. The results suggest that the structural intervention of cash transfer alone was enough to affect behavior. Specifically, baseline schoolgirls in the intervention group were more likely to choose younger partners and report less frequent sex with those partners, even though the study found no effect on the frequency of unprotected sex. In the baseline dropout cohort, the intervention group was also found to be more likely to report less frequent sex compared to the control group. Cluver and others (2013) found in an observational study in South Africa that receipt of a cash transfer was associated with reduced incidence and prevalence of transactional sex and age-disparate sex in girls aged 12–17, which agrees with the current study. Hallfors and colleagues (2015) conducted a similar three-year intervention but subsidized school costs specifically for 328 orphan adolescent girls. While other beneficial effects were found, such as improved likelihood to stay in school, socioeconomic status and reduced likelihood to marry in the intervention group versus controls, there was no difference in HIV or HSV-2 prevalence after five years. Pettifor and others (2016) completed a randomized clinical trial of young women in rural South Africa, looking at the effect of CCT on HIV incidence. Interestingly, they found that HIV incidence did not differ between girls who received the cash transfer and those who did not. However, they did find that school attendance significantly reduced the risk of HIV infection, regardless of whether the girl was in the intervention or control group.

In this paper, we perform a replication analysis of the study conducted by Baird and others (2012a) described above. The replication includes three objectives: perform a pure replication, a measurement and estimation analysis (MEA) and a theory of change analysis. As part of the replication process, the replication plan was finalized and approved prior to conducting any analysis. The full replication plan is published online (Smith 2016). The background and necessity of measurement and estimation and theory of change analyses are described in detail by Brown and others (2014).

The first objective of this replication study was to complete a pure replication. This pure replication attempts to reproduce the results presented in the paper using the author's cleaned data set and reported statistical methods from the original paper. In addition to the pure replication, push button replication was conducted and results are presented in Appendixes B and C. The push button replication takes the data analysis code provided by the original authors along with the data to reproduce the tables in the original paper.

The second objective was to conduct an MEA. Alternative estimation strategies exist for the analysis of complex survey data like those obtained in the study under replication. In the MEA section of the report, we explore some of the alternate strategies to determine the robustness of the results to the analysis method chosen. The MEA focuses on outcomes that were statistically significant or close to being statistically significant in the original paper – specifically, HIV and HSV-2 prevalence at 18 months, enrolled in school during 2008, sexual debut, had unprotected sexual intercourse, had sexual intercourse once per week, and had a sexual partner aged ≥25 years are considered for reanalysis. The baseline schoolgirls and baseline dropouts are treated as two separate cohorts for analysis, as in the original paper. An additional analysis for the baseline schoolgirls is conducted with the intervention classified as CCT, UCT and control.

The third objective is to conduct theory of change analyses to extend the study. Three different theory of change analyses are considered: 1) using principal component analysis, a composite HIV awareness variable was created that can be used to examine the effects of the treatment on HIV awareness; 2) a wealth index was constructed to determine if the cash transfer intervention would be more effective in poorer households; and 3) a causal pathway was explored to determine the direct effect of being enrolled in school and risky sexual behaviors (sexual debut, had unprotected sexual intercourse, had sexual intercourse once per week and had a sexual partner aged ≥25 years) on HIV and HSV-2 prevalence at 18 months.

2. The pure replication

2.1 The data

The database was downloaded from the World Bank website on January 18, 2016, and included the round 1 baseline data and the round 2 outcome data (World Bank 2012, 2015). The original study includes three data sets: baseline, follow-up and test results for HIV, HSV-2 and syphilis. The original authors pooled the relevant survey questions from baseline, follow-up and the test results into one deidentified merged public data set in which any individual identifiers were removed. The original authors have included user guides for the baseline and follow-up data sets and Stata code used to analyze the data for the original paper on the World Bank website. For our analysis, we used the deidentified merged public data set[1] and reconstructed the tables from the original paper using the described statistical analysis methods. There are strengths and weaknesses of this approach. By using the deidentified merged data, we ensure that the same sampling was used as employed by the original authors, but at the cost of independently constructing the sampling weights.

The baseline and follow-up data sets are standard survey data sets. They include information on 176 geographic enumeration areas with a sample of 4,051 individuals, of whom 3,796 were enrolled and completed a baseline interview at the end of 2007. Twelve months later, the same individuals completed a follow-up interview, and 18 months from the initial interview, 1,706 individuals were tested for HIV, HSV-2 and syphilis. The baseline data is broken down into 16 different data sets, each accounting for part of the survey instrument. In addition to these 16 data sets, there is a data set of identifiers. The follow-up data are structured in a similar manner, except there are 23 data sets and a data set of identifiers and baseline and controls. These data cover a broad range of topics that include schooling, income and health that may have an impact on individual risk behaviors.

2.1.1 Sampling procedure

A cluster randomized trial design was used to assess the effect of a cash transfer intervention on outcomes, including HIV, HSV-2 and syphilis prevalence; school enrollment; HIV knowledge; and risky sexual behaviors. One hundred seventy-six geographic EAs were selected out of a total of 550 in the Zomba district of Malawi. EAs are defined by the National Statistical Office of Malawi (Baird et al. 2012a). Each EA contains an average of 250 households that span several villages. These 176 selected

[1] File name: “Lancet_HIV-HSV2_dataset_PUBLIC_DEIDENTIFIED.dta”

EAs came from three geographic strata: urban (Zomba City), near rural (<16 kilometers from Zomba City) and far rural (≥16 kilometers from Zomba City). The use of a 16-kilometer radius was arbitrarily selected based on the cost of transportation (Baird et al. 2012a).

Of the 176 selected EAs, 29 were within Zomba City, 119 were near rural and the remaining 28 were far rural. After selecting the 176 EAs, a two-stage listing procedure was used to identify participants. All households were listed within the 176 EAs. The first form, form A, was sent to all households listed and asked the following question: “Are there any never-married girls in the household who are between the ages of 13 and 22?” This question allowed the researchers to identify possible participants for the investigation quickly and efficiently. If the answer from form A was yes, then form B was given to members of the household to collect data on age, marital status, current school and other characteristics.

Two cohorts were then defined: those enrolled in school at baseline (baseline schoolgirls) and those not enrolled in school at baseline (baseline dropouts). Because of the small number of baseline dropouts – on average 5.1 individuals who were dropouts per EA – all baseline dropouts were selected for the study. When selecting baseline schoolgirls, the percentage selected varied by age group and geographic location. Fourteen percent to 45 percent were selected from EAs within Zomba City and 70 percent to 100 percent from EAs within near or far rural. This sampling procedure yielded 889 baseline dropouts and 2,907 baseline school girls (Baird et al. 2012a).

Written informed consent was obtained from all participants, as well as consent from parents or legal guardians of girls younger than 18 years. At the time of screening, informed consent was obtained for HIV, HSV-2 and syphilis testing. The original study design was approved by the ethics review committees at the University of California at San Diego (USA) and the National Health Sciences Research Council (Malawi).

2.1.2 Randomization and masking

Following completion of baseline surveys, EAs were randomly assigned to either the intervention group (cash transfer program) or the control group (no cash transfer program). A 1:1 randomization was used. EAs within each stratum were assigned an identification number from a computer-generated list of random numbers. The three lists of EAs were then sorted in ascending order by their identification number. EAs in the first half of the list were assigned to the intervention group and those in the second half of the list were assigned to the control group, based on the random numbers.

Based on the randomization, all girls within an EA were either in the intervention group or control group. This method of randomization helped to reduce the possibility of crossover. Baseline schoolgirls who were selected for the intervention group were further randomized to two possible interventions, CCT or UCT. Again, a computer-generated list of random numbers was used to assign participants to either conditional or unconditional cash transfer program. All baseline dropouts were assigned to receive CCTs.

The percentage of schoolgirls selected to receive cash transfers in intervention EAs were randomly selected with computer-generated random numbers. None were selected in 15 EAs; 33 percent in 15 conditional EAs and 9 unconditional EAs; 66 percent in 16

conditional EAs and 9 unconditional EAs; or 100 percent in 15 conditional EAs and 9 unconditional EAs. By varying the percentage of school girls selected in treatment EAs, the potential effects of the program on untreated school girls in intervention areas could be measured (Baird et al. 2012a).

2.1.3 Statistical methods

The original paper conducted analyses in an intent to treat manner, separately for baseline schoolgirls and baseline dropouts. Unadjusted and adjusted ORs were computed with logistic regression models, with robust standard errors, which allows for intraclass correlation. Sampling weights were utilized to account for probability of inclusion that varied by age group and stratum. Adjusted ORs were calculated, including age group and geographical stratum as fixed effects, as well as the baseline values for any behavioral outcomes. Heterogeneity of intervention effects was assessed for prevalence of HIV and HSV-2 for the conditional and unconditional cash transfer in the baseline schoolgirls cohort. Stata version 10.1 (StataCorp, College Station, Texas, USA) was used for the original analysis.

The replication analysis was conducted using the same methods as the original analysis, using SAS/STAT software version 9.4 (SAS Institute Inc., Cary, North Carolina, USA) and Stata version 14.1. The SAS SURVEY procedures used for this analysis including PROC SURVEYFREQ, PROC SURVEYLOGISTIC, PROC SURVEYMEANS and PROC MEANS. The survey procedures include domain statements that allow for subgroup analysis, weight statements for weighting of observations according to the study design and clustering for the primary sampling unit (EA) and stratum (urban, near rural and far rural). This methodology allows for the computation of unadjusted and adjusted OR, with robust standard errors and the inclusion of sampling weights. The domain analysis in SAS is equivalent to the subpopulation analysis in Stata; it allows for estimation in a subpopulation while taking into account the study design, whereas a strict subgroup analysis does not.

Based on reading the methods from the paper, we analyzed the data using SAS software and the SURVEY procedures. The original authors provided their code to us as a courtesy, so we know exactly how the original analysis was conducted. There are two ways to analyze clustered data with weights in Stata. The authors analyzed the data using estimation commands that allow for weight and robust cluster options. The other way to analyze clustered survey data in Stata is using the svy set of commands. Based on our reading of the paper we expected that the svy commands would have been used because they allow for clustering, sampling weights, stratification and subpopulation analysis (which is like the domain analysis in SAS). In general, we would expect the standard errors to be smaller with the svy command, because it takes into account the strata, whereas the other method does not (University of North Carolina, Carolina Population Center).

2.1.4 Formatting the data

The data provided by the authors was obtained in Stata and converted to SAS using Stata software version 14.1 (StataCorp 2015). If there was a discrepancy in the replication results using SAS software, we reanalyzed the results using the Stata software version 14.1 using the code provided by the authors, as well as the svy commands in Stata. Overall, there were some differences in the confidence intervals

(CIs) depending on whether the analysis accounted for stratum and subpopulation analysis/domain analysis (described further below). Most of these are in the hundredths decimal place, are not considered to be discrepancies and do not change the results of the paper. If there was a difference in the sample size or in the point estimate, we did recognize this as a discrepancy. If the CIs for ORs differed by more than a hundredth, then these are also highlighted in the tables.

A summary of the pure replication results is included in Appendix A, Table A1, as a courtesy for the reader.

2.2 Reproducing the summary statistics

The pure replication begins by reproducing the baseline characteristics table of the schoolgirls and dropouts. Baseline characteristics are presented by randomization group, control group, pooled intervention group and CCT and UCT groups. The results of the original paper and the pure replication are found in Table 1A for the baseline schoolgirls and Table 1B for the baseline dropouts. Differences are presented in bold text with gray highlight and sample sizes are presented for all survey questions.

Since the original authors did not test all participants for HIV, HSV-2 and syphilis, we had to use the original authors' randomization and sampling weights. Therefore, it is not surprising that our results were almost identical to the original paper. However, there were some sample size discrepancies. The first piece of information presented in the baseline characteristic table is the number of EAs sampled for biological data collection. The numbers presented in the original table showed 52 EAs in the control group and intervention groups and 25 in the CCT and 27 in the UCT groups. The pure replication results found the number of EAs in the CCT arm to be 26, not 25 as reported in the original paper (Table 1A). Upon further investigation, it was determined that one of the EAs appears in both CCT and UCT groups in the data set (specifically, eaid=218). Some of the girls sampled in that EA are listed as CCT and some are listed as UCT.

Additional sample size discrepancies were observed. Depending on the survey question being analyzed, the sample size (denominator) was either larger or smaller than what was stated in the original paper. A smaller sample size is not of a concern, since participants do not always answer all questions of a survey. Sample sizes that were larger than the listed sample sizes were disconcerting, and these differences can be observed in Tables 1A and 1B. For example, when looking at the first panel of the original data in Table 1A, the number of individuals in the pooled intervention group is 501; however, the pure replication showed 503 individuals in the pooled intervention group denominator. This discrepancy became clear when looking at specific questions for female-headed household, household owns a radio, household owns a television, household has access to a mobile telephone, and electricity available in dwelling.

Similar discrepancies were found in the control group and UTC groups, as seen in Table 1A and in the intervention group in Table 1B. These differences are difficult to explain, since the numerators and row percentages match. In a personal communication, the first author of the original paper, Dr. Baird, wrote, "at baseline a few girls either did not get the Part I (Household) or Part II (Girl) survey. Thus, when you summarize, for example, ever married at baseline, you will get the sample sizes in the paper, as it is missing for a few

girls." Therefore, it can be concluded that the number of individuals shown in the second row of the table was taken from the marginal totals from one of the variables analyzed, which had a few missing data points.

When the authors reported a response that was a continuous variable, a weighted mean and standard deviation were reported. It should be noted that the standard deviation reported in the original tables is an asymptotic estimate of an observation with average weight (SAS Institute Inc. 2015). Therefore, the standard error of the mean may have been a more appropriate statistic to report. In keeping with the pure replication, we reported standard deviation. A couple of discrepancies in standard deviation estimates are noted and are highlighted in Tables 1A and 1B but are within rounding error.

One further discrepancy was noted for the survey question about whether the mother was alive. We note that for the school girls in the intervention group, the original paper reported that 423/501 (84%) indicated the mother being alive. Our analysis found that 420/501 (84%) reported the mother as alive.

In summary, for the pure replication results for the baseline characteristic data, we assume that the denominator issues can be explained by missing data in the survey as well as the discrepancy in the second row where the number of individuals are shown (Tables 1A and 1B). We determined that one of the EAs appears in both CCT and UCT groups, clarifying the reason for the discrepancy in the number of EAs sampled. With these issues aside, baseline characteristics were similar between the original study and the pure replication.

Table 1A: Baseline characteristics of participants among baseline schoolgirls, original versus replication results

	Panel A: original paper				Panel B: replication results			
	Control group	**Intervention group**			**Control group**	**Intervention group**		
		Pooled	CCT	UTC		Pooled	CCT	UTC
Enumeration areas sampled for biological data collection	52	52	25	27	52	52	**26**	27
Number of individuals	827	501	236	265	**828**	**503**	236	**267**
Ever had sexual intercourse	182 (19%)	130 (22%)	70 (22%)	60 (22%)	182/827 (19%)	130/501 (22%)	70/236 (22%)	60/265 (22%)
Ever pregnant	21 (3%)	16 (3%)	9 (3%)	7 (3%)	21/827 (3%)	16/500 (3%)	9/235 (3%)	7/265 (3%)
Age (years)	15.3 (1.9)	15.1 (1.9)	14.9 (1.8)	15.4 (1.9)	15.3 (1.9)	15.1 (1.9)	14.9 (1.8)	15.4 (1.9)
Age at sexual debut (years)	15.7 (1.7)	15.8 (1.8)	15.7 (2.0)	15.9 (1.7)	15.7 (1.7)	15.8 (1.8)	15.7 **(1.9)**	15.9 (1.7)
Highest grade attended	7.6 (1.6)	7.4 (1.7)	7.1 (1.7)	7.9 (1.6)	7.6 (1.6)	7.4 (1.7)	7.1 (1.7)	7.9 (1.6)
Mother alive	707 (85%)	423 (84%)	198 (85%)	222 (83%)	707/827 (85%)	**420/**501 (84%)	198/236 (85%)	222/265 (83%)
Father alive	601 (74%)	367 (75%)	176 (74%)	191 (76%)	601/826 (74%)	367/499 (75%)	176/236 (74%)	191/263 (76%)
Female-headed household	275 (32%)	141 (25%)	63 (26%)	78 (24%)	275/826 (32%)	141/503 (25%)	63/236 (26%)	78/267 (24%)
Household owns a radio	479 (59%)	309 (58%)	143 (53%)	166 (65%)	479/826 (59%)	309/503 (58%)	143/236 (53%)	166/267 (65%)
Household owns a television	130 (24%)	110 (30%)	40 (27%)	70 (34%)	130/826 (24%)	110/503 (30%)	40/236 (27%)	70/267 (34%)
Household has access to a mobile telephone	464 (61%)	303 (60%)	145 (60%)	158 (61%)	464/826 (61%)	303/503 (60%)	145/236 (60%)	158/267 (61%)
Electricity available in dwelling	86 (20%)	80 (26%)	31 (28%)	49 (24%)	86/825 (20%)	80/503 (26%)	31/236 (28%)	49/267 (24%)
Piped water available in dwelling	277 (47%)	183 (49%)	48 (41%)	135 (60%)	277/822 (47%)	183/500 (49%)	48/233 (41%)	135/267 (60%)

Note: Data are n/N (weight %) or mean (standard deviation (SD)). Sampling weights were used to account for variation in the probability of being selected and varied by geographic location and age.

Table 1B: Baseline characteristics of participants among baseline dropouts, original versus replication results

	Panel A: original paper		Panel B: replication results	
	Control	**Intervention**	**Control**	**Intervention**
Number of individuals	223	226	223	**227**
Ever had sexual intercourse	151 (68%)	154 (68%)	151/223 (68%)	154/226 (68%)
Ever pregnant	98 (44%)	90 (40%)	98/223 (44%)	90/226 (40%)
Age (years)	17.6 (2.2)	16.8 (2.4)	17.6 (2.2)	16.8 **(2.5)**
Age at sexual debut (years)	16.4 (1.8)	15.9 (2.2)	16.4 (1.8)	15.9 (2.2)
Highest grade attended	6.2 (2.9)	5.8 (2.9)	6.2 (2.9)	5.8 (2.9)
Mother alive	175 (78%)	180 (80%)	175/223 (78%)	180/226 (80%)
Father alive	146 (66%)	144 (65%)	146/222 (66%)	144/223 (65%)
Female-headed household	93 (42%)	90 (39%)	93/223 (42%)	90/227 (39%)
Household owns a radio	118 (53%)	107 (47%)	118/223 (53%)	107/227 (47%)
Household owns a television	16 (7%)	24 (11%)	16/223 (7%)	24/227 (11%)
Household has access to a mobile telephone	103 (46%)	110 (49%)	103/223 (46%)	110/227 (49%)
Electricity available in dwelling	16 (7%)	24 (11%)	16/223 (7%)	24/227 (11%)
Piped water available in dwelling	64 (29%)	63 (25%)	64/221 (29%)	63/226 (25%)

Note: Data are n (weight %) or mean (SD). Sampling weights were used to account for variation in the probability of being selected and varied by geographic location and age.

2.3 Reproducing the main results

The main results of the original paper are presented in panel A of Tables 2A and 2B and the pure replication results are presented in panel B. In these tables, summary statistics for baseline and follow-up responses are presented. The primary outcome was the prevalence of HIV and HSV-2 at 18 months for individuals selected for biological testing. Secondary outcomes are also reported, including syphilis prevalence, school enrollment, self-reported marriage, pregnancy, sexual behavior and knowledge of HIV/AIDS. Baseline schoolgirl and baseline dropout cohorts were analyzed separately. The conditional and unconditional treatment arms were pooled and analyzed as the intervention group.

From the original paper, unadjusted ORs were calculated using logistic regression with treatment status as the only covariate. A weight statement was used to adjust for the probability of being selected and robust standard errors were calculated by clustering by enumeration area to relax the assumption of independent and identically distributed errors within an EA. Adjusted ORs were calculated in a similar manner, except baseline measurements for behavior outcomes, age and geographic location were included as fixed effects. Age was a binary variable that indicated whether the individual was 13–14 years old at baseline.

Our results for the effects of the intervention on outcome measures are presented in Table 2A for the baseline schoolgirls cohort and Table 2B for the baseline dropouts cohort. Panel A contains the original results and panel B the pure replication results. The replication results generally follow the original results. However, there were a few notable differences. As in Table 1A, there were some sample size issues that are highlighted in

bold; one appeared to be a typo. The original paper reported a sample size of 299 for the survey question “had sexual intercourse once per week” for the baseline school girls at follow-up. The sample size should have been 499. The percentage matches, which indicates this is most likely a typo. Tables 3A and 3B show the original and pure replication results presenting unadjusted and adjusted ORs and CIs for the intervention versus control groups. There were two point estimates that did not match and could not be explained by rounding errors. The point estimates correspond to the outcomes of syphilis prevalence and “had unprotected sexual intercourse.” The discrepancies in these point estimates did not affect the significance of the results or the interpretations of the results. The differences seen in the upper confidence limits can be explained by whether or not the analysis accounts for strata and subpopulation/domain. This was determined by a second analysis, which did not include strata and subpopulation/domain; in this analysis, the confidence limits matched.

It was found that the intervention lowered the odds of HIV and HSV-2 prevalence in baseline schoolgirls but did not have a significant effect for baseline dropouts. The adjusted odds of HIV prevalence in the intervention group of baseline schoolgirls is 0.36 (95% CI = [0.14, 0.92]) of that of a baseline school girl in the control group when controlling for age and geographic location, whereas the adjusted OR and 95% CI for the baseline dropout was 1.37 (95% CI = [0.72, 2.60]). The adjusted OR of HSV-2 prevalence in the intervention group of baseline schoolgirls is 0.24 (95% CI = [0.09, 0.65]) of that of a baseline school girl in the control group when controlling for age and geographic location, whereas the OR and 95 percent CI for the baseline dropout was 1.03 (95% CI = [0.47, 2.25]).The remaining results can be seen in Tables 3A and 3B.

The authors of the original paper then examined the heterogeneity of the treatment arms. Examining the same outcomes as previously described, summary statistics are presented in Tables 4A and 4B for the control group, CCT group and UCT group at 12-month and 18-month follow-up for baseline schoolgirls. Adjusted ORs and 95 percent CIs (using the method described above) are calculated, comparing each treatment arm to the control group. The two ORs are then compared and a p-value was reported. The asymptotic Wald statistic is used for the comparison of the OR and the p-value is calculated using the chi-square distribution.

When examining the heterogeneity of the treatment arms, it was found that the only significant difference between treatments was whether a baseline schoolgirl was currently pregnant. The UCT arm had lower odds of being pregnant than the CCT arm, p-value = 0.0129. There were no discrepancies between the original paper and our results other than the upper confidence limits that are explained by accounting for strata and subpopulation/domain, shown in Tables 4A and 4B. The differences in the upper confidence limits that are more than a hundredth are highlighted in Table 4B. The CI for the HIV prevalence in the CCT versus control changed enough that the borderline significant difference shown in the original analysis is now non-significant by a small margin. In our opinion, we do not think these differences change the original authors’ findings.

Table 2A: Results of effects of cash transfer intervention on outcome measures – baseline schoolgirls, original versus replication results (SAS)

	Panel A: original paper				Panel B: replication results			
	Baseline		Follow-up		Baseline		Follow-up	
	Intervention group	Control group	Intervention group	Control group	Intervention group	Control group	Intervention group	Control group
Enrolled during 2008 school year			419/484 (90%)	669/801 (84%)			419/484 (90%)	669/801 (84%)
Ever married	0/501 (0%)	0/827 (0%)	19/501 (3%)	45/827 (4%)	0/**503** (0%)	0/**828** (0%)	19/501 (3%)	45/827 (4%)
Currently pregnant	3/501 (<1%)	2/827 (<1%)	15/501 (2%)	35/827 (4%)	3/501 (<1%)	2/827 (<1%)	15/501 (2%)	35/827 (4%)
Sexual debut*			39/371 (8%)	100/645 (13%)			39/371 (8%)	100/645 (13%)
Had unprotected sexual intercourse	91/500 (16%)	107/825 (11%)	49/500 (8%)	63/826 (7%)	91/500 (15%)	107/825 (11%)	49/500 (8%)	63/826 (7%)
Had sexual intercourse once per week	16/500 (3%)	22/825 (2%)	22/299 (3%)	62/826 (7%)	16/500 (3%)	22/825 (2%)	22/**499** (3%)	62/826 (7%)
Had a sexual partner aged ≥ 25 years†			4/502 (<1%)	20/827 (2%)			4/**500** (<1%)	20/**826 (2%)**
Had an HIV test	121/501 (22%)	174/827 (19%)	307/501 (54%)	470/826 (52%)	121/501 (22%)	174/827 (19%)	307/501 (54%)	470/826 (52%)
Knows that a healthy looking person can have HIV	443/501 (88%)	752/827 (90%)	454/501 (91%)	768/826 (92%)	443/501 (88%)	752/827 (90%)	454/501 (91%)	768/826 (92%)
Knows that HIV can be transmitted through breastfeeding	466/500 (93%)	785/827 (95%)	481/501 (97%)	786/827 (96%)	466/500 (93%)	785/827 (95%)	481/501 (97%)	786/827 (96%)
Received health training about HIV/AIDS†			398/501 (78%)	657/827 (80%)			398/501 (78%)	657/827 (80%)
HIV prevalence†,‡			7/490 (1%)	17/799 (3%)			7/490 (1%)	17/799 (3%)
HSV-2 prevalence†,‡			5/488 (<1%)	27/796 (3%)			5/488 (<1%)	27/796 (3%)
Syphilis prevalence†,‡			1/491 (<1%)	4/800 (<1%)			1/491 (<1%)	4/800 (<1%)

Note: * Data are n/N (weighted %) unless otherwise stated. Sampling weights were used to account for variation in the probability of inclusion in the study according to age and stratum.

Table 2B: Results of effects of cash transfer intervention on outcome measures – baseline dropouts, original versus replication results (SAS)

	Panel A: original paper				Panel B: replication results			
	Baseline		Follow-up		Baseline		Follow-up	
	Intervention group	Control group	Intervention group	Control group	Intervention group	Control group	Intervention group	Control group
Enrolled during 2008 school year			124/219 (57%)	27/220 (12%)			124/219 (57%)	27/220 (12%)
Ever married	0/227 (0%)	0/223 (0%)	37/226 (17%)	64/223 (29%)	0/227 (0%)	0/223 (0%)	37/226 (17%)	64/223 (29%)
Currently pregnant	11/226 (5%)	10/223 (5%)	16/226 (7%)	26/223 (12%)	11/226 (5%)	10/223 (4%)	16/226 (7%)	26/223 (12%)
Sexual debut*			18/72 (26%)	27/72 (38%)			18/72 (26%)	27/72 (38%)
Had unprotected sexual intercourse	133/222 (61%)	128/223 (57%)	59/225 (25%)	64/222 (29%)	133/222 (61%)	128/223 (57%)	59/225 (25%)	64/222 (29%)
Had sexual intercourse once per week	31/222 (14%)	28/223 (13%)	43/225 (19%)	66/223 (30%)	31/222 (14%)	28/223 (13%)	43/225 (19%)	66/223 (30%)
Had a sexual partner aged ≥ 25 years†			20/225 (8%)	23/223 (10%)			20/225 (8%)	23/223 (10%)
Had an HIV test	98/225 (43%)	104/223 (47%)	163/225 (72%)	169/223 (76%)	98/225 (43%)	104/223 (47%)	163/225 (72%)	169/223 (76%)
Knows that a healthy looking person can have HIV	198/225 (88%)	201/223 (90%)	204/226 (90%)	212/223 (95%)	198/225 (88%)	201/223 (90%)	204/226 (90%)	212/223 (95%)
Knows that HIV can be transmitted through breastfeeding	198/223 (89%)	210/223 (94%)	213/226 (94%)	214/223 (96%)	198/223 (89%)	210/223 (94%)	213/226 (94%)	214/223 (96%)
Received health training about HIV/AIDS†			130/226 (57%)	94/223 (42%)			130/226 (57%)	94/223 (42%)
HIV prevalence†,‡			23/210 (10%)	17/207 (8%)			23/210 (10%)	17/207 (8%)
HSV-2 prevalence†,‡			17/211 (8%)	17/208 (8%)			17/211 (8%)	17/208 (8%)
Syphilis prevalence†,‡			3/211 (2%)	2/208 (1%)			3/211 (2%)	2/208 (1%)

Note: Data are n/N (weighted %) unless otherwise stated. Sampling weights were used to account for variation in the probability of inclusion in the study according to age and stratum.

Table 3A: Results of effects of cash transfer intervention on outcome measures – baseline schoolgirls, original versus replication results (SAS)

	Panel A: original paper		Panel B: replication results	
	Unadjusted odds ratio (95% CI)	**Adjusted odds ratio (95% CI)**	**Unadjusted odds ratio (95% CI)**	**Adjusted odds ratio (95% CI)**
Enrolled during 2008 school year	1.67 (1.09-2.56)	1.62 (1.07-2.45)	1.67 (1.09-2.57)	1.62 (1.07-2.46)
Ever married	0.63 (0.31-1.28)	0.68 (0.37-1.28)	0.63 (0.31-1.29)	0.68 (0.36-1.29)
Currently pregnant	0.66 (0.33-1.30)	0.71 (0.36-1.41)	0.66 (0.33-1.31)	0.71 (0.35-1.42)
Sexual debut*	0.59 (0.33-1.05)	0.64 (0.38-1.07)	0.59 (0.33-1.05)	0.64 (0.38-1.07)
Had unprotected sexual intercourse	1.28 (0.78-2.09)	1.08 (0.67-1.75)	1.28 (0.78-2.10)	1.08 (0.66-1.75)
Had sexual intercourse once per week	0.44 (0.23-0.85)	0.46 (0.26-0.82)	0.44 (0.23-0.85)	0.46 (0.26-0.82)
Had a sexual partner aged ≥ 25 years†	0.20 (0.07-0.59)	0.21 (0.07-0.62)	0.20 (0.07-0.59)	0.21 (0.07-0.62)
Had an HIV test	1.12 (0.74-1.68)	1.18 (0.83-1.69)	1.12 (0.74-1.69)	1.18 (0.83-1.69)
Knows that a healthy looking person can have HIV	0.94 (0.58-1.53)	1.00 (0.61-1.62)	0.94 (0.58-1.53)	1.00 (0.61-1.63)
Knows that HIV can be transmitted through breastfeeding	1.70 (0.84-3.43)	1.72 (0.89-3.34)	1.70 (0.83-3.44)	1.72 (0.88-3.35)
Received health training about HIV/AIDS†	0.89 (0.61-1.30)	0.90 (0.63-1.30)	0.89 (0.61-1.30)	0.90 (0.63-1.30)
HIV prevalence†,‡	0.39 (0.15-1.02)	0.36 (0.14-0.91)	0.39 (0.15-**1.04**)	0.36 (0.14-0.92)
HSV-2 prevalence†,‡	0.23 (0.08-0.66)	0.24 (0.09-0.66)	0.23 (0.08-0.66)	0.24 (0.09-0.66)
Syphilis prevalence†,‡	1.20 (0.15-9.68)	0.92 (0.12-6.85)	**1.19 (0.14-9.87)**	**0.89 (0.12-6.72)**

Note: Sampling weights were used to account for variation in the probability of being selected and varied by geographic location and age. Adjusted ORs are calculated with a logistic regression model where the fixed effects are baseline measurements, treatment status, age category and geographic location unless otherwise stated. 95% CIs are calculated using robust standard errors, clustered by enumeration area. *Cumulative risk, so no baseline adjustment. †Data not collected at baseline, so no baseline adjustment made. ‡Measured at 18 months, all others measured at 12 months.

Table 3B: Results of effects of cash transfer intervention on outcome measures – baseline dropouts, original versus replication results (SAS)

	Panel A: original paper		Panel B: replication results	
	Unadjusted odds ratio (95% CI)	**Adjusted odds ratio (95% CI)**	**Unadjusted odds ratio (95% CI)**	**Adjusted odds ratio (95% CI)**
Enrolled during 2008 school year	9.31 (5.31-16.3)	8.77 (5.07-15.1)	9.31 (5.30-16.3)	8.77 (5.07-15.1)
Ever married	0.50 (0.29-0.83)	0.48 (0.29-0.80)	0.50 (0.29-0.84)	0.48 (0.29-0.80)
Currently pregnant	0.59 (0.28-1.25)	0.55 (0.27-1.18)	0.59 (0.28-1.26)	0.56 (0.27-1.18)
Sexual debut*	0.57 (0.28-1.16)	0.70 (0.33-1.45)	0.57 (0.28-1.16)	0.70 (0.33-1.45)
Had unprotected sexual intercourse	0.75 (0.46-1.22)	0.74 (0.44-1.23)	**0.83 (0.51-1.35)**	0.74 (0.44-1.24)
Had sexual intercourse once per week	0.56 (0.34-0.91)	0.53 (0.32-0.86)	0.56 (0.34-0.91)	0.53 (0.32-0.87)
Had a sexual partner aged ≥ 25 years†	0.73 (0.40-1.34)	0.79 (0.42-1.50)	0.73 (0.40-1.35)	0.79 (0.41-1.51)
Had an HIV test	0.84 (0.51-1.38)	1.00 (0.64-1.57)	0.84 (0.51-1.38)	1.00 (0.64-1.57)
Knows that a healthy looking person can have HIV	0.45 (0.19-1.05)	0.51 (0.23-1.15)	0.45 (0.19-1.05)	0.51 (0.23-1.16)
Knows that HIV can be transmitted through breastfeeding	0.63 (0.25-1.56)	0.69 (0.26-1.81)	0.63 (0.25-1.57)	0.69 (0.26-1.82)
Received health training about HIV/AIDS†	1.82 (1.23-2.69)	1.91 (1.29-2.83)	1.82 (1.23-2.70)	1.91 (1.29-2.84)
HIV prevalence†,‡	1.30 (0.69-2.48)	1.37 (0.72-2.61)	1.30 (0.68-**2.50**)	1.37 (0.72-**2.63**)
HSV-2 prevalence†,‡	0.99 (0.46-2.10)	1.03 (0.47-2.24)	0.99 (0.46-2.11)	1.03 (0.47-2.25)
Syphilis prevalence†,‡	1.59 (0.27-9.50)	1.63 (0.27-9.95)	1.59 (0.26-**9.70**)	1.63 (0.26-**10.18**)

Note: Sampling weights were used to account for variation in the probability of being selected and varied by geographic location and age. Adjusted ORs are calculated with a logistic regression model where the fixed effects are baseline measurements, treatment status, age category and geographic location unless otherwise stated. 95% CIs are calculated using robust standard errors, clustered by enumeration area. *Cumulative risk, so no baseline adjustment. †Data not collected at baseline, so no baseline adjustment made. ‡Measured at 18 months, all others measured at 12 months.

Table 4A: Results of effect of conditional or unconditional cash transfers on baseline schoolgirls by outcome measures, original versus replication results (SAS)

	Panel A: original paper			Panel B: replication results		
	Control group	**CCT group**	**UCT group**	**Control group**	**CCT group**	**UCT group**
Enrolled during the 2008 school year	669/801 (84%)	207/229 (92%)	212/255 (87%)	669/801 (84%)	207/229 (92%)	212/255 (87%)
Ever married	45/827 (4%)	14/236 (4%)	5/265 (2%)	45/827 (4%)	14/236 (4%)	5/265 (2%)
Currently pregnant	35/827 (4%)	13/236 (4%)	2/265 (1%)	35/827 (4%)	13/236 (4%)	2/265 (1%)
Sexual debut†	100/645 (13%)	18/166 (7%)	21/205 (10%)	100/645 (13%)	18/166 (7%)	21/205 (10%)
Unprotected sexual intercourse	63/826 (7%)	30/235 (9%)	19/265 (8%)	63/826 (7%)	30/235 (9%)	19/265 (8%)
Had sexual intercourse once per week	62/826 (7%)	14/235 (3%)	8/264 (3%)	62/826 (7%)	14/235 (3%)	8/264 (3%)
Had a sexual partner aged ≥ 25 years‡	20/826 (2%)	1/235 (<1%)	3/235 (1%)	20/826 (2%)	1/235 (<1%)	3/235 (1%)
HIV prevalence‡	17/799 (3%)	3/235 (1%)	4/255 (2%)	17/799 (3%)	3/235 (1%)	4/255 (2%)
HSV-2 prevalence‡	27/796 (3%)	4/233 (1%)	1/255 (<1%)	27/796 (3%)	4/233 (1%)	1/255 (<1%)
Syphilis prevalence‡	4/800 (<1%)	1/235 (<1%)	0/256 (0%)	4/800 (1%)	1/235 (1%)	0/256 (0%)

Note: Data are n/N (weighted %) unless otherwise stated.

Table 4B: Results of effect of conditional or unconditional cash transfers on baseline schoolgirls by outcome measures, original versus replication results (SAS)

	Panel A: original paper			Panel B: replication results		
	CCT vs. control (adjusted odds ratio [95% CI])	**UCT vs. control (adjusted odds ratio [95% CI])**	**Heterogeneity of odds ratio* (p value)**	**CCT vs. control (adjusted odds ratio [95% CI])**	**UCT vs. control (adjusted odds ratio [95% CI])**	**Heterogeneity of odds ratio* (p value)**
Enrolled during the 2008 school year	2.08 (1.14-3.82)	1.22 (0.77-1.96)	0.14	2.08 (1.13-3.83)	1.22 (0.76-1.96)	0.14
Ever married	0.93 (0.47-1.86)	0.36 (0.12-1.07)	0.11	0.93 (0.47-1.87)	0.36 (0.12-1.07)	0.11
Currently pregnant	1.17 (0.56-2.43)	0.16 (0.04-0.68)	0.0121	1.17 (0.56-2.44)	0.16 (0.04-0.68)	**0.0129**
Sexual debut†	0.58 (0.29-1.15)	0.72 (0.37-1.40)	0.62	0.58 (0.29-1.15)	0.72 (0.37-1.41)	0.61
Unprotected sexual intercourse	1.17 (0.67-2.05)	0.96 (0.50-1.83)	0.59	1.17 (0.67-2.06)	0.96 (0.50-1.83)	0.59
Had sexual intercourse once per week	0.53 (0.26-1.07)	0.37 (0.16-0.85)	0.49	0.53 (0.26-1.07)	0.37 (0.16-0.85)	0.49
Had a sexual partner aged ≥ 25 years‡	0.08 (0.01-0.60)	0.36 (0.11-1.19)	0.19	0.08 (0.01-0.61)	0.36 (0.11-1.20)	0.19
HIV prevalence‡	0.29 (0.09-0.98)	0.47 (0.14-1.59)	0.57	0.29 (0.09-**1.00**)	0.47 (0.13-**1.62**)	0.57
HSV-2 prevalence‡	0.37 (0.13-1.03)	0.08 (0.01-0.58)	0.16	0.37 (0.13-**1.05**)	0.08 (0.01-**0.60**)	0.17
Syphilis prevalence‡	1.37 (0.20-9.41)			1.37 (0.20-**9.62**)		

Note: Adjusted ORs calculated with a logistic regression model of individual data with independent variables that include treatment status, age group, indicators for near rural and far rural strata, and baseline measure unless otherwise stated. Sampling weights were used to account for variation in the probability of inclusion in the study according to age and stratum. *Wald test of equality of adjusted ORs for the conditional and unconditional cash transfer interventions. †Cumulative risk measure, so no adjustment made for baseline status. ‡No adjustment for baseline measure because data not collected at baseline.

2.4 Pure replication conclusions

The pure replication reproduced the results of the original paper very well, with just a few minor discrepancies. There were some discrepancies in the reported numbers of EAs and group sample sizes, and two point estimates did not match. However, the difference in these point estimates did not change their significance. The pure replication leads to the same conclusions as the original authors, after accounting for strata and subpopulation/domain. The cash transfer program was effective in reducing the prevalence of HIV and HSV-2 in unmarried school-aged girls currently attending school in Malawi. There was no significant reduction of HIV or HSV-2 prevalence for school-aged girls who dropped out of school. Additionally, there were no significant differences in the type of intervention, conditional versus unconditional cash transfer for school-aged girls currently attending school, except in one outcome, whether the schoolgirl was currently pregnant.

3. Measurement and estimation analysis

The MEA explores the robustness of the findings through additional analyses. The design strategy applied in the paper for recruiting patients is reasonably convincing, as is the analysis presented in the article. The original analysis calculated unadjusted and adjusted ORs by fitting logistic regression models. Robust standard errors were calculated, allowing for intracluster correlation, and sampling weights were included to adjust for the probability of inclusion based on age and EA stratum. Despite this, there is not a gold standard analysis method for this type of data, and alternate methods can be utilized to determine if the results are dependent on the analysis method chosen. Here, we pre-specified three analysis methods to explore – generalized linear mixed models (GLMMs), permutation methods and a bivariate generalized estimating equation model.

3.1 Generalized linear mixed model

3.1.1 GLMM methods

GLMMs (also known as hierarchical or multilevel models), can be used successfully in this type of study design. Rabe-Hesketh and Skrondal (2006) and Pfeffermann and others (1998) describe a multilevel model for complex survey data. These models can be implemented in Stata with the GLLAMM package or in SAS with PROC GLIMMIX. This model allows for estimation of random effects as well as fixed effects. In keeping with the original paper, baseline measurements for behavior outcomes, age and geographic location were included as fixed effects. EA was included as a random effect. Random effects take into account the clustered nature of the data.

As in the original paper, a weight statement was included to adjust for the probability of being selected. However, Pfeffermann and others (1998) assert that when the sample selection probabilities are related to the response variable even after adjusting for covariates, the estimators of the model parameters may be biased. Therefore, Pfeffermann and colleagues (1998) recommend scaling the weights using one of two methods. Scaling method 1 uses a scaling factor such that the apparent cluster size equals the effective sample size. Scaling method 2 sets the apparent cluster size equal to the actual cluster size. Pfeffermann and others (1998) tentatively recommend using scaling method 1 when the level 1 weights are non-informative and using scaling method

2 when the level 1 weights are informative. To test whether the sampling weights are informative, the method described by Pfeffermann was used (1993). Briefly, the ignorability of the design is tested by examining the significance of the difference between the estimated coefficients when sampling weights are included and excluded. The test statistic is constructed as

$$(\hat{\theta}_w - \hat{\theta}_0)'[V(\hat{\theta}_w) - V(\hat{\theta}_0)]^{-1}(\hat{\theta}_w - \hat{\theta}_0)$$

where $\hat{\theta}_w$ is the optimal estimate of the parameters when the sampling weights are included and $\hat{\theta}_0$ is the best estimate of the parameters when the sampling weights are excluded. $V(\hat{\theta})$ is the variance-covariance matrix. By comparing the test statistic to the appropriate chi-square distribution, one can determine whether incorporating the sampling design is necessary. The results from either scaling method 1 or scaling method 2 are reported, depending on if the level 1 weights are informative and the scaling method chosen is indicated in the table. The alternative scaling method was examined, as well as an analysis with no weights, and the results compared with the chosen scaling method, as a sensitivity analysis. The selection of the scaling method for the weights was not pre-specified in the replication plan, though the GLMM model was pre-specified.

The amount of variability due to both the EA and the individual were estimated. As prespecified, if the GLMM estimated ORs differ from those initially reported by more than 10 percent for the primary outcome variables (HIV and HSV-2), it is concluded that the results are somewhat sensitive to the model choice.

This model can address some concerns that Webb and others' *Lancet* commentary (2012) makes regarding this article. They make an important point that cluster-level baseline characteristics are not reported in the paper as recommended by the CONSORT guidelines (Campbell et al. 2006). Cluster-level statistics were calculated from the GLMM analysis, including intercluster variability and other cluster-level statistics, such as median and range for the number of subjects. It is important to report cluster-level characteristics because the randomization occurred at the cluster level. In randomized clinical trials where individuals are randomized, the randomization ensures that differences at baseline are a result of chance and not systematic. However, this assumption does not hold in cluster randomized trials; therefore, reporting as much baseline data as possible for both individuals and clusters is important for interpretability and assessment of potential bias (Campbell et al. 2006).

The intraclass correlation coefficient (ICC) is the proportion of the total variance that can be explained by variance between clusters. ICC can also be interpreted as the degree of correlation within clusters (Campbell et al. 2006). Based on the revised CONSORT guidelines for cluster randomized trials, in addition to reporting cluster-level baseline characteristics, ICC should be reported for each outcome (Campbell et al. 2006). These ICC values will aid in the design of future similar studies and allow readers to assess how much correlation exists at the cluster level. ICC was calculated using the ANOVA estimator along with 95 percent CIs using a modified Wald test.

3.1.2 GLMM results

Table 5 provides cluster-level summary statistics for the baseline schoolgirls and baseline dropouts. The median, minimum and maximum numbers of subjects per cluster, per intervention group for each of the original baseline characteristics are reported. The cluster-level baseline characteristics are similar for the control and intervention groups for the baseline dropouts and the two treatment arms of the baseline schoolgirls; however, there are some dissimilarities when comparing the clusters of the control and pooled intervention of baseline schoolgirls. The median number of individuals per cluster is quite different between the control and combined intervention groups of the baseline schoolgirls. The number of individuals per cluster has a wide range as well, with as few as one girl per cluster and a maximum of 41 girls per cluster. See Table 5 for the full results.

Table 1: Cluster characteristics of participants

	Median # individuals in a cluster (min, max)			
		Intervention group		
	Control	Pooled	Conditional cash transfer program	Unconditional cash transfer program
Baseline schoolgirls				
Number of individuals	15 (3, 41)	7.5 (1, 37)	6.5 (1, 37)	7 (2, 26)
Ever had sexual intercourse	3 (0, 12)	2 (0, 15)	2 (0, 15)	2 (0, 5)
Ever pregnant	0 (0, 3)	0 (0, 2)	0 (0, 2)	0 (0, 1)
Mother alive	13.5 (1, 38)	7 (1, 31)	6 (1, 31)	7 (2, 25)
Father alive	11 (1, 32)	6 (1, 31)	5 (1, 31)	6 (1, 18)
Female-headed household	4 (0, 18)	2 (0, 15)	1.5 (0, 9)	2 (0, 15)
Household owns a radio	8 (1, 25)	5 (1, 24)	4 (0, 24)	5 (1, 16)
Household owns a television	1 (0, 17)	1 (0, 15)	0 (0, 15)	2 (0, 12)
Household has access to mobile telephone	8 (0, 25)	4 (0, 23)	4 (0, 23)	5 (0, 17)
Electricity available in dwelling	0 (0, 15)	0 (0, 16)	0 (0, 16)	0 (0, 10)
Piped water available in dwelling	2.5 (0, 30)	0 (0, 20)	0 (0, 18)	2 (0, 20)
Baseline dropouts				
Number of individuals	3 (1, 14)	3 (1, 21)		
Ever had sexual intercourse	2.5 (0, 11)	3 (0, 17)		
Ever pregnant	2 (0, 8)	1 (0, 10)		
Mother alive	3 (0, 10)	3 (0, 18)		
Father alive	2 (0, 10)	2 (0, 14)		
Female-headed household	1 (0, 9)	1 (0, 11)		
Household owns a radio	2 (0, 7)	1 (0, 10)		
Household owns a television	0 (0, 3)	0 (0, 5)		
Household has access to mobile telephone	1 (0, 10)	2 (0, 11)		
Electricity available in dwelling	0 (0, 3)	0 (0, 5)		
Piped water available in dwelling	0 (0, 10)	0 (0, 9)		

Table 6 displays the ICC for each outcome, stratified by baseline schoolgirls and dropout cohorts. The ICC values tend to be very small, most near zero, or even negative in some cases. Some ICCs had negative variance values, which did not allow for the calculation of CIs. When negative values of the ICC occur, according to Wu and others (2012), it is

common practice to use zero or a small positive value for the ICC when performing sample size and power calculations. These extremely small ICCs show that individuals within a cluster are not correlated in terms of outcome. These results may not be unexpected in light of the results in Table 5, where it can be seen that some clusters contain very few individuals. However, the baseline dropout cohort had two ICCs that were somewhat higher, at 0.24 for enrolled during the 2008 school year and 0.17 for HSV-2 prevalence; for those two variables, the girls within a cluster showed a higher level of correlation.

Table 2: Intracluster correlation

	ICC (95% CI)
Baseline schoolgirls	
Enrolled during 2008 school year	0.02 (-0.11, 0.15)
Ever_married	0.02 (-0.23, 0.27)
Currently pregnant	-0.01 (., .)
Sexual debut	0.03 (-0.18, 0.23)
Had unprotected sexual intercourse	0.04 (-0.29, 0.36)
Had sexual intercourse once per week	0.02 (-0.19, 0.22)
Had sexual partner aged >= 25	-0.03 (-0.33, 0.28)
Had an HIV test	0.08 (-0.28, 0.45)
Knows that a healthy looking person can have HIV	0.01 (-0.07, 0.08)
Knows that HIV can be transmitted through breastfeeding	-0.01 (., .)
Received health training about HIV/AIDS	0.01 (-0.05, 0.07)
HIV prevalence	0.05 (-0.87, 0.97)
HSV-2 prevalence	-0.01 (-0.07, 0.04)
Syphilis prevalence	-0.03 (-0.71, 0.66)
Baseline dropouts	
Enrolled during 2008 school year	0.24 (-1.04, 1.53)
Ever_married	0.04 (-0.24, 0.32)
Currently pregnant	0.05 (-0.44, 0.54)
Sexual debut	0 (-0.23, 0.23)
Had unprotected sexual intercourse	0.03 (-0.16, 0.22)
Had sexual intercourse once per week	0.02 (-0.12, 0.16)
Had sexual partner aged >= 25	0.06 (-0.54, 0.67)
Had an HIV test	0.04 (-0.2, 0.28)
Knows that a healthy looking person can have HIV	0.04 (-0.39, 0.46)
Knows that HIV can be transmitted through breastfeeding	0.06 (-0.75, 0.88)
Received health training about HIV/AIDS	0.07 (-0.28, 0.41)
HIV prevalence	0.07 (-0.57, 0.7)
HSV-2 prevalence	0.17 (-1.43, 1.76)
Syphilis prevalence	-0.02 (-0.41, 0.37)

Note: ICC is calculated using the ANOVA estimator and 95% CIs are calculated using the modified Wald test.

Next, GLMM methodology was used to examine the robustness of the estimation method as described in the methods section. Table 7 provides the adjusted ORs and 95 percent CIs using the GLMMs. Along with the adjusted ORs, also provided are the p-value for the OR, the estimate of the random intercept and its standard error, the ICC and the scaling method used. We have also included the original adjusted OR in the table for comparison. The ICC that is included in Table 7 cannot be directly compared to the ICC previously reported in Table 6, since the ICC in Table 7 uses the log scale and the ICC in

Table 6 uses the proportion scale. Additionally, the ICC in Table 7 is the ICC after adjusting for baseline measurements, age and geographic location.

The baseline schoolgirls displayed more sensitivity to the model selection than did the baseline dropouts. When modeling with GLMMs, the statistical significance of the effects of intervention for the examined outcomes tended to coincide with the original results. However, there were two outcomes where the effects of the intervention did change statistical significance. Most notable was the significance of the intervention when examining HIV prevalence for the baseline schoolgirls. In the original analysis, the intervention significantly lowered the odds of HIV prevalence compared to control. However, in a multilevel model, the intervention effect was not statistically significant for individuals after adjusting for baseline characteristics. The point estimate increased by slightly more than 50 percent, from 0.36–0.54, and the 95 percent CI widened from 0.14–0.91 to 0.19–1.54 when using a GLMM. One of the main results of the paper appeared to be especially sensitive to model selection for the cohort of baseline schoolgirls. The narrower CI in the original analysis may be due to the fact that the logistic regression with robust standard errors included the observed variability in the data which could be smaller by chance from sampling variability. The wider CIs in the GLMM analysis are constructed by including random effects where the variability measure was model based. HSV-2 point estimates are also sensitive, changing by 40 percent, but they do not change statistical significance. The other outcome that changed statistical significance was sexual debut. In the GLMM analysis, the intervention was found to reduce the likelihood of sexual debut during the course of the study, compared to the control group, by 39 percent (OR=0.61, 95% CI=0.40–0.93), whereas the original analysis did not find a significant reduction. The ORs were similar in the two analysis methods, but the CI narrowed in the GLMM analysis. See Table 7 for full results. The results that changed statistical significance are highlighted.

The individual intervention arms (CCT and UCT groups) of the baseline schoolgirls showed similar sensitivity. Again, the effects of treatment on the outcome HIV prevalence were especially sensitive to the model selection. When modeling using GLMM, the adjusted ORs are not significant when comparing HIV prevalence of a baseline schoolgirl in the CCT group versus the control group with the same baseline characteristics. In the original paper, the adjusted OR was 0.29 with a 95 percent CI of 0.09 to 0.98. When using a GLMM to model the odds of HIV prevalence, the point estimate is OR=0.42, with 95 percent CI of 0.12 to1.51. HSV-2 results are also sensitive in the CCT group, changing by 59 percent, but they do not change statistical significance. Of the sexual behavior outcomes, had sexual partner ≥25 years changed to be non-significant in the CCT arm and sexual debut changed to significant in the UCT arm. Both the ORs and confidence limits changed substantially for both these outcomes. The OR for had sexual partner ≥25 years changed from 0.07 in the original analysis to 0.17 in the GLMM analysis and the OR for sexual debut changed from 0.72 to 0.56 using GLMM. Upon GLMM analysis, the 95 percent CI widened for the outcome had sexual partner ≥25 years and narrowed for sexual debut. The full analysis is included in Table 8. The original results have been incorporated in the table for comparison. Outcomes that changed statistical significance are highlighted.

Table 3: Effects of cash transfer intervention on outcome measures modeled using GLMM

	Original Adjusted odds ratio (95% CI)	GLMM Adjusted odds ratio (95% CI)	P-value	Random effects estimate (std. err.)	ICC	Scaling method	Percent difference
Baseline schoolgirls							
HIV prevalence	0.36 (0.14-0.91)	0.54 (0.19, 1.54)*	0.25	0.49 (1.08)	0.13	2	50.7%
HSV-2 prevalence	0.24 (0.09-0.65)	0.34 (0.14, 0.83)	0.02	0 (--)	0	2	40.7%
Enrolled during 2008 school year	1.62 (1.07-2.45)	1.29 (0.86, 1.92)	0.22	0.15 (0.09)	0.04	1	-20.5%
Had sexual partner >25 years	0.21 (0.07-0.62)	0.31 (0.11, 0.87)	0.03	0 (--)	0	2	46.4%
Had unprotected sexual intercourse	1.08 (0.67-1.75)	1.07 (0.69, 1.67)	0.76	0 (--)	0	1	-0.8%
Had sexual intercourse once per week	0.46 (0.26-0.82)	0.56 (0.33, 0.97)	0.04	0.16 (0.13)	0.05	1	22.5%
Sexual debut	0.64 (0.38-1.07)	0.61 (0.40, 0.93)*	0.02	0.04 (0.09)	0.01	1	-4.1%
Baseline dropouts							
HIV prevalence	1.37 (0.72-2.61)	1.44 (0.76, 2.74)	0.27	0 (--)	0	1	5.0%
HSV-2 prevalence	1.03 (0.47-2.24)	1.08 (0.45, 2.55)	0.87	0.77 (0.91)	0.19	1	4.6%
Enrolled during 2008 school year	8.77 (5.07-15.1)	10.02 (5.40, 18.58)	<.0001	0.18 (0.16)	0.05	1	14.3%
Had sexual partner >25 years	0.79 (0.42-1.50)	0.93 (0.48, 1.80)	0.82	0 (--)	0	1	17.2%
Had unprotected sexual intercourse	0.74 (0.44-1.23)	0.76 (0.46, 1.26)	0.28	0.12 (0.15)	0.03	1	2.7%
Had sexual intercourse once per week	0.53 (0.32-0.86)	0.53 (0.33, 0.85)	0.01	0 (--)	0	1	0.7%
Sexual debut	0.70 (0.33-1.45)	0.67 (0.32, 1.37)	0.27	0 (--)	0	1	-4.8%

Note: * Result changed from statistically significant to non-significant or non-significant to statistically significant when compared to the original paper.

Table 4: Effects of conditional or unconditional cash transfers on baseline schoolgirls by outcome measures modeled using GLMM

	CCT vs. control Original adjusted odds ratio (95% CI)	CCT vs. control GLMM adjusted odds ratio (95% CI)	UCT vs. control Original adjusted odds ratio (95% CI)	UCT vs. control GLMM adjusted odds ratio (95% CI)	Original Heterogeneity of odds ratios (p-value)	GLMM Heterogeneity of odds ratios (p-value)
HIV prevalence	0.29 (0.09-0.98)	0.42 (0.12, 1.51)*	0.47 (0.14-1.59)	0.65 (0.17, 2.40)	0.57	0.60
HSV-2 prevalence	0.37 (0.13-1.03)	0.59 (0.23, 1.50)	0.08 (0.01-0.58)	0.12 (0.02, 0.81)	0.16	0.12
Enrolled during 2008 school year	2.08 (1.14-3.82)	1.81 (1.13, 2.90)	1.22 (0.77-1.96)	1.01 (0.61, 1.67)	0.14	0.06
Had sexual partner >25 years	0.08 (0.01-0.60)	0.17 (0.02, 1.14)*	0.36 (0.11-1.19)	0.42 (0.13, 1.37)	0.19	0.40
Had unprotected sexual intercourse	1.17 (0.67-2.05)	1.39 (0.84, 2.30)	0.96 (0.50-1.83)	0.79 (0.43, 1.45)	0.59	0.10
Had sexual intercourse once per week	0.53 (0.26-1.07)	0.76 (0.41, 1.42)	0.37 (0.16-0.85)	0.40 (0.18, 0.86)	0.49	0.16
Sexual debut	0.58 (0.29-1.15)	0.68 (0.38, 1.22)	0.72 (0.37-1.40)	0.56 (0.33, 0.96)*	0.62	0.60

Note: * Significance level is in a different direction from the original result.

3.2 Permutation test

In Webb and others' *Lancet* commentary, the authors state that "the point estimate without clustering had a very wide CI and was not significant and only after significant adjustment was there a significant finding" (2012). In that same *Lancet* commentary, the original authors reply that "sampling weights are used to account for the fact that younger girls and those living in urban areas were sampled at a lower rate in the study design" (Baird et al. 2012b). Since the design of the study incorporates multistage sampling and unequal sampling probabilities, the analysis must include those components to have unbiased results. Crude ORs that are not adjusted for the sampling design can be calculated based on the data provided, but will be biased. Group permutation-based methods that account for the cluster randomization and the intraclass correlation of EAs are used to explore the critique by Webb and others that the results are sensitive to the adjustment of weights and cluster size. Group permutation methods account for the dependent nature of outcomes among study participants in the same area. An advantage of group permutation testing is that no distributional or modeling assumptions need to be specified. Note that Peterson and others (2000a) did not use covariates in their primary analysis to maintain the model-free nature of randomization-based permutation methods.

In permutation testing, the EA is considered to be the experimental unit, and thus accounts for the intraclass correlation within EAs by permuting the areas rather than individuals. The permutation test statistic used is the difference in overall average between the control and experimental groups, but others can be utilized as well (Peterson et al. 2000b). Permutation tests were used by the well-known statistician Sir R.A. Fisher (1935). These methods can be employed when asymptotic theory does not apply, for example with small sample sizes. The real advantage is that they require few distributional assumptions, as mentioned earlier. Although these methods may not be as powerful as parametric methods, there are instances where they have greater power (Anderson and Legendre 1999).

3.2.1 Permutation test methods

In general, hypothesis testing begins with the assumption that the null hypothesis of no treatment effect is correct, and the test statistic and the sampling distribution of the test statistic are derived under the null hypothesis. For permutation tests, the procedure is essentially reversed. For permutation testing, the procedure is as follows:

1. Define a test statistic that is large if the treatment effect is large and small if the treatment effect is small.
2. Define the null hypothesis.
3. Create a new data set consisting of your data, randomly rearranged (permutations) keeping the clusters intact.
4. Calculate the test statistic for the randomly arranged data set and compare it to the observed test statistic.
5. Repeat steps 3 and 4 several hundred times.
6. If the observed test statistic is greater than 95 percent of the randomly generated test statistics, then reject the null hypothesis at p-value<0.05.

The effects of the cash transfer program on HIV and HSV-2 prevalence for the baseline schoolgirls was examined by performing an approximate clustered permutation test as

outlined by Stedman and others (2009). To get an approximation of the achieved significance level (ASL), which can be interpreted as a permutation p-value, a sample of 100 permutations of the Wald chi-square statistics was taken from a logistic regression that adjusted for the probability of unequal selection. The number of permutations chose to reduce Monte Carlo error was based on recommendations in Stedman and colleagues (2009).

3.2.2 Permutation test results

To demine the effect of intervention on HIV prevalence in baseline schoolgirls in an unadjusted model, a sample of 100 permutations of the Wald chi-square statistics was taken, and the ASL was approximated to be 0.08 with a coefficient of variation of 33.9 percent. To reduce the Monte Carlo error around the ASL from 33.9 percent to less than 5 percent, we needed at least 4,600 permutations. We ran 10,000 permutations that resulted in an ASL of 0.11 with a coefficient of variation of 3.4 percent. This was repeated adjusting for age and stratum. Ten thousand permutations resulted in an ASL of 0.07 with a coefficient of variation of 4.0 percent. Therefore, the effect of the cash transfer program on HIV prevalence was not statistically significant by permutation test.

A similar procedure was followed to examine the effects of the intervention on HSV-2 prevalence in baseline schoolgirls. Based on 500 permutations, the approximate ASL was 0.008 with a coefficient of variation of 49.8 percent. The Monte Carlo error around the ASL can be reduced to be 11.1 percent by running 10,000 permutations. Based on these 10,000 permutations, the ASL was 0.009. Again, the procedure was repeated using an adjusted model. Ten thousand permutations resulted in an ASL of 0.01 with a coefficient of variation of 8.4 percent. Therefore, the effect of the cash transfer program on HSV-2 prevalence was highly statistically significant by permutation test.

The results of the permutation test further highlight the significance of the distributional assumptions when examining the effects of the cash transfer program on HIV prevalence for baseline schoolgirls. As previously stated, permutations tests may be more conservative than parametric methods. However, clustered permutations tests tend to preserve Type I error, while parametric methods may seem more powerful due to a shift in the power curve that inflates the Type I error (Stedman et al. 2009). A decrease in power with a nominal Type I error may be preferable to an increase in power with an inflated Type I error. A non-significant result from a permutation test indicates that additional studies need to be performed.

3.3 Bivariate outcome estimation

The primary outcomes of this study include the prevalence of HIV and HSV-2 at 18 months, and prevalence of syphilis is a secondary outcome. It is possible that these binary outcomes are correlated with each other through risky sexual behavior; however, HIV prevalence rates can include girls who were perinatally infected (mother-to-child transmission of HIV, which can occur in utero, during labor and delivery, or postnatally through breastfeeding) which would not be associated with risky behaviors, and in that case, they would not be correlated (Shetty 2005). We model the interrelationship between risk factors (including the intervention), and between the bivariate outcomes HIV and HSV-2 prevalence, simultaneously in a multivariate model.

By employing a multivariate model, it is possible to gain precision compared to estimating separate models for each outcome. This could be a significant advantage when event rates are small, such as in this study, by increasing precision. This method could be of particular use for examining the difference in the rates of HIV/HSV-2 in the conditional versus unconditional cash transfer groups. The study was underpowered to detect a difference between those groups, and if the two outcomes are highly correlated, then by modeling the covariance structure we can gain precision in the estimates. If there is an actual gain in precision, then the bivariate model would automatically increase the power for that comparison (Lu and Yang 2012). One disadvantage to this type of design is that the models do not always converge to a solution.

3.3.1 Bivariate outcome methods

A multivariate approach is applied, using alternating logistic regression (ALR) with generalized estimating equation methodology to fit a simultaneous survey logistic regression to multiple binary outcomes, specifically HIV and HSV-2 prevalence (Lu and Yang 2012). This method allows for the complete modeling of the data in one analysis, testing correlations between multiple outcomes and directly estimating the difference in the association between risk factors and multiple outcomes. It can test the effect of the intervention on HIV and HSV-2 prevalence separately and the effect of the intervention on HIV and HSV-2 prevalence combined, and allows for an interaction term to test whether the effect of the intervention on HIV and HSV-2 prevalence is the same or different. This analysis method also permits the testing of correlations between multiple binary outcomes. The ALR was fit using PROC GENMOD and modeling the association between pairs of responses using log ORs.

3.3.2 Bivariate outcome results

A bivariate model was fit to model HIV and HSV-2 prevalence simultaneously. Each cohort, i.e. baseline schoolgirls and baseline dropouts, was analyzed separately. Unfortunately, using the ALR approach to model the correlated binary responses using log odds did not converge, i.e. find a valid solution, for the baseline schoolgirls. The results for baseline dropouts are presented in Table 9, along with the original adjusted ORs and 95 percent CIs. The baseline dropouts showed slight deviation in their ORs and 95 percent CIs based on the model specification. The generalized estimating equation model estimated the OR for within-cluster dependence to be 7.56, p-value<0.0001 after adjusting for covariates. This indicates that the odds of HIV positivity given that HSV-2 is positive are 7.56 times that of HIV positivity given that HSV-2 is negative. The outcomes are highly dependent on each other.

Table 5: Effects of cash transfer intervention on HIV and HSV-2 prevalence modeled using ALR, in baseline dropouts

	Original adjusted odds ratio (95% CI)	ALR adjusted odds ratio (95% CI)
HIV prevalence	1.37 (0.72-2.61)	1.33 (0.68, 2.62)
HSV-2 prevalence	1.03 (0.47-2.24)	1 (0.48, 2.07)

4. Theory of change analysis

The study was extended in a theory of change analysis in three ways: 1) by directly evaluating the effects of the intervention on improving the HIV awareness, i.e. having an HIV test, or gaining of HIV knowledge; 2) a wealth index for the participants was computed using principal component analysis based on available data at baseline and then we evaluated whether it influenced the effect of the intervention through an interaction (Filmer and Pritchett 2001); and 3) the causal pathway implied by the study was tested.

4.1 Theory of change methods

4.1.1 HIV awareness methods

Principal component analysis (PCA) is a commonly used data reduction technique that can be employed in constructing composite variables for subsequent analyses. PCA takes a set of correlated variables and transforms them into a set of uncorrelated linear components. The first principal component explains the most variance, the second principal component explains less variance than the first component and so on. Each variable that makes up the component has a positive or negative weight associated with it that gives the direction of the association, so variables that are expected to be positively associated with higher HIV awareness or higher wealth, for example, would have a positive weight and variables associated with lower HIV awareness or wealth would have a negative weight. PCA is one method used to construct wealth indices and has been adopted by the World Bank and Macro International Inc. (now ICF International) for use with Demographic and Health Surveys as a proxy for socioeconomic position (Howe et al. 2008). Howe and others studied alternate methods for wealth index construction and found them to be equally associated with per capita consumption expenditure. We use PCA in this study for construction of a wealth index and likewise in the construction of an HIV awareness variable.

A composite HIV awareness variable was created using PCA, based on several of the survey variables, including had an HIV test, knows a healthy-looking person can have HIV, knows that HIV can be transmitted through breastfeeding and received health training about HIV/AIDS. The intervention effect on this composite HIV variable was examined using a linear regression model with PROC SURVEYREG, adjusting for baseline levels of knowledge, the subject's age, and geographic area (urban, rural and far rural). Received health training about HIV is not included in the baseline awareness, since it was measured only at 12 months. A possible interaction between age and HIV awareness is considered (this interaction was not prespecified in the replication plan, though the rest of the model was prespecified).

4.1.2 HIV awareness results

The intervention did not have a significant effect on HIV awareness for either baseline schoolgirls or baseline dropouts, p-value 0.40 and 0.56, respectively. The regression coefficient for baseline schoolgirls was 0.04 with a standard error of 0.04, while the baseline dropouts had a regression coefficient of 0.04 and standard error of 0.07. Neither cohort showed a significant interaction between age and HIV awareness. The best predictor of HIV awareness at 12 months was baseline knowledge. Both groups revealed that baseline knowledge was a significant predictor of HIV awareness at 12 months, p-

value 0.0009 for baseline schoolgirls and <0.0001 for baseline dropouts. With the baseline schoolgirls, being older was also a significant predictor of HIV awareness at 12 months, p-value <0.0001. Geographic location was not significant for either cohort.

4.1.3 Wealth index methods

A wealth index was constructed using variables collected at baseline, which are shown in Tables 1A and 1B and include mother alive, father alive, female-headed household, household owns a radio, television, access to a mobile telephone, electricity and piped water available. From these variables, a PCA was conducted to produce the two wealth indexes, such as described by Wamani and others (2004). The wealth index variables were tested in a multiple logistic regression model, along with the intervention, the interaction between the wealth indexes and the intervention variable, the age of the girl, and geographic area (taking into account the design of the study as well as the sampling weights with PROC SURVEYLOGISTIC).

Models for HIV and HSV-2 prevalence at 18 months were run separately, as were separate models for baseline schoolgirls and baseline dropouts. One might expect that the cash transfer intervention would be most effective in poorer households. As Pettifor and others (2012) point out, “conditioning payments on school attendance may only be relevant in settings where there is a financial barrier to schooling.” By looking for interactions with the wealth indexes, we can begin to determine if this type of intervention is unequally effective based on the wealth of the individual. This may be most interesting in the baseline dropouts cohort of the study. This group of subjects may be in most need of the cash transfers to attend school and, by definition, is most at risk. With the wealth indexes, it can be determined if the effect of the intervention on the outcome is affected by wealth, i.e. is there less of an effect in higher wealth groups and more of an effect in the lower wealth groups.

4.2 Theory of change results

4.2.1 Wealth index results

Based on the PCA, it was natural to create two wealth variables. The first wealth index, “wealth item,” included the variables household owns a television, access to a mobile telephone, electricity available in dwelling, and piped water available in dwelling. While the second wealth index, “wealth family,” included the variables mother alive, father alive, female-headed household. The variable household owns a radio was not correlated with either index, so was excluded.

Neither wealth index influenced the effect of the cash transfer program on HIV prevalence or HSV-2 prevalence for either the baseline schoolgirls or baseline dropouts. The interaction between wealth item x intervention and wealth family x intervention had p-values of 0.81 and 0.63, respectively, for the cohort of baseline dropouts when examining the outcome HSV-2 prevalence. When considering HIV prevalence for baseline dropouts, the respective p-values were 0.44 and 0.61. However, when examining the intervention arms separately in the baseline schoolgirls, there was a significant interaction between the family structure and intervention for the UCT arm for both HIV and HSV-2 prevalence. The interaction between “wealth family” and intervention was further examined by categorizing the wealth family variable into low or high (divided at median value of schoolgirls). The odds of HIV or HSV-2 was less in the

UCT group compared to control in both the wealth family high and low groups, based on point estimates and CIs (Table 11B). However, the odds of HIV or HSV-2 were even smaller in the UCT group when the wealth family level is low, indicating that the UCT intervention was highly effective when the wealth family is low. Looking at the intervention effect in either low or high wealth index levels was not prespecified in the replication plan, but is a natural extension when a significant interaction is found. See Tables 10, 11A and 11B for full results.

Table 6: Effects of wealth indexes on HIV and HSV-2 prevalence

	Estimate	Std. error	P-value
HIV Prevalence			
Baseline schoolgirls			
Wealth item*intervention	-0.1802	0.6345	0.78
Wealth family*intervention	0.5788	0.6983	0.41
Baseline dropouts			
Wealth item*intervention	0.5315	0.6949	0.44
Wealth family*intervention	-0.2431	0.4776	0.61
HSV-2 Prevalence			
Baseline schoolgirls			
Wealth item*intervention	-1.8003	1.0969	0.10
Wealth family*intervention	1.6957	1.2539	0.18
Baseline dropouts			
Wealth item*intervention	0.165	0.6814	0.81
Wealth family*intervention	-0.2386	0.4906	0.63

Table 11A: Effects of wealth indexes on HIV and HSV-2 prevalence by treatment arm

	Estimate	SE	P Value
HIV Prevalence			
CCT Schoolgirls			
Wealth item*intervention	-0.2267	0.5516	0.68
Wealth family*intervention	-0.6977	1.5386	0.65
UCT Schoolgirls			
Wealth item*intervention	-0.3383	1.0145	0.74
Wealth family*intervention	1.4207	0.5398	0.0085
HSV-2 Prevalence			
CCT Schoolgirls			
Wealth item*intervention	-2.0758	1.4347	0.15
Wealth family*intervention	1.4800	1.3251	0.26
UCT Schoolgirls			
Wealth item*intervention	0.8309	0.4836	0.086
Wealth family*intervention	3.9652	1.5342	0.0098

Table 11B: Effects of categorized wealth index on HIV and HSV-2 prevalence by intervention arm in baseline schoolgirls

	Odds Ratio*	95% CI	P Value
Wealth_Family (HIV)			
CCT low^	0.1602	(0.0150, 1.7111)	0.1282
CCT high	0.1090	(0.0118, 1.0094)	0.0510
UCT low	0.2005	(0.0365, 1.1027)	0.0644
UCT high	0.2934	(0.0557, 1.5448)	0.1462
Wealth_Family (HSV2)			
CCT low	0.0723	(0.0048, 1.0949)	0.0580
CCT high	0.2847	(0.0710, 1.1414)	0.0757
UCT low	0.0070	(0.0023, 0.0211)	<.0001
UCT high	0.1335	(0.0165, 1.0776)	0.0586

Note: *All categories are compared to control group. ^low indicates lower levels of wealth

4.3 Causal pathway

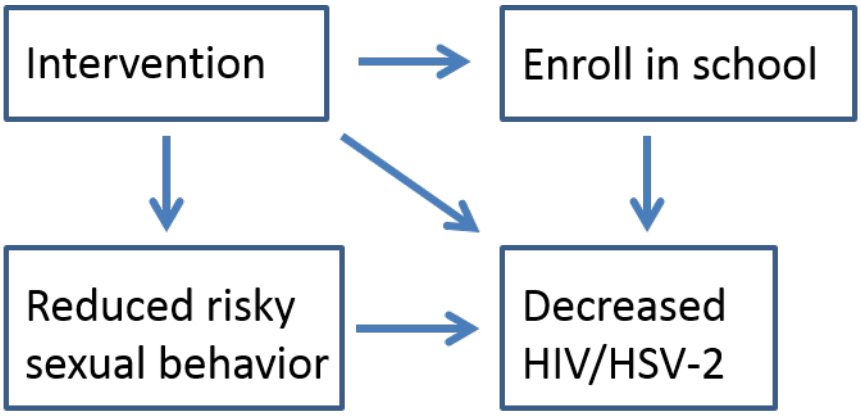
The authors of the paper examined whether the intervention had an effect on whether the participant enrolled in school in 2008, on the prevalence of risky sexual behaviors and on the prevalence of HIV and HSV-2 at 18 months in univariate and multivariate models. Here, the direct relationship between school enrollment and risky sexual behaviors (sexual debut, had unprotected sexual intercourse, had sexual intercourse once per week, and had a sexual partner aged ≥25 years) with HIV and HSV-2 prevalence at 18 months are examined (Figure 1).

The intervention of cash transfers lasted from baseline to 24 months; enrolled in school in 2008 and sexual behaviors are measured at 12 months during the intervention period; and prevalence of HIV and HSV-2 are both measured at 18 months. Since enrolled in school in 2008 and sexual behaviors are measured before HIV and HSV-2, it should be valid to look at the association between these variables. Baird and colleagues have looked extensively at the connection between the intervention and school enrollment, but the direct link between enrollment in school and risky behaviors and HIV/HSV-2 prevalence has not been assessed in this study.

Associations between enrolled in school in 2008 and risky behaviors can also be examined, but the direction of the relationship cannot, since they were measured at the same time point. In addition, it is difficult to measure the direction of the relationship between sexual behavior and school enrollment because these two variables are both likely influenced by the intervention. As a technical point, when two variables are measured simultaneously, we cannot say that one caused the other, only that they are associated. The cause and effect aspect cannot be truly assessed in this study other than the intervention effect on each variable individually. Therefore, the cause and effect relationship cannot be directly determined for enrolled in school in 2008 and sexual behaviors. Another potential pathway would be intervention affects school enrollment, which in turn affects risky behaviors and then HIV/HSV-2 prevalence, but it is likely that they are all interconnected and difficult to tease out. Based on the timing of the variables described above and the fact that the intervention is likely affecting both variables, the mediator effect for enrolled in school in 2008 and sexual behaviors is difficult to resolve.

There are many potential pathways for how the intervention effects HIV and HSV-2 prevalence; however, we prespecified the pathway shown in Figure 1 in the replication proposal. In this report, we test only the prespecified relationships shown in Figure 1.

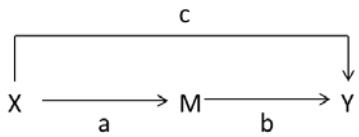
Figure 1: Pathway for reduced HIV/HSV-2 prevalence



4.3.1 Causal pathway methods

Pathway-specific effects are explicitly investigated to see how much of effect of the intervention are mediated through reduced sexual behavior and enrollment in school. According to Baron and Kenny (1986), "a given variable may be said to function as a mediator to the extent that it accounts for the relationship between the predictor and the criterion." To do this, a four-step approach proposed by Baron and Kenny was used. This method involves a series of four regression models shown pictorially in Figure 2. X is the intervention variable; M is the mediator variable (school enrollment or risky sexual behaviors); Y is the outcome variable; a and b are direct effects; and c is the direct effect of X on Y.

Figure 2: Mediator pathway



To test this, the following models were run:

	Analysis
Step 1	Predict Y with X to test for path c. $E[Y]=B_0+B_1X$
Step 2	Test for path a, the effect of X on M. $E[M]=B_0+BX$
Step 3	Test for path b, the effect of M on Y. $E[Y]=B_0+B_1M$
Step 4	Multiple regression with X and M predicting Y. $E[Y]=B_0+B_1X+B_2M$

If one or more of these relationships are not significant, then we can conclude that mediation is not likely in this case. If relationships exist in steps 1−3, then step 4 is considered. If the effect for M in the multiple regression model is significant, then the

conclusion is there is some form of mediation; if X is not significant, then it is full mediation; if both are significant, then the model supports partial mediation. Full mediation would indicate that the intervention affects the mediator variable (school enrollment or risky sexual behaviors), then these mediators in turn affect the outcome (HIV/HSV-2 prevalence), and after controlling for the mediator variable, the relationship between the intervention and outcome is no longer significant, i.e. the effect goes to zero. Partial mediation indicates that the relationship between intervention and outcome is still significant, but may be reduced. This can occur when an outcome can have multiple causes (there are multiple mediating factors). This means that the mechanism of the intervention on HIV/HSV-2 prevalence cannot be fully explained by the mediating variable and that the relationship between the two is complex (Baron and Kenny 1986). The models will employ the survey weights and clustering, as in the logistic regression models described in the original paper, adjusting for baseline characteristics and geographical location using PROC SURVEYLOGISTIC.

4.3.2 Causal pathway results

When examining possible mediators, the analysis concentrated on the baseline schoolgirls. Baseline dropouts were excluded from the analysis, since the effects of the cash transfer program on HIV and HSV-2 prevalence were not statistically significant and therefore not subject to mediation. School enrollment, had unprotected sexual intercourse, had sexual intercourse once per week, and had sexual partner aged ≥25 years were considered as possible mediators of the effects of intervention on HIV prevalence or HSV-2 prevalence, as prespecified in Figure 1.

Step 1 results show the effect of intervention on HIV and HSV-2 prevalence, adjusting for age group and stratum, and are the same as found in the original manuscript; there is a significant intervention effect on outcome in baseline schoolgirls. Step 2 examines the effect of intervention on the various mediators. The intervention is significantly associated with school enrollment, had a sexual partner aged ≥25 years, and had sexual intercourse once per week. Intervention was not significantly associated with had unprotected sexual intercourse; therefore, it is no longer considered a mediating variable. Step 3 looks at the effect of the mediators on outcome. All the variables considered are associated with both HIV and HSV-2 prevalence.

Finally, we evaluate the actual mediator status in Step 4. For HIV prevalence, the mediator conclusion is challenging to assess because for both the mediator variable and the intervention, the p-values testing the effect hover near 0.05, either just above or below it. If we use a strict 0.05 interpretation for significance, then we find that having a sexual partner ≥25 years of age results in full mediation, indicating that the intervention works through affecting the mediator variable. This result is shown in Table 12. When both the intervention and having a sexual partner aged ≥25 years are in the model, the intervention effect is no longer statistically significant but the mediating variable is significant. Enrolled in school and had sexual intercourse once per week are not mediators based strictly on a 0.05 cutoff, since the mediator variable is not significantly associated with outcome in Step 4 (Table 12). However, these p-values, at 0.052 and 0.062, are close to significant, so we would not rule them out as potential mediators of the intervention.

Looking at HSV-2, enrollment in school, sexual partner aged ≥25 years, and had sexual intercourse once per week resulted in partial mediation (both the intervention and mediator were significantly associated with HSV-2 in Step 4) between the effects of intervention and HSV-2 prevalence, indicating that the intervention effect on HSV-2 prevalence is not fully explained by school enrollment or by sexual behaviors.

Table 7: Mediator analysis of HIV and HSV-2 prevalence

	P-value					
	Step 1	Step 2	Step 3	Step 4		Conclusion
				Intervention	Mediator	
HIV prevalence						
Enrolled during 2008 school year	0.033	0.023	0.041	0.066	0.052	None
Had a sexual partner aged ≥25 years	0.033	0.005	0.023	0.051	0.048	Full
Had unprotected sexual intercourse	0.033	0.759	<0.001	-	-	None
Had sexual intercourse once per week	0.033	0.009	0.027	0.050	0.062	None
HSV-2 prevalence						
Enrolled during 2008 school year	0.006	0.023	0.004	0.016	0.008	Partial
Had a sexual partner aged ≥25 years	0.006	0.005	<0.001	0.009	<0.001	Partial
Had unprotected sexual intercourse	0.006	0.759	<0.001	-	-	None
Had sexual intercourse once per week	0.006	0.009	0.005	0.009	0.009	Partial

Notes:
Step 1: Intervention effect outcome (HIV or HSV-2 prevalence)
Step 2: Intervention effect on mediator (enrolled in school, sexual behavior)
Step 3: Mediator (enrolled in school, etc.) effect on outcome (HIV or HSV-2 prevalence)
Step 4: Intervention and mediator effect on outcome

5. Discussion

In this paper, a replication and sensitivity check of Baird and colleagues' *Effect of a cash transfer programme for schooling on prevalence of HIV and HSV-2 in Malawi* was performed. The paper clearly defined the research and analysis methodology and the original results were replicated in the pure replication portion of the study.

The measurement and estimation portion of the replication study examined alternative methodology for estimation in cluster randomized trials. The original analysis used logistic regression models, with robust standard errors, which allows for intraclass correlation, along with sampling weights to account for probability of inclusion that varied by age group and stratum. Also provided in the original paper are ORs adjusted for age group, geographical stratum and the baseline value if available. Alternative methods include GLMM, permutation tests and generalized estimating equations in a bivariate model. When examining the sensitivity of model selection on various outcomes in the MEA portion, the baseline schoolgirls showed sensitivity to model selection using GLMM and permutation testing. Specifically, the outcome HIV prevalence for baseline schoolgirls was particularly sensitive to the model selection in both analyses. HSV-2

prevalence results also showed sensitivity in the model estimates in the GLMM analysis, but statistical significance and the conclusion did not change. It is unclear as to why the point estimates changed depending on the estimation method. By permutation test methodology, which has no distributional assumptions, but preserve the clustered design, it was found that the cash transfer intervention had a highly significant effect on HSV-2 prevalence in the baseline schoolgirls cohort, but the HIV prevalence results were no longer statistically significant.

At this point, one might ask, which is the best method? Which analysis should be preferred? A literature search provided no comprehensive comparisons of the competing methods for cluster randomized trials with binary outcome and sampling weights. Green and Vavreck (2007) examine the robust standard errors methodology in a simulation study comparing to a GLMM random effects methodology. They found that robust standard errors are biased downward (too small) when the number of clusters are small and that GLMM random effects standard errors are closer to the empirical standard errors (the truth). However, as the number of clusters increase, the robust standard errors methodology becomes less biased. Green and Vavreck conclude that GLMM random effects models "provide the most accurate estimates and [standard errors]," even though the advantage is slight; however, they do not consider sampling weights in their comparisons. Peters and others (2003) compared robust standard errors methodology to GLMM random-effects logistic regression in a sensitivity analysis with actual trial data. Since this was not a simulation, the true parameter estimates are unknown, but we can compare the effects of covariate adjustment on the parameter estimates and standard errors. They found that the parameter estimates were similar between the two methods; however, adjustment for covariates drastically decreased the standard errors in the robust standard errors methodology. In the GLMM random effects models, the effect of covariate adjustment had only a minor effect on the standard error estimates. Pfeffermann and colleagues (1998) describe utilization of sampling weights in GLMM with random effects. They recommend caution using GLMM when within cluster sample sizes are small, because variance component parameters can become biased, although scaling the weights reduces the bias.

Given the large number of clusters randomized in this trial, all the methods accounting for the study design and weighting should give valid and similar results. Consistent results were found across methods for HSV-2 prevalence but not for HIV prevalence in the baseline schoolgirls. The likely cause for inconsistency in estimation results for HIV prevalence is the extremely low number of baseline schoolgirls with HIV at the end of the study, only 7 in the intervention group and 17 in the control group. This means that there were many clusters with no events. The original paper also points out this finding as a limitation to their study and recommends interpreting the results with caution (Baird et al. 2012a). This problem was also seen in the HSV-2 outcome, but the effect size was so large for this comparison that the significance did not change based on estimation method. Based on the lack of large-scale simulation studies comparing the various estimation techniques for cluster randomized trials, it is a good idea to compare the results of all the valid techniques and report these results as a sensitivity analysis (Thabane et al. 2013).

In the theory of change analyses, the intervention did not have a significant effect on HIV awareness for either baseline schoolgirls or baseline dropouts, awareness at 12 months

was found to be highly associated with baseline level awareness. No interaction was found to exist between the constructed wealth indexes and overall intervention on HIV or HSV-2 in baseline schoolgirls or dropouts. However, when looking at the UCT group in the baseline schoolgirls, there was a significant interaction between the wealth index for family and the UCT for both HIV and HSV-2 prevalence. “Wealth family” offered different levels of protection against HIV and HSV-2 depending on whether it was high or low. UCT in individuals with high wealth family offered less protection against HIV and HSV-2 than UCT in individuals with low wealth family. A larger value of wealth family is correlated with a schoolgirl’s mother and father being alive and the mother’s not being the head of the household. However, there were only four positive results for HIV and one positive result for HSV-2 for UCT schoolgirls, so these results are based on a very small number of events.

In the analysis of the causal pathway, had a sexual partner aged ≥25 years was found to be a full mediator of HIV prevalence. After adjusting for had a sexual partner aged ≥25 years, the intervention was no longer a significant predictor of HIV prevalence in baseline schoolgirls, if using a strict 0.05 cut point for statistical significance. School enrollment, had sexual intercourse once per week, and had sexual partner aged ≥25 years were found to be partial mediators of HSV-2 prevalence in baseline schoolgirls. The intervention effect was still significant for HSV-2 after adjusting for each of those variables. When treated as an outcome, had unprotected sexual intercourse was not related to the intervention. The pathway analysis for HIV prevalence was ambiguous, based on the borderline significant results in Step 4 of the mediation analysis. It appears that partial mediation of the intervention is occurring with school enrollment and sexual behaviors, had a sexual partner aged ≥25 years, and had sexual intercourse once per week with both HSV-2 prevalence and HIV prevalence. Based on these results, we can infer that the intervention is affecting HIV and HSV-2 prevalence partially through school enrollment and selected sexual behaviors.

6. Limitations

The original authors supplied a cleaned data set with deidentified individuals. Additionally, not all subjects were tested at 18 months for HIV, HSV-2 or syphilis, and no individuals were tested for these at baseline. Because of these limitations, the original sampling design was used and it is possible that the results are due to features of the sampling methodology. To overcome this limitation, a clustered permutation test that does not have any distributional assumptions was performed. However, permutation tests have their own limitations. The clustered permutation test is not as powerful as parametric tests that use distribution assumptions. Therefore, a permutation test may not provide sufficient power to detect small effect sizes.

Continuing the extended analysis, a hierarchal model was fit to examine the effects of cash transfer program on outcomes that were significant or close to being significant in the original paper. Using the hierarchal model (GLMM analysis), one of the main results – effect of intervention on HIV prevalence – was not statistically significant. However, the data set that was used did not contain the level 2 weights, and therefore, a finite population correction to the model could not be applied. The finite population correction would have reduced the variance and possibly kept the primary results significant. However, the finite population correction did not change the point estimate, in which we

also saw changes. Regardless of the availability of the level 2 weights, the baseline schoolgirls showed sensitivity to the model choice for both HIV and HSV-2 prevalence. Additionally, the hierarchal model is trying to measure a latent variable, the correlation between individuals within EAs. This correlation is specified in the model as a random intercept that is assumed to have a normal distribution with mean zero and variance σ^2. Should the distribution of this random intercept be misspecified, then the model is uninterpretable.

Finally, an ALR model was used to examine the primary outcome of the paper in a bivariate model, estimating the effect of intervention on HIV and HSV-2 prevalence simultaneously. The ALR model using generalized estimating equations is less sensitive to model misspecification and is a population-based model. Unfortunately, the model did not converge for the baseline schoolgirls.

7. Conclusions

The few cases of HIV in the baseline schoolgirls made the primary results especially sensitive to the model being used; however, the HSV-2 results were found to be more robust. Further, it could not be determined if the main results were influenced by the sampling design. Therefore, it is recommended that additional research is performed in assessing the effectiveness of cash transfer programs in reducing HIV and HSV-2 prevalence. Future studies should include attempts to increase the number of individuals included in each cluster (such as EA) as well as a longer intervention period and a follow-up period to determine whether the effect is sustained once the intervention is removed. Ideally, HIV and HSV-2 incidence would be measured rather than prevalence. In the present study, existing HIV and HSV-2 infections at baseline were not determined, so we do not have a measure of new infections that occurred following the intervention.

Appendix A: Summary table of the pure replication

This appendix contains a summary table of the differences found between the original paper and the replication analysis during the pure replication for the convenience of the reader.

Table A1: Summary of the pure replication

	Discrepancy	Replication	Original	Comments
Table 1				
	Row 1 totals (number of enumeration areas)	CCT group (n=26)	CCT group (n=25)	One enumeration area is in both the CCT and UTC groups in dataset
	Row 2 totals (denominator issues)	Larger in several cases	Smaller in several cases	Missing values for variables?
	Mother Alive	420/501	423/501	Typographical?
	Several SDs	Age at sexual debut		Within rounding
		Age (dropouts)		Within rounding
Table 2				
	Differences in denominators	Several		Typographical?
	Had sexual intercourse once per week (follow-up-Intervention)	22/499	22/299	%'s match therefore possibly typo
	Confidence limit differences	Several		No impact; explained by inclusion of stratum and subpop/domain
	Syphilis prevalence unadjusted OR	1.19 (0.14-9.87)	1.20 (0.15-9.68)	No impact; explained by inclusion of stratum and subpop/domain
	Syphilis prevalence adjusted OR	0.89 (0.12-6.72)	0.92 (0.12-6.85)	No impact; explained by inclusion of stratum and subpop/domain
Table 3				
	Confidence limit differences	Several		No impact; explained by inclusion of stratum and subpop/domain
	Had unprotected sexual intercourse	0.83 (0.51-1.35)	0.75 (0.46-1.22)	No impact; explained by inclusion of stratum and subpop/domain
Table 4				
	Confidence limit differences	Several		No impact; explained by inclusion of stratum and subpop/domain

Appendix B: Push button replication, final report – *The Lancet*: Baird et al. 2012

Section 1: Basic information

- Original paper citation:
  Baird, SJ, Garfein, RS, McIntosh, CT and Özler, B, 2012. Effect of a cash transfer programme for schooling on prevalence of HIV and herpes simplex type 2 in Malawi: a cluster randomised trial. *The Lancet,* 379(9823), pp.1320–1329.
- Original authors and contact email addresses:
  Sarah J Baird, George Washington University, sbaird@gwu.edu
  Richard S Garfein, University of California, San Diego, rgarfein@ucsd.edu
  Craig T McIntosh, University of California, San Diego, ctmcintosh@ucsd.edu
  Berk Özler, World Bank, bozler@worldbank.org
- PBR researchers: Lynette M Smith and Nicholas Hein
- List of materials obtained: readme document, a .do files, two data sets, and log file
- Classification: *Comparable and incomplete replication*

Section 2: Replication process

Our goal was to perform a push button replication (PBR) of Baird et al. (2012a) using the supplied Stata .do file, data sets and user guides that were downloaded from the World Bank website: https://sites.google.com/site/decrgberkozler/papers-by-topic, http://microdata.worldbank.org/index.php/catalog/1005, http://microdata.worldbank.org/index.php/catalog/2338. The datasets included baseline, follow-up, deidentified merged outcome data.

We begin our PBR by using the supplied Stata, Lancet_HIV-HSV2_analysis_public.do file in Stata 14. In the .do file, we renamed appropriate directories and ran the .do file. The original authors had commented the code to indicate which tables in the manuscript it corresponded. Upon running the file, we encountered two errors; the delimiter was missing from the end of two procedures. After including the delimiter, the .do file executed with no further glitches. We compared the coefficients and significance level of each table and reported the results.

Section 3: PBR classification justification

Comparable and incomplete replication. The vast majority of the results were comparable. Only in a few cases did we have to write some code to get the comparable results. The Stata .do file did not contain code to replicate the number of enumeration areas sampled or to calculate the standard deviation for the continuous variables as reported in Table 2 of the published results. Tables 3 and 4 had complete code and the results were comparable. All results from the PBR matched the original results exactly, except for a few rounding errors, what appear to be two typos, minor differences in four-point estimates and CIs, and five differences in sample size. The rounding errors resulted in differences less than a hundredth of a unit for ORs and CIs and a percent for proportions. We do not consider the rounding errors differences; therefore, these inconsistencies are not highlighted in Appendix C. The differences in the PBR compared to the published results did not change the significance of any of the results and therefore did not alter the interpretation of the main finding – i.e. the cash transfer program was effective in reducing the prevalence of HIV and HSV-2 in the baseline schoolgirls cohort.

Appendix C: Push button replication comparison tables and description

A) Ineligible tables:

Tables from the paper that are not subject to replication because they are not data driven:

- Table 1: Eligibility and follow-up criteria
- Figure 1: Intervention groups
- Figure 2: Trial profile

B) Description of PBR table comparisons:

Table 2
Comparable and incomplete replication

- The code to identify the *number of enumeration areas* sampled for biological data collection was not included in the Stata .do file.
- In the baseline schoolgirls cohort, the *number of individuals* in the pooled intervention group was found to be 503 compared to 501 reported. This also affected the baseline schoolgirls UCT program total number of individuals, which we found to be 267 compared to 265 reported.
- In the baseline dropouts cohort, the *number of individuals* in the pooled intervention group was found to be 227 compared to 226 reported in the original results.
- The code for the standard deviation of continuous variables was not included in the Stata .do file. *Age*, *age at sexual debut* and *highest grade attended* were the continuous variables.
- For the pooled intervention group of baseline schoolgirls, the original paper reported that 423 baseline schoolgirls' *mother was still alive*, while the PBR calculated that number to be 420.

Table 3
Comparable replication

- Inconsistencies in sample sizes for baseline schoolgirls in the follow-up period.
  - *Had sexual intercourse once per week*: PBR 499 compared to a sample size of 299 in the original results, in the intervention group.
  - *Had sexual partner aged ≥25 years*: PBR sample size is 500 compared to 502 in the original results, in the intervention group.
  - *Had sexual partner aged ≥25 years* is 826 as opposed to 827 in the PBR and original results, respectively, in the control group.
- Point estimates and CIs that are discrepant for baseline schoolgirls.
  - *Syphilis prevalence* Unadjusted: PBR 1.19 (0.15 – 9.62) versus original 1.20 (0.15 – 9.68)
  - *Syphilis prevalence* Adjusted: PBR 0.89 (0.12 – 6.56) versus original 0.92 (0.12 – 6.85)
- Point estimates and CIs that are discrepant for baseline dropouts.

- *Currently pregnant* Adjusted: PBR 0.56 (0.27 – 1.15) versus original 0.55 (0.27 – 1.13).
- *Had unprotected sexual intercourse* Unadjusted: PBR 0.83 (0.15 – 1.35) compared to original 0.75 (0.46 – 1.22).

Table 4
Comparable replication

Exactly same coefficients and p-values.

C) PBR tables

	Comparable
	Minor differences
	Major differences
	No access to data
	Information not reported in table

Table A2: Baseline characteristics of participants

	Control group	Intervention group		
		Pooled	Conditional cash transfer program	Unconditional cash transfer program
Enumeration areas sampled for biological data collection	52	52	25	27
Baseline schoolgirls				
Number of individuals	827	503	236	267
Ever had sexual intercourse	182 (19%)	130 (22%)	70 (22%)	60 (22%)
Ever pregnant	21 (3%)	16 (3%)	9 (3%)	7 (3%)
Age (years)	15.3 (1.9)	15.1 (1.9)	14.9 (1.8)	15.4 (1.9)
Age at sexual debut (years)	15.7 (1.7)	15.8 (1.8)	15.7 (2.0)	15.9 (1.7)
Highest grade attended	7.6 (1.6)	7.4 (1.7)	7.1 (1.7)	7.9 (1.6)
Mother alive	707 (85%)	420 (84%)	198 (85%)	222 (83%)
Father alive	601 (74%)	367 (75%)	176 (74%)	191 (76%)
Female-headed household	275 (32%)	141 (25%)	63 (26%)	78 (24%)
Household owns a radio	479 (59%)	309 (58%)	143 (53%)	166 (65%)
Household owns a television	130 (24%)	110 (30%)	40 (27%)	70 (34%)
Household has access to a mobile telephone	464 (61%)	303 (60%)	145 (60%)	158 (61%)
Electricity available in dwelling	86 (20%)	80 (26%)	31 (28%)	49 (24%)
Piped water available in dwelling	277 (47%)	183 (49%)	48 (41%)	135 (60%)
Baseline dropouts				
Number of individuals	223	227		
Ever had sexual intercourse	151 (68%)	154 (68%)		
Ever pregnant	98 (44%)	90 (40%)		
Age (years)	17.6 (2.2)	16.8 (2.4)		
Age at sexual debut (years)	16.4 (1.8)	15.9 (2.2)		
Highest grade attended	6.2 (2.9)	5.8 (2.9)		
Mother alive	175 (78%)	180 (80%)		
Father alive	146 (66%)	144 (65%)		
Female-headed household	93 (42%)	90 (39%)		
Household owns a radio	118 (53%)	107 (47%)		
Household owns a television	16 (7%)	24 (11%)		
Household has access to a mobile telephone	103 (46%)	110 (49%)		
Electricity available in dwelling	16 (7%)	24 (11%)		
Piped water available in dwelling	64 (29%)	63 (25%)		

Note: Data are n (weighted %) or mean (SD). Sampling weights were used to account for variation in the probability of inclusion in the study according to age and stratum. All baseline dropouts received the conditional cash transfer intervention, irrespective of the treatment status of baseline schoolgirls in their enumeration areas; thus, the conditional and unconditional cash transfer columns are not applicable to this cohort.

Table A3: Effects of cash transfer intervention on outcome measures

	Baseline		Follow-up					
	Intervention group	Control group	Intervention group	Control group	Unadjusted odds ratio (95% CI)	P-value	Adjusted odds ratio (95% CI)	P-value
Baseline school girls								
Enrolled during 2008 school year			419/484 (90%)	669/801 (84%)	1.67 (1.09-2.56)	0.018	1.62 (1.07-2.45)	0.022
Ever married	0/501 (0%)	0/827 (0%)	19/501 (3%)	45/827 (4%)	0.63 (0.31-1.28)	0.20	0.68 (0.37-1.28)	0.24
Currently pregnant	3/501 (<1%)	2/827 (<1%)	15/501 (2%)	35/827 (4%)	0.66 (0.33-1.30)	0.23	0.71 (0.36-1.41)	0.33
Sexual debut*			39/371 (8%)	100/645 (13%)	0.59 (0.33-1.05)	0.071	0.64 (0.38-1.07)	0.09
Had unprotected sexual intercourse	91/500 (15%)	107/825 (11%)	49/500 (8%)	63/826 (7%)	1.28 (0.78-2.09)	0.34	1.08 (0.67-1.75)	0.76
Had sexual intercourse once per week	16/500 (3%)	22/825 (2%)	22/499 (3%)	62/826 (7%)	0.44 (0.23-0.85)	0.014	0.46 (0.26-0.82)	0.008
Had a sexual partner aged ≥ 25 years†			4/500 (<1%)	20/826 (2%)	0.20 (0.07-0.59)	0.004	0.21 (0.07-0.62)	0.005
Had an HIV test	121/501 (22%)	174/827 (19%)	307/501 (54%)	470/826 (52%)	1.12 (0.74-1.68)	0.60	1.18 (0.83-1.69)	0.36
Knows that a healthy looking person can have HIV	443/501 (88%)	752/827 (90%)	454/501 (91%)	768/826 (92%)	0.94 (0.58-1.53)	0.80	1.00 (0.61-1.62)	0.99
Knows that HIV can be transmitted through breastfeeding	466/500 (93%)	785/827 (95%)	481/501 (97%)	786/827 (96%)	1.70 (0.84-3.43)	0.14	1.72 (0.89-3.34)	0.11
Received health training about HIV/AIDS†			398/501 (78%)	657/827 (80%)	0.89 (0.61-1.30)	0.55	0.90 (0.63-1.30)	0.59
HIV prevalence†,‡			7/490 (1%)	17/799 (3%)	0.39 (0.15-1.02)	0.056	0.36 (0.14-0.91)	0.031
HSV-2 prevalence†,‡			5/488 (<1%)	27/796 (3%)	0.23 (0.08-0.66)	0.006	0.24 (0.09-0.65)	0.005
Syphilis prevalence†,‡			1/491 (<1%)	4/800 (<1%)	1.19 (0.15-9.62)	0.87	0.89 (0.12-6.56)	0.91
Baseline dropouts								

	Baseline		Follow-up					
	Intervention group	Control group	Intervention group	Control group	Unadjusted odds ratio (95% CI)	P-value	Adjusted odds ratio (95% CI)	P-value
Enrolled during 2008 school year			124/219 (57%)	27/220 (12%)	9.31 (5.31-16.3)	<0.001	8.77 (5.07-15.1)	<0.001
Ever married	0/227 (0%)	0/223 (0%)	37/226 (17%)	64/223 (29%)	0.50 (0.29-0.83)	0.008	0.48 (0.29-0.80)	0.005
Currently pregnant	11/226 (5%)	10/223 (4%)	16/226 (7%)	26/223 (12%)	0.59 (0.28-1.25)	0.17	0.56 (0.27-1.15)	0.12
Sexual debut*			18/72 (26%)	27/72 (38%)	0.57 (0.28-1.16)	0.12	0.70 (0.33-1.45)	0.34
Had unprotected sexual intercourse	133/222 (61%)	128/223 (57%)	59/225 (25%)	64/222 (29%)	0.83 (0.51-1.35)	0.45	0.74 (0.44-1.23)	0.24
Had sexual intercourse once per week	31/222 (14%)	28/223 (13%)	43/225 (19%)	66/223 (30%)	0.56 (0.34-0.91)	0.018	0.53 (0.32-0.86)	0.011
Had a sexual partner aged ≥ 25 years†			20/225 (8%)	23/223 (10%)	0.73 (0.40-1.34)	0.31	0.79 (0.42-1.50)	0.47
Had an HIV test	98/225 (43%)	104/223 (47%)	163/225 (72%)	169/223 (76%)	0.84 (0.51-1.38)	0.49	1.00 (0.64-1.57)	0.99
Knows that a healthy looking person can have HIV	198/225 (88%)	201/223 (90%)	204/226 (90%)	212/223 (95%)	0.45 (0.19-1.05)	0.064	0.51 (0.23-1.15)	0.11
Knows that HIV can be transmitted through breastfeeding	198/223 (89%)	210/223 (94%)	213/226 (94%)	214/223 (96%)	0.63 (0.25-1.56)	0.32	0.69 (0.26-1.81)	0.46
Received health training about HIV/AIDS†			130/226 (57%)	94/223 (42%)	1.82 (1.23-2.69)	0.003	1.91 (1.29-2.83)	0.001
HIV prevalence†,‡			23/210 (10%)	17/207 (8%)	1.30 (0.69-2.48)	0.42	1.37 (0.72-2.61)	0.33
HSV-2 prevalence†,‡			17/211 (8%)	17/208 (8%)	0.99 (0.46-2.10)	0.97	1.03 (0.47-2.24)	0.94
Syphilis prevalence†,‡			3/211 (2%)	2/208 (1%)	1.59 (0.27-9.50)	0.61	1.63 (0.27-9.95)	0.59

Notes: Data are n/N (weighted %) unless otherwise stated. Sampling weights were used to account for variation in the probability of inclusion in the study according to age and stratum. Adjusted ORs were calculated with a logistic regression model of individual data with independent variables that include treatment status, age group, indicators for near rural and far rural strata and baseline measure, unless otherwise stated. HSV=herpes simplex virus. *Cumulative risk measure, so no adjustment made for baseline status. †No adjustment for baseline measure because data not collected at baseline. ‡Measured at 18 months, all others measured at 12 months.

Table A4: Effects of conditional or unconditional cash transfers on baseline schoolgirls by outcome measures

	Control group	CCT group	UCT group	CCT vs control (adjusted odds ratio [95% CI])	P Value	UCT vs control (adjusted odds ratio [95% CI])	P Value	Heterogeneity of odds ratio* (p value)
Enrolled during the 2008 school year	669/801 (84%)	207/229 (92%)	212/255 (87%)	2.08 (1.14-3.82)	0.018	1.22 (0.77-1.96)	0.40	0.14
Ever married	45/827 (4%)	14/236 (4%)	5/265 (2%)	0.93 (0.47-1.86)	0.85	0.36 (0.12-1.07)	0.065	0.11
Currently pregnant	35/827 (4%)	13/236 (4%)	2/265 (1%)	1.17 (0.56-2.43)	0.67	0.16 (0.04-0.68)	0.012	0.0121
Sexual debut†	100/645 (13%)	18/166 (7%)	21/205 (10%)	0.58 (0.29-1.15)	0.12	0.72 (0.37-1.40)	0.34	0.62
Unprotected sexual intercourse	63/826 (7%)	30/235 (9%)	19/265 (8%)	1.17 (0.67-2.05)	0.58	0.96 (0.50-1.83)	0.90	0.59
Had sexual intercourse once per week	62/826 (7%)	14/235 (3%)	8/264 (3%)	0.53 (0.26-1.07)	0.075	0.37 (0.16-0.85)	0.019	0.49
Had a sexual partner aged ≥ 25 years‡	20/826 (2%)	1/235 (<1%)	3/235 (1%)	0.08 (0.01-0.60)	0.014	0.36 (0.11-1.19)	0.094	0.19
HIV prevalence‡	17/799 (3%)	3/235 (1%)	4/255 (2%)	0.29 (0.09-0.98)	0.047	0.47 (0.14-1.59)	0.22	0.57
HSV-2 prevalence‡	27/796 (3%)	4/233 (1%)	1/255 (<1%)	0.37 (0.13-1.03)	0.057	0.08 (0.01-0.58)	0.013	0.16
Syphilis prevalence‡	4/800 (<1%)	1/235 (<1%)	0/256 (0%)	1.37 (0.20-9.41)	0.75			

Notes: Data are n/N (weighted %) unless otherwise stated. Adjusted ORs were calculated with a logistic regression model of individual data with independent variables that include treatment status, age group, indicators for near rural and far rural strata and baseline measure, unless otherwise stated. Sampling weights were used to account for variation in the probability of inclusion in the study according to age and stratum. HSV=herpes simplex virus. CCT=conditional cash transfer. UCT=unconditional cash transfer. *Wald test of equality of adjusted ORs for the conditional and unconditional cash transfer interventions. †Cumulative risk measure, so no adjustment made for baseline status. ‡No adjustment for baseline measure, because data not collected at baseline.

Appendix D: List of files received from the authors

List of files received from the authors. These can be found on the World Bank website.
https://sites.google.com/site/decrgberkozler/papers-by-topic
http://microdata.worldbank.org/index.php/catalog/1005
http://microdata.worldbank.org/index.php/catalog/2338

Data (Round 1)

- MWI_2007_SIHR_v01_A_PUF_ASCII_SAS.zip
- MWI_2007_SIHR_v01_A_PUF_ASCII_SPSS.zip
- MWI_2007_SIHR_v01_A_PUF_SPSS.zip
- MWI_2007_SIHR_v01_A_PUF_STATA8.zip
- ddi-documentation-english-microdata-1005.pdf

Data (Round 2)

- MWI_2008_SIHRIE-R2_V01_M_ASCII_SPSS.zip
- MWI_2008_SIHRIE-R2_V01_M_SPSS.zip
- MWI_2008_SIHRIE-R2_V01_M_STATA8.zip
- ddi-documentation-english_microdata-2338.pdf

Replication files

- Lancet_HIV-HSV2_analysis_Public.do
- Lancet_HIV-HSV2_dataset_PUBLIC_DEIDENTIFIED.dta
- HIV-HSV2.dta
- Lancet_HIV-HSV2_Public_Dataset_and_Analysis_Notes_151223.docx

References

Abdool Karim, Q and Dellar, R, 2014. Inclusion of adolescent girls in HIV prevention research – an imperative for an AIDS-free generation. *Journal of the International AIDS Society,* 17(1).

Anderson, MJ and Legendre, P, 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation*, 62(3), pp.271−303.

Baird, SJ, Garfein, RS, McIntosh, CT and Özler, B, 2012a. Effect of a cash transfer programme for schooling on prevalence of HIV and herpes simplex type 2 in Malawi: a cluster randomised trial. *The Lancet,* 379(9823), pp.1320−1329.

Baird, SJ, Garfein, RS, McIntosh, CT and Özler, B, 2012b. Authors' reply. *The Lancet,* 380(9844), pp.802.

Baron, RM and Kenny, DA, 1986. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), pp.1173−1182.

Blankenship, KM, Bray, SJ and Merson, MH, 2000. Structural interventions in public health. *AIDS,* 14, pp.S11−S21.

Brown, AN, Cameron, D and Wood, B, 2014. Quality evidence for policymaking: I'll believe it when I see the replication. *Journal of Development Effectiveness*, 6(3), pp.215−235.

Campbell, MK, Elbourne, DR and Altman DG, 2006. CONSORT statement: extension to cluster randomized trials. *BMJ*, 328(7441), pp.702−708.

Cluver, L, Boyes, M, Orkin, M, Pantelic, M, Molwena, T and Sherr, L, 2013. Child-focused state cash transfers and adolescent risk of HIV infection in South Africa: a propensity-score-matched case-control study. *The Lancet Global Health,* 1(6), pp.e362−e370.

Dellar, RC, Dlamini, S and Karim, QA, 2015. Adolescent girls and young women: key populations for HIV epidemic control. *Journal of the International AIDS Society,* 18(2 Suppl 1), p.19408.

Duflo, E, Dupas, P, Kremer, M and Sinei, S, 2006. Education and HIV/AIDS prevention: evidence from a randomized evaluation in western Kenya. *World Bank Policy Research Working Paper*, 4024.

Dunbar, MS, Maternowska, MC, Kang, MJ, Laver, SM, Mudekunye-Mahaka, I and Padian, NS, 2010. Findings from SHAZ!: a feasibility study of a microcredit and life-skills HIV prevention intervention to reduce risk among adolescent female orphans in Zimbabwe. *Journal of Prevention & Intervention in the Community,* 38(2), pp.147−161.

Erulkar, A, Bruce, J, Dondo, A, Sebstad, J and Matheka, J, 2006. Tap and Reposition Youth (TRY) providing social support savings and microcredit opportunities for young

women in areas with high HIV prevalence. *Seeds* [online]. Available at: <https://www.popcouncil.org/uploads/pdfs/seeds/SEEDS23.pdf> [Accessed 24 September 2017].

Filmer, D and Pritchett, LH, 2001. Estimating wealth effects without expenditure data—or tears: an application to educational enrollments in states of India. *Demography*, 38(1), pp.115−132.

Fisher, RA, 1935. *The design of experiments*. Edinburgh: Oliver and Boyd.

Green, DP and Vavreck, L, 2007. Analysis of cluster-randomized experiments: a comparison of alternative estimation approaches. *Political Analysis*, 16(2), pp.138−152.

Hallfors, DD, Cho, H, Rusakaniko, S, Mapfumo, J, Iritani, B, Zhang, L, Luseno, W and Miller, T, 2015. The impact of school subsidies on HIV-related outcomes among adolescent female orphans. *Journal of Adolescent Health,* 56(1), pp.79−84.

Howe, LD, Hargreaves, JR and Huttly SRA, 2008. Issues in the construction of wealth indices for the measurement of socio-economic position in low-income countries. *Emerging Themes in Epidemiology,* 5(1), p. 3 doi:10.1186/1742-7622-5-3.

Lu, M and Yang, W, 2012. Multivariate logistic regression analysis of complex survey data with application to BRFSS data. *Journal of Data Science*, 10(2), pp.157−173.

Peters, TJ, Richards, SH, Bankhead, CR, Ades, AE and Sterne, JAC, 2003. Comparison of methods for analysing cluster randomized trials: an example involving a factorial design. *International Journal of Epidemiology,* 32(5), pp.840−846.

Peterson, A, Kealey KA, Mann SL, Marek PM and Sarason IG, 2000a. Hutchinson smoking prevention project: long-term randomized trial in school-based tobacco use prevention-results on smoking. *Journal of the National Cancer Institute*, 92, pp.1979–1991.

Peterson A, Mann SL, Kealey KA and Marek PM, 2000b. Experimental design and methods for school-based randomized trials: experience from the Hutchinson smoking prevention project (HSPP). *Controlled Clinical Trials*, 21, pp.144-165.

Pettifor, A, MacPhail C, Hughes JP, Selin A, Wang J, Gómez-Olivé FX, Eshleman SH, Wagner RG, Mabuza W, Khoza N, Suchindran C, Mokoena I, Twine R, Andrew P, Townley E, Laeyendecker O, Agyei Y, Tollman S and Kahn K, 2016. The effect of a conditional cash transfer on HIV incidence in young women in rural South Africa (HPTN 068): a phase 3, randomized controlled trial. *The Lancet Global Health*, 4(12), pp.e978−e988.

Pettifor, A, MacPhail C, Nguyen N and Rosenberg M, 2012. Can money prevent the spread of HIV? A review of cash payments for HIV prevention. *AIDS and Behavior*, 16(7), pp.1729−1738.

Pfeffermann, D, 1993. The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique*, pp.317−337.

Pfeffermann, D, Skinner, CJ, Holmes, DJ, Goldstein, H and Rasbash, J, 1998. Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1), pp.23−40.

Pronyk, PM, Kim, JC, Abramsky, T, Phetla, G, Hargreaves, JR, Morison, LA, Watts, C, Busza, J and Porter, JD, 2008. A combined microfinance and training intervention can reduce HIV risk behaviour in young female participants. *AIDS (London, England),* 22(13), pp.1659−1665.

Rabe-Hesketh, S and Skrondal A, 2006. Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(4), pp.805−827.

SAS Institute Inc., 2015. *Base SAS® 9.4 Procedures Guide, Fifth Edition*. Cary, North Carolina, USA: SAS Institute Inc.

Shetty, AK, 2005. Perinatally Acquired HIV-1 Infection: Prevention and Evaluation of HIV-Exposed Infants. *Seminars in Pediatric Infectious Diseases*, 16(4), pp.282–295.

Smith, LM, 2016. Effect of a Cash Transfer Programme for Schooling on Prevalence of HIV and Herpes Simplex Type 2 in Malawi, A Replication Study of Baird (2012) [online]. Available at: <http://3ieimpact.org/media/filer_public/2016/04/09/smith-revised-replication-plan.pdf>

StataCorp. 2017. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC

Stedman, MR, Gagnon, DR, Lew, RA, Solomon, DH, Losina, E and Brookhart, MA, 2009. A SAS macro for a clustered permutation test. *Computer methods and programs in biomedicine*, 95(1), pp.89−94.

Sumartojo, E, Doll, L, Holtgrave, D, Gayle, H and Merson, M, 2000. Enriching the mix: incorporating structural factors into HIV prevention. *AIDS,* 14, pp.S1−S2.

Thabane, L, Mbuagbaw, L, Zhang, S, Samaan, Z, Marcucci, M, Ye, C, Thabane, M, Giagregorio, L, Dennis, B, Kosa, D and Debono, VB, 2013. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. *BMC medical research methodology*, 13(1), p.92.

UNC Carolina Population Center. *Analyzing data from sample surveys*. Available at: <http://www.cpc.unc.edu/research/tools/data_analysis/statatutorial/sample_surveys> [Accessed 17 June 2016].

Wamani, H, Tylleskär, T, Åstrøm, AN, Tumwine, JK and Peterson, S, 2004. Mothers' education but not fathers' education, household assets or land ownership is the best predictor of child health inequalities in rural Uganda. *International Journal for Equity in Health*, 3(1), pp.9.

Webb, EL, Hayes, RJ and Glynn, JR, 2012. Cash transfer scheme for reducing HIV and herpes simplex type 2. *The Lancet,* 380(9844), pp.802.

World Bank, 2012. Malawi - Schooling, Income, and Health Risk Impact Evaluation Household Survey 2007-2008, Round I (Baseline), [online]. Available at: <http://microdata.worldbank.org/index.php/catalog/1005/related_materials> [Accessed 18 January 2016].

World Bank, 2015. Malawi - Schooling, Income, and Health Risk Impact Evaluation Household Survey 2008-2009, Round 2 (Midline), [online]. Available at: <http://microdata.worldbank.org/index.php/catalog/2338> [Accessed 18 January 2016].

Wu, S, Crespi, CM and Wong, WK, 2012. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. *Contemporary Clinical Trials*, 33(5), pp.869−880.

Other publications in the 3ie Replication Paper Series

The following papers are available from http://www.3ieimpact.org/en/publications/3ie-replication-paper-series/

Power to the people?: a replication study of a community-based monitoring programme in Uganda, 3ie Replication Paper 11. Donato, K and Garcia Mosqueira, A (2016)

Fighting corruption does improve schooling: a replication study of a newspaper campaign in Uganda, 3ie Replication Paper 10. Kuecken, M, and Valfort, MA (2016)

The effects of land titling on the urban poor: a replication of property rights, 3ie Replication Paper 9. Cameron, Drew B, Whitney, Edward M and Winters, Paul C (2015)

Male circumcision and HIV acquisition reinvestigating the evidence from young men in Kisumu, Kenya, 3ie Replication Paper 8. Djimeu, EW, Korte, JE and Calvo, FA (2015)

Walking on solid ground: a replication study on Piso Firme's impact, 3ie Replication Paper 7. Basurto, MP, Burga, R, Toro, JLF and Huaroto, C (2015)

The impact of India's JSY conditional cash transfer programme: A replication study, 3ie Replication Paper 6. Carvalho, N and Rokicki, S (2015)

Recalling extra data: A replication study of finding missing markets, 3ie Replication Paper 5. Wood, BDK and Dong, M (2015)

The long and short of returns to public investments in fifteen Ethiopian villages, 3ie Replication Paper 4. Bowser, WH (2015)

Reanalysis of health and educational impacts of a school-based deworming program in western Kenya Part 2: Alternative analyses, 3ie Replication Paper 3, part 2. Aiken, AM, Davey, C, Hayes, RJ and Hargreaves, JR (2014)

Reanalysis of health and educational impacts of a school-based deworming program in western Kenya Part 1: A pure replication, 3ie Replication Paper 3, part 1. Aiken, AM, Davey, C, Hargreaves, JR and Hayes, RJ (2014)

TV, female empowerment and demographic change in rural India, 3ie Replication Paper 2. Iversen, V and Palmer-Jones, R (2014)

Quality evidence for policymaking: I'll believe it when I see the replication, 3ie Replication Paper 1. Brown, AN, Cameron, DB and Wood, BDK (2014)

Replication Paper Series

International Initiative for Impact Evaluation
1029 Vermont Avenue, NW
Suite 1000
Washington, DC 20005
USA

replication@3ieimpact.org
Tel: +1 202 629 3939

www.3ieimpact.org