Evaluation of Wyman's Teen Outreach Program® in Florida

Final Impact Report for

Florida Department of Health
August 10, 2015

Prepared by

Ellen Daley
University of South Florida

Eric Buhi
San Diego State University

Wei Wang	Helen Mahony
Ashley Singleton	Elizabeth Powers
Rita Debate	Shireen Noble
Stephanie Marhefka	Saba Rahman
Kay Perrin	Markku Malmi
Charlotte Noble	Kristin Hall
Sarah Maness	Robert Ziemba

Citation:

Acknowledgements:

Contents

EVALUATION OF WYMAN'S TEEN OUTREACH PROGRAM ® (TOP) IN FLORIDA: FINDINGS FROM THE REPLICATION OF AN EVIDENCE-BASED TEEN PREGNANCY PREVENTION PROGRAM

I.   Introduction

A.   Introduction and study overview

Florida, the 3rd most populous state in the United States,[1] ranks poorly on most adolescent education and health indicators. Compared with other states, Florida ranked 25th (50th being the worst) in adolescents (16 to 19 years) not going to school and not working ("drop-outs")[2] and 45th in public high school graduation rates[3] in 2009. In 2010, Florida ranked 31st for adolescent (15 to 19 years) live birth rates,[2] 30th for rates of infection with Chlamydia, 34th for rates of infection with Gonorrhea, and 35th for rates of infection with Syphilis.[4] Additionally, Florida had the 3rd highest adolescent (15 to 19 years) HIV diagnosis rate in the nation.[5] One contributor to these poor outcomes is the extent to which Florida's substantial number of non-metropolitan geographic areas are medically underserved.[6, 7] In addition, from 2009 to 2013, 22.3% of the rural population[1] had not completed high school as compared to 13.6% of those living in urban areas.[8] Further, in 2013, 24.3% of the population in rural areas lived in poverty, compared to 16.8% of the population living in urban areas.[8]

The U.S. Department of Health and Human Services (HHS) responded to the national issue of adolescent sexual risk behaviors by organizing a rigorous review of evidence-based programs that have documented impacts for reducing risk behaviors related to teen pregnancy and sexually transmitted infections (STIs). HHS contracted Mathematica Policy Research and its partner, Child Trends, to conduct this assessment and compile a thorough list of programs with strong evidence supporting their effectiveness.[9] As a result, HHS and the Office of Adolescent Health (OAH) released several grant announcements supporting further implementation and evaluation of these evidence-based programs.

The Florida Department of Health (FDOH) was awarded a Tier 1 grant to replicate (with minimal adaptations) and evaluate Wyman's Teen Outreach Program® (TOP), an HHS evidence-based program.[10]

---

[1] The definition of 'rural' is based on Florida's statutory definition: "an area with a population density of less than 100 individuals per square mile or an area defined by the most recent United States Census as rural."

FDOH contracted an independent evaluation team from the University of South Florida (USF) College of Public Health to examine the TOP model and determine the scope of its impacts in a school-based setting. This evaluation study was a pair-matched cluster randomized controlled trial, meaning that naturally occurring groups or "clusters" of individuals (in this case, schools) were paired and randomly assigned to an intervention (treatment) or comparison (control) group. In addition, this study used a longitudinal design, meaning data were gathered for the same participants repeatedly over time (in this case, at 3 time points).

Figure I.1: TOP Evaluation Map

Counties	# Schools	Total # Classes	Treatment Classes	Control Classes
Hardee	1	11	-	11
DeSoto	1	10	10	-
Union	1	5	-	5
Gadsden	1	2	2	-
Gulf	2	7	4	3
Holmes	4	11	8	3
Manatee	6	40	17	23
Pasco	2	13	5	8
Santa Rosa	6	45	20	25
Suwannee	2	10	4	6
Totals	26	154	70	84
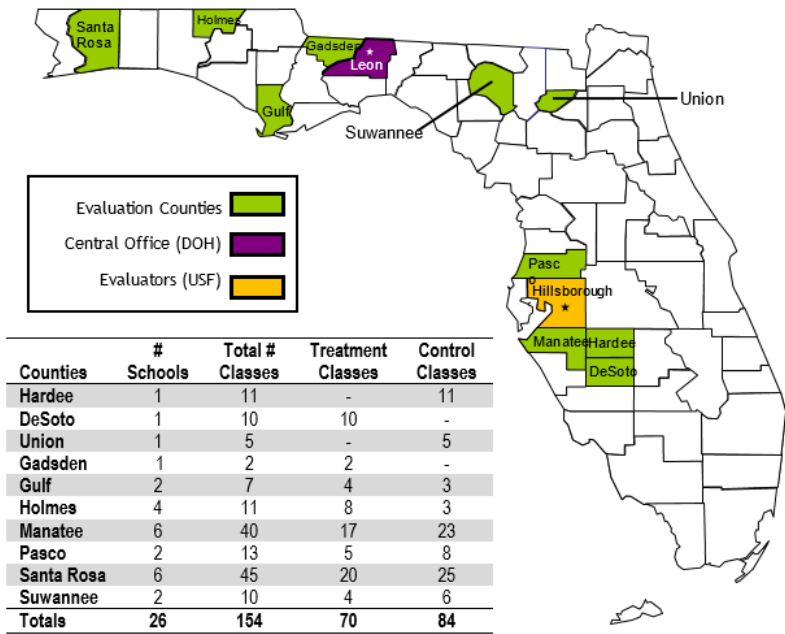
Designed to promote healthy choices, reduce teen pregnancy and increase school success, TOP models an asset-based approach by helping youth develop positive skills, attitudes, and knowledge related to school, problem solving, community engagement, goal setting, and social relationships.[11]

Since TOP's inception in 1978,[12] researchers have examined how, why, and with whom TOP makes the greatest impact. Multiple studies have documented that TOP reduces teen pregnancy,[12, 13, 14] academic suspension,[12-14, 15] and course failure[12, 14, 15] for youth participants. However, after a thorough assessment of all prior study findings, Mathematica and HHS determined that only one study presented strong evidence supporting the effectiveness of TOP.[16] That particular study found that teen pregnancy rates were reduced specifically for adolescent females who participated in the program; though that same impact (causing pregnancy) was not observed for adolescent males.

One distinct different between that study and this one is that TOP was implemented in a traditional school-based setting in the state of Florida. In each Florida county, a school board, headed by a

superintendent, governs the local public school system. FDOH selected 26 non-metropolitan counties[2] to participate in this replication study. The selected counties were assessed for appropriateness and need, based on 6 key indicators: 1) birth rate per female population aged 15-19 years, 2) repeat birth rate per female population aged 15-19 years, 3) combined Chlamydia and Gonorrhea rates per female population aged 15-19 years, 4) high school dropout rates, 5) graduation rates, and 6) out-of-school suspension rates. If a county had poor outcomes for at least 1 indicator, then that county was eligible to participate in the study. Selection of counties also considered their local health department and school district's capacity to implement the program. With funding from the Office of Adolescent Health (OAH) Teen Pregnancy Prevention (TPP) program, the USF evaluation team conducted a rigorous evaluation of TOP in Florida. This report will contribute to evidence on the program's effectiveness by outlining implementation as well as impact findings (Section IV) and reviewing their implications (Section V).

B. Primary and secondary research questions

Research questions focus on outcomes observed after the course of the intervention, also referred to as the end of the program. The primary research questions consider 2 key outcomes—ever having sexual intercourse, and ever having been pregnant or gotten someone pregnant—at approximately 10 months after the end of the program. The secondary research questions consider the same 2 outcomes immediately after the end of the program, which is approximately 8 months after the program began.

Figure I.1: Primary and Secondary Research Questions
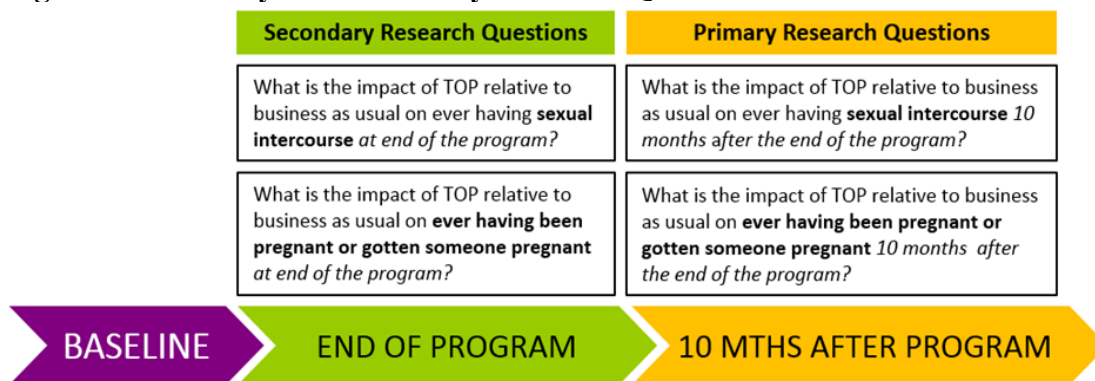
II. Program and comparison programming

A. Description of program as intended

TOP is a positive youth development program that uses weekly educational peer group sessions, Community Service Learning (CSL), and positive adult guidance to help youth in grades 6-12 build healthy behaviors, life skills, and a sense of purpose (see Appendix A).[17] The program curriculum consists of 4 levels tailored for age appropriateness for youth ages 12-17.[18] The TOP Changing Scenes Curriculum (CSC) incorporates topics such as goal setting, communication/assertiveness, sexuality, and human development. The curriculum also features a CSL Guide that provides structured exercises to identify community needs, brainstorm service project ideas, identify and develop group skills, and celebrate community service accomplishments. Exercises from the guide are used for community service preparation, and the exercises can be tailored to specific service projects. Flexible in nature, TOP can be implemented in regular school settings, after-school programs, or within community organizations.[18]

As intended, TOP should be implemented over 9 consecutive months with a minimum of 25 weekly sessions. Sessions can comprise CSC lessons, CSL exercises, CSL hours (meaning hours spent planning the project, the act of service, and reflecting on the project) and/or general meetings which may include guest speakers or group discussions. Sessions do not have a prescribed length requirement but the suggested length for most lessons and exercises is 40-50 minutes. The majority of weekly sessions (80% or more) should be CSC lessons, CSL exercises, or CSL hours; general meetings should not exceed 20% of the total program sessions offered.[19] A minimum of 20 CSL hours should be offered; the majority of these hours should be dedicated to the act of service. CSL hours can take place during a regularly scheduled weekly session or during separate supplemental time that would also contribute to the overall session count[3]. One or 2 certified adult program facilitators should lead a group of 10 to 25 youth and develop session plans based on the group's needs and interests.

---

[3] For example, suppose a class normally meets for TOP on Thursday from 1-2pm. If on a particular Thursday this class made cards for sick children from 1-1:30pm and listened to a guest speaker from 1:30-2pm that day would count as 1 program session and 1/2 hour of CSL. If the class met on Saturday from 3-4pm to deliver the cards to a local hospital then that day would count as 1 program session and 1 hour of CSL.

For this study, TOP was implemented in traditional public high schools in Florida and delivered by local health department staff, who were trained and certified as TOP facilitators. These facilitators delivered CSC lessons from Level 2, which is appropriate for 14-year-old youth (target population). To increase the likelihood of reaching the target population, TOP was delivered in classes required for graduation in which mostly 9th grade students enroll. These classes included Health Opportunities through Physical Education (HOPE) and the HOPE/PE variation (see Table II.1, next page).

Youth enrolled in these classes received the TOP intervention in addition to, not as a replacement for, all business-as-usual curriculum content. To accommodate this supplementation, classroom teachers often condensed their regular lesson plans into other class days. The TOP Changing Scenes Curriculum is well suited to implementation as supplemental education in health and physical education classes because it covers many of the competencies outlined for these classes in the Florida Sunshine State Standards. Program facilitators were not required to coordinate their lesson plans with the class curriculum. Instead, the program facilitators were encouraged to choose lessons and exercises based on their students' needs and interests. Youth in intervention schools may have also had access to reproductive health content through services or programs typically available at school, including other courses or guest speakers. The extent to which they received reproductive health content is described in Section IV.A and a description of the source for these data can be found in Section III.C2.

Schools implementing TOP do not require, but often promote, community service so that students will be eligible for Florida Bright Futures Scholarships (which has a 30-100 hour community service requirement).[20] TOP CSL activities meet the Florida Bright Futures requirements only if the activities occur off-campus. Therefore, the CSL component of TOP generally does not satisfy Florida Bright Futures community service requirements.

B. Description of counterfactual condition

Youth enrolled in the comparison, or counterfactual, setting for this evaluation also received business-as-usual curriculum content. The majority of youth in comparison schools were enrolled in HOPE or HOPE-PE classes (74%) or similar classes such as PE-Fitness (13%); the remaining youth were

enrolled in a Leadership Skills Development course (13%). Classroom teachers delivered the required content in all comparison classes (Table II.1), and youth received the standard number of hours of content depending on each school's semester schedule (ranging approximately 50-90 minutes per day). Similar to intervention schools, comparison schools also often promote, but do not require community service. In addition, youth in comparison schools may have access to reproductive health content through services or programs typically available at school, like other courses or guest speakers.

Table II.1. Florida Department of Education course descriptions

Course	Purpose	Content includes, but not limited to:
HOPE (#3026010) http://www.cpalms.org/Public/PreviewCourse/Preview/4051	Develop and enhance healthy behaviors that influence lifestyle choices and student health and fitness	Mental/social health; physical activity/fitness; nutrition/wellness; diseases/disorders; health advocacy; first aid/CPR; alcohol, tobacco, and drug prevention; human sexuality including abstinence, HIV education; Internet safety
HOPE/PE Variation (#1506320) http://www.cpalms.org/Public/PreviewCourse/Preview/4058	Develop and enhance healthy behaviors that influence lifestyle choices and student health and fitness. Students will combine the learning of principles and background information in a classroom setting with physical application of the knowledge. A majority of class time should be spent in physical activity.	Mental/social health; physical activity; components of physical fitness; nutrition and wellness planning; diseases and disorders; health advocacy
Leadership Skills Development (#2400300) http://www.cpalms.org/Public/PreviewCourse/Preview/4222	Teach leadership skills, parliamentary procedure, problem solving, decision making, communication skills, group dynamics, time and stress management, public speaking, human relations, public relations, team building, and other group processes	Self-understanding; goal setting, self-actualization, and assertiveness; organizational theories and management
Personal Fitness (#1501300) http://www.cpalms.org/Public/PreviewCourse/Preview/4082	Provide students with the knowledge, skills, and values they need to become healthy and physically active for a lifetime. This course addresses both the health and skill-related components of physical fitness which are critical for students' success.	Not listed

All of the classes in intervention schools and the majority of classes in comparison schools received the same or very similar course content, so the difference between the evaluation conditions is that youth enrolled in the intervention schools received the aforementioned TOP components (CSC curriculum, CSL hours, etc.) in addition to all business-as-usual curriculum content. Leadership Skills Development and PE-Fitness class types were represented in comparison schools, but were not represented in the intervention schools. The youth in these particular class types in comparison schools (26%) received

different course content than those in the intervention condition. Despite class type, the TOP CSL experience was not provided for students in the comparison sites.

III. Study design

A. Sample recruitment

Recruitment of schools for the study was completed in 2 phases. In phase 1, FDOH assessed the appropriateness and need of the study in all non-metropolitan counties in Florida using 6 indicators: 1) birth rate per female population aged 15-19 years, 2) repeat birth rate per female population aged 15-19 years, 3) combined Chlamydia and Gonorrhea rates per female population aged 15-19 years, 4) high school dropout rates, 5) graduation rates, and 6) out-of-school suspension rates. Counties with higher rates and/or rankings overall, for 1 or more indicator, were then selected for the second phase of assessment. In phase 2, FDOH conducted a community needs assessment by meeting with local health department and school district key informants. If the local health department and school district expressed an interest in the study, and both entities had the capacity to adhere to FDOH's Memorandum of Understanding terms, then that particular county was recruited for study participation. If the county was recruited for study participation, all public schools within the particular county, serving 9th – 12th grade students, were eligible to participate.

Using this recruitment process, FDOH identified and selected 26 counties throughout northwest, northeast, and central Florida for the study. Within the first grant year a total of 17 counties were lost due to various reasons: school board did not approve the study due to schedule constraints or concerns about youth survey questions (n=6), schools with semester-long classes did not fit the requirements of the TOP fidelity model (n=6); and counties were not allowed to participate in the study due to involvement in the OAH National Evaluation for the *Live the Life* program (n=5). Three counties were added, resulting in a total of 12 counties and 28 high schools for the final evaluation sample. One school was later lost due to its lack of continued interest in the program; that school (the only one in its county) and its matched pair (in a different county) were removed from the study. The final sample comprised 26 schools within 10 counties (see Figure I.1).

B. Study design

This evaluation is a school-based longitudinal cluster randomized controlled trial. In summer 2011, schools were matched into pairs based on 5 characteristics in order from most important to least important: county, courses offered, school size, region/proximity, and block or non-block schedule. Ideally schools would be matched on all criteria, but where perfect matches were not possible, schools were matched following the list as prioritized above. If possible, schools within counties were matched so that potential confounding variables between counties could be controlled.

Matched pairs were assigned to an evaluation conditions (intervention or comparison) in summer 2011. The USF evaluation team's biostatistician randomly assigned matched pairs of schools to conditions A or B. Independent of this process, another USF evaluation team member made the decision of whether condition A or B will receive TOP, and then this information was merged. Random assignment of the schools took place once (summer 2011), and the program was implemented a full year before enrolling youth in the evaluation in the fall of 2012. The year of implementation allowed the program facilitators and evaluation staff to implement and examine procedures, related to lesson plan development and survey administration, and make adjustments as needed. Schools were not re-randomized prior to drawing the evaluation sample, making it possible for parents and students to be aware of the treatment or control status of each school. However, the evaluation staff took precautions to ensure that parents and students were blind to each school's condition (more detail can be found in the Data Collection – Impact Evaluation section below).

Youth in both intervention and comparison schools completed a survey at 3 time points: baseline, which was administered pre-program delivery in fall 2012; first follow-up, administered immediately post-program delivery in spring 2013; and second follow-up, administered approximately 10 months post-program delivery in spring 2014. While baseline and first follow-up survey administration took place in participating classrooms (HOPE, HOPE/PE, etc.), second follow-up survey administration took place in cafeterias, libraries, and media centers. This was the most practical approach for administering surveys to participants who were then enrolled in a variety of classes (math, English, etc.). Because cafeterias,

libraries, and media centers were also used for mandated end-of-year standardized tests, second follow-up was scheduled 10 months rather than 12 months post-program to accommodate school testing schedules (details in Appendix B).

Prior to baseline survey administration, in large schools where the number of classes was in excess of what could be surveyed in the time allotted, random assignment occurred at the class level to determine which classes would be surveyed, using an algorithm in S-PLUS.[20] Classes were randomly selected out of all eligible classes to decrease data collection burden. Program facilitators did not begin program implementation until after baseline data collection procedures (described below) were completed.

C. Data collection

1. Impact evaluation

In both intervention and comparison schools, trained evaluation data collectors from the USF evaluation team, in cooperation with school personnel, collected data from participants. Youth were asked to complete paper-and-pencil surveys consisting of questions about attitudes, behaviors, and intentions, including sexual health behavior. All data collection procedures and instruments outlined below were reviewed and approved by the FDOH Institutional Review Board; Protocol H11180.

Prior to baseline data collection, in the first 3 weeks of the school year, the evaluation consent process was carried out and youth eligibility to participate was determined. Youth were deemed eligible if the following conditions were met: 1) enrolled in a course selected for evaluation, 2) had parental consent, 3) were proficient in English and, and 4) capable of independently taking a paper and pencil survey.

In all schools, the parental consent process began with sending a consent form to the mailing address of eligible students during the first week of school. The same consent form was given to youth and they were asked to deliver the form to their parents or guardians. In addition, the evaluation team requested that school administrators record a robo-call—a brief automated phone message, usually employed to make important school-related announcements—for all parents and guardians with phone numbers on file at the school. The robo-call served as an announcement of the replication study and to inform parents and

guardians about the consent process. Successful completion of the robo-calls could not be independently confirmed by the evaluation team.

This evaluation employed a passive (i.e., opt-out) consent process; that is, if parents did not want their child to participate in the evaluation, then they would sign and date the form to indicate refusal of permission. Parents were given 1 week to return the form, though they could withdraw their permission at any point in the study by contacting a school administrator or the evaluation staff directly. After this process was complete, student eligibility was determined and baseline data collection was initiated by evaluation data collectors. While it cannot be confirmed whether parents, guardians, or students were aware of the school's condition status prior to the consent process and baseline data collection, neither FDOH nor the evaluation team made any announcements or distributed any materials identifying the treatment or comparison status of each school.

Data collection consisted of 3 parts: a) assent, b) student information sheets, and c) youth surveys. During the assent process, the youth were given time to read the assent form and then decide whether or not to participate in the baseline survey. Signed assent forms were collected by the evaluation data collectors. All eligible students (assenting and non-assenting youth) were then asked to fill out a student information sheet with contact information—youth were advised this information would only be used for follow-up purposes if needed. All assenting youth then completed the survey after being given standardized survey instructions. To protect the confidentiality of responses, youth were required to sit apart from one another during survey administration. Youth were also instructed to separate their surveys from their student information sheets upon returning their materials. If a participant was absent during data collection, an evaluation data collector returned to the school up to 3 times, before program implementation began, to administer the survey on make-up days arranged with school personnel.

At first and second follow-up, the 3-part data collection procedure was repeated. During second follow-up additional procedures were implemented to contact participants that no longer attended the same school or who were absent for data collection and all make up days. During this out-of-school follow-up, trained members of the USF evaluation team reached out to students via the contact

information they provided on their student information sheet. These students were given the opportunity to complete an abridged version of the in-school survey over the phone, online through a confidential link, or by paper-and-pencil through mail. Because participants were asked to complete a survey out of school during their personal time, the USF evaluation team offered them a $10 incentive.

2. Implementation evaluation

Using several sources of data (see Appendix C), the implementation evaluation examined 4 domains: adherence, quality, counterfactual, and context. Adherence measures program fidelity or in other words the extent to which programmatic elements were delivered to youth as intended. Quality measures the quality of staff-participant interaction and youth engagement with the program. Counterfactual examines the *experiences* of the comparison condition, as opposed to what was *intended* for the comparison condition. Context considers other TPP programming offered to intervention and comparison youth, external events that affected implementation, and substantial unplanned adaptations, if any.

To assess adherence, the USF evaluation team collected Youth Attendance Records and Facilitator Logs that program staff completed for every regularly scheduled TOP session. Youth Attendance Records captured session dates and daily youth attendance. Facilitator Logs captured details about the TOP session, including 1) what lessons or CSL exercises were taught, 2) how much time spent was spent on Lessons, CSL or other activities, 3) the degree of fidelity to curriculum (all of it, most of it, some of it), and 4) open-ended responses to questions about adaptions, what went well, challenges, and out of the ordinary events affecting the session. The USF evaluation team also analyzed CSL Records, which facilitators used to document the details of various CSL activities throughout the year that result in a completed CSL project(s).

To assess quality, the USF evaluation team analyzed a section of questions from the first follow-up survey pertaining to implementation quality, including the Learning Climate Questionnaire.[21] To contextualize the survey findings regarding quality, the USF evaluation team typically conducted 2 Youth Program Quality Assessment (YPQA)[22] observations for each program facilitator in separate classes once

during the school year. In addition, FDOH provided information regarding the qualifications and training of program facilitators.

To assess both the counterfactual and context elements, program facilitators conducted structured interviews with teachers and school administrators. Additional information about the counterfactual condition was obtained from the Florida Department of Education website. The USF evaluation team also analyzed questions from the first follow-up survey about other TPP programming youth received outside of TOP or the equivalent business-as-usual course.

D. Outcomes for impact analyses

For the primary and secondary research questions, the outcomes of interest for this study were ever having sexual intercourse and ever having been pregnant or caused a pregnancy (Table III.1). The outcome of ever having sexual intercourse was defined by a single question: "Have you ever had sexual intercourse?" The outcome of ever been pregnant or caused a pregnancy was assessed by the previously mentioned questionnaire item (ever having sexual intercourse) as well as the item: "To the best of your knowledge, have you ever been pregnant or gotten someone pregnant, even if no child was born?". The pregnancy outcome relies on the response to the intercourse question because one should not be able to endorse the ever having been pregnant or gotten someone pregnant question if one has not ever had sexual intercourse. Response options were yes/no for each item. These measures were assessed at baseline, first follow-up, and second follow-up. Both the primary and secondary questions focus on the same measures, but at 10 months post-programming and immediate post-programming, respectively.

Table III.1. Behavioral outcomes used for primary research questions

Outcome name	Description of outcome	Timing of measure relative to program
Ever had sexual intercourse	The variable is a yes/no measure of whether a person has ever had sexual intercourse. The measure is taken directly from the following item on the survey: • "Have you ever had sexual intercourse?"	*10 months after the program has ended*
Ever been pregnant or ever gotten someone pregnant	The variable is a yes/no measure of whether a person has ever been or gotten someone pregnant. The measure is taken directly from the following item on the survey * : • "Have you ever had sexual intercourse?" • "To the best of your knowledge, have you ever been pregnant or gotten someone pregnant, even if no child was born?"	*10 months after the program has ended*

* Since those who answered no to "Have you ever had sexual intercourse?" question were supposed to skip the pregnancy question, their answers to the pregnancy question were imputed as 'no' if they did skip it.

E. Study sample

A sample flow table describing how the analytic samples were created for this study is included in Appendix D. Among the 26 schools that contributed at least one youth at baseline (13 intervention and 13 comparison schools), 4,327 youth were eligible to participate in the evaluation (1,950 in the intervention condition and 2,377 in the comparison condition). Passive parental consent was obtained after random assignment for a total of 4,063 youth throughout the study (1,845 intervention and 2,218 comparison). The total response rates for this study's primary and secondary research questions are listed in Table III.2.

Of respondents included in the impact analysis for the first primary research question (2,106), most were in the 9th grade at baseline 87% (1,788). Seven percent (144) of all respondents were in the 10th grade, 4% (188) in the 11th grade, and 1% (30) in the 12th grade. Among all respondents in the study sample, youths' average age at baseline was approximately 14 years old. The majority of respondents were 14 years old or younger (69%; 1,460), approximately a fifth were 15 years old (22%; 458), and 9% (188) were 16 years or older. Fifty-three percent identified as female (1,115), and the majority identified as Non-Hispanic White (63%; 1333), 7% (148) were Non-Hispanic Black. Non-Hispanic Other made up 9% (199), and the other 20% (426) were Hispanic. The demographics for the other primary outcome are similar. See Tables III.3, III.4 in Section III.F and additional tables in Appendix E for the demographic

breakdowns of the intervention and comparison conditions at baseline, first follow-up, and second follow-up.

Table III.2 Response Rates for Primary/Secondary Research Questions

Research questions	Total response rate	Total number of responses	Total number of youth
Primary Research Questions			
What is the impact of TOP relative to business as usual on ever having sexual intercourse 10 months after the end of the program?	48.7%	2,106	4,327
What is the impact of TOP relative to business as usual on ever having been pregnant or gotten someone pregnant 10 months after the end of the program?	47.6%	2,058	4,327
Secondary Research Questions			
What is the impact of TOP relative to business as usual on ever having sexual intercourse at end of the program?	56.3%	2,438	4,327
What is the impact of TOP relative to business as usual on ever having been pregnant or gotten someone pregnant at end of the program?	55.5%	2,401	4,327

F. Baseline equivalence

Baseline equivalence was first assessed between treatment and control groups for demographic and pre-intervention measure of the outcomes for the 4 analytic samples. We found no significant difference between the two conditions with respect to demographic and pre-intervention measures of pregnancy at either follow-up. However, youths in the comparison group were more likely to answer 'yes' to the ever had sex question than youths in the treatment group at both follow-up surveys (see Tables III.3 and III.4). Baseline measure of the outcomes were included in the impact analysis models to minimize the influence of these differences on the estimates of treatment effects. In addition, a sensitivity analysis that uses inverse probability of treatment weighting was conducted to further account for the baseline difference.

Equivalence was also tested for youths who, after completing the baseline survey, did or did not complete a follow-up survey. For both follow-up surveys, youths who were present for the surveys were significantly more likely to be younger, females and more likely to report being white or 'other' rather than black compared to youths who were not present. Youths who were present were also less likely to

answer 'yes' to both the ever had sex question and the ever been pregnant or gotten someone pregnant

question at baseline (see more details in Appendix E).

Table III.3 Summary statistics of key baseline measures for youth who completed baseline and responded to "ever had sex" question at second follow-up

Baseline measure	Intervention mean or % (standard deviation)	Comparison mean or % (standard deviation)	Intervention versus comparison % difference	Intervention versus comparison *p*-value of difference
Demographic Characteristics	-	-	-	-
Age (in years)	12.38 (0.78)	14.43 (0.76)	N/A	.85
≤ 14	71.78%	67.35%	4.43%	-
15	19.17%	23.82%	4.65%	-
≥ 16	9.05%	8.83%	0.22%	-
Gender (female)	52.29%	53.47%	1.18%	.69
Race/ethnicity	-	-	-	.40
White, Non-Hispanic	63.58%	63.07%	0.51%	-
Black, Non-Hispanic	7.35%	6.77%	0.58%	-
Hispanic/Latino	20.55%	29.97%	9.42%	-
Other, Non-Hispanic	8.52%	10.20%	1.68%	-
Outcomes	-	-	-	-
Ever had sex	15.02%	19.54%	4.52%	<.001
Total Sample Size	939	1,167	N/A	N/A

Table III.4. Summary statistics of key baseline measures for youth who completed baseline and responded to "ever been pregnant or gotten someone pregnant" question at second follow-up

Baseline measure	Intervention mean or % (standard deviation)	Comparison mean or % (standard deviation)	Intervention versus comparison % difference	Intervention versus comparison p-value of difference
Demographic Characteristics	-	-	-	-
Age (in years)	12.38 (0.78)	14.42 (0.74)	N/A	.65
≤ 14	72.13%	67.34%	4.79%	-
15	18.87%	24.21%	5.34%	-
≥ 16	9.00%	8.45%	0.55%	-
Gender (female)	52.71%	53.26%	0.55%	.92
Race/ethnicity	-	-	-	.34
White, Non-Hispanic	63.56%	63.03%	0.53%	-
Black, Non-Hispanic	7.27%	6.69%	0.58%	-
Hispanic/Latino	20.82%	20.16%	0.66%	-
Other, Non-Hispanic	8.35%	10.12%	1.77%	-
Outcomes	-	-	-	-
Ever been pregnant or gotten someone pregnant	0.65%	0.62%	0.03%	.45
Total Sample Size	922	1,136	N/A	N/A

G. Methods

1. Impact evaluation

The method used to assess the program was a linear probability model;[23] by calculation of a percentage point difference. This allowed the USF evaluation team to ascertain the impact of the intervention compared to the comparison group on ever having had sex and ever being pregnant for both the primary and secondary research questions. Participants were included in the impact analyses if outcome data were available for them at baseline, first follow-up, and second follow-up. The t-statistic, which is the ratio of the regression parameter to its standard error, was used to determine statistical significance. To account for the two contrasts presented by the primary research questions, a multiple

comparison adjustment was conducted using the Bonferroni method[24] with the family error rate set at .05. All data were analyzed using SAS software[25] version 9.4.

Due to the cluster randomized controlled trial design of this evaluation, students from within the same school were not assumed to be independent of one another. In order to account for this lack of independence, a multi-level model was estimated to adjust for clustering at the school level. Depending on the research question, the model contained the dependent variables of ever having sexual intercourse or ever having been pregnant. All models contained baseline measures of the following: participants' age, gender, and race/ethnicity; an indicator delineating a schools' matched pair; whether youth were in the intervention or comparison group; and the corresponding outcome at baseline. This was done in order to account for any variation that may have been produced due to these variables.

If demographic information for a participant was missing from baseline, then their demographic information was taken from a subsequent time point if available. If a participant was missing data on the outcomes of interest at any time point, then that participant was removed from the analysis, per guidance from OAH. Equations for estimating impacts, model specification, and missing data approaches for the impact analysis are described in detail in Appendix F. All data were cleaned prior to analysis; the procedures are discussed in detail in Appendix G. Three sensitivity analyses were conducted to explore alternative approaches to handling missing data and model specifications. Specifics of the sensitivity analyses methods and results are described in Appendix H.

2. Implementation evaluation

To assess the degree to which each element of the intervention was implemented as intended, descriptive statistics were calculated for all of the implementation elements listed in Section III.C.2 and Table C.1 in Appendix C. For a detailed description of how each implementation element was addressed, refer to Table I.1 in Appendix I. For many of the implementation elements, frequencies were analyzed, and proportions and averages are reported.

Five adherence measures for this study were based on TOP's own fidelity benchmarks[18] (see Section II.A); these measures are the proportion of classes that offered 1) at least 25 program sessions[4] 2) sessions at least once per week, 3) sessions over the span of 9 months, 4) CSC/CSL in at least 80% of sessions and 5) at least 20 hours of CSL[5]. Other adherence measures, such as the percentage of participants who received at least 75% of the intended 25 sessions, were drawn from OAH fidelity monitoring guidance.[26] The key limitation of the adherence measures is incomplete attendance data, facilitator logs, and CSL records (see Footnote D, Footnote E, and Appendix I – Adherence for details)

For both quality elements, averages and proportions are reported from youth-reported survey data and are supplemented by measures created from the YPQA observation tool (see Appendix I, Quality). The YPQA measures used are averages of the sub-scales that are most relevant to these implementation elements in a school setting: Warm Welcome, Encouragement, Skill-Building, Adult Partners, Engagement, and Active Engagement. The YPQA measures are limited because the 2 observations were conducted at 1 time point (i.e., during the same day). Furthermore, the measures do not represent the whole validated tool because some of the items are irrelevant to this implementation evaluation and are non-applicable to school settings. Despite these limitations, the analyses of the measures are useful in contextualizing the results from the survey data.

We examined counterfactual and context elements by reviewing the school staff interviews and facilitator logs for prominent and recurring themes. Despite the limitation of when the structured interviews were completed (2 years post-intervention), they provide important data to compare with the youth-reported survey data related to context. Because the counterfactual condition is business-as-usual, other data for the adherence and quality measures discussed above were not available for further analyzing the counterfactual condition.

---

[4] For this evaluation the count of program sessions did not include supplemental sessions (outside of normal class time) in which only CSL hours were offered. The individual dates of these sessions were not consistently reported; often only the cumulative number of hours accrued by the class was reported.

[5] The number of CSL hours offered per class and the proportion of classes with 20 or more hours was analyzed because individual student attendance at CSL sessions was not consistently reported.

IV. Study findings

A. Implementation study findings

Adherence: In the 13 schools receiving the intervention, TOP was implemented in 70 individual classes. TOP facilitators implemented sessions throughout the school year in each of the 70 classes, with 51 (73%) classes receiving at least 25 sessions as prescribed (range: 23-57 sessions). All classes received sessions at least weekly, as prescribed; 29 classes (41%) received more than 1 session per week at least once. Sessions ranged from 30-150 minutes; the average duration was 58 minutes. While sessions do not have a prescribed length, the suggested time for most lessons and exercises is 40-50 minutes. Due to the lack of student-level CSL data (see Appendix I), adherence to prescribed CSL hours had to be evaluated at the class level. Facilitators reported that a limited number of classes—8 (or 11.4%)—received at least 20 CSL project hours. Overall 29 classes (41.4%) offered at least 75% of the prescribed hours. On average, 14.9 CSL hours were offered per class. No classes received programming over the course of 9 months, as prescribed. Due to variations in school calendars and survey administration schedules, programming took place over a range of 203-266 days or approximately 6.7-8.7 months (median: 239 days or approximately 7.9 months).

Attendance data were collected for 1,630 intervention participants, of which 45% received 25 or more program sessions as prescribed. Overall 89% received at least 75% of the intended 25 program sessions. The average number of sessions received by youth was 24.1 sessions (range: 1-46 sessions, median: 24 sessions; see Appendix J).

All intervention classes used the TOP CSC or CSL activities in 80% or more of their sessions as prescribed (range: 84-100% of sessions, median: 97% of sessions). All intervention schools included lessons from the sexuality section of the curriculum (see Appendix K for the list of lessons): TOP facilitators at 5 schools (34 classes) implemented all 3 sexuality lessons (23, 24, 25), and 8 schools (41 classes) implemented 2 of them (lessons 23, 24). Appendix L details the frequency of all the lessons taught from the CSC. Lesson fidelity was high overall, with 1598 lessons (90%) being reported as "all" or

"most" of the lesson being covered. Only 157 (9%) of the reported lessons fell into the "some of it"

fidelity category (see Figure IV.1).

Figure IV.1: TOP Lessons by Fidelity

Qualitative thematic coding of the Facilitator Logs revealed that the "all of it" category is characterized by the fewest adaptations to the TOP CSC lessons, with the most class time spent on lesson content. Adaptations to lessons categorized with this level of fidelity report tend to be additions to CSC rather than omissions. With a gradient of fidelity ranging from "all of it" to "most" to "some of it," the "some of it" category is categorized by the most CSC adaptations, the most class time lost to

As reported in facilitator logs in response to question: "To what extent did you implement today's lesson/CSL Exercise as it was set out in the guide?"

other activities or school events, and more adaptations related to CSC omissions. TOP facilitators

reported completing 98 CSL projects; each class did 1-4 projects. Use of the CSC and CSL Guide

resources were high; 73 projects (74%) used CSC planning lessons, 82 (84%) used CSL planning

exercises, and 64 (65%) used at least 1 of the CSL reflection exercises. The majority of projects (n=77,

79%) consisted of indirect service (fundraising, creating items to donate to various groups); 19 (19%)

were direct service (youth interacted with intended recipients of action), and 2 (2%) were advocacy

(awareness raising).

The program staff consisted of 1 project coordinator, 3 regional coordinators, and 12 program

facilitators who supervised and delivered the intervention. All staff members were trained and annually

re-certified in TOP. Furthermore, all program facilitators met these minimum qualification requirements:

1) A bachelor's degree in youth development, social work, psychology, education, or related field, or

equivalent work experience; 2) a minimum of 1 year experience in teen program delivery; and 3) the

ability to become trained as a TOP facilitator.

Quality: The quality of the intervention was assessed using observations and student reported

engagement with the intervention. Overall, the quality of staff-participant interactions were rated highly,
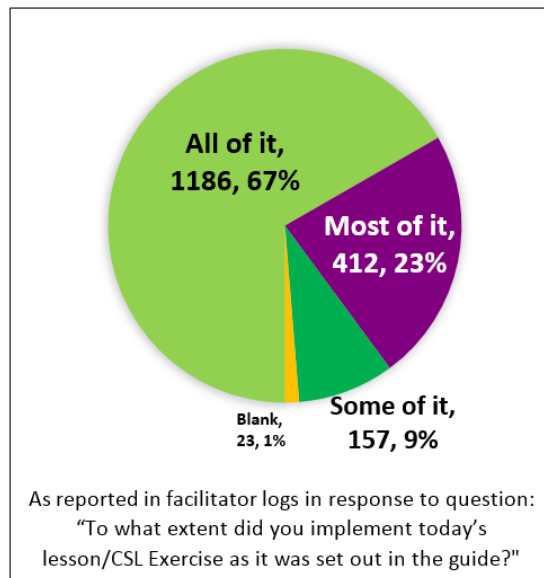
while the quality of youth engagement with the program offered some mixed results. Based on YPQA facilitator observations, the overall quality of staff-participant interactions was 3.90 out of 5 (n=21, SD=0.66) for the following subscales: Warm Welcome, Encouragement, Skill-Building, and Adult Partners. The overall quality of staff-participant interactions, as reported by youth, was assessed as 4.95 out of 7 (n=1,280, SD=1.63) on the Learning Climate Questionnaire. In addition, the majority of youth perceived that their TOP facilitators "care about me" (78%, n=1,259), "understand me" (76%, n=1,245) and "support and accept me" (80%, n=1,258).

The observed overall quality of youth engagement with the program was assessed as 2.58 out of 5 (n=24, SD=.85), for the following YPQA subscales: Planning, Choice, Reflection, and Active Engagement. Despite this moderate rating, the majority of youth affirmed in their survey responses that during their TOP CSL project they: learned new skills (69%, n=1,220), helped plan their service project (66%, n=1,202), learned how to deal with challenges (69%, n=1,221), enjoyed their community service (71%, n=1,207), and 72% (n=1,208) reported that the CSL project they did helped them make a positive difference in the lives of others.

Experiences of Comparison Group: With the business-as-usual counterfactual condition, the majority of youth in comparison classes were enrolled in HOPE or HOPE-PE classes (74%; n=84) or similar classes such as PE-Fitness (13%, n=84). The remainder of comparison classes were enrolled in Leadership Skills Development (13%, n=84). Classroom teachers deliver the required content for each course (see Table II.1 in Section II.B, page 6), and youth receive the standard number of hours of content depending on each school's semester schedule (ranging approximately 50-90 minutes per day). Community service is not required as part of the course content in any comparison class but in all Florida schools is generally encouraged for scholarship eligibility.

Context: According to structured interviews with teachers and administrators, the majority of both intervention and comparison schools did not include content or programming related to reproductive health in business-as-usual courses or in other school programs. When reproductive health content was offered, it was largely limited to 1-8 hours of content in business-as-usual courses (HOPE, HOPE/PE,

etc.) related to sexually transmitted diseases (STDs). Only one comparison school and one intervention school reported delivering lesson content related to contraception. In intervention schools, such business-as-usual curriculum content was in addition to the TOP CSC lessons.

According to the first follow-up youth survey data (see Appendix M), in the 2012-2013 school year, more than half to three-quarters of both intervention and comparison respondents reported receiving information in school (out of TOP or the comparison class) about abstinence, sexuality, pregnancy prevention and STDs/HIV. Less than one-half of respondents reported receiving such information through community organizations (such as Boys Club or Girls Club, Scouts, or YMCA), pamphlets or flyers, or their house of worship. The exact nature of content received in these topic areas was not collected.

Other contextual factors examined were external events and substantial unplanned adaptations. No external events such as school closures affected implementation, and there were no substantial unplanned adaptations reported in facilitator logs.

B. Impact study findings

For the outcome of ever having sexual intercourse 10 months after the program ended, the intervention group's rates were not significantly different from the comparison group's rates. The differences in rates of ever being pregnant or causing a pregnancy did not differ significantly between the treatment and comparison group after applying a correction multiple comparisons (Table IV.1). These conclusions were reached after the Bonferroni multiple comparisons correction.[24] That is, the typical level of .05 to determine statistical significance was divided by two to adjust for two comparisons being made.

Table IV.1. Estimated effects of complete case analysis using data from second follow-up surveys to address the primary research questions

Outcome measure	Intervention %[a]	Comparison %[b]	Intervention compared to comparison % difference (p-value of difference)
Ever had sex (n = 2,106)	37.83%	39.87%	2.04% (.27)[2]
Ever been pregnant or gotten someone pregnant (n = 2,058)	2.83%	5.38%	2.55% (.04)[2]

Source: Follow-up surveys administered 10 months after the program.

Notes: See Table III.1 for a more detailed description of each measure and Section III.G.1 for a description of the impact estimation methods.

[a]Regression adjusted means that projected on the whole sample.

[b]We adjusted for multiple comparisons, so the p-value considered to be statistically significant is $p \langle .025$. After adjustment, the outcome is not statistically significant.

For the outcome of ever having had sexual intercourse at the end of the program, TOP was found to be effective in reducing the number of youth who reported engaging in sexual intercourse relative to the comparison group. Exposure to the intervention reduced the number of youth having sex by approximately 3.7% (Table IV.2). For the outcome of ever being pregnant or causing a pregnancy, the intervention was found to be statistically significantly effective, compared to the comparison group, at the end of the program. A multiple comparisons adjustment was not made for secondary research questions.

Table IV.2. Estimated effects of complete case analysis using data from first follow-up surveys to address the secondary research questions

Outcome measure	Intervention %[a]	Comparison %[a]	Intervention compared to comparison % difference (p-value of difference)
Ever had sex (n=2,438)	28.07%	31.74%	3.67% (.009)*
Ever been pregnant or gotten someone pregnant (n=2,401)	1.60%	2.74%	1.14% (0.04)*

Source: Follow-up surveys immediately after the program.

Notes: See Table III.1 for a more detailed description of each measure and Section III.G.1 for a description of the impact estimation methods.

[a]Regression adjusted means that projected on the whole sample.

*Statistically significant result at $p \langle 0.05$.

Three sensitivity analyses were conducted to assess whether the findings were robust. Detailed descriptions of the methods and results of the sensitivity analyses are presented in Appendix H. Only one aspect of the analytical approaches was altered in each sensitivity analysis, thus providing a better chance at pinpointing the sources of discrepancy had it occurred.

The first sensitivity analysis employed a different approach for handling inconsistent survey responses, where the responses given in the first survey taken by a student were treated as the correct response for that student. All later responses were adjusted if they conflicted with this first response in order to adjust for inconsistent data, as opposed to categorizing all inconsistent responses as missing as done in the benchmark analysis. The results of this sensitivity analysis matched those from the benchmark analysis for the primary outcomes and for the secondary outcome of ever having had sexual intercourse at the end of the program with respect to the statistical significance of the treatment effect. This analysis did not present significant findings for the secondary outcome of ever being pregnant or causing a pregnancy at the end of the program.

The second sensitivity analysis employed a different approach for handling missing data. Multiple imputation was applied to impute (or substitute) values for missing data, as opposed to simply deleting cases with missing information (listwise deletion). Similar to the first sensitivity analysis, the results matched those from the benchmark analysis for the primary outcomes and for the secondary outcome of ever having had sexual intercourse at the end of the program with respect to the statistical significance of the treatment effect. This analysis did not present significant findings for the secondary outcome of ever being pregnant or causing a pregnancy at the end of the program.

The last sensitivity analysis attempted to address whether the differences in the baseline measure of "ever had sex" between the two conditions (see Tables III.4 and E.5 in Appendix E) may have brought bias to the analysis results by re-running the linear probability model using inverse probability of treatment weighting (IPTW)[27] that was originally proposed by Rosenbaum (1987). We first conducted a logistic regression where the treatment assignment was regressed on the demographic covariates and the baseline measure of "ever had sex" to obtain propensity scores. We then ran the regression models

incorporating IPTW, where the weight is equal to the inverse of the estimated probability of receiving the treatment that the subject actually received. The findings were similar to the benchmark analysis.

V.   Conclusion

The Office of Adolescent Health awarded the Florida Department of Health and University of South Florida a grant to replicate and evaluate Wyman's Teen Outreach Program®, an evidence-based program designed to promote healthy choices and reduce teen pregnancy.

Implementation Findings: Figure V.1 is a depiction of the program's fidelity benchmarks and how they were maintained in this study. Overall, youth in this study were exposed to fewer program sessions over a shorter span of time than recommended by the curriculum developers (Wyman), receiving an average of 24.2 sessions over 7.8 months. Study participants within a class were offered fewer CSL hours (14.9 hours) in comparison to youth in previous studies.[12, 14] The number of CSL hours completed on a student-level is unknown, due to the lack of data on this measure. Nevertheless, the overall attendance for program sessions was high, and the majority of students (89%) attended at least 75% of the 25 sessions prescribed. The program facilitators used CSC lessons or CSL exercises for nearly all sessions (97%), far surpassing the 80% fidelity

Figure V.1: Fidelity Benchmarks

TOP BENCHMARK	WHAT WAS RECEIVED (AVERAGE)	% CLASSES MET BENCHMARK
25 sessions	24 sessions	73%*
At least 1 weekly session	1.1 weekly sessions	100%
80% of sessions are CSC or CSL	96% sessions were CSC or CSL	100%
20 Hours CSL	14.9 hours/class	11%
9 months programming	7.8 months programming	0%

* The range of the number of sessions per youth was 1-46 sessions.

benchmark; additionally, facilitators made little to no adaptations to the curriculum lessons.

Throughout program implementation, facilitators provided high quality staff interactions with participants as reported by youth and observed by independent evaluators. During observations, youth appeared to be moderately engaged in program lessons, but reported that they enjoyed community service and felt that the experience helped them make a positive difference for others. Although there are no

benchmarks for staff interaction or quality components, positive adult guidance and youth's skills in community engagement are fundamental elements of the program's framework.

TOP was implemented in ninth grade high school classes as an addition to, not replacement of, course content. For the purposes of this study, business-as-usual is the term used to describe course content in both intervention and comparison class types. With the exception of Leadership Skills Development and PE-Fitness class types, which were only represented in the comparison group, business-as-usual was the same in both conditions. In both conditions, sexuality content, which primarily focused on sexually transmitted infections, was taught by classroom teachers but only minimally (1-8 hours of content per academic year). Apart from the information presented in business-as-usual and TOP, the majority of youth reported receiving additional information in school related to pregnancy, sexually transmitted infections/HIV, and sexuality education (see Appendix M).

Impact Findings: The impact of TOP was assessed over time – at the conclusion of the program (first follow-up) and 10 months after the completion of the program (second follow-up). Before the program began, the intervention and comparison groups were similar in composition, meaning participants in both groups reported similar demographics (e.g., age, race/ethnicity). However we did find that there was a higher prevalence of those who responded "yes" to the "Ever had sex?" question at baseline in the comparison group than in the intervention group among those who completed the post-program survey, and among those who completed the 10-month follow up (second follow-up) survey. Baseline prevalence of "Ever had sex" was included as a covariate in the corresponding impact analyses and the residual analysis indicated that baseline difference was appropriately adjusted. Not all youth participated in survey administration at the first and second follow-up; losing these youth from the sample (attrition) means this study's results are not as applicable to the general population of schools that participated in the study.

At the end of programming, intervention youth were *less* likely to report having had sex than their comparison counterparts. This measure has not been analyzed in previous studies; however, becoming pregnant or causing a pregnancy has been analyzed in multiple studies.[12-15] Like other studies,

intervention youth were *less* likely to report causing a pregnancy or becoming pregnant than their comparison counterparts, however this finding was not confirmed by the sensitivity analyses.

Explanation of Outcomes: Though previous studies have reported that TOP reduces teen pregnancy,[12-14] academic suspension,[12-15] and course failure[12, 14, 15] for youth participants, an HHS evidence review determined that only one study presented strong evidence supporting the effectiveness of TOP.[16] That study found that teen pregnancy rates were reduced for adolescent females who participated in the program, though that same impact (causing pregnancy) was not observed for adolescent males. This study showed there were differences between intervention and comparison youth only on the initiation of sex.

Explanations for the dissimilarity in findings are potentially related to: 1) differences in the study population, 2) length of follow up, and 3) limitations related to program setting. First, the background characteristics of this study's participants were more representative of gender and race/ethnicity compared to previous studies. Participants in other studies of TOP were primarily African American females, [13] white females, [15] or on average older adolescents, [13-14] in comparison to the participants this study served. In addition, this study had a longitudinal follow up component that was not part of previous studies. Finally, this study took place in a school-based setting, which introduced several challenges.

Meeting fidelity benchmarks in a school setting was often challenging due to scheduling conflicts with classroom teachers and/or school events. Additional scheduling challenges were introduced by the logistics of evaluation data collection. Baseline data collection, including the consent process, occurred throughout August, delaying the start of program implementation to September. While the academic school year generally ends in May, a full 9 months of program implementation was not possible because post-program data collection had to be scheduled as to not interfere with state standardized testing or end-of-year exams. Despite the shortened length of implementation, it should be noted that previous studies report that the quality of service projects, the youth's perception of input when selecting service projects, and the perception of an emotionally safe environment, are more crucial than the raw number of CSL hours completed and the number of sessions attended.

Restrictions such as access to transportation often limited service projects selected by youth, possibly influencing the youth's perception of input and development of autonomy; both of these factors are closely associated with positive program outcomes.[28] Due to these restrictions, the majority of CSL projects were indirect (76%), while the Allen et al. study participants were offered a broader range of service learning types.[13] There is growing evidence that direct service learning has a greater impact on social-emotional development for adolescents, and this is vital to TOP's framework.[29]

Limitations: Selection bias is a potential limitation in this study. Programming occurred in the intervention schools 1 year prior to drawing the evaluation cohort. This means that knowledge of the program from the previous year may have influenced youth or parent decisions to enroll or delay enrollment in health classes.

Given the extensive research previously conducted on TOP and the differences in results, further research on the program impacts is warranted. Specifically, further investigation of 1) the long-term program impacts on first sex, 2) developing benchmarks for quality of service learning experiences, staff-participant interactions, and youth engagement, and 3) the level of association of an emotionally safe environment and types of service learning on program impacts.

VI. References

[1] United States Census Bureau. Florida passes New York to become the nation's third most populous state, Census Bureau reports. 2014; http://www.census.gov/newsroom/press-releases/2014/cb14-232.html. Accessed March 8, 2015.

[2] Annie E. Casey Foundation. The 2010 kids count data book: state profiles of child well-being; 2010.

[3] National Center for Higher Education Management Systems (NCHEMS). State profile reports: Florida, public high school graduation rates 2009; 2015.

[4] Centers for Disease Control and Prevention (CDC), National Center for HIV, STD and TB Prevention (NCHSTP), Division of STD/HIV Prevention. Sexually transmitted disease morbidity for selected STDs by age, race/ethnicity and gender 1996-2011, CDC WONDER Online Database; 2015.

[5] Centers for Disease Control and Prevention (CDC), National Center for HIV, STD and TB Prevention (NCHSTP), Division of STD/HIV Prevention. HIV surveillance in adolescents and young adults; 2011.

[6] National Women's Law Center. People in Medically Underserved Areas (%). 2010.

[7] Robert Graham Center. Access denied: A look at America's medically disenfranchised. National Association of Community Health Centers, Incorporated; 2007.

[8] United States Department of Agriculture Economic Research Service (USDA ERS). State fact sheets: Florida. 2015. http://www.ers.usda.gov/data-products/state-fact-sheets/state-data.aspx?StateFIPS=12&StateName=Florida#.VOTly_nF_nh. Accessed February 17, 2015.

[9] Goesling B, Colman S, Trenholm C, Terzian M, Moore K. Programs to Reduce Teen Pregnancy, Sexually Transmitted Infections, and Associated Sexual Risk Behaviors: A Systematic Review. Journal of Adolescent Health. 2014;54(5):499-507.

[10] Office of Adolescent Health (OAH), Office of Public Health and Science (OPHS), U.S. Department of Health and Human Services (DHHS). Teenage pregnancy prevention: replication of evidence-based programs. Funding opportunity announcement and application instructions; 2010.

[11] Wyman Center, Inc. Following the TOP approach: Wyman's Teen Outreach Program facilitator training participant's guide. Eureka, MO; 2010.

[12] Allen JP, Philliber S. Evaluating why and how the Teen Outreach Program works: years 3 – 5 of the Teen Outreach national replication (1986/87 – 1988/89): Association of Junior Leagues Inc.; 1991.

[13] Allen JP, Philliber S, Herrling S, Kuperminc GP. Preventing teen pregnancy and academic failure: experimental evaluation of a developmentally based approach. Child Development. 1997;68(4):729-742.

[14] Allen JP, Kuperminc G, Philliber S, Herre K. Programmatic prevention of adolescent problem behaviors: The role of autonomy, relatedness, and volunteer service in the teen outreach program. Am J Commun Psychol. 1994;22(5):617-638.

[15] Allen JP, Philliber S, Hoggson N. School-based prevention of teen-age pregnancy and school dropout: Process evaluation of the national replication of the Teen Outreach Program. Am J Commun Psychol. 1990;18(4):505-524.

[16] Office of Adolescent Health (OAH), Office of Public Health and Science (OPHS), U.S. Department of Health and Human Services (DHHS). Research evidence for Teen Outreach Program (TOP); 2012. http://tppevidencereview.aspe.hhs.gov/pdfs/TeenOutreachProgram.pdf. Accessed February 23, 2015.

[17] Wyman Center, Inc. Wyman's Teen Outreach Program® logic model. 2012.

[18] Wyman Center, Inc. Wyman's Teen Outreach Program (TOP). 2013; http://wymancenter.org/nationalnetwork/top/.

[19] Florida Department of Education. Florida Bright Futures Scholarship program: Bright Futures Student Handbook. n.d.; http://www.floridastudentfinancialaid.org/ssfad/bf/. Accessed March 3, 2015.

[20] TIBCO Spotfire S+® 8.2 Guide to Packages [computer program]. TIBCO Software Inc.

[21] Bartram D, Foster, J., Lindley, P. A., Brown, A. J., & Nixon, S. Learning climate questionnaire (LCQ): background and technical information. Oxford: Employment Service and Newland Park Associates Limited; 1993.

[22] High/Scope Educational Research Foundation. Youth program quality assessment. Ypsilanti, MI: High/Scope Press; 2005.

[23] Office of Adolescent Health (OAH), Office of Public Health and Science (OPHS), U.S. Department of Health and Human Services (DHHS). Using the linear probability model to estimate impacts on binary outcomes in randomized controlled trials. Brief 6 (December). 2014. http://www.hhs.gov/ash/oah/oah-initiatives/assets/lpm-tabrief.pdf. Accessed March 2, 2015.

[24] Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilita. Libreria internazionale Seeber; 1936.

[25] The SAS system for Windows, Version 9.4 [computer program]. Cary, NC: SAS Institute; 2011.

[26] Office of Adolescent Health (OAH), Office of Public Health and Science (OPHS), U.S. Department of Health and Human Services (DHHS). Fidelity monitoring guidance for TPP grantees; 2011.

[27] Rosenbaum P.R. Model-based direct adjustment. The Journal of the American Statistician. 1987; 82:387–394.

[28] Teen Outreach Program, Wyman's Teen Outreach Program © Research Brief.

[29] Wilczenski FL, Coomey SM. Basics of Service Learning. A Practical Guide to Service Learning. New York, NY: Springer Science+Business Media, LLC; 2007. http://books.google.com/books?id=I6KfxGxI3GUC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false. Accessed March 23, 2015.

Appendices list

Appendix A: Wyman's Teen Outreach Program logic model

Appendix B: Data collection efforts

Appendix C: Implementation evaluation data collection

Appendix D: Study sample

Appendix E: Baseline Equivalence Methods and Results

Appendix F: Missing data, model specification, and estimating impacts

Appendix G: Data cleaning procedures

Appendix H: Sensitivity analyses

Appendix I: Implementation evaluation methods

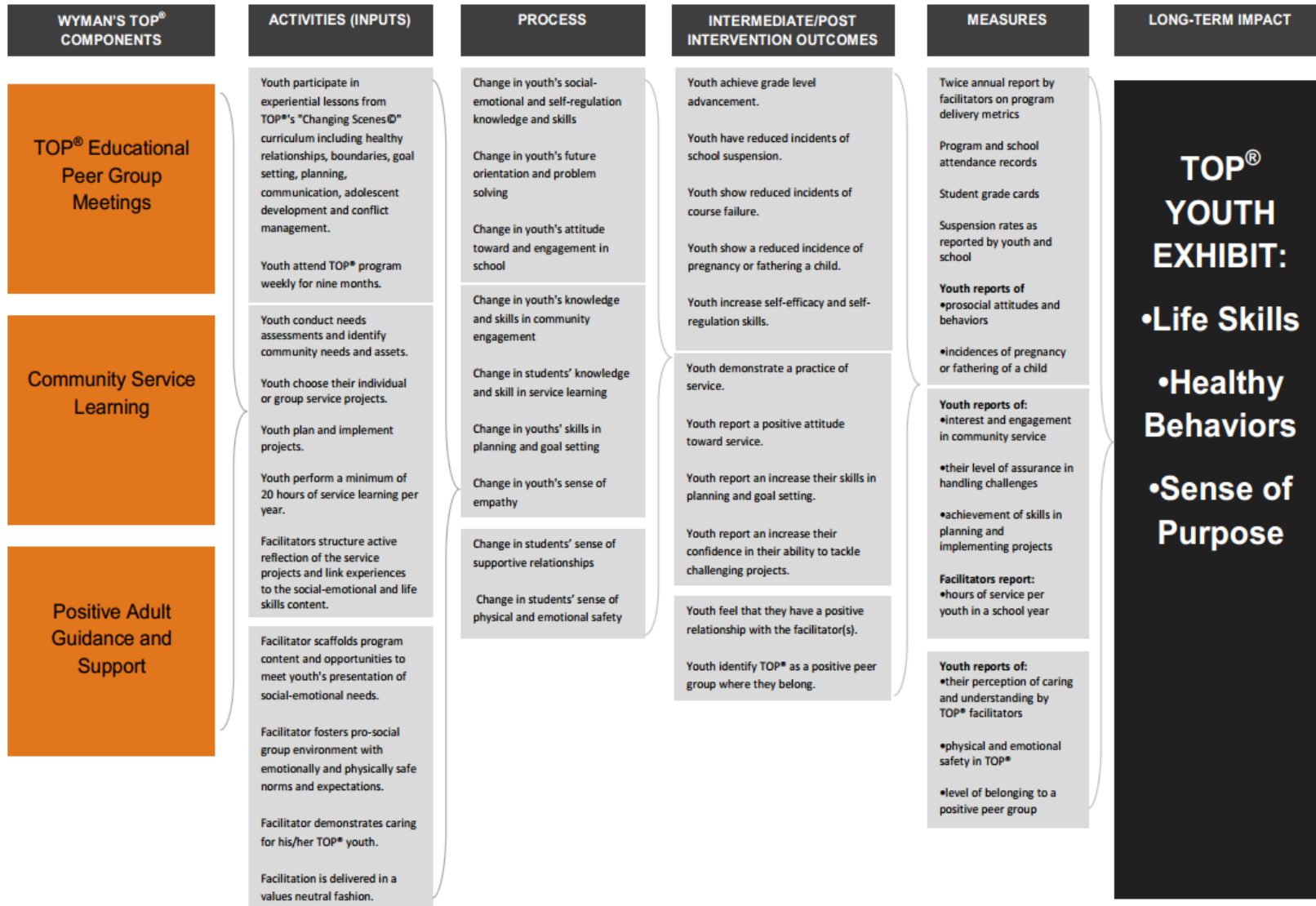Appendix J: Number of sessions received as reported by attendance records

Appendix K: TOP Changing Scenes Curriculum© Lesson Titles

Appendix L: Frequency and fidelity of lessons taught from Changing Scenes Curriculum©

Appendix M: Implementation findings – context

Appendix A: Wyman's Teen Outreach Program Logic Model

Figure A.1. Wyman's Teen Outreach Program Logic Model

WYMAN'S TOP® COMPONENTS	ACTIVITIES (INPUTS)	PROCESS	INTERMEDIATE/POST INTERVENTION OUTCOMES	MEASURES	LONG-TERM IMPACT
TOP® Educational Peer Group Meetings	Youth participate in experiential lessons from TOP®'s "Changing Scenes©" curriculum including healthy relationships, boundaries, goal setting, planning, communication, adolescent development and conflict management. Youth attend TOP® program weekly for nine months.	Change in youth's social-emotional and self-regulation knowledge and skills Change in youth's future orientation and problem solving Change in youth's attitude toward and engagement in school Change in youth's knowledge and skills in community engagement Change in students' knowledge and skill in service learning Change in youths' skills in planning and goal setting Change in youth's sense of empathy Change in students' sense of supportive relationships Change in students' sense of physical and emotional safety	Youth achieve grade level advancement. Youth have reduced incidents of school suspension. Youth show reduced incidents of course failure. Youth show a reduced incidence of pregnancy or fathering a child. Youth increase self-efficacy and self-regulation skills. Youth demonstrate a practice of service. Youth report a positive attitude toward service. Youth report an increase their skills in planning and goal setting. Youth report an increase their confidence in their ability to tackle challenging projects. Youth feel that they have a positive relationship with the facilitator(s). Youth identify TOP® as a positive peer group where they belong.	Twice annual report by facilitators on program delivery metrics Program and school attendance records Student grade cards Suspension rates as reported by youth and school **Youth reports of** •prosocial attitudes and behaviors •incidences of pregnancy or fathering of a child **Youth reports of:** •interest and engagement in community service •their level of assurance in handling challenges •achievement of skills in planning and implementing projects **Facilitators report:** •hours of service per youth in a school year **Youth reports of:** •their perception of caring and understanding by TOP® facilitators •physical and emotional safety in TOP® •level of belonging to a positive peer group	**TOP® YOUTH EXHIBIT:** •Life Skills •Healthy Behaviors •Sense of Purpose
Community Service Learning	Youth conduct needs assessments and identify community needs and assets. Youth choose their individual or group service projects. Youth plan and implement projects. Youth perform a minimum of 20 hours of service learning per year. Facilitators structure active reflection of the service projects and link experiences to the social-emotional and life skills content.				
Positive Adult Guidance and Support	Facilitator scaffolds program content and opportunities to meet youth's presentation of social-emotional needs. Facilitator fosters pro-social group environment with emotionally and physically safe norms and expectations. Facilitator demonstrates caring for his/her TOP® youth. Facilitation is delivered in a values neutral fashion.				

Rev. January 2012

Appendix B: Data collection efforts

Table B.1. Data collection efforts used in the impact analysis of the Teen Outreach Program and timing

Data Collection Effort	Timing
Start date of programming	September 2012
Baseline survey	August 27, 2012 – September 7, 2012
First follow-up survey	May 6, 2013 – May 31, 2013
Second follow-up survey	March 3, 2014 – September 3, 2014*

*During second follow-up, an in-school data collection took place from March 3, 2014 to April 3, 2014. An out-of-school data collection was conducted from June 24, 2014- to September 3, 2014 (see Data Collection – Impact Evaluation in the main text for details). The start date was influenced by the amount of time required to obtain IRB approval, secure incentives, and resolve survey software issues. The end date was influenced by low initial participation due to invalid/out of date contact information and the recommendation of OAH and Mathematica.

Appendix C: Implementation evaluation data collection

Table C.1. Data used to address implementation research questions

Implementation element	Types of data used to assess whether the element of the intervention was implemented as intended	Frequency/sampling of data collection	Party responsible for data collection
Adherence: How often were sessions offered? How many were offered?	*The count, weekly frequency, and fidelity of all TOP sessions offered as captured in the combined attendance and facilitator log records*	*Participant attendance data recorded daily for every session*	*Program staff*
	The duration of program sessions as captured in the facilitator logs	*Facilitator logs submitted for every session*	*Program staff*
	The sum of CSL project hours per class as captured in the CSL Records	*CSL records for all CSL projects submitted at the completion of each project*	*Program staff*
Adherence: What and how much was received?	*Student daily attendance as captured in the attendance records*	*Participant attendance data recorded daily for every session*	*Program Staff*
Adherence: What content was delivered to youth?	*The number of sessions covering TOP Changing Scenes lessons and CSL activities as captured in the Facilitator Logs*	*Facilitator logs submitted weekly for every session*	*Program Staff*
	The number of CSL resources used for each CSL project as captured in the CSL records	*CSL records for all CSL projects submitted at the completion of each project*	*Program Staff*
	The types of CSL projects as captured in the CSL Records	*CSL records for all CSL projects submitted at the completion of each project*	*Program Staff*
Adherence: Who delivered material to youth?	*List of staff members hired and trained to implement program*	*Data on all staff members are maintained throughout the program*	*Program Staff*
	List of program staff who attended Recertification Training	*Data on all staff members are maintained throughout the program*	*Program Staff*
	List of program staff who met minimum qualification requirements	*Data on all staff members are maintained throughout the program*	*Program Staff*
Quality: Quality of staff-participant interactions	*Observations of interaction quality using the YPQA tool*	*All facilitators were observed at least once, and most (75%) were observed twice at one time-point*	*Evaluation Staff*
Quality: Quality of staff-participant interactions	*Youth survey items on the first follow-up assessment*	*Youth were surveyed on these items once at Immediate Post-intervention*	*Evaluation Staff*

Implementation element	Types of data used to assess whether the element of the intervention was implemented as intended	Frequency/sampling of data collection	Party responsible for data collection
Quality: Quality of youth engagement with program	*Observations of engagement using the YPQA tool*	*All facilitators were observed twice at one time-point*	*Evaluation Staff*
	Survey items on the first follow-up assessment	*Youth were surveyed on these items once at Immediate Post-intervention*	*Evaluation Staff*
Counterfactual: Experiences of comparison condition	*Florida Department of Education website describing course curriculums*	*Curriculum information was retrieved from website once, 10 months post-intervention*	*Program and Evaluation staff*
	Structured Interviews with comparison group school personnel	*One-time interviews 2-years post-intervention with teachers and school administrators from TOP classrooms during the 2012-2013 school year*	*Program and Evaluation staff*
Context: Other TPP programming available or offered to study participants (both intervention and comparison)	*Survey items on first follow-up assessment*	*Youth were surveyed on these items once at Immediate post-intervention*	*Program and Evaluation staff*
	Structured Interviews with treatment and comparison group school personnel	*One-time interviews 2-years post-intervention with teachers and school administrators from TOP classrooms during the 2012-2013 school year*	*Program and Evaluation staff*
Context: External events affecting implementation	*News sources or reports*	*As discovered*	*Program Staff*
	Events as captured in the Facilitator Logs	*Facilitator logs submitted weekly for every session*	*Program Staff*
Context: Substantial unplanned adaptation(s)	*Adaptations as captured in the Facilitator Logs*	*Facilitator logs submitted weekly for every session*	*Program Staff*

Appendix D: Study sample

Table D.1. Cluster and youth sample sizes by intervention status

Number of:	Time period	Total sample size	Intervention sample size	Comparison sample size	Total response rate	Intervention response rate	Comparison response rate
Clusters/Schools: At beginning of study		28	14	14			
Clusters/Schools: Contributed at least one youth at baseline	Baseline	26	13	13	.929	.929	.929
Clusters/Schools: Contributed at least one youth at follow-up #1	Immediately post-programming	26	13	13	.929	.929	.929
Clusters/Schools: Contributed at least one youth at follow-up #2	10-months post-programming	26	13	13	.929	.929	.929
Youth: In non-attriting clusters/ schools at time of assignment		4,327	1,950	2,377			
Youth: Who consented		4,063	1,845	2,218	.939	.946	.933
Youth: Contributed a baseline survey "Ever having sexual intercourse"	Baseline	3,196	1,454	1,742	.739	.746	.733
Youth: Contributed a baseline survey "Ever pregnant or causing pregnancy"	Baseline	3,166	1,449	1,717	.732	.743	.722
Youth: Contributed a follow-up #1 survey "Ever having sexual intercourse"	Immediately post-programming	2,438	1,070	1,368	.563	.549	.576
Youth: Contributed a follow-up #1 survey "Ever pregnant or causing pregnancy"	Immediately post-programming	2,401	1,064	1,337	.555	.546	.562
Youth: Contributed a follow-up #2 survey "Ever having sexual intercourse"	10-months post-programming	2,106	939	1,167	.487	.482	.491
Youth: Contributed a follow-up #2 survey "Ever pregnant or causing pregnancy"	10-months post-programming	2,058	922	1,136	.476	.473	.478

Appendix E: Baseline Equivalence Methods and Results

The equality of demographic variables and baseline prevalence of the main outcomes of interest across treatment groups were tested for everybody that completed the baseline survey, as well as in the analytical samples corresponding to the primary and secondary research questions. That is, overall baseline equivalence tests were conducted, and then were repeated for 4 subsets that were used in the impact analysis models; i) for those who complemented the 10 month follow up (second follow-up) survey and responded to the ever had sex question, ii) for those who complemented the 10 month follow up (second follow-up) survey and responded to the ever been pregnant or gotten someone pregnant question, iii) for those who completed the end of program (first follow-up) survey and responded to the ever had sex question, iv) for those who completed the end of program (first follow-up) survey and responded to the ever been pregnant or gotten someone pregnant question.

Linear probability models were used for the binary variables, cumulative logit models were used for age, and generalized logit models were used for race. Pair-matched cluster indicator variables were included as main effects, and clustering effects at the school level were accounted for as random effects. Results for the equivalence tests on the 4 analytic samples are shown in Tables III.3 to III.7 below, along with sample sizes, percentages for each baseline measure, and percent differences between the intervention and comparison groups. Also included are results for equivalence tests between those present at baseline and presented for post survey for the outcome and time point of interest compared to those who were present at baseline and then were not present for the post survey of the same outcome and time point of interest; this was done for each analytic sample.

Table E.1. Summary statistics of key baseline measures for youth who completed baseline and responded to "ever had sex" question at first follow-up

Baseline measure	Intervention mean or % (standard deviation)	Comparison mean or % (standard deviation)	Intervention versus comparison % difference	Intervention versus comparison *p*-value of difference
Demographic Characteristics	-	-	-	-
Age (in years)	14.49 (0.90)	14.52 (0.90)	N/A	.52
≤ 14	67.94%	64.25%	3.69%	-
15	20.00%	23.68%	3.68%	-
≥ 16	12.06%	12.06%	0.00%	-
Gender (female)	52.62%	52.56%	0.06%	.86
Race/ethnicity	-	-	-	.68
White, Non-Hispanic	62.42%	64.47%	2.05%	-
Black, Non-Hispanic	7.57%	7.31%	0.26%	-
Hispanic/Latino	21.12%	18.57%	2.55%	-
Other, Non-Hispanic	8.88%	9.65%	0.77%	-
Outcomes	-	-	-	-
Ever had sex	17.48%	21.13%	3.65%	.02
Total Sample Size	1,070	1,368	N/A	N/A

Table E.2. Summary statistics of key baseline measures for youth who completed baseline and responded to "ever been pregnant or gotten someone pregnant" question at first follow-up

Baseline measure	Intervention mean or % (standard deviation)	Comparison mean or % (standard deviation)	Intervention versus comparison % difference	Intervention versus comparison *p*-value of difference
Demographic Characteristics	-	-	-	-
Age (in years)	14.48 (0.89)	14.52 (0.89)	N/A	.45
≤ 14	67.95%	64.55%	3.40%	-
15	20.39%	23.93%	3.54%	-
≥ 16	11.65%	11.52%	0.13%	-
Gender (female)	52.73%	52.51%	0.22%	.95
Race/ethnicity	-	-	-	.72
White, Non-Hispanic	62.41%	63.95%	1.54%	-
Black, Non-Hispanic	7.52%	7.63%	0.11%	-
Hispanic/Latino	21.24%	18.85%	2.39%	-
Other, Non-Hispanic	8.83%	9.57%	0.74%	-
Outcomes	-	-	-	-
Ever been pregnant or gotten someone pregnant	1.13%	0.60%	0.53%	.72
Total Sample Size	1,064	1,337	N/A	N/A

Table E.3. Summary statistics of key baseline measures for youth who completed baseline to compare between those who completed the second follow-up survey and those who missed the second follow-up survey for "ever had sex" question

Baseline measure	Present mean or % (standard deviation)	Not Present mean or % (standard deviation)	Present versus Not Present % difference	Present versus Not Present *p*-value of difference
Demographic Characteristics	-	-	-	-
Age (in years)	14.41 (0.77)	14.90 (1.14)	N/A	<.001
≤ 14	69.33%	48.71%	20.62%	-
15	21.75%	27.26%	5.51%	-
≥ 16	8.93%	24.02%	15.09%	-
Gender (female)	52.94%	47.57%	5.37%	<.01
Race/ethnicity	-	-	-	.01
White, Non-Hispanic	63.30%	60.72%	2.58%	-
Black, Non-Hispanic	7.01%	10.68%	3.67%	-
Hispanic/Latino	20.23%	21.07%	0.84%	-
Other, Non-Hispanic	9.45%	7.53%	1.92%	-
Outcomes	-	-	-	-
Ever had sex	17.52%	37.94%	20.42%	<.001
Total Sample Size	2,106	1,049	N/A	N/A

Table E.4. Summary statistics of key baseline measures for youth who completed baseline to compare between those who completed the second follow-up and those who missed the second follow-up survey for "ever been pregnant or gotten someone pregnant" question

Baseline measure	Present mean or % (standard deviation)	Not Present mean or % (standard deviation)	Present versus Not Present % difference	Present versus Not Present *p*-value of difference
Demographic Characteristics	-	-	-	-
Age (in years)	14.40 (0.76)	14.88 (1.13)	N/A	<.001
≤ 14	69.48%	49.35%	20.13%	-
15	21.82%	27.66%	5.84%	-
≥ 16	8.70%	22.99%	14.29%	-
Gender (female)	53.01%	47.57%	5.44%	<.01
Race/ethnicity	-	-	-	.01
White, Non-Hispanic	63.27%	60.65%	2.62%	-
Black, Non-Hispanic	6.95%	10.75%	3.80%	-
Hispanic/Latino	20.46%	20.93%	0.47%	-
Other, Non-Hispanic	9.33%	7.66%	1.67%	-
Outcomes	-	-	-	-
Ever been pregnant or gotten someone pregnant	0.63%	3.36%	2.73%	<.001
Total Sample Size	2,058	1,070	N/A	N/A

Table E.5. Summary statistics of key baseline measures for youth who completed baseline to compare between those who completed the first follow-up survey and those who missed the first follow-up survey for "ever had sex" question

Baseline measure	Present mean or % (standard deviation)	Not Present mean or % (standard deviation)	Present versus Not Present % difference	Present versus Not Present *p*-value of difference
Demographic Characteristics	-	-	-	-
Age (in years)	14.50 (0.90)	14.79 (1.02)	N/A	<.001
≤ 14	65.87%	50.91%	14.96%	-
15	22.01%	28.73%	6.72%	-
≥ 16	12.06%	20.36%	8.30%	-
Gender (female)	52.58%	46.30%	6.28%	<.01
Race/ethnicity	-	-	-	.03
White, Non-Hispanic	63.58%	58.58%	5.00%	-
Black, Non-Hispanic	7.42%	11.02%	3.60%	-
Hispanic/Latino	19.69%	23.29%	3.60%	-
Other, Non-Hispanic	9.31%	7.11%	2.20%	-
Outcomes	-	-	-	-
Ever had sex	19.52%	40.59%	21.07%	<.001
Total Sample Size	2,438	717	N/A	N/A

Table E.6. Summary statistics of key baseline measures for youth who completed baseline to compare between those who completed the first follow-up survey and those who missed the first follow-up survey for "ever been pregnant or gotten someone pregnant" question

Baseline measure	Present mean or % (standard deviation)	Not Present mean or % (standard deviation)	Present versus Not Present % difference	Present versus Not Present *p*-value of difference
Demographic Characteristics	-	-	-	-
Age (in years)	14.50 (0.89)	14.78 (1.02)	N/A	< .001
≤ 14	66.06%	51.17%	14.89%	-
15	22.37%	28.61%	6.24%	-
≥ 16	11.58%	20.22%	8.64%	-
Gender (female)	52.60%	46.35%	6.25%	< .01
Race/ethnicity	-	-	-	.01
White, Non-Hispanic	63.27%	59.42%	3.85%	-
Black, Non-Hispanic	7.58%	10.45%	2.87%	-
Hispanic/Latino	19.91%	22.97%	3.06%	-
Other, Non-Hispanic	9.25%	7.15%	2.10%	-
Outcomes	-	-	-	-
Ever been pregnant or gotten someone pregnant	0.83%	3.99%	3.16%	< .001
Total Sample Size	2,401	727	N/A	N/A

Appendix F: Missing data, model specification, and estimating impacts

- Missing data

  For the benchmark analysis we conducted a complete case analysis. We excluded any cases that had a missing response at anyone of the demographic variables, baseline risk, or response variables of interest from the analysis for that research question. Alternative missing data approaches conducted as sensitivity analysis are described in Appendix H.

- Equations for estimating impact

  A linear mixed effects model was used to estimate the treatment impact. Here is the equation that links the outcome variable and the predictors for student i in school j:

  $$P(Y_{ij}=1)=\beta_{0j}+\beta_1 T_{ij}+\sum \beta_k X_{(k)ij}+\varepsilon_{ij}$$

  $$\beta_{0j}=\gamma_{00}+u_{0j}$$

  $$u_{0j}=\text{random effect for school } j$$

- where $Y_{ij}=1$ indicates the presence of the outcome of interest; T_{ij} is the treatment indicator $(T_{ij}=1,\text{ if treatment; 0 otherwise})$. The estimated β_1 coefficient is the estimated impact of the treatment on the probability of the event of interest. Covariates $\{X_{(k)ij,k}=1,2...\}$ is a group of baseline characteristics. The term ε_{ij} denotes the random error at the individual level and u_{0j} is the random effect of school j on the intercept.

- Model Specification

  More specifically, the baseline characteristics included in the impact analysis models are comprised of indicator variables for baseline measure of the outcome, gender, age, and race as well as the treatment blocks. The exact specification for this analysis was:

  $$\sum \beta_k x_{(k)ij}=\sum_{k=2}^{8}\beta_k x_{(k)ij}+\sum_{l=1}^{12}\beta_{l+8}block_{(l)ij}$$

  where

  $$x_2=\begin{cases}1 & \text{if event of interest presented at baseline}\\0 & \text{otherwise}\end{cases}$$

  $$x_3=\begin{cases}1 & \text{if female}\\0 & \text{male}\end{cases}$$

$$x_4 = \begin{cases} 1 & \text{if age} = 15 \\ 0 & \text{otherwise} \end{cases}$$

$$x_5 = \begin{cases} 1 & \text{if age} \geq 16 \\ 0 & \text{otherwise} \end{cases}$$

$$x_6 = \begin{cases} 1 & \text{if black} \\ 0 & \text{otherwise} \end{cases}$$

$$x_7 = \begin{cases} 1 & \text{if Hispanic} \\ 0 & \text{otherwise} \end{cases}$$

$$x_8 = \begin{cases} 1 & \text{if other race} \\ 0 & \text{otherwise} \end{cases}$$

$$block_l = \begin{cases} 1 & \text{if in block } l \\ 0 & \text{otherwise} \end{cases}$$

The random errors $\{\varepsilon_{ij}\}$ are assumed to be independent and identically distributed following normal distribution and so as the random effects $\{u_{0j}\}$. It is also assumed that the random errors and the random effects are independent to each other.

Appendix G: Data cleaning procedures

In preparation for data analysis, data were cleaned at all time points (baseline through second follow-up). Our data cleaning focused on preparing the demographic and outcome variables for analysis. Outcome variables were ever had sex and ever been pregnant or caused a pregnancy and are described in Table III.1. Demographics variables included in the analysis were age, race, ethnicity, and gender. Demographic measures used are listed in Table G.1.

Table G.1. Demographic variables used as covariates

Question	Response Options
In what month were you born?	January through December
In what year were you born?	1991 through 2002
What is your age?	Open
Are you Hispanic or Latino?	• Yes • No
What is your race? You may mark more than one answer.	• American Indian or Alaskan Native • Asian • Black or African-American • Native Hawaiian or other Pacific Islander • White
Are you male or female?	• Male • Female • Refuse to answer

Covariates: Three questions were asked of youth to determine their age. Some youth chose to answer all, none, or some of these questions. To account for this missing data and to be as consistent as possible, age was calculated from the month and year of birth. The cutoff date for calculating age was the date of baseline survey administration and the day of birth used was the 15th of each month.

Two questions on the survey assessed a youth's race and ethnicity. The questions which collected data on race allowed for multiple responses. Hispanic ethnicity was determined by a yes or no response. These 2 variables were merged to form a single variable termed race/ethnicity. The collapsed categories were as follows: Non-Hispanic White, Non-Hispanic Black, Hispanic (of any race), and an Other category which included American Indian or Alaskan Native, Native Hawaiian or other Pacific Islander,

and Asian. For the variable gender, youth had the option of selecting one of the following: male, female, refuse to answer. Responses of "refuse to answer" were removed from the analysis.

The above described demographic variables were collected at each time point. Demographic variables from the baseline survey were preferred for use as covariates. However, if the participant did not provide demographic information at baseline, then this information was adopted from the next available time point. For example, if no demographic information was provided at baseline, then demographic information from the immediate post-program time point was used. If demographic information from this time point was also missing, then the information provided at second follow-up was used. If age at baseline was missing, then month and year of birth provided at either of the follow-up time points was used to calculate age at baseline.

Outcomes: The outcomes included and the response options are listed in Tables III.1. These questions contained a logical skip pattern whereby youth who answered "no" to the question "Have you ever had sexual intercourse?" (ever had sex) were then instructed to skip the question "To the best of your knowledge, have you ever been pregnant or gotten someone pregnant, even if no child was born?" (ever been pregnant). However, it can logically be concluded that if a youth answered "no" to ever having had sex, then it would be biologically impossible for the youth to have ever been pregnant. Therefore, for youth who answered "no" to ever having sex, a "no" answer was imputed for the ever pregnant question.

Inconsistent Responses: Some youth answered inconsistently to the outcome questions. The simplest type of inconsistent response is when youth answered questions inconsistently across variables, but within the same time point. As in the example above, some youth answered "no" to the question ever had sex and then answered "yes" to the question ever been pregnant, rather than skipping the question as instructed (Table G.2). Another form of inconsistent data was when youth answered "yes" to the ever had sex or ever been or gotten someone pregnant questions on one survey, then responded "no" to the same question on a later survey. This is an example of an across time point inconsistent response (Table G.3). The final example of inconsistent responses is when a youth had a combination of the above 2 scenarios. That is, they answered inconsistently both across variable and across time points (Table G.4). For the

benchmark analyses, all inconsistent responses were set to missing. An alternative strategy in which the

first response was considered the correct response was considered in the sensitivity analysis.

Table G.2. Frequencies of within time point inconsistent responses

Time Point	N
Baseline	2
First follow-up	42
Second follow-up	2

Table G.3. Frequencies of within Variable and Across Time Points Inconsistent Responses

Variable: **Ever Had Sex** (Response Chosen by Youth)

Time 0	Time 1	Time 2	N
Yes	No	No	35
No	Yes	No	53
Yes	Yes	No	39
Yes	No	Yes	31

Variable: **Ever Been Pregnant** (Response Chosen by Youth)

Time 0	Time 1	Time 2	N
Yes	No	No	8
No	Yes	No	9
Yes	Yes	No	6
Yes	No	Yes	5

Table G.4. Frequencies of Across Variable and Across Time Points Inconsistent Responses

Scenario 1 (n = 8)

Variable	Time 0	Time 1	Time 2
Ever Had Sex (Response Chosen by Youth)	Yes	No	No
Ever Been Pregnant (Response Chosen by Youth)	Yes	Missing	Missing

Scenario 2 (n = 0)

Variable	Time 0	Time 1	Time 2
Ever Had Sex (Response Chosen by Youth)	Yes	Yes	No
Ever Been Pregnant (Response Chosen by Youth)	Yes	Yes	Missing

Scenario 3 (n = 8)

Variable	Time 0	Time 1	Time 2
Ever Had Sex (Response Chosen by Youth)	No	Yes	No

Appendix H: Sensitivity analyses

The benchmark approach presents the main analytical method used to estimate the program impacts. However, certain methodological decisions were made in order to complete this analysis. Three sensitivity analyses were carried out to evaluate the sensitivity of the benchmark estimates to these various methodological decisions. Provided in this appendix is the summary of the benchmark approach, descriptions of each of the sensitivity analyses, a summary of the results of the sensitivity analyses in comparison to the benchmark analysis results, and tables detailing the results of each sensitivity analysis.

The benchmark analysis used the linear probability model to assess the percent differences between the intervention group and the comparison group for complete cases only. In the benchmark analysis, inconsistent responses were removed from the analysis and the following variables were controlled for: gender, race, age, matched pairs, and baseline risk. School level was inputted into the model as a random effect to account for clustering. The purpose of the benchmark analysis was to determine if there was a statistically significant difference in the percentages between the treatment and comparison groups for the primary and secondary research questions.

In order to confirm the robustness of the results of our benchmark analysis, three separate sensitivity analyses were conducted. In this first sensitivity analysis, for inconsistent responses, the first response the youth provided was considered to be the correct response and responses provided at subsequent time points were altered accordingly.

The benchmark analysis was a complete case analysis based on listwise deletion. That is, only youth who completed baseline and at least one follow-up time point were included in the analysis of that research question. For the second sensitivity analysis, multiple imputation was used to impute likely values in place of the missing values. The demographic variables were imputed first separately by treatment conditions, independently of the sexual behavior variables. The outcome variables at baseline, first follow-up and second follow-up were imputed sequentially afterwards. Twenty imputed data sets were generated and each was analyzed using the linear probability model that contained the same

covariates and options as the benchmark analysis. The SAS procedure MIANALYZE was used to combine the results to generate generalizable estimates.

The last sensitivity analysis utilized inverse probability of treatment weighting (IPTW) to minimize any potential bias that may have caused by the difference in baseline measure of "ever had sex". The linear probability model was conducted in conjunction with IPTW.

Results of Sensitivity Analyses

The first sensitivity analysis found that the intervention was not effective compared to the comparison group for the outcomes of ever having sex 10 months after the program, ever being pregnant or causing a pregnancy 10 months after the program, and at the end of the program. This analysis found the intervention to be effective in reducing the number of youth who had engaged in sex relative to the comparison group at the end of the program. Treatment reduced the number of youth having sex by approximately 3.8%.

The second sensitivity analysis that is based on multiple imputation also obtained the same findings as the benchmark analysis with respect to statistical significance of treatment effects on primary outcomes. The intervention effects were found to be statistically significant for reporting ever having had sex at the end of program, where the treatment group had a rate that was 4.02% lower than that of their comparison counterpart. No significant differences were detected between the treatment and comparison groups on any other outcomes.

The third sensitivity analysis did not find that the intervention was effective compared to the comparison group for either of the primary outcomes 10 months after the program. It confirmed that the intervention was effective in reducing the number of youth who engaged in sex related to the comparison group at the end of the program. Treatment reduced the number of youth having sex by approximately 3.8%. The significant finding on ever being pregnant or causing a pregnancy from the benchmark analysis was not replicated.

Table H1. Sensitivity of impact analyses using data from follow-up surveys administered 10 months after program to address the primary research questions

Intervention compared with comparison	Benchmark difference[a]	Benchmark *p*-value	First response difference	First response *p*-value	Multiple imputation difference	Multiple imputation *p*-value	IPTW difference	IPTW *p*-value
Ever Had Sex (n=2,106)	2.04%	.27	2.56% (n=2,273)	.19	1.92% (n=3827)	.31	2.10% (n=2106)	.16
Ever been pregnant or gotten someone pregnant (n=2,058)	2.55%	.04	1.56% (n=2,265)	.13	2.27% (n=3827)	.03	2.62% (n=2058)	.03[1]

Source: Follow-up surveys administered 10 months after the program.

Notes: See Table III.1 for a more detailed description of each measure and Section III.G.1 for a description of the impact estimation methods.

[a]The outcome is not statistically significant after adjusting for multiple comparisons.

Table H.2. Sensitivity of impact analyses using data from follow-up surveys immediately after the program to address the secondary research questions

Intervention compared with comparison	Benchmark difference	Benchmark *p*-value	First response difference	First response *p*-value	Multiple imputation difference	Multiple imputation *p*-value	IPTW difference	IPTW *p*-value
Ever Had Sex (n=2,438)	3.67%	.009*	3.83% (n=2621)	.01*	4.02% (n=3827)	<.001*	3.76% (n=2438)	<.001*
Ever been pregnant or gotten someone pregnant (n=2,401)	1.14%	0.04*	1.34% (n=2614)	.84	1.08% (n=3827)	.11	1.31% (n=2401)	.07

Source: Follow-up surveys immediately after the program.

Notes: See Table III.1 for a more detailed description of each measure and Section III.G.1 for a description of the impact estimation methods.

*Statistically significant result at *p* < 0.05

Appendix I: Implementation evaluation methods

**Table I.1. Methods used to address implementation research questions**

Implementation element	Methods used to address each implementation element
Adherence: How often were sessions offered? How many were offered?	*Total number of sessions is a sum of the sessions captured by the attendance data and facilitator logs. The proportion of classes achieving intended 25 sessions is calculated as the number of classes with at least 25 sessions reported in the attendance/log data divided by the total number of classes (n=70). The span of programming is the number of days between the first and last date of programming reported in attendance/log data divided by the average number of days in a month (30.43). The number of programming weeks per class is a count of the weeks for which attendance data was recorded per class. Average weekly frequency of sessions is calculated as the number of sessions divided by the number of weeks when programming was offered per class, summed for all of the classes and then divided by the number of classes (n=70). Average session duration is calculated as the average of the reported session lengths (class length), measured in minutes. Total number of CSL activities is a sum of the projects captured by the CSL Records submitted. The proportion of classes achieving intended CSL hours is calculated as the number of classes with at least 20 hours of CSL activities reported in CSL Records divided by the total number of classes.* *(Note: These data have some limitations because of incomplete reporting; a total of 2,012 program sessions were captured by facilitator logs and/or attendance records, 6% were captured only in facilitator logs and 9% were captured only in attendance data. CSL records were not submitted for 21.5% of classes (15/70). In addition the total number of sessions does not include sessions that consisted only of active service; these sessions were documented in CSL Records which did not report individual session dates but rather the total number of project hours completed.)*
Adherence: What and how much was received?	*Average number of sessions attended is calculated as the average of the number of sessions that each student attended. The percentage of participants who received at least 75% of the intended 25 sessions is calculated as the total number of participants who received at least 19 sessions divided by the total number of participants.* *(Note: These data have some limitations because attendance data were not recorded for all students in the treatment condition of this evaluation, and student-level data CSL data was not obtained for hours of CSL completed; only class-level average data was available.)*
Adherence: What content was delivered to youth?	*The percentage of classes that cover lessons/CSL in at least 80% of their sessions is calculated as the number of lesson sessions plus the number of CSL sessions divided by the total number of sessions offered in each class. The percentage of lesson sessions that cover all, most, or some of the content is calculated as the number of lesson sessions identified by facilitators as each divided by total number of lesson sessions offered. Topic frequency is calculated as the number of sessions covering each lesson divided by the total number of lesson sessions. Topic fidelity is calculated as the proportion of each lesson's frequency that covers all, most or some of the content as it was set out in the curriculum guide. The proportion of intervention classes that would allow TOP reproductive health lessons to be taught from the sexuality section of the curriculum is calculated as the number of intervention classes that would allow lessons 23, 24, and 25 to be taught divided by the total number of intervention classes. The proportions of reported CSL projects (98) utilizing TOP CSL resources are calculated to include: the proportion of projects that utilized Level II curriculum lessons 6-9 to prepare for planning CSL projects, the proportion utilized Planning Lessons from the Wyman Community Service Learning Guide, and the proportion that utilized Reflection Lessons from the Wyman Community Service Learning Guide. The prevalence of different types of CSL projects is calculated as the proportion of reported CSL projects (108) that were classified by facilitators as direct, indirect, or advocacy CSL.* *(Note: These data have some limitations because sessions that consisted only of active service were not counted as CSL sessions. These sessions were documented in CSL records which did not report individual session dates but rather the total number of project hours completed.)*

Implementation element	Methods used to address each implementation element
Adherence: Who delivered material to youth?	*Total number of staff delivering the program is a simple count of staff members implementing the program. Percentage of staff trained is calculated as the number of staff members who were trained divided by the total number of staff who delivered the program. Percentage of staff recertified is calculated as the number of staff members who were annually recertified divided by the total number of staff who delivered the program. Percentage of staff who met the minimum qualification requirements is calculated as the number of staff members who met the minimum requirements divided by the total number of staff who delivered the program.* *(Note: These data are limited because the specific qualifications of staff members were unavailable to the evaluation team.)*
Quality: Quality of staff-participant interactions	*Observed quality of staff-participant interactions is calculated as the average of 4 subscales from the YPQA observation tool: Warm Welcome (3 items), Encouragement (3 items), Skill-Building (5 items), and Adult Partners (2 items). The items are measured using a 1-5 scale, where 3 is average and higher scores reflecting higher program quality for that measure. Youth-reported quality of staff-participant interactions is calculated using the Learning Climate Questionnaire (6-item scale) which poses questions to youth about whether or not the TOP facilitator/teacher: 1) provides them with choices and options, 2) conveyed confidence in their ability to do well in the course, 3) encouraged them to ask questions, 4) listens to how they would like to do things, 5) tries to understand how youth see things before suggesting a new way to do things, and 6) Youth felt understood by their TOP facilitator/teacher. The scale is scored 1-7, with higher scores reflecting higher perceived level of support for youth autonomy (vs. a controlling environment). As a second measure of the quality of youth-reported staff-participant observations, the proportion of youth who perceived supportive adult presence is calculated as the percentage of youth answering "Yes! Very Much" or "Yes, Somewhat" in response to 3 questions: 1) TOP facilitators care about me 2) TOP facilitators understand me 3) TOP facilitators support and accept me.* *(Note: There are limitations to these data because the YPQA observation tool includes some items non-applicable in a school setting, the YPQA observations were conducted at only one time point, and one item of the Learning Climate Questionnaire (4) is missing data. Simple mean imputation was used to derive these missing data.)*
Quality: Quality of youth engagement with program	*Observed quality of youth engagement with the program is calculated as the average of 2 subscales from the YPQA observation tool: Engagement (8 items) which includes Planning, Choice, and Reflection, and Active Engagement (4 items). The items are measured using a 1-5 scale, where 3 is average and higher scores reflecting higher program quality for that measure. Youth-reported quality of youth engagement with the program is calculated as the proportion of youth answering "Yes! Very Much" or "Yes, Somewhat" in response to 5 questions: 1) I learned new skills during my TOP community service project. 2) I helped plan my TOP community service project 3) I learned how to deal with challenges during my TOP community service project 4) I enjoyed the community service part of TOP 5) The community service project I did during TOP helped me make a positive difference in the lives of others.* *(Note: A limitation of the observed data is that the YPQA tool includes items non-applicable in a school setting. Furthermore, YPQA observations were conducted at only one time point.)*
Counterfactual: Experiences of counterfactual condition	*The school curriculum and requirements of business as usual (BAU) courses are described in Section II as part of the intervention as well as the counterfactual conditions. Based on the structured interview question "In your 2012-2013 [Health/HOPE/Critical Thinking/Etc.] course, was any sexuality education content besides TOP delivered to your students?," the BAU programming available to both intervention and comparison groups described by school personnel is listed in Section IV.* *(Note: Because the counterfactual condition is BAU, data for the adherence and quality measures used above for the intervention condition are unavailable for the counterfactual condition.)*

Implementation element	Methods used to address each implementation element
Context: Other TPP programming available or offered to study participants (both intervention and counterfactual)	*Youth-reported survey data on other TPP information received is presented as frequency counts and percentages. School-reported data on other TPP programming available to both intervention and comparison groups is presented descriptively in Section IV of the final report based on responses to the structured interview question, "In 2012-2013, besides in the [Health/HOPE/etc.] or TOP classes, was any sexuality education content/programming delivered to students?"* *(Note: A limitation of these school-reported data is the sample size and time point at which they were collected, 2 years post-program.)*
Context: External events affecting implementation	*The number of schools that were closed as a result of district turnaround initiatives (unrelated to the TPP programming that occurred in this project) is reported in Section IV.*
Context: Substantial unplanned adaptation(s)	*Basic frequencies of substantial unplanned adaptations are calculated from qualitative content analysis of the facilitator logs and reported in Section IV.*

Appendix J: Number of sessions received as reported by attendance records

**Figure J.1. Number of program sessions received as reported by attendance records**

Total number of sessions received
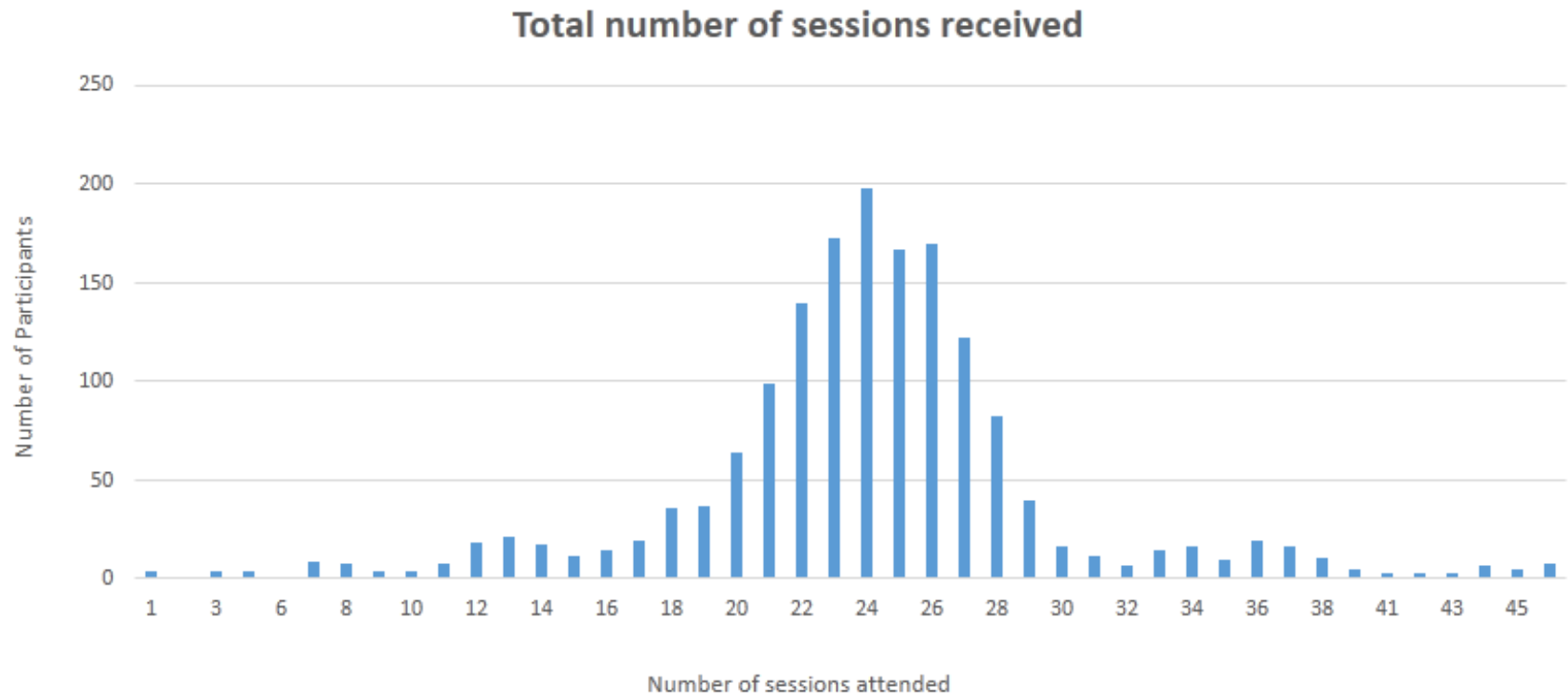
Appendix K: TOP Changing Scenes Curriculum© Lesson Titles

Figure K.1 TOP Changing Scenes Curriculum© lesson titles

Section 1 - Welcome to TOP
1. Spinning a Yarn
2. Autograph Hunt
3. Intro to TOP
4. Compiling group rules
5. Cooperation, Trust Building, Ice Breakers, Energizers

Section 2 - Preparing for Service
6. Defining the circles of your community
7. Identifying Resources: What can you offer?
8. Using Community Resources
9. What our community needs: Community survey

Section 3 - Values
10. Intro to Values: What does your family say?
11. Intro to Gender: Cross-Gender dialogue
12. Changing Places

Section 4 - Development
13. Health and Hygiene
14. Looking Back
15. Interview with a family member

Section 5 - Relationships
16. Intro to relationships: Making Friends
17. Intro to romantic relationships
18. Love or infatuation
19. Dealing with pressure situations
20. Pressure role plays

Section 6 - Influence
21. Television viewing: The choice is up to you
22. Social Pressures

Section 7 - Sexuality
23. Intro: Sexual Myths or Facts
24. STD Jeopardy
25. Pregnancy Probability

Section 8 - Communication/Assertiveness
26. Intro: Active Listening
27. Speaking for Yourself
28. Refusing a Request

Section 9 - Goal Setting
29. Intro: 10-year class reunion
30. Setting and Achieving Goals
31. Examining teenage parenthood

Section 10 - Decision making
32. Intro: Examining recent decisions
33. Getting it together: Thee C's to good decision making

Section 11 - Wrap-up
34. This time next year
35. Someone who...

Appendix L: Relative Frequency of Lessons Offered from Changing Scenes Curriculum©

Figure L.1. Relative frequency of lessons offered from Changing Scenes Curriculum© across all intervention classes
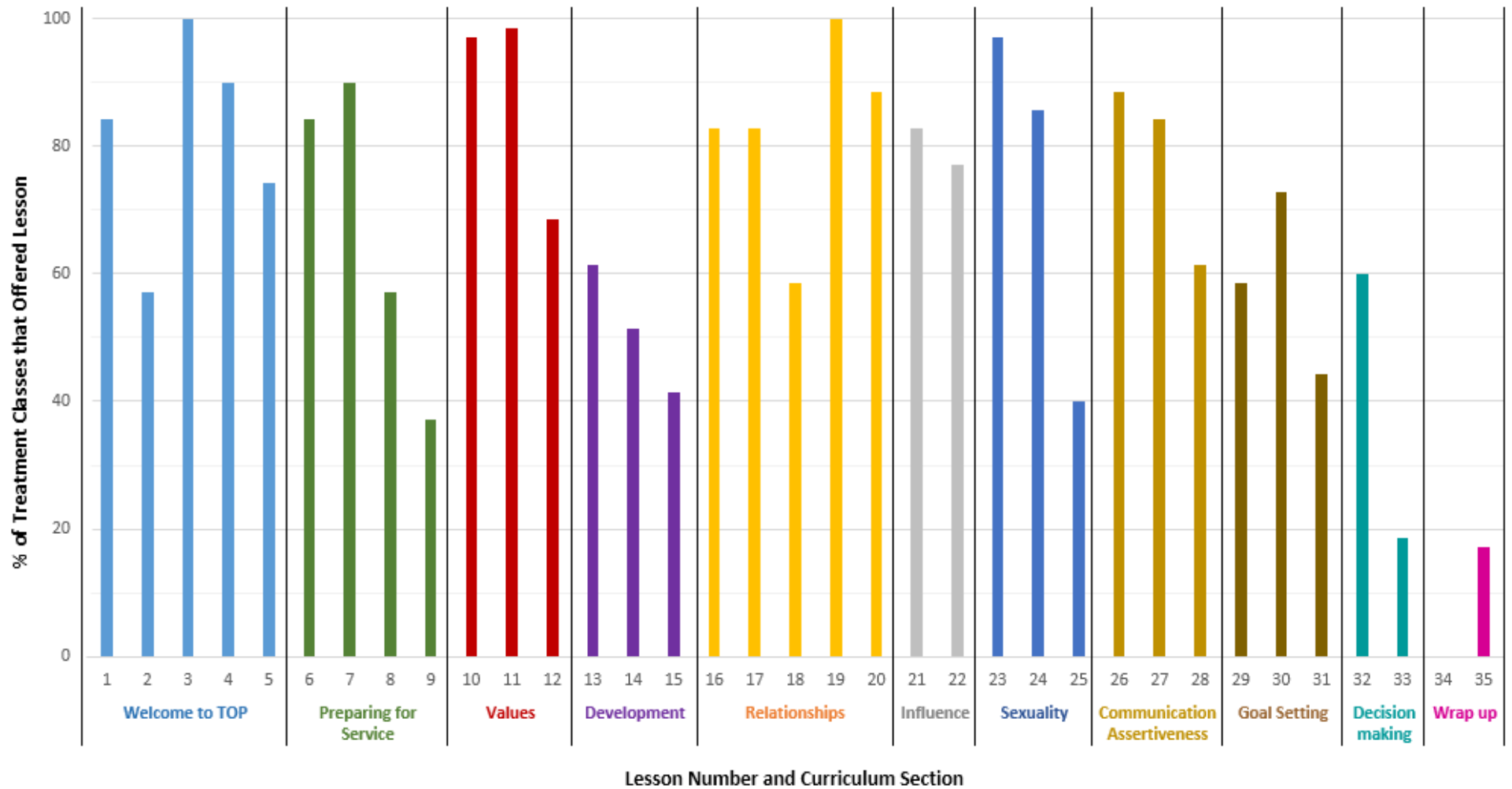
Total number of Treatment classes = 70

Lessons were considered "offered" to a class if a facilitator log was submitted indicating such

Appendix M: Implementation findings – context

Table M.1. Implementation Context: Results of the first follow-up youth survey data

In the 2012-2013 school year . . .		Information type by intervention/comparison							
		Abstinence		Sexuality		Pregnancy Prevention		STD/HIV	
		Intervention	Comparison	Intervention	Comparison	Intervention	Comparison	Intervention	Comparison
…outside of TOP/this class, did you receive information about the following in school?	YES	52%	52%	75%	64%	77%	64%	79%	67%
	NO	48%	48%	25%	36%	23%	36%	21%	33%
		(n=949)	(n=1,220)	(n=1,019)	(n=1,250)	(n=1,015)	(n=1,258)	(n=1,024)	(n=1,273)
…outside of TOP/this class, have you heard any guest speakers	YES	52%	52%	52%	53%	53%	51%	58%	58%
	NO	48%	48%	48%	47%	47%	49%	42%	42%
		(n=949)	(n=1,220)	(n=963)	(n=1,242)	(n=965)	(n=1,246)	(n=959)	(n=1,247)
…in your community, such as through the Boys Club or Girls Club, Scouts, or YMCA (but not including church), did you receive any of the following information?	YES	22%	21%	22%	22%	22%	22%	21%	21%
	NO	78%	79%	78%	78%	78%	78%	79%	79%
		(n=896)	(n=1,115)	(n=891)	(n=1,121)	(n=888)	(n=1,126)	(n=885)	(n=1,113)
…have you received any pamphlets or flyers on the following?	YES	37%	28%	37%	28%	39%	29%	41%	30%
	NO	63%	72%	63%	72%	61%	71%	59%	70%
		(n=929)	(n=1,193)	(n=932)	(n=1,202)	(n=931)	(n=1,208)	(n=921)	(n=1,189)
…have you heard any announcements or seen any ads on the following?	YES	50%	44%	51%	47%	61%	54%	57%	52%
	NO	50%	56%	49%	53%	39%	46%	43%	48%
		(n=936)	(n=1,193)	(n=930)	(n=1,202)	(n=941)	(n=1,208)	(n=928)	(n=1,202)
…through your church/temple/mosque, did you receive any of the following information?	YES	49%	48%	44%	44%	37%	40%	35%	36%
	NO	51%	52%	56%	56%	63%	60%	65%	64%
		(n=629)	(n=833)	(n=635)	(n=842)	(n=989)	(n=1,201)	(n=928)	(n=1,201)