The Heritage Foundation

BACKGROUNDER

No. 3230 | JULY 31, 2017

The Failure of the Teen Pregnancy Prevention Program: Advocates of Evidence-Based Policymaking Ignore the Evidence

David B. Muhlhausen, PhD

Abstract

The federal government has a poor record of replicating local social programs. The federal Teen Pregnancy Prevention (TPP) grant program is intended to fund "evidence-based" sex-education programs that reduce the number of teen pregnancies. Evidence-based-policy advocates mistakenly believe that these "evidence-based" grants will be effective because they are replicating program models that were previously thought to be successful. The evidence of effectiveness underlying the TPP grants is not nearly as robust as the federal government and evidence-based-policy advocates claim. Overwhelmingly, evaluations of TPP grants replicating "evidence-based" models have been demonstrated to be ineffective. Yet, the evidence-based policymaking community is virtually silent on this failure. Clearly, replicating an "evidenced-based" model does not guarantee success. The funding for ineffective TPP programs should be eliminated.

Policymakers frequently assume that when an intervention was found effective in one setting, the same results can be repeated elsewhere. However, the history of social programs is replete with examples of programs that, while effective in one location, simply failed to work elsewhere. The federal government has a poor record of replicating effective social programs.[1] Examples include the Center for Employment Training (CET) replication,[2] the Head Start CARES Demonstration,[3] and Hawaii's Opportunity Probation with Enforcement (HOPE) program.[4]

A more recent example is the federal government's Teen Pregnancy Prevention (TPP) grants, created by the Consolidated Appropriations Act of 2010.[5] TPP grants are administered by the Office of

Key Points

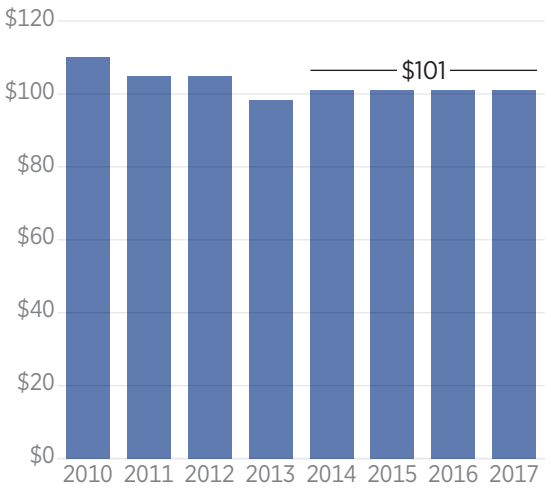
- Policymakers frequently assume that when an intervention was found effective in one setting, the same results can be repeated elsewhere—but the federal government has a poor record of replicating social programs.
- The federal Teen Pregnancy Prevention (TPP) grant program funds "evidence-based" sex-education programs intended to reduce teen pregnancy.
- Evidence-based-policy advocates mistakenly believe that "evidence-based" grants will be effective because they are replicating program models that were previously thought to be successful.
- The evidence of effectiveness underlying the TPP grants is not nearly as robust as the federal government and evidence-based-policy advocates claim.
- Overwhelmingly, evaluations of TPP grants demonstrate that replications of "evidence-based" models were ineffective. Clearly, replicating an "evidenced-based" model does not guarantee similar results. Funding for the ineffective TPP should be eliminated.

This paper, in its entirety, can be found at **http://report.heritage.org/bg3230**

The Heritage Foundation
214 Massachusetts Avenue, NE
Washington, DC 20002
(202) 546-4400 | heritage.org

CHART 1

Teen Pregnancy Prevention Funding Remains Stable



SOURCE: Carmen Solomon-Fears, "Teenage Pregnancy Prevention: Statistics and Programs," Congressional Research Service *Report* RS20301, October 26, 2016, Appendix A, pp. 24–25, https://fas.org/sgp/crs/misc/RS20301.pdf (accessed June 12, 2017).

BG3230 heritage.org

Adolescent Health (OAH) within the Department of Health and Human Services (HHS). The OAH "invests in the implementation of evidence-based TPP programs, and provides funding to develop and evaluate new and innovative approaches to prevent teen pregnancy."[6]

Funded with approximately $100 million per fiscal year (FY) since its inception, the TPP is supposed to award "competitive contracts and grants to public and private entities to fund medically accurate and age appropriate programs that reduce teen pregnancy."[7] Chart 1 provides the amount of funding for TPP from FY 2010 to FY 2017. To date, Congress has spent more than $820 million on TPP.[8]

As shown in Chart 2, the trend in births to girls between 15 years and 19 years of age has steadily declined for decades. Commenting on the decline in the teen birth rates since the implementation of TPP in 2010, Results for America—an evidence-based-policy advocacy group—concluded, *without* any evidence: "While it is not realistic to associate all the success to TPP alone, it has contributed significantly to the use of proven approaches to reduce teen pregnancy."[9] Similarly, a TPP-funded supporter, Associate Professor Christine Dehlendorf of the University of California, San Francisco, recently wrote:

> Teen birth rates have been declining since the 1990s. New data reveal an even sharper drop in the five years following the inception of the TPP program, from about 34 births per 1,000 girls in 2010 to 22 per 1,000 in 2015—a 35 percent decrease. This unprecedented decline suggests that the Office of Adolescent Health's funding strategy for teen pregnancy prevention has been highly effective.[10]

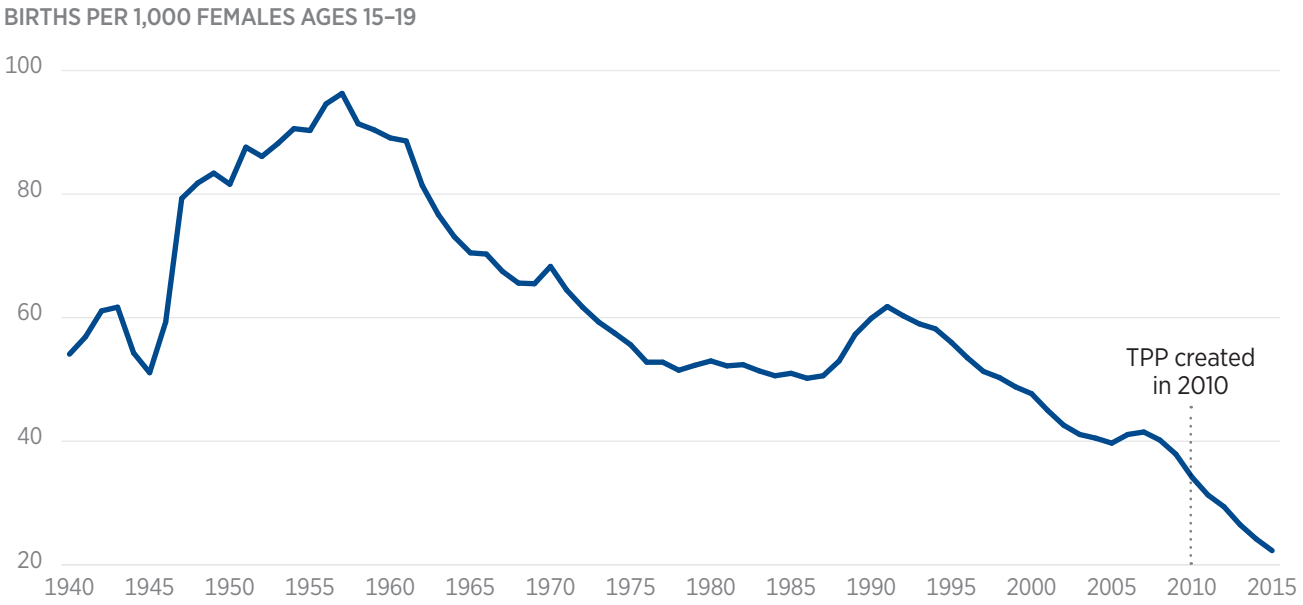
Readily apparent in Chart 2 is the fact that the beginning of the decline in teen births began decades before the creation of TPP. Advocates of evidence-based policymaking should be above confusing correlation with causation.

In other cases, proponents of TPP causally assert that the program is effective while conveniently ignoring the actual, publicly available, evaluations that conclusively demonstrate the program's *in*effectiveness. For example, Robert Gordon, the acting Deputy Director of the Office of Management and Budget during the Obama Administration, recently criticized the Trump Administration's plan to cut funding for the program. Gordon concluded that TPP "works" and that President Trump's "decision to terminate the program was based on ideology rather than evidence."[11]

The statements by Results for America, Associate Professor Dehlendorf, and Gordon raise a significant flaw in the nearly automatic assumption by the evidence-based-policy community that the replications of program models labeled "evidence-based" are effective. Advocates of evidence-based policymaking, especially those in Washington, DC, pay little regard to the difficulty of replicating program models. As Amy Feldman Farb and Amy Margolis of the OAH wisely caution, "Programs that were effective at one point in time, particularly decades ago, may no longer be effective today, nor in new settings and populations of young people."[12] In addition, the quality of the staff replicating the program may not be the same as that of the original staff. A particularly good instructor may have

CHART 2

Teen Birthrate in Steady Decline



SOURCE: Data Brief 259: Centers for Disease Control, National Center for Health Statistics, "Continued Declines in Teen Births in the United States, 2015," Data table for Figure 1. Birth rates for females aged 15–19, by age group: United States, 1991–2015, and Stephanie J. Ventura, T.J. Matthews, and Brady E. Hamilton, "Births to Teenagers in the United States, 1940-2000," *National Vital Statistics Reports*, Vol. 49, No. 10 (September 25, 2001), Table 1, p. 10.

BG3230 ☎ heritage.org

certain "intangibles" that influence participant outcomes far more than the faithful implementation of the curriculum. This conclusion is highly relevant to the evaluation literature used to identify program models labeled "evidence-based" and, thus, qualified for federal funding.

Teen Pregnancy Prevention Grants

TPP has two funding streams: Tier I and Tier II grants. According to HHS, Tier I grants are awarded to grantees replicating programs that "have been shown, in at least one program evaluation, to have a positive impact on preventing teen pregnancies, sexually transmitted infections, or sexual risk behaviors."[13] Thus, Tier I grants are supposed to be "evidence-based." The majority of TPP funding is dedicated to "effective program models" funded by the Tier I grants.[14] The other set of TPP grants, Tier II, fund demonstration programs that do not meet the OAH's evidence-based definition, but are considered by the OAH to be innovative programs worthy of funding.

In June 2016, Ron Haskins, a research fellow at the Brookings Institution and co-chair of the Commission on Evidence-Based Policymaking, testified before Congress that HHS requires "high-quality evidence showing that the programs produced significant impacts on important measures of teen sexual activity or teen pregnancy for the TPP program."[15] According to Results for America, the "tiered-evidence framework enables more dollars to be directed towards programs that have demonstrated success and are ready to be scaled for wider impact, while also directing lesser amounts of funding toward interventions that need to be tested and proven."[16] Further, Results for American claims that "Tier 1 grants support the replication of evidence-based programs that are *proven* to reduce teenage pregnancy or related risk behaviors."[17] (Emphasis added.)

Results for America and others believe that these grants will be effective because they are replicating programs labeled "evidence-based." Is this assumption correct? Ron Haskins wisely acknowledges that

most of the TPP Tier I models "had been evaluated only once by rigorous methods, leaving open the question of whether they could be successfully replicated."[18] As will be discussed later, the evidence of effectiveness underlying the Tier I grants is not nearly as robust as evidence-based-policy advocates have claimed. Many of the reviewed evaluations are not rigorous at all. Further, the evaluations of the evidence-based replications overwhelmingly find failure. Yet, the evidence-based policymaking community is virtually silent on this failure.

Unlike many federal grants that award funding with little regard to ensure that grantees faithfully implement the intended programs, the OAH places high standards on reporting measures on grantees to ensure that the "evidence-based" models were administered as intended. These requirements are intended to ensure implementation fidelity—the degree to which programs follow the theory underpinning the program, and how correctly the program components are put into practice.

The evaluated TPP grants "were required to engage in a phased-in implementation period lasting up to one year to allow time for thorough needs assessments and partner development."[19] Further,

> [i]mplementations were required to maintain fidelity to the program model and be of high quality as rated by an independent observer, high levels of youth retention and engagement were expected, and programs had to be medically accurate and age appropriate.[20]

Performance measurement data was reported to the OAH every six months to ensure implementation fidelity.[21]

Each of the Tier I grantees is supposed to evaluate the impact of the evidence-based model they are replicating. So far, from 2015 to May 2017, 13 experimental evaluations of nine "evidence-based" models have been published by HHS or in the *American Journal of Public Health*.[22] This review of this literature focuses on the Tier I grants that have undergone randomized experiments to assess effectiveness. Overwhelmingly, these evaluations demonstrate that replicating "evidence-based" models to affect the sexual behaviors of participants fails to produce the intended results. Clearly, replicating an "evidenced-based" model does not guarantee similar results.

Table 1 summarizes the results of the TPP Tier I experimental replication evaluations. Due to their methodological weaknesses, quasi-experimental evaluations are excluded from Table 1. First, the level of random assignment is classified as individual or cluster. Experimental evaluations that use random assignment are the "gold standard" of evaluation designs.[23] Randomized experiments attempt to demonstrate causality by holding constant all other possible causes of the outcome, isolating the program intervention as the only possible cause of differing outcomes, and observing whether the outcomes differ between the intervention and control groups. This methodology works best when the unit of analysis is randomly assigned to intervention and control groups. For the TPP I replication evaluations, the unit of analysis is the individual (for instance, student or youth).

However, a drawback to the scientific rigor of several TPP Tier I experimental evaluations is that while the unit of analysis is the individual, random assignment was, instead, based on clusters of individuals (such as schools and classrooms). Groups of students in classrooms or schools were randomly assigned to intervention and control groups. As will be seen from the literature review, several of these evaluations had intervention and control groups that were not equivalent on characteristics that can bias the results.[24] Therefore, these cluster randomization evaluations do not provide results that are as definitive as evaluations that randomly assigned individuals to intervention and control groups.

Second, Table 1 provides the sample size for each of the evaluations. The benefits of random assignment are most likely to occur with large sample sizes. Randomized evaluations using small sample sizes do not have the same scientific rigor as randomized evaluations using large sample sizes. Random assignment helps to ensure that the control group is equivalent to the intervention group in composition, predisposition, and experience. The groups are composed of the same types of individuals in terms of program-related and outcome-related characteristics. In addition, members of both groups should be similarly disposed toward the program. Further, the intervention and control groups should have the same experiences regarding time-related variables, such as their maturity level and history.[25]

Randomized experiments have the highest internal validity when sample sizes are large enough to

TABLE 1

The Failure of TPP Tier I Replication Evaluations

Program Name	Study	Random Assignment Level	Sample Size	Locations	SEXUAL BEHAVIOR OUTCOMES		
					Beneficial	**None**	**Harmful**
Becoming A Responsible Teen (BART)	Jenner et al. (2016)	Individual	850	Single-site (New Orleans, LA)	0	2	0
Children's Aid Society (CAS)-Carrera Program	Herrling (2016)	Individual	600	Single-site (3 schools, Chicago, IL)	0	6	0
¡Cuídate!	Kelsey, Layzer, et al. (2016)	Individual	2,169	Multi-site (small city, Southern CA; Phoenix, AZ; and Boston, MA)	0	9	0
It's Your Game... Keep It Real (IYG)	Potter et al. (2016)	Cluster	3,143	Multi-site (24 middle schools in rural South Carolina)	0	4	1
	Coyle et al. (2016)	Cluster	2,403	Single-site (20 middle schools in Houston, TX)	0	3	0
Promoting Health Among Teens! Abstinence-Only Intervention	Walker et al. (2016)	Individual	1,319	Single-site (8 middle schools in Yonkers, NY)	0	3	0
Reducing the Risk	Barbee et al. (2016)	Cluster	1,365	Single-site (Louisville, KY)	4	4	0
	Kelsey, Blocklin at al (2016)	Individual	3,314	Multi-site (17 schools in St. Louis, MO; Austin TX; and San Diego, CA	0	7	0
Safer Sex Intervention	Jenner et al. (2016)	Individual	268	Single-site (New Orleans, LA)	0	3	0
	Kelsey, Walker, et al. (2016)	Individual	2,108	Multi-site (38 clinics in Minnesota, Tennessee, and Florida)	1	9	0
Seventeen Days	Eichner et al. (2015)	Individual	1,317	Multi-site (20 clinics in Ohio, Pennsylvania, and West Virginia)	0	6	0
Teen Outreach Program (TOP)	Francis et al. (2016)	Cluster	17,194	Multi-site (Hennepin, MN; Northwest states; Kansas City, MO; Nonmetropolitan counties, FL; and Chicago, IL)	0	3	0
	Robinson et al. (2016)	Individual and cluster	3,252	Multi-site (Louisiana and Rochester, NY)	0	4	0
				Total	**5**	**63**	**1**
				Share of Total	**7.2%**	**91.3%**	**1.5%**

SOURCE: Individual programs. See Appendix for details.

BG3230 ☎ heritage.org

ensure that idiosyncrasies that can affect outcomes are evenly distributed between the program and control groups. With small sample sizes, disparities in the program and control groups can influence the findings. For this reason, evaluations with large samples are more likely to yield scientifically valid impact estimates.

Third, Table 1 classifies the evaluations as single or multi-site evaluations. The evaluations are classified as single site if the study takes place in a single county, city, town, or school district. When the evaluations take place in more than one county, city, town, or school district, these studies are classified as multi-site evaluations. This means that evaluations that take place in several schools in a single school district, for example, are classified as single-site evaluations.

Large-scale experimental evaluations based on multiple sites avoid problems of simplistic generalizations. A multitude of confounding factors that vary by location can influence the performance of social programs.[26] What works in Tulsa, Oklahoma, may not work in Baltimore, Maryland. Thus, the larger the size of the evaluation (for instance, the sample size and number of sites), the more likely the social program will be assessed under all of the conditions under which it operates. For TPP, the multi-site evaluations are an attempt to "scale-up" the OAH's evidence-based models to determine if these models can be successful when applied in multiple settings. However, "[r]esearch across many fields has demonstrated that when programs are scaled up, as in effectiveness or replication studies, they often don't find the same positive outcomes the original studies found."[27]

Fourth, Table 1 summarizes the results of the evaluations by classifying outcomes for sexual behaviors as "beneficial," "no effect," and "harmful." A statistically significant impact where the intervention group fared better than the control group is classified as beneficial. For example, if the intervention group reports statistically lower rates of sexual activity than the rates reported by the control group, this outcome is considered beneficial. However, a statistically significant impact where the intervention group did worse than the control group is classified as harmful. A finding of no effect occurs when the difference in outcomes for the intervention and control groups is statistically indistinguishable—meaning that the intervention failed to influence the outcome being assessed in either a beneficial or harmful way.

As becomes immediately clear from Table 1, the replications of TPP Tier I "evidence-based" models overwhelmingly find failure. Of 69 main outcomes, 63 (91.3 percent) were statistically insignificant—meaning that these "evidence-based" replications had no meaningful effect on sexual behaviors. Only five (7.2 percent) of the main outcomes were found to have beneficial impacts that were statistically significant, while one (1.4 percent) outcome was a statistically significant harmful impact.

Commenting on the effectiveness of TTP, Russell Cole, a senior researcher at Mathematica Policy Research, understatedly wrote, "Despite these investments, many of the evaluations did not show favorable, statistically significant results on behavioral outcomes."[28] These results should not be surprising. The federal government does not have a successful track record of funding effective sex-education programs. For example, a multi-site experimental evaluation of abstinence-education programs found that this approach had no effect on the sexual activities of youth.[29]

In addition to the low likelihood that programs that worked in one setting, would work in other circumstances, another reason for the failure of TPP may be the inconsistent and methodologically weak evidence used to label the program models as evidence-based. For example, the OAH used contradictory evidence of the effectiveness of Becoming A Responsible Teen (BART) program to label this model "evidence-based." Of the three randomized experiments that were classified with a "high" ranking for scientific rigor, two found the model to be ineffective.[30] Labeling BART an "evidence-based" model contradicts the body of research evaluating the program.

The results for the Tier II grants are similar to the failure of the Tier I grants. From 2015 to May 2017, the OAH has released 12 final reports based on experimental evaluations of Tier II grant programs.[31] These evaluations overwhelmingly find that these programs fail to affect the sexual behavior outcomes.[32]

A Review of the Evidence

The following sections review the evidence-based literature used by HHS to label specific models as evidence-based, and the results of the replications of these models through Tier I grants:[33]

- Becoming A Responsible Teen (BART);
- Children's Aid Society, Carrera Adolescent Pregnancy Prevention Program;
- ¡Cuídate!;
- It's Your Game: Keep It Real (IYG);
- Promoting Health Among Teens! Abstinence-Only Intervention;
- Reducing the Risk;
- Safer Sex Intervention;
- Seventeen Days; and
- Teen Outreach Program (TOP).

The programs reviewed are limited to replications that have undergone experimental evaluations that have been released to the public. First, the original evaluations that the OAH reviewed to identify program models as "evidence-based" are described. "High" quality ratings are assigned to random assignment studies with attrition rates that were not considered problematic.[34] Quasi-experimental studies received "moderate" quality ratings, along with random assignment studies with high attrition. Studies with ratings of "low" quality did not meet either of the high or moderate quality criteria. Second, the results of Tier I replication findings for each evidence-based model are presented.

Becoming A Responsible Teen (BART)

Prior Evaluations. The evidence-based classification used by the OAH for the BART model is based on five evaluations of "low" to "high" in scientific rigor that have inconsistent findings of success.[35] The first evaluation, published in 1995, received a "high" quality rating by the OAH for its random assignment design.[36] The small-scale evaluation assessed the effect of the eight-week education and behavioral skills program implemented in an after-school community-based setting that served black youth with an average age of 15.3 in an undisclosed Southern city with 400,000 residents.[37] Self-reported sexual behavior was assessed during six-month and 12-month follow-ups. Averaged over the entire length of the follow-up period, the treatment group reported lower incidents of unprotected oral sex and anal intercourse, and higher incidents of condom-protected intercourse, than the control group.[38]

Similar to the 1995 study, the 1999 study received a "high" rating for scientific rigor.[39] This evaluation attempted to assess the effectiveness of BART applied to incarcerated males in a state reformatory in the Southern state.[40] A total of 428 young men entering a juvenile correctional facility were randomly assigned to an intervention and control group. At the six-month follow-up after release from the facility, members of the intervention group fared no better or worse on all sexual outcomes assessed.[41]

A 2002 study failed to use a control group and only assessed the before-and-after participation effect of BART on attitude and knowledge of the risks of sexual activity.[42] Correctly, the OAH gave this study a "low" ranking for scientific rigor because of its weak scientific methodology and failure to assess behavioral changes.[43] Similarly, the OAH classified a 2009 study as not meeting their review criteria.[44] While this particular study used random assignment, the evaluators did not assess any outcomes related to actual changes in sexual behavior.[45]

Last, a 2011 random assignment study, classified with a "high" ranking, assessed the effectiveness of BART when applied to incarcerated female youths.[46] With an average follow-up of nine months post-release, the small-scale experiment found that the program had no effect on contraceptive use, the frequency of sexual intercourse while under the influence of alcohol or drugs, or acquiring sexually transmitted infections or HIV.[47]

Thus, the majority (two out of three) of random assignment evaluations with high rankings of scientific rigor found BART to be ineffective. Despite more consistent evidence of failure than success, the OAH misleadingly labeled BART as an "evidence-based" model.

TPP Tier I Replication. The OAH funded an evaluation that attempted to replicate the inconsistent impacts of BART in a different setting.[48] Performed by The Policy & Research Group, the evaluators randomly assigned 850 minority teens, ages 14 to 18, participating in a summer youth program in New Orleans, to a control group and an intervention group.[49] As seen in Table 1, this replication failed to have an impact on both of the measured sexual behavior outcomes: At the six-month follow-up, BART had no effect on the inconsistency of condom use or the frequency of sex.[50]

Despite the ineffective replication, the evaluators report that their program "appears to have been implemented with reasonable fidelity" to the BART model.[51] Thus, poor implementation of the model cannot be used as an excuse for the replication's ineffectiveness.

The authors appropriately acknowledge that they were attempting to replicate the beneficial impact of a single study published over 20 years ago that may be no longer relevant to today's youth.[52] They elaborate that "[i]t is conceivable that any historical change in adolescents' social, normative, educational, and informational environments now as compared with then could help explain differences in findings."[53] Not only was the BART replication based on high-quality evidence that found more failure than success, but the grant award was based on an outdated study.

Children's Aid Society (CAS)–Carrera Program

Prior Evaluation. The OAH classified the Carrera program as "evidenced-based" based on a single-site "high-quality" randomized experiment published in 2002.[54] The 2002 study assessed the effectiveness of a three-year multifaceted intervention that served primarily black and Hispanic teens ages 13 to 15. The multifaceted intervention included job-related training, academic assistance, sex education, art instruction, sports activities, and mental health and health care services.[55] At the time of the three-year follow-up, 484 intervention and control group members were assessed on several sexual and reproduction outcomes.

Overall, the intervention and control group members had self-reported rates of 63 percent and 72 percent for ever having had sex, respectively—a statistically significant difference of 9 percent.[56] However, this beneficial effect was primarily the result of females reporting statistically lower occurrences of ever having had sex, compared to no effect for males.

Generally, the program failed to affect reported use of condoms and hormonal methods during most recent intercourse.[57] This inconclusive finding was the result of the different responses by gender. For females, the intervention and control group members had self-reported rates of using condoms and hormonal methods of 36 percent and 20 percent, respectively—a statistically significant difference of 16 percent. However, the program had a harmful impact for males, with reported outcomes of 9 percent and 20 percent for male intervention and control group members, respectively—a statistically significant harmful impact of 11 percent. When the usage of only condoms was assessed, the intervention failed to affect reported use, even when the outcomes were reported by gender.

The intervention reduced self-reported incidences of becoming pregnant or causing a pregnancy with reported rates of 10 percent and 17 percent for the intervention and control groups, respectively.[58] The impact was a statistically significant difference of 7 percent. However, this effect was driven entirely by the impact on females. Similarly, the intervention had no overall impact on reports of giving birth or becoming a father. However, when the sample itself was limited to females, 3 percent of the intervention group reported giving birth, compared to 10 percent for the control group—a statistically significant difference of 7 percent. The program had no effect on male self-reports of becoming fathers.

TPP Tier I Replication. The OAH funded two evaluations of the Carrera program that have been released to the public—an experimental evaluation and a quasi-experimental evaluation. The experimental evaluation assessed the impact of the program using the random assignment of 600 students ages 13 to 15 from three schools in the Englewood neighborhood of Chicago.[59] The Chicago replication was implemented over four years.

There was no evidence that the replication affected any of the measures of sexual activity after four years of programing.[60] As detailed in Table 1, this replication produced no statistically significant effects on any of the six sexual behavior outcomes. Specifically, the random assignment replication failed to have statistically measurable effects on self-reports of ever having sex and sexual intercourse without contraception.[61] Further, the replication failed to affect any of these outcomes when analyzed by gender.[62]

The evaluators report that their program "was not delivered with fidelity, due in large part to the instability of the Chicago Public School (CPS) system."[63] For example, only 12 percent of the intervention group attended at least 75 percent of the scheduled sessions. This issue may reflect the failure of the program administrators in getting students interested in participating in the provided services.

The less scientifically rigorous quasi-experimental evaluation attempted to replicate the Carrera program in rural, urban, and "micropolitan" (popu-

lation of at least 10,000 and less than 50,000) communities in Georgia.[64] Due to the quasi-experimental design, this study is not summarized in Table 1. The intervention group consisted of youth participating in the Carrera services provided by three community-based organizations, while the comparison group consisted of youth participating in three Boys and Girls Clubs. Each of the rural, urban, and micropolitan locations were represented with an intervention and comparison group site. The initial sample size was 400 adolescents, but dwindled to 204 by the time of the three-year follow-up.

The evaluators reported that the "Carrera Model was implemented with fidelity and quality, particularly with program components and staffing: however, attendance was a challenge."[65] Over the course of the three-year intervention, intervention group members increasingly dropped out of the program.

The intervention and control group members were assessed over the course of three years on measures of ever having had sex, and sex without a condom or other birth control. In each of these annual assessments, this replication had no statistically measurable impacts on the outcomes.[66] Further, no effect occurred when these outcomes were assessed by gender in year three.

¡Cuídate!

Prior Evaluation. The OAH categorizes the ¡Cuídate! program as an evidenced-based model due to a single "high quality" randomized experiment published in 2006.[67] The OAH also reviewed two other studies that did not meet its criteria for an evidence-based classification because program impacts were not assessed.[68]

For the highly rated study, 553 Hispanic adolescents in Philadelphia with an average age of nearly 15 were randomly assigned to the ¡Cuídate! program—an HIV prevention program that is an adaption of Be Proud! Be Responsible!—and to a health promotion program that served as the control group.[69] Over the three-month, six-month, and 12-month follow-ups, the evaluation found that participation in ¡Cuídate! was associated with declines in self-reported sexual intercourse and number of sexual partners, and an increase in consistent use of condoms.[70] However, there was no effect for the outcomes of condom use at last time of sex and the proportion of days of unprotected sex.[71] Thus, ¡Cuídate! had beneficial effects on only four of seven outcomes.

TPP Tier I Replication. In an attempt at replication, the OAH funded a large-scale multi-site replication of ¡Cuídate! in a small city in southern California, in Phoenix, and in Boston.[72] This replication attempt is crucial to the potential of evidence-based policymaking because the "study was designed to address important research and policy questions about the effectiveness of an evidence-based program taken to scale and replicated with different populations and in different settings."[73] The rigorous evaluation randomly allocated 2,169 adolescents, primarily Hispanic, to the intervention and control groups.[74] Outcomes were assessed at the six-month follow-up.

For the entire sample, nine outcomes were assessed.[75] As summarized in Table 1, ¡Cuídate! had no statistically meaningful effect on any of the outcomes. For sexual behavior, the program had no effect on ever being sexually active, sexually active within the past 90 days, sexual intercourse in the past 90 days, oral sex in the past 90 days, or anal sex in the past 90 days. For sexual risk within past 90 days, the program had no effect on sexual intercourse without birth control, sexual intercourse without a condom, oral sex without a condom, or anal sex without a condom.

When effectiveness was assessed by subgroups, several harmful effects were found. For teens who were sexually active at the beginning of the study, intervention group members were 7 percentage points *more likely* to report having recently had sexual intercourse than similar teens in the control group.[76] White teens participating in the Hispanic-focused program were about 9 percentage points more likely to report having recently had oral sex and oral sex without a condom, than similar teens in the control group. For Hispanic and black teens, the program had no effect on all outcomes.

According to the authors, "Each of the grantees successfully delivered the program with fidelity (adherence to its core elements and without modifications that threatened those core elements)."[77] Thus, the failure of this replication cannot be blamed on a lack of implementation fidelity. Further, this replication provides more evidence that scaling up "evidence-based" models is unlikely to produce successful results.

It's Your Game: Keep It Real

Prior Evaluation. Initially, the OAH categorized the It's Your Game: Keep It Real (IYG) program as an evidenced-based model based on a few "modera-

te quality" randomized experiments.[78] A 2010 study randomly assigned 10 middle schools from a large urban school district in Texas to intervention and control conditions.[79] The study suffered from high attrition, so the OAH gave the study a "moderate quality" rating.[80] The IYG curriculum consists of multiple group-based classroom lessons during the seventh and eighth grades.

The sexual activities of students were assessed during the ninth grade. Students attending the intervention schools were statistically less likely to report initiating sex as well as engaging in oral or anal sex.[81] For example, 23.4 percent and 29.9 percent of the students attending the intervention and control school, respectively, reported initiating sexual activities by the ninth grade. After adjusting for the background characteristics of the students, members of the control group were 29 percent more likely to initiate sexual activities than their peers in the intervention group. However, IYG had no statistically measurable effect on participants engaging in vaginal sex. Overall, 22.3 percent and 26.9 percent of the intervention and control groups self-reported engaging in vaginal sex, respectively—a statistically insignificant difference.

Another pair of random-assignment evaluations of IYG published in 2012 and 2014 were classified as "moderate" in scientific rigor based on high attrition problems.[82] The 2012 study assessed the effectiveness of IYG in 15 urban middle schools.[83] More than 1,200 predominately minority seventh-grade students were followed until the ninth grade. The 15 schools were randomly assigned to a risk-avoidance (RA) program that fulfilled federal abstinence education guidelines, a risk-reduction (RR) program that stressed abstinence along with condom usage for those deciding against abstinence (abstinence-plus), and a control group.

When the RA group was compared to their peers in the control group, RA had no effect on self-reports of any sexual initiation, oral sex, vaginal sex, or anal sex.[84] However, the RA students were 30 percent less likely to engage in unprotected vaginal sex than their peers in the control group. On the contrary, the RA students were 69 percent *more likely* to have two or more vaginal sex partners than one or no vaginal sex partners, than their peers in the control group.

When the RR group was compared to the peer control group, RR had mixed effects on self-reports of any sexual initiation, oral sex, vaginal sex, or anal sex.[85] For the initiation of any sexual activity, students in the RR group were 35 percent less likely to engage in such activities than their peers in the control group. While RR had no effect on the likelihood of engaging in oral and anal sex, the program was associated with a 36 percent decrease in the likelihood of having vaginal sex. Further, students with access to RR instruction were 33 percent less likely to engage in unprotected vaginal sex than members of the control group.

While the 2012 study had problems with attrition, the authors also warned that "baseline imbalances in demographics and prevalence of sexual behavior between study conditions may have biased outcomes away from the null hypothesis."[86] For example, students in the control schools had higher rates of previously engaging in sexual activity than students in the RA and RR schools.[87] Thus, the underlying biases in the study may cause the effects—beneficial and harmful—to be overstated. This bias may be the result of cluster randomization used by the evaluators.

In a follow-up to the 2012 study, the 2014 study updates the findings for the 10th grade.[88] Again, the 2014 study suffers from the same attrition and selection bias that afflicted the 2012 study. By the 10th grade, students in the RA and RR schools were just as likely to report engaging in any sexual activity, oral sex, or vaginal sex.[89] However, students in the RA and RR schools were 36 percent and 35 percent less likely to report engaging in anal sex than their peers in the control schools, respectively. Students in the RA schools were 39 percent less likely to have unprotected vaginal intercourse, while there was no effect for students in the RR schools. In contrast, the RA and RR groups were 180 percent and 114 percent more likely to have two or more vaginal sex partners than one or no vaginal sex partners, respectively, compared to their peers in the control group.

TPP Tier I Replication. In an attempt at replication, the OAH funded two large-scale replications of IYG in two locations in South Carolina and Texas.[90] Published in 2016, the South Carolina multi-site study randomly assigned 24 rural middle schools across the state, representing 3,143 students, to provide IYG services or the usual non-evidence-based sex education programming. Except for age, students in the IYG and control schools did not statistically differ in baseline characteristics.[91] On average, students in the control schools were 0.1 years older.

To assess the effectiveness of IYG, the self-reported behavioral outcomes were assessed in the eighth

and ninth grades. As presented in Table 1, the evaluation found that IYG had one harmful impact and four statistically insignificant impacts. Students in the IYG schools were no more, and no less, likely to initiate vaginal intercourse by the end of the eighth grade than students in the control schools.[92] By the end of the ninth grade, however, the students in the IYG schools were 27 percent more likely to engage in vaginal intercourse than similar peers in the control schools. This harmful impact, when translated into an effect size (Cohen's *d*) is 0.10, which is extremely small.[93] Interpreting this harmful effect, the authors write that the "usual programming outperformed IYG, although the magnitude of the difference was small."[94] Additionally, within the last three months at the time of the ninth-grade follow-up, the IYG replication failed to affect incidences of vaginal intercourse, sex without effective birth control, and sex without the use of condoms.[95]

Could a failure in faithfully implementing the IYG model have led to the replication's failure? The authors do not seem to think so: "Fidelity and quality of implementation by IYG facilitators was high, as was students' exposure to the curriculum."[96]

The authors raise two important issues that may explain why scaling up and replication may not work. First, the South Carolina replication "was an effectiveness trial that used classroom teachers for implementation rather than an efficacy trial more tightly controlled by the original researchers; existing literature suggested effectiveness trials often yield smaller effects than efficacy trials."[97] Efficacy trials test whether a social program is effective under optimal conditions, while effectiveness trials test the effectiveness of social programs delivered in real-world conditions.[98] Second, the authors acknowledge that replicating supposedly effective models in different settings and with dissimilar demographic groups does not mean that the same results should be expected.

The single-site replication of the IYG in Houston evaluation randomly assigned 10 middle schools to the IYG group, and 10 middle schools to the control group.[99] The control group schools implemented regular school-based health education programming. The baseline sample consisted of 2,403 students.[100] The final sample of students for assessing program impact was limited to students who reported having had no vaginal or oral sex at baseline.[101]

Students in the IYG and control schools did not statistically differ in demographic characteristics.[102] However, there was an important difference between the intervention and control groups at baseline. The school-level rate of seventh-graders reporting ever having had sex was 12.14 percent in the IYG schools, and 7.02 percent in the control schools—a statistically significant difference of 5.12 percent.[103] This difference may reflect different cultures and underlying characteristics in the schools that may bias the impact estimates. This bias is another reason why cluster randomization does not have the same scientific rigor as individual randomization.

Despite being implemented in an urban setting like the original evaluations of IYG, the Houston replication failed to produce any impacts on three sexual behavior outcomes during the follow-up in the ninth grade.[104] (See Table 1.) Students in the IYG schools did not differ on self-reported measures of the initiation of vaginal or oral sex.

Similar to the authors of the South Carolina replication, the authors of the Houston replication offer the use of school teachers for IYG curriculum instruction, instead of outside experts, as a possible explanation for the replication's failure.[105] Program models are less likely to succeed when implemented under real-world conditions. Further, the authors did not provide evidence that the IYG model was poorly implemented in Houston.

Promoting Health Among Teens! Abstinence-Only Intervention

Prior Evaluation. Based on small, single-site "high quality" randomized evaluation, the OAH classified Promoting Health Among Teens! Abstinence-Only Intervention as an "evidence-based" model.[106] In all, 662 black sixth-grade and seventh-grade students from four public middle schools in a city in the Northeast were randomly assigned to five groups that received different educational services:

- Abstinence-only intervention;
- Safer sex-only intervention;
- Comprehensive intervention (short duration);
- Comprehensive intervention (long duration); or
- Health-promotion control intervention.

The abstinence-only intervention offered participants eight hours of instruction on the risks of sexual activity and benefits of abstinence, while the safer sex-only intervention offered similar instruction on the risks of sexual activity, but differed from the abstinence-only instruction by encouraging the use of condoms.[107] The comprehensive interventions offered eight hours and 12 hours of instruction on the risk of sexual activity and encouraged abstinence. However, this intervention offered instruction on condom usage to students deciding to have sex.[108] The health-promotion control intervention "focused on behaviors associated with risk of heart disease, hypertension, stroke, diabetes, and certain cancers. It was designed to increase knowledge and motivation regarding healthful dietary practices, aerobic exercise, and breast and testicular self-examination, and to discourage cigarette smoking."[109]

Outcomes were assessed over a 24-month period. Students in the abstinence-only intervention had a 33.5 percent probability of ever having sexual intercourse, compared to 48.5 percent for similar peers in the control group.[110] The risk ratio for this effect is 0.67, which means that members of the abstinence-only intervention group were 33 percent less likely to engage in sexual intercourse, compared to similar students in the control group.[111] Members of the abstinence-only intervention were also slightly less likely to engage in sexual intercourse within the last three months.[112] The risk ratio for this outcome was 0.94, which translates into a decrease of 6 percent. As for the other comparisons, the "safer sex and comprehensive interventions did not differ from the control group in sexual initiation."[113]

The authors of the evaluation add context to the findings by cautioning that the "results of this trial should not be taken to mean that all abstinence-only interventions are efficacious."[114] Further, and perhaps most important for federal policy, "[t]his trial tested a theory-based abstinence-only intervention that would not have met federal criteria for abstinence programs."[115]

TPP Tier I Replication. The OAH awarded a grant to replicate the Promoting Health Among Teens! Abstinence-Only Intervention in Yonkers, New York.[116] The single-site evaluation randomly assigned more than 1,300 sixth-grade and seventh-grade students to the intervention and control groups in eight middle schools in sections of the city with the highest occurrences of births to teens. Members of the control group were offered the Promoting Health Among Teens! Health Intervention that offered educational programming regarding the benefits of exercise and healthy eating habits.

Baseline characteristics of the treatment and control group did not differ at the time of the 12-month follow-up.[117] As summarized in Table 1, the replication failed to affect all three of the sexual behavior outcomes. The intervention failed to yield statistically significant results on the self-reported outcome of ever having had sex during the three-month, six-month, and 12-month follow-ups.[118] For example, 1.3 percent and 2.1 percent of the intervention and control group self-reported ever having had sex, respectively, at the 12-month follow-up—a statistically insignificant difference of 0.8 percent.[119]

The failure of the replication cannot be blamed on poor implementation because "the results of this evaluation also suggest that implementation fidelity is a necessary but not sufficient condition for attaining successful replication. This replication attained a high level of fidelity and yet failed to reproduce the original findings."[120] The authors also caution against the assumption that replicating program models based on outdated studies will produce the same results, Further,

> [i]t is perhaps the case that evidenced-based interventions from a decade or so ago may lose their relevancy in more contemporary times. Human behavior is dynamic and subject to broader changes and influences from a myriad of sources. Thus, when consideration is being given to testing the effectiveness of an intervention where there has been some time lag, situating that intervention in the present reality and adapting it to meet this reality may be one of the decisions potential implementers need to make.[121]

Reducing the Risk

Prior Evaluation. Based on several experimental and quasi-experimental studies rated as "moderate quality" to "low quality" in scientific rigor, the OAH classified the Reducing the Risk program as an evidence-based model.[122] All of these studies earned rankings lower than "high quality" due to the methodological shortcomings of these studies, so the results need to be interpreted with great skepticism.[123] Further, two studies with moderate ratings did not provide consistent evidence of effectiveness.[124]

A 2008 cluster random assignment study of Reducing the Risk that suffered from high attrition study was classified by the OAH as having no impact on sexual outcomes.[125] The study was rated as "moderate quality" in scientific rigor by the OAH. The evaluators randomly assigned 17 schools, consisting of 1,944 students, to three curricula: Reducing the Risk, modified version of Reducing the Risk, and the standard curriculum. The schools were located in Cleveland, Ohio, and Louisville, Kentucky. Reducing the Risk is a curriculum

> designed to enhance students' skills to resist unprotected sex by modeling those skills and then providing students opportunities for practice. The curriculum emphasizes that youth should avoid unprotected intercourse; that the best way to do this is to abstain from sex; and that if they do not abstain from sex, they should use contraceptives (especially condoms) to guard against pregnancy and STDs, especially HIV.[126]

The modified version of Reducing the Risk was "specifically designed for high sensation-seeking and impulsive students" and the standard curriculum offered in the schools served as the control.[127] All of the services provided had the goal of preventing pregnancy and HIV. Students were assessed from the beginning of the ninth grade and the end of the 10th grade.

Attrition rates at the three-month and six-month follow-ups were statistically different for the three groups.[128] Further, students who reported being more sexually experienced were less likely to complete the follow-up surveys. Only 52 percent of the original sample completed the 12-month and 18-month follow-up surveys. Not only did the 2008 evaluation suffer from attrition, but the three groups of students were not statistically equivalent on gender, race, or educational aspirations.[129] This problem means that the individual students are not equivalent on these factors—a problem not uncommon with cluster randomization.

Overall, participation in Reducing the Risk or the modified version had no effect on initiating sexual intercourse, compared to students in the standard curriculum.[130] However, when the samples of both Reducing the Risk interventions are combined, students that received the standard curriculum were less likely to engage in sexual intercourse.

A quasi-experiment that resulted in two publications published in 1991 and 1992 was rated as "moderate quality" in scientific rigor by the OAH.[131] The 1991 study tried to assess the effect of Reducing the Risk by non-randomly allocating more than 1,000 high school students from 13 California schools to intervention and comparison groups.[132] Only 758 students responded to the 18-month follow-up survey.

Questionably, the OAH assigned the 1991 study a "moderate quality" scientific-rigor rating, even though the quasi-experiment only tested the statistical differences in outcomes between the intervention and comparison groups without controlling for any variables that could influence the outcomes.[133] At the six-month follow-up, the difference between the self-reported initiation of intercourse for the intervention and comparison groups did not differ.[134] At the 18-month follow-up, the intervention group had a statistically significant lower rate of self-reported intercourse.

Interestingly, the authors performed a logistic regression which would presumably control for some factors that could influence self-reported outcomes.[135] The presumably more rigorous logistic regression found that Reducing the Risk failed to affect the initiation of sexual intercourse.

When the outcome of unprotected intercourse was estimated for all of the students in the study, Reducing the Risk failed to affect this outcome.[136] Further, Reducing the Risk had no effect on whether female students reported becoming pregnant or male students reported getting a girl pregnant.

The 1992 study of the same sample of students reports findings only from the six-month follow-up for the same evaluation.[137] It used the same weak methodology of the 1991 study. The 1992 study found no differences in rates of sexual intercourse and pregnancy between the intervention and comparison groups at the six-month follow-up.[138]

After Reducing the Risk was designated an "evidence-based" model by the OAH, a multi-site evaluation of over 700 adolescents drawn from high schools and community youth groups was published in 2014.[139] The study was originally intended to use random assignment to assess the impact of Reducing the Risk and a revised Reducing the Risk curriculum (RTR+) in three states (Arizona, New York, and Texas). After the initial random assignment, however, the evaluators non-randomly reassigned some of the sample to intervention and control groups, so the

study is not a true randomized experiment.[140] The OAH classified the study as moderate in scientific rigor and concluded that the program has no effect on relevant outcomes.[141]

Self-reported sexual activities of the sample were followed up at three months, six months, and 12 months.[142] There were no differences between the control group and the Reducing the Risk group in the likelihood of sexual initiation during all three of the follow-ups.[143] The pattern of ineffectiveness was almost similar for the revised Reducing the Risk curriculum. The revised curriculum had no effect on sexual initiation for the first two follow-up periods, while members of this intervention group were less likely to engage in sexual activity at the 12-month follow-up.

The number of self-reported sexual partners and number of unprotected sexual acts were also assessed. The regular Reducing the Risk curriculum did not have statistically meaningful effects on either outcome.[144] While the revised Reducing the Risk curriculum was associated with a decrease in the number of sexual partners, the intervention had no effect on unprotected sex acts. Thus, both of these "evidence-based" interventions failed to affect the majority of outcomes.

TPP Tier I Replication. The OAH awarded grants to fund two replications of Reducing the Risk.[145] The first single-site replication study used cluster randomization to assess the impact of Reducing the Risk and another intervention, Love Notes, in Louisville, Kentucky.[146] At the time of the award for this Tier I grants, Love Notes was not classified as an evidence-based model. According to the authors, Love Notes "embeds pregnancy and disease prevention messages in a curriculum that emphasizes the importance of forming healthy relationships and avoiding intimate partner control or violence for individuals to reach their life goals."[147] The control curriculum was The Power of We (POW) curriculum—a program for teaching adolescents to "learn more about assets in their neighborhoods and ways to bring about positive change."[148] However, "POW did not include any mention of individual planning, self-esteem, sexual health, healthy relationships, or intimate partner violence, and thus had zero overlap with content in either" Reduce the Risk or in Love Notes.[149] The interventions implemented in Louisville were performed by highly trained academics, so this replication can be considered an efficacy evaluation as the programs were implemented under optimal conditions.

Students ages 14 to 19 who were thought to be of high risk for pregnancy and were participating in a community-based organization were recruited for participation in the study.[150] Once the teens were randomly assigned to clusters, the clusters were randomly assigned to three conditions.[151]At baseline, 1,365 teens were involved in the evaluation. Because the technique does not randomly assign individuals, cluster randomization may not yield equivalent groups.[152] Members of the Reducing the Risk and Love Notes groups were slightly more likely to be non-Hispanic blacks than members of the control group.[153]

Of the eight sexual outcomes measured, Reducing the Risk had no effect on half, while the intervention had beneficial impacts on the other half. (See Table 1.) At the time of the three-month follow-up, Reducing the Risk had no effect on two of the four outcomes assessed.[154] Compared to control group teens, those in the Reducing the Risk group were no more or less likely to report having sex without a condom, or ever having sex. However, teens in this intervention group were less likely to report having sex without any type of birth control and had fewer reports of several sexual partners.

The results for Reducing the Risk at the six-month follow-up are similar.[155] The intervention had no impact on condom usage and ever having sex, while the program was associated with decreased self-reports of having sex without any form of birth control and the number of sexual partners.

At the time of the three-month follow-up, Love Notes had no effect on any of the four outcomes assessed.[156] Compared to the control group, those in the Loves Notes group were no more or less likely to use condoms, use any form of birth control, or have sex. The number of sex partners of this intervention group was not statistically different from what was reported by the control group. By the time of the six-month follow-up, however, the results for Love Notes changed completely—the intervention was associated with beneficial outcomes on all four measures.[157]

In regards to Love Notes, the evaluators caution that a "replication of these results is needed to increase the strength of the evidence for the intervention."[158] For the OAH and others to label Love Notes an "evidence-based" model would be premature because the results are based on a single-site evaluation that was not implemented under real-world conditions.

A more relevant evaluation for policymakers is the large-scale, multi-site replication of Reducing the Risk in six schools in St. Louis, Missouri, five schools in Austin, Texas, and six schools in San Diego, California.[159] Like the other replication evaluation, the evaluators of this study randomly assigned classes to intervention and control groups. The control classrooms received the "business as usual" curriculum. In all three sites, the intervention was implemented in public school classrooms that ranged from the eighth to tenth grades. At the start of the study, 3,314 students in 150 classrooms participated in the study.[160] At the time of the 12-month follow-up, 2,689 (81 percent) of the original sample completed the self-reported survey.

According to the evaluators, the intervention "was well implemented across the 3 replication sites" and the "program was delivered with fidelity."[161] Despite the successful implementation of the intervention, members of the intervention classrooms in the three sites did not differ on seven outcomes of sexual behavior and risk at the 12-month follow-up, in contrast to similar members in the control classrooms.[162] For sexual behavior, Reducing the Risk failed to affect being "ever sexually active," "currently sexually active," having "sexual intercourse," and having "oral sex." Further, the intervention had no effect on sexual intercourse without any birth control, sexual intercourse without a condom, or oral sex without a condom.

When the results were analyzed by the three sites, Reducing the Risk failed to have any impact on all seven measures in the Austin and San Diego sites.[163] In the St. Louis site, the intervention failed to affect six of the seven outcomes. The only measure that had a statistically significant effect was the self-reported decrease in engaging in sexual intercourse.

The findings of this replication provide caution for expecting similar results of "evidence-based" models taken to scale. As the authors acknowledge, "As an examination of the effectiveness of evidence-based programs and what happens when they are taken to scale, replicated with different populations, and offered in different settings, this study provides important information on the effectiveness of *Reducing the Risk*."[164] Further, the "evidence for the effectiveness of this program is from a single quasi-experimental study completed 25 years ago in rural and urban areas of northern California with primarily White high school students."[165] Thus, what worked in one setting did not work in other settings.

Safer Sex Intervention

Prior Evaluation. The OAH assigned the Safer Sex Intervention (SSI) model an evidenced-based classification based on a single-site "moderate quality" randomized experiment published in 2001.[166] Suffering from high attrition, the 2001 study randomly assigned 60 and 63 youth to the intervention and control group, respectively. The sample of sexually active female participants were less than 24 years old, and were either attending a hospital-based clinic for treatment for cervicitis or were admitted to a hospital for management of pelvic inflammatory disease. During patient visitations, participants were asked about their sexual activities in one-month, six-month, and 12-month follow-ups. Only 33 percent of participants attended all follow-up visits.

Dealing with a population already infected with a sexually transmitted disease (STD), the SSI curriculum imparted information on how to change sexual behavior to reduce risks that also involved individualized sessions tailored to the participants.[167] Members of the control group received standard STD education.

- The following seven outcomes were assessed during each of the three follow-up visits:
- Condom usage with last sexual encounter;
- Currently have a main sexual partner;
- Frequency of condom use with main partner in last five sexual encounters;
- Consistent use of condoms ("Every time") with main partner;
- Another partner in the last six months;
- Frequency of condom use with another partner in last five sexual encounters; and
- Consistent use of condoms ("Every time") with another partner.[168]

At the one-month follow-up, SSI failed to affect all seven of the outcome measures.[169] At the six-month follow-up, participation in the intervention had no statistically measurable effect on six of seven outcomes. Members of the intervention group were less likely to

report sexual partners in addition to their main partner than similar members in the control group. The intervention failed to have any statistically significant effect on all seven outcomes at the 12-month follow-up. Thus, of a total of 21 outcomes, the program had only one (4.8 percent) statistically significant outcome.

TPP Tier I Replication. The OAH awarded grants to fund two replications of SSI.[170] The first replication was a small single-site evaluation that assessed the impact of SSI implemented in New Orleans, Louisiana.[171] Girls ages 14 to 19 were referred by clinicians, and clinic staff were asked to participate in the study. Individuals were randomly assigned to intervention (SSI) and control groups. The results are based on 268 participants with 133 in the SSI group and 135 in the control group.

For all three sexual behavior outcomes, the intervention failed to have statistically measurable impacts. (See Table 1.) According to the evaluators, the primary outcome for judging the effectiveness of SSI was the inconsistency of condom usage at the six-month follow-up.[172] According to the author, the "Safer Sex intervention had no significant effect on participants' inconsistency of condom use."[173] Fifty percent of the SSI group reported inconsistent use, compared to 46 percent for the control group—a statistically insignificant difference of 4 percent.[174]

The same pattern of ineffectiveness occurred with the secondary outcomes. SSI had no statistically meaningful impact on the inconsistency of contraceptive use and the frequency of sex.[175] Thus, SSI, as implemented in New Orleans, failed to affect all three outcome measures.

The second replication evaluation used random assignment to assess the impact of SSI in multiple sites.[176] More than 1,200 female adolescents attending 38 clinics in Minnesota, Tennessee, and Florida were randomly assigned to SSI and a control group. Control group members received the standard, less-intensive care provided by the clinics. Thus, this replication attempted to scale-up SSI.

As summarized in Table 1, the replication failed to affect nine of 10 sexual behavior outcomes. At the nine-month follow-up, 86 percent of the study participants completed self-reported surveys.[177] According to the evaluators, the main indicators of effectiveness was sexual activity in the past 90 days, and sexual intercourse without birth control in the past 90 days.[178] SSI failed to affect whether participants were sexually active, but did decrease self-reported sexual intercourse without birth control.[179] For this measure, 22.05 percent of the SSI and 27.82 percent of the control group reported having sexual intercourse without using birth control—a statistically significant difference. According to the evaluators

> SSI had no impact on any other measures of sexual activity. The program was not effective in reducing sexual intercourse, oral sex, or anal sex in the past 90 days. It did not affect rates of condom use during sexual intercourse, oral sex, or anal sex, nor did it affect the likelihood of having sexual intercourse with more than 1 partner or more than 5 partners in one's lifetime.[180]

Seventeen Days

Prior Evaluation. Similar to other classifications, the OAH assigned the Seventeen Days (formerly What Could You Do?) model an evidenced-based classification based on a single-site "high quality" randomized experiment published in 2004.[181] Seventeen Days employs an interactive video intervention that attempts to increase the aptitude of participants in making less-risky sexual decisions. In the area of Pittsburgh, Pennsylvania, 300 urban adolescent girls were randomly assigned to the Seventeen Days intervention group and two control groups. The first control group used books to offer the same information that Seventeen Days delivered in interactive videos. The second control group was provided commercially available brochures covering the same topics.

The outcomes consisted of self-reported questions regarding sexual behavior and the acquisition of STDs.[182] More important, medical tests for chlamydia trachomatis were administered. In the realm of sex education, outcomes are almost exclusively based on self-reported data that can be susceptible to false or misleading answers. For this reason, the use of a medical test generates more reliable data than self-reported measures. This is an important advancement in the evaluation literature.

Despite using random assignment, the predispositions of the intervention group were statistically different from members of the control groups on a key factor that may have substantially affected the outcomes.[183] Woman assigned to the Seventeen Days group were more likely to be sexually abstinent than members of both control groups. Unsurprisingly, the intervention group was more likely to report being abstinent during the three-month and six-month follow-ups.[184]

Participation in Seventeen Days had no effect on condom use at the three-month and six-month follow-ups.[185] However, those in the intervention group reported fewer condom failures than their counterparts in the control groups. As for STDs, the intervention group self-reported lower rates of acquiring any type of STD. Except for chlamydia, the number of participants reporting specific types of other STD infections were too small to conduct valid statistical tests.

According to self-reports, members of the Seventeen Days group were significantly less likely to have chlamydia than their counterparts. However, this result demonstrates the unreliability of self-reported data, because the results of the chlamydia medical test found that Seventeen Days failed to yield statistically significant results. Thus, the results of self-reported outcomes need to be taken with a healthy dose of skepticism.

TPP Tier I Replication. The OAH attempted to replicate and scale up Seventeen Days in Ohio, Pennsylvania, and West Virginia.[186] The large-scale, multi-site evaluation randomly assigned more than 1,300 sexually active girls ages 14 to 19 who attended 20 clinics in the three states to intervention and control groups.

The samples for the three-month and six-month follow-ups had 52 percent and 43 percent response rates, respectively, suggesting that attrition was a problem.[187] However, the evaluators report that there was no difference in attrition between the intervention and control groups. The resulting samples for the follow-ups did not differ in baseline characteristics, despite attrition.[188]

Of the six sexual behavior outcomes, Seventeen Days failed to affect all. (See Table 1.) The evaluators found "no evidence that viewing *Seventeen Days* [the video] impacted engaging in safe sexual behavior compared to the comparison group."[189] At the three-month follow-up, the intervention failed to have any effect on sexual behaviors or abstinence.[190] At the six-month follow-up, the results were similar. The intervention failed to affect any sexual behavior or abstinence.[191]

Notably, the evaluators did not completely rely on self-reported outcomes. In addition to pregnancy tests, the evaluators tested for chlamydia and gonorrhea. At the time of the six-month follow-up, participation in Seventeen Days failed to effect positive test results for pregnancy and STD infection.[192]

Teen Outreach Program

Prior Evaluation. The OAH assigned the Teen Outreach Program (TOP) model an "evidenced-based" classification based on a single-site, "high quality" randomized experiment published in 1997, and a 2001 quasi-experiment with a "low quality" ranking.[193] The 2001 quasi-experiment failed to establish that members of the intervention and comparison groups were similar enough to ensure that the impact estimates were scientifically valid, so the results are not discussed.[194] As detailed later, the 1997 study had similar flaws.

TOP has the goal of reducing teenage pregnancy, academic failure, and school suspension.[195] For the purposes of TPP, the OAH only considered the teen-pregnancy-related outcomes for assessing whether to classify the model as evidence-based. The main emphasis of TOP "is to engage young people in a high level of structured, volunteer community service that is closely linked to class-room-based discussions of future life options, such as those surrounding future career and relationship decisions."[196]

The 1997 multi-site study assessed the impact of TOP in 25 sites nationwide using a sample of 695 high school students.[197] The majority of the sample was randomly assigned on the level of the individual, however, it was not possible to use individual random assignment for all the sites. In these cases, classrooms were randomly assigned. The sample consisted of students in the ninth to 12th grades. Consequently, the study is not a true individual-level randomized experiment.

Despite the mixed method of random assignment (or perhaps because of it), the intervention and control groups were not equivalent on key factors at the beginning of the study. Members of the control group had statistically higher incidents of prior course failure, school suspension, and pregnancy.[198] Therefore, the classification of this study as "high" in scientific rigor is extremely questionable.

Unfortunately, the impact of TOP over time is unknown, because the evaluators only assessed outcomes at the time of program exit.[199] This means that this evaluation cannot inform policymakers about the effectiveness of TOP after students left the program. The evaluators found that the risks of self-reporting a pregnancy were greatly reduced for the TOP group, compared to the risks of the control group.[200] In fact, the "[r]isk of teen pregnancy was only 41% as large as in the control group."[201] Due

to the intervention and control groups not being equivalent on key factors that likely affected the outcomes, the results of this evaluation are highly suspect.

TPP Tier I Replication. Through two large-scale, multi-site evaluations, the OAH tried to replicate the TOP model in sites throughout the nation.[202] Both evaluations used cluster randomization.

The first evaluation assessed the effect of TOP and consisted of an ethnically diverse sample of middle and high school students drawn from varied locations:

- Hennepin County, Minnesota;
- Northwestern states (Idaho, Montana, Oregon, Washington, and Alaska);
- Kansas City, Missouri;
- Nonmetropolitan counties in Florida; and
- Chicago, Illinois.[203]

In the Northwestern states and Kansas City, the randomization procedure assigned classes to intervention and control groups.[204] In Chicago and Florida, schools were randomly assigned, while in Hennepin County, teachers were assigned to the groups. In Florida, schools were matched in pairs based on similar characteristics, and the pairs were then randomly allocated to the groups. In all, 17,194 students participated in the study.

The outcomes are based on self-reported data.[205] Varying across the geographic sites, the follow-up periods ranged from nine to 24 months.[206] The first outcome assessed whether sexually inexperienced students at baseline ever had sex. The other two outcomes assessed whether the entire sample ever had sex and had sex without contraception.

When the combined impact of the program across all the geographic locations was assessed, the program failed to affect any of the three sexual behavior outcomes. (See Table 1.) When the results were analyzed by geographic locations, the program failed to affect any of the 27 sexual behavior outcomes. In each of the sites, TOP had no effect on sexually inexperienced students ever having sex. Further, TOP failed to affect the self-reported outcomes in each of the geographic locations.[207]

The evaluators concluded that

> [b]ased on data from 5 studies that, together, included more than 17,000 youths in 5 diverse geographic settings, we found little evidence to support the effectiveness of TOP in reducing sexual risk-taking behaviors that should, in turn, reduce adolescent pregnancy. Because most programs identified by the TPP Evidence Review as of 2016 are based on evidence from single studies, the extent to which these programs will be effective in different settings and with different populations over time is a critical question as the evidence base continues to evolve.[208]

The other evaluation assessed the impact of TOP in Louisiana communities and Rochester, New York.[209] Using individual-level random assignment, three cohorts consisting of 2,428 Louisiana and 824 Rochester teens were randomly assigned to intervention and control groups. In both locations, TOP failed to delay sexual onset and having sex without birth control at the three-month follow-up.[210] The cumulative effect of the program implemented in both sites is not presented by the authors. Of the two sexual behavior outcomes for each site, the program failed to affect either of the outcomes. (See Table 1.) The authors conclude that the "results of these 2 community-based trials did not demonstrate that TOP had an immediate impact on sex with no form of effective birth control in the past 3 months, nor did it demonstrate an impact on delay of sexual onset among youths who reported never having had sex at baseline."[211]

Lessons Learned

If the evidence-based-policymaking community is serious about funding what works and defunding what does not work, terminating funding for TPP should be an easy decision. A close review of the scientific literature to identify "evidence-based" TPP program models and their replications reveals several lessons for the evidence-based-policymaking community and policymakers.

Weak Evidence, Weak Replication Findings. Some of the studies used to classify the programs as evidence-based had serious flaws. First, inconclusive evidence was used to label program models, such as BART, as "evidence-based." Since two of three "high quality" experimental evaluations found that BART

was ineffective, the program model should not be labeled as "evidence-based." Additionally, the OAH classified programs as "evidence-based" if they found at least one beneficial outcome. Thus, a program with a single beneficial outcome, and many outcomes of no effect, would be labeled "evidence-based."

Second, some of the studies had intervention and control groups that were not equivalent on key variables that may have affected the outcomes. Cluster randomization should not be considered to have the same methodological rigor as individual-level randomization. For example, the race and ethnicity composition of the intervention and control groups for Reducing the Risk replications in Louisville, Kentucky, were not equivalent.[212] In addition, the intervention group for the Houston IYG replication came from schools that were more likely to have students already engaged in sexual activity by the seventh grade than students attending the control group schools. Given these crucial differences on observable variables, the intervention and control groups are very likely to also differ on critical unobserved factors that can influence the impact estimates. Thus, the results of these studies should be taken with great caution.

Single-Instance Fallacy. Just because an evidence-based program appears to have worked in one location does not mean that the program can be effectively implemented on a larger scale (scaled up), in different locations, or with different populations. Proponents of evidence-based policymaking should not automatically assume that pumping taxpayer dollars into programs attempting to replicate previously "successful" findings will yield the same results. Failure is the norm.

The faulty reasoning that drives such failed expansions of social programs is known as the "single-instance fallacy."[213] This fallacy means believing that a single-site social program that works in one instance will yield the same results when scaled up, or replicated elsewhere. Additionally, programs thought to be effective based on decades of old research may not be relevant today. What worked in the 1980s or 1990s may not work in 2017. The TPP Tier I replications certainly prove this point.

Compounding the effects of this fallacy, one often does not truly know why a certain program worked in the first place. In particular, the dedication and entrepreneurial enthusiasm of a program's founder and the quality of original instructors are difficult to quantify or duplicate. The single-instance fallacy, is perhaps, the most overlooked problem when the evidence-based-policymaking community generalizes the results of the scientific literature.

The OAH's definition that defines a program model as "evidence-based" based on a single evaluation with a single beneficial outcome is faulty. A more meaningful way to deem program models "evidence-based" only occur after they have been found by experimental evaluations to have consistent statistically significant effects that ameliorate the targeted social problem in at least three different settings.[214] Once a program model has been found to produce meaningful results in multiple settings, the likelihood of its successful replication elsewhere should increase significantly.

Implementation Fidelity No Guarantee for Success. For many of the replication evaluations, the program models were well implemented, so lack of implementation fidelity cannot be to blame for the consistent failure of these programs to change sexual behavior outcomes. In many cases, the evaluator and administrators do not know why the program worked. The exact combination of program ingredients, such as intangible qualities of the staff, that lead to success is often unknown.

Efficacy Trials Do Not Show How Programs Perform in the Real World. A further complicating issue that evidenced-based-policy advocates need to address is the difficulty of replicating and scaling up programs based on efficacy trials. Efficacy trials test whether a social program is effective under optimal conditions and implemented by highly trained professionals.[215] These programs are carefully monitored to ensure that the participants receive the intended level of treatment. In the real world, program conditions are often much less than optimal.

On the other hand, effectiveness trials test the effectiveness of social programs delivered in real-world conditions.[216] Under real-world circumstances, staff training and other resource inputs are frequently less than optimal. The distinction between efficacy and effectiveness trials is particularly important when the federal government attempts to replicate and scale up an "evidence-based" model that was deemed effective based on an efficacy trial.

Effectiveness trials provide more valid information about the actual prospects of replicating social programs. For example, the multi-site replication of Reducing the Risk in St. Louis, Austin, and San Diego

tells policymakers more about the real potential of the program model than the efficacy trial implemented in Louisville.

Social Programs Can Cause Harm. Evidence-based-policymaking advocates too frequently concentrate on any beneficial, even if only modest, impacts that have been identified. These same advocates need to recognize that social programs can produce harmful impacts, too.[217] These harmful effects are rarely mentioned by program advocates.

An evaluation used to classify the Carrera program as evidence-based found that males participating in the program were less likely to use condoms than their peers who did not participate in the program.[218] An evaluation used to label IYG as an evidence-based model found that the RA group members were much *more* likely to have multiple vaginal sex partners by the ninth grade than their peers in the control group.[219] By the 10th grade, the RA and RR groups were both more likely to have multiple vaginal sex partners than their peers in the control group.[220]

The Tier I replications also found harmful impacts. The replication of IYG in South Carolina found that students in the IYG schools were more likely to have sexual intercourse than their peers in the control schools.[221] The Tier I replication of ¡Cuídate! also had harmful impacts. Among sexually active teens at the beginning of the study, intervention group members became more sexually active.[222] Further, white teens participating in the ¡Cuídate! were more likely to report having recently had oral sex and oral sex without a condom than similar teens in the control group. Evidence-based-policymaking advocates must not ignore the evidence that social programs sometimes cause harm.

Conclusion

The replications of TPP evidence-based program models demonstrate conclusively that the federal government has a dismal record of replicating social programs thought to be effective. The scientific rigor of the evidence used to identify "evidence-based" teen pregnancy prevention programs funded through TPP Tier I grants is highly flawed. Further, evidence-based-policy advocates mistakenly believe that these grants will be automatically effective because they are replicating previously successful program models.

Overwhelmingly, evaluations of TPP grants replicating "evidence-based" models have been demonstrated to be ineffective. Yet, the evidence-based-policy community is virtually silent on this failure. Clearly, replicating an "evidenced-based" model does not guarantee success. When programs that fail to produce results receive reduced funding or are terminated altogether, and when programs that generate results continue to receive funding, the result is a better allocation of scarce resources. Given the overpowering evidence of TPP ineffectiveness, funding for this program should be terminated.

*—**David B. Muhlhausen, PhD,** is a Research Fellow for Empirical Policy Analysis in the Thomas A. Roe Institute for Economic Policy Studies, of the Institute for Economic Freedom, at The Heritage Foundation.*

Endnotes

1. David B. Muhlhausen, *Do Federal Social Programs Work?* (Santa Barbara, CA: Praeger, 2013).
2. Cynthia Miller et al., "The Challenge of Replicating Success in a Changing World: Final Report on the Center for Employment Training Replication Cites," Manpower Demonstration Research Corporation, September 2005, http://www.mdrc.org/publication/challenge-repeating-success-changing-world (accessed September 2, 2015).
3. Pamela Morris et al., "Impact Findings from the Head Start CARES Demonstration: National Evaluation of Three Approaches to Improving Preschoolers' Social and Emotional Competence," U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation *OPRE Report* No. 2014-44, August 2014, http://www.acf.hhs.gov/programs/opre/resource/impact-findings-from-the-head-start-cares-demonstration-national-evaluation-of-three-approaches-to-improving-preschoolers-social (accessed November 8, 2016), and JoAnn Hsueh et al., "Impacts of Social-Emotional Curricula on Three-Year-Olds: Exploratory Findings from the Head Start CARES Demonstration," U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation, *OPRE Report* No. 2014-78, December 2014, http://www.acf.hhs.gov/programs/opre/resource/exploratory-impacts-of-three-social-emotional-curricula-on-three-year-olds-in-the-head-start-cares-demonstration (accessed November 8, 2016). For a summary of the demonstration, see David B. Muhlhausen, "The Head Start CARES Demonstration: Another Failed Federal Early Childhood Education Program," Heritage Foundation *Backgrounder* No. 3040, August 6, 2015, http://www.heritage.org/research/reports/2015/08/the-head-start-cares-demonstration-another-failed-federal-early-childhood-education-program.
4. Pamela K. Lattimore et al., "Outcome Findings from the HOPE Demonstration Field Experiment: Is Swift, Certain, and Fair an Effective Supervision Strategy?" *Criminology & Public Policy*, Vol. 15, No. 4 (2016), pp. 1103–1141, and Daniel J. O'Connell, John J. Brent, and Christy A. Visher, "Decide Your Time: A Randomized Trial of a Drug Testing and Graduated Sanctions Program for Probationers," *Criminology & Public Policy*, Vol. 15, No. 4 (2016), pp. 1073–1102.
5. Public Law 111–117.
6. U.S. Department of Health and Human Services, Office of Adolescent Health, "Teen Pregnancy Prevention," http://www.hhs.gov/ash/oah/oah-initiatives/tpp_program/about/ (accessed July 22, 2016).
7. Ibid.
8. Calculation based on data provided in Carmen Solomon-Fears, "Teenage Pregnancy Prevention: Statistics and Programs," Congressional Research Service *Report* RS20301, October 26, 2016, X, Appendix A, pp. 24 and 25.
9. Results for America, "Federal Evidence-Based Innovation Programs," Invest in What Works Fact Sheet, October 21, 2015, http://results4america.org/tools/invest-works-fact-sheet-federal-evidence-based-innovation-programs (accessed January 4, 2017).
10. Christine Dehlendorf, "Successful Teen Pregnancy Prevention Program Threatened by Funding Cuts," STAT, April 20, 2017, https://www.statnews.com/2017/04/20/successful-teen-pregnancy-prevention-program-threatened-funding-cuts/ (accessed May 25, 2017).
11. Robert Gordon, "What Happened to Teen Pregnancy Prevention? A Trump Budget Mystery," *Democracy: A Journal of Ideas*, May 25, 2017, http://democracyjournal.org/briefing-book/what-happened-to-teen-pregnancy-prevention/ (accessed May 26, 2017).
12. Amy Feldman Farb and Amy L. Margolis, "The Teen Pregnancy Prevention Program (2010–2015)," p. S14.
13. U.S. Department of Health and Human Services, Office of Adolescent Health, "Evidence-Based TPP Programs," http://www.hhs.gov/ash/oah/oah-initiatives/tpp_program/db/ (accessed July 22, 2016).
14. Ron Haskins and Greg Margolis, *Show Me the Evidence: Obama's Fight for Rigor and Results in Social Policy* (Washington, DC: Brookings Institution Press, 2015), p. 101.
15. Ron Haskins, "Renewing Communities and Providing Opportunities Through Innovative Solutions to Poverty," testimony before the Committee on Homeland Security and Governmental Affairs, U.S. Senate, June 22, 2016, http://www.brookings.edu/research/testimony/2016/06/22-renewing-communities-and-providing-opportunities-through-innovative-solutions-to-poverty-haskins (accessed July 22, 2016).
16. Results for America, "Federal Evidence-Based Innovation Programs," *Invest in What Works Fact Sheet*, October 21, 2015, http://results4america.org/tools/invest-works-fact-sheet-federal-evidence-based-innovation-programs (accessed January 4, 2017).
17. Ibid.
18. Ron Haskins and Greg Margolis, *Show Me the Evidence: Obama's Fight for Rigor and Results in Social Policy* (Washington, DC: Brookings Institution Press, 2015), pp. 75 and 76.
19. Farb and Margolis, "The Teen Pregnancy Prevention Program (2010–2015)," p. S9.
20. Ibid.
21. Ibid.
22. U.S. Department of Health and Human Services, Office of Adolescent Health, "Grantees FY 2010–2014," http://www.hhs.gov/ash/oah/oah-initiatives/evaluation/grantee-led-evaluation/grantees-2010-2014.html (accessed May 25, 2017). (Note: When this website was accessed on September 26, 2016, Reducing the Risk and ¡Cuídate! were listed as Tier I models.); U.S. Department of Health and Human Services, Office of Adolescent Health, "Teen Pregnancy Prevention Replication Study," https://www.hhs.gov/ash/oah/evaluation-and-research/federal-

led-evaluation/teen-pregnancy-prevention-program-replication-study/index.html (accessed May 25, 2017); and "Building the Evidence to Prevent Adolescent Pregnancy: Office of Adolescent Health Impact Studies (2010–2015)," *American Journal of Public Health*, Vol. 106, No. S1 (October 2016).

23. For a detailed discussion of evaluation methodology, see Muhlhausen, *Do Federal Social Programs Work?*, pp. 41–79.
24. Some may assert that the problem of intervention and control group differences in observed characteristics in cluster randomization can be solved by appropriately adjusting the standard errors of the impact estimates. However, this statistical correction cannot account for the unobserved differences between the clustered intervention and control groups. When members of the intervention and control groups are randomly assigned on an individual basis, the potentially unobserved differences between the groups are more likely to balance out. Thus, impact estimates based on individual-level random assignment are superior to cluster randomization for determining causal impact.
25. The internal validity threat of history occurs when events taking place concurrently with the intervention could cause the observed effect, while maturation occurs when natural changes in participants that occur over time could be confused with an observed outcome. For a more detailed discussion of threats to internal validity, see Muhlhausen, *Do Federal Social Programs Work?*, pp. 50–62.
26. Muhlhausen, *Do Federal Social Programs Work?*
27. Farb and Margolis, "The Teen Pregnancy Prevention Program (2010–2015)," p. S10.
28. Russell P. Cole, "Comprehensive Reporting of Adolescent Pregnancy Prevention Programs," *American Journal of Public Health*, Vol. 106, No. S1 (October 2016), p. S15.
29. Christopher Trenholm et al., *Impacts of Four Title V, Section 510 Abstinence Education Programs: Final Report* (Princeton, NJ: Mathematica Policy Research, 2007). For a review of this evaluation, see Muhlhausen, *Do Federal Social Programs Work?*, pp. 138–144.
30. The two studies that found BART to be ineffective are Angela R. Robertson et al., "The Healthy Teen Girls Project: Comparison of Health Education and STD Risk Reduction Intervention for Incarcerated Adolescents Females," *Health Education & Behavior*, Vol. 38, No. 3 (2011), pp. 241–250, and Janet S. St. Lawrence et al., "Sexual Risk Reduction and Anger Management Interventions for Incarcerated Male Adolescents: A Randomized Controlled Trial of Two Interventions," *Journal of Sex Education and Therapy*, Vol. 24, No. 1–2 (1999), pp. 9–17. The study that found at least one beneficial effect is Janet S. St. Lawrence et al., "Cognitive-Behavioral Intervention to Reduce African American Adolescents' Risks for HIV Infection," *Journal of Consulting and Clinical Psychology*, Vol. 63, No. 2 (1995), pp. 221–237.
31. U.S. Department of Health and Human Services, Office of Adolescent Health, "Grantees FY 2010–2014," https://www.hhs.gov/ash/oah/oah-initiatives/evaluation/grantee-led-evaluation/grantees-2010-2014.html (accessed May 25, 2017).
32. Stephanie Martin et al., "Evaluation of Alaska Promoting Health Among Teens, Comprehensive Abstinence and Safer Sex (AKPHAT) in Alaska," University of Alaska–Anchorage Institute of Social and Economic Research, October 6, 2015; Holli Slater and Diane Mitschke, "Evaluation of the Crossroads Program in Arlington, TX: Findings from an Innovative Teen Pregnancy Prevention Program," Arlington, TX, University of Texas at Arlington, December 20, 2015; Traci Schwinn et al., "Evaluation of Circle of Life in Tribes of the Northern Plains: Findings from an Innovative Teen Pregnancy Prevention Program," final behavioral impact report submitted to the Office of Adolescent Health, August 18, 2015; Amita N. Vyas et al., "The Evaluation of Be Yourself/Sé Tú Mismo in Maryland," Washington, DC, The George Washington University Milken Institute School of Public Health, October 30, 2015; Patricia Kissinger, Norine Schmidt, and Jakevia Green, "Evaluation of BUtiful: An Internet Pregnancy Prevention for Older Teenage Girls in New Orleans, Louisiana," Tulane University School of Public Health and Tropical Medicine, November 11, 2015; Tamara Vehige Calise et al., "Evaluation of Healthy Futures in Public Middle Schools in Three Northeastern Massachusetts Cities," *Final Impact Report* for the Black Ministerial Alliance of Greater Boston, Inc., December 30, 2015; Mathilda B. Ruwe et al., "Evaluation of Haitian-American Responsible Teen," final impact report for Boston Medical Center, August 2016; Sheana Salyers Bull et al., "Evaluation of Youth All Engaged (YAE) in Denver, CO," *Final Impact Report* for Denver Public Health, August 31, 2015; Yasuyo Abe et al., "Early Findings from the Evaluation of the Pono Choices Program—A Culturally Responsive Teen Pregnancy and Sexually Transmitted Infection Prevention Program for Middle School Youth in Hawai'i," IMPAQ International, March 23, 2016; Robert G. LaChausse, "Evaluation of the Positive Prevention PLUS Teen Pregnancy Prevention Program," final impact report for San Bernardino County Superintendent of Schools, February 22, 2016; American Empirical Solutions, "Evaluation of Will Power/Won't Power in Los Angeles County," *Final Impact Report* for Volunteers of America–Girls, Inc. of Greater Los Angeles, October 31, 2015; Anita P. Barbee et al., "Impact of Two Adolescent Pregnancy Prevention Interventions on Risky Sexual Behavior: A Three-Arm Cluster Randomized Control Trial," *American Journal of Public Health*, Vol. 1, No. S1 (2016), pp. S85–S90.
33. Originally, OAH listed Reducing the Risk and ¡Cuídate! as Tier I models. For confirmation, see U.S. Department of Health and Human Services, Office of Adolescent Health, "Summary of Findings from TPP Program Grantees (FY2010–2014), https://www.hhs.gov/ash/oah/sites/default/files/ash/oah/oah-initiatives/evaluation/grantee-led-evaluation/summary-ebps.pdf (accessed May 25, 2017).
34. Mathematica Policy Research and Child Trends, "Identifying Programs that Impact Teen Pregnancy, Sexually Transmitted Infections, and Associated Sexual Risk Behaviors: Review Protocol–Version 2.0," 2012, https://www.hhs.gov/ash/oah/oah-initiatives/teen_pregnancy/db/eb-programs-review-v2.pdf (accessed February 27, 2017).
35. U.S. Department of Health and Human Services, Office of Adolescent Health, "TPP Resource Center: Evidence-Based Programs," http://www.hhs.gov/ash/oah/oah-initiatives/teen_pregnancy/db/ (accessed February 21, 2017).
36. St. Lawrence et al., "Cognitive-Behavioral Intervention to Reduce African American Adolescents' Risks for HIV Infection," and U.S. Department of Health and Human Services, Office of Adolescent Health, "Study Details," http://tppevidencereview.aspe.hhs.gov/StudyDetails.aspx?id=47 (accessed October 20, 2016).

37. St. Lawrence et al., "Cognitive-Behavioral Intervention to Reduce African American Adolescents' Risks for HIV Infection."
38. Ibid., pp. 227 and 228.
39. St. Lawrence et al., "Sexual Risk Reduction and Anger Management Interventions for Incarcerated Male Adolescents," and U.S. Department of Health and Human Services, Office of Adolescent Health, "Study Details," http://tppevidencereview.aspe.hhs.gov/StudyDetails.aspx?id=45 (accessed October 20, 2016).
40. St. Lawrence et al., "Sexual Risk Reduction and Anger Management Interventions for Incarcerated Male Adolescents."
41. Ibid., pp. 13 and 14.
42. Janie B. Butts and Sherry Hartman, "Effectiveness of a Behavioral Intervention to Reduce HIV Risk in Adolescents," *American Journal of Maternal/Child Nursing*, Vol. 27, No. 3 (2002), pp. 163–169.
43. U.S. Department of Health and Human Services, Office of Adolescent Health, "Study Details."
44. Robert M. Malow et al., "Effects of a Culturally Adapted HIV Prevention Intervention in Haitian Youth," *Journal of the Association of Nurses in AIDS Care*, Vol. 20, No. 2 (2009), pp. 110–121, and U.S. Department of Health and Human Services, Office of Adolescent Health, "Study Details."
45. Malow et al., "Effects of a Culturally Adapted HIV Prevention Intervention in Haitian Youth."
46. Robertson et al., "The Healthy Teen Girls Project," and U.S. Department of Health and Human Services, Office of Adolescent Health, "Study Details."
47. Robertson et al., "The Healthy Teen Girls Project," pp. 246 and 247.
48. Eric Jenner et al., "Impact of an Intervention Designed to Reduce Sexual Health Risk Behaviors of African American Adolescents: Results of a Randomized Controlled Trial," *American Journal of Public Health*, Vol. 106, No. S1 (2016), pp. S78–S84.
49. Ibid.
50. Ibid., p. S83, Table 3.
51. Ibid., p. S83.
52. Ibid.
53. Ibid.
54. Susan Philliber et al., "Preventing Pregnancy and Improving Health Care Access Among Teenagers: An Evaluation of the Children's Aid Society-Carrera Program," *Perspectives on Sexual and Reproductive Health*, Vol. 34, No. 5 (2002), pp. 244–251, and U.S. Department of Health and Human Services, Office of Adolescent Health, "Study Details."
55. Philliber et al., "Preventing Pregnancy and Improving Health Care Access Among Teenagers," p. 245.
56. Ibid., p. 249, Table 3.
57. Ibid.
58. Ibid.
59. Scott Herrling, "Evaluation of the Children's Aid Society (CAS)–Carrera Adolescent Pregnancy Prevention Program in Chicago, IL," *Final Impact Report* for Children's Home + Aid, February 29, 2016, https://collections.nlm.nih.gov/catalog/nlm:nlmuid-101683798-pdf (accessed June 9, 2017).
60. Ibid., p. 26.
61. Ibid, p. 27, Table IV.4.
62. Ibid, p. 27, Table IV.5.
63. Ibid., p. 22.
64. Tressa Tucker, "Evaluation of CAS–Carrera Program in Georgia," *Final Impact Report* for Morehouse School of Medicine," December 30, 2015, https://collections.nlm.nih.gov/catalog/nlm:nlmuid-101683800-pdf (accessed June 9, 2017).
65. Ibid., p. 20.
66. Ibid., p. 26, Table 8.
67. Antonia M. Villarruel, John B. Jemmott, and Loretta S. Jemmott, "A Randomized Controlled Trial Testing an HIV Prevention Intervention for Latino Youth." *Archives of Pediatrics and Adolescent Medicine*, Vol. 160 (August 2006), pp. 772–777, and U.S. Department of Health and Human Services, Office of Adolescent Health, "Study Details."
68. Trisha E. Mueller et al., "The Implementation of a Culturally Based HIV Sexual Risk Reduction Program for Latino Youth in a Denver Area High School," *AIDS Education and Prevention*, Vol. 21, Supplement B (2009), pp. 164–170, and Kim L. Larson et al., "Testing the Feasibility of ¡Cuídate! With Mexican and Central American Youth in a Rural Region of a Southern State," *Research in Nursing & Health*, Vol. 37 (2014), pp. 409–422.
69. Villarruel, Jemmott, and Jemmott, "A Randomized Controlled Trial Testing an HIV Prevention Intervention for Latino Youth."
70. Ibid., pp. 774 and 775.
71. Ibid., p. 775.

72. Meredith Kelsey et al., "Replicating ¡Cuídate!: 6-Month Impact Findings of a Randomized Controlled Trial," *American Journal of Public Health*, Vol. 106, No. S1 (2016), pp. S70–S77.
73. Ibid., p. S76.
74. Ibid., p. S70.
75. Ibid., p. S76, Table 3.
76. Ibid., p. S75.
77. Ibid., p. S73.
78. Susan R. Tortolero et al., "It's Your Game. Keep It Real: Delaying Sexual Behavior with an Effective Middle School Program," *Journal of Adolescent Health*, Vol. 46, No 2 (2010), pp. 1–19, and U.S. Department of Health and Human Services, Office of Adolescent Health, "It's Your Game: Keep It Real (IYG): Reviewed Studies," https://tppevidencereview.aspe.hhs.gov/document.aspx?rid=3&sid=86&mid=5 (accessed January 23, 2017).
79. Tortolero et al., "It's Your Game. Keep It Real: Delaying Sexual Behavior with an Effective Middle School Program."
80. U.S. Department of Health and Human Services, Office of Adolescent Health, "Study Details."
81. Tortolero et al., "It's Your Game. Keep It Real: Delaying Sexual Behavior with an Effective Middle School Program," p. 14, Table 2.
82. Christine M. Markham et al., "Sexual Risk Avoidance and Sexual Risk Reduction Interventions for Middle School Youth: A Randomized Controlled Trial," *Journal of Adolescent Health*, Vol. 50 (2012), pp. 279–288; Christine M. Markham et al., "Behavioral and Psychosocial Effects of Two Middle School Sexual Health Education Programs at Tenth-Grade Follow-Up," *Journal of Adolescent Health*, Vol. 54 (2014), pp. 151–159; and U.S. Department of Health and Human Services, Office of Adolescent Health, "Study Details."
83. Markham et al., "Sexual Risk Avoidance and Sexual Risk Reduction Interventions for Middle School Youth: A Randomized Controlled Trial."
84. Ibid., p. 284, Table 3.
85. Ibid.
86. Ibid., p. 287.
87. Ibid., p. 283, Table 2.
88. Markham et al., "Behavioral and Psychosocial Effects of Two Middle School Sexual Health Education Programs at Tenth-Grade Follow-Up."
89. Ibid., p. 156, Table 2.
90. Susan C. Potter et al., "It's Your Game…Keep It Real in South Carolina: A Group Randomized Trial Evaluating the Replication of an Evidence-Based Adolescent Pregnancy and Sexually Transmitted Infection Prevention Program," *American Journal of Public Health*, Vol. 106, No. S1 (2016), pp. 560–569, and Karin Coyle et al., "Evaluation of It's Your Game…Keep It Real in Houston, TX: Final Report," *Final Report* for University of Texas Health Science Center–Houston, February 10, 2016.
91. Ibid., p. 564.
92. Ibid., p. 566, Table 3.
93. For a review of how to interpret effect sizes, see David B. Muhlhausen, "The Head Start CARES Demonstration: Another Failed Federal Early Childhood Education Program," Heritage Foundation *Backgrounder* No. 3041, August 6, 2015, http://www.heritage.org/research/reports/2015/08/the-head-start-cares-demonstration-another-failed-federal-early-childhood-education-program.
94. Potter et al., "It's Your Game…Keep It Real in South Carolina," p. 566.
95. Ibid., p. 566, Table 3.
96. Ibid., p. 562.
97. Ibid., p. 560.
98. Brian R. Flay, "Efficacy and Effectiveness Trials (and Other Phases of Research) in the Development of Health Promotion Programs," *Preventive Medicine*, Vol. 15 (1986), pp. 451–474.
99. Coyle et al., "Evaluation of It's Your Game…Keep It Real in Houston, TX."
100. Ibid., p. 15.
101. Ibid., p. 16.
102. Ibid., pp. 17 and 18, Table III.3; p. 19, Table III.4; and p. 20, Table III.5.
103. Ibid., pp. 17 and 18, Table III.3.
104. Ibid., p. 27, Table IV.1, and p. 28, Table IV.2.
105. Ibid., p. 29.
106. John B. Jemmott, Loretta S. Jemmott, and Gregory T. Fong, "Efficacy of a Theory-Based Abstinence-Only Intervention over 24 Months," *Archives of Pediatrics and Adolescent Medicine*, Vol. 164, No. 2 (February 2010), pp. 152–159, and U.S. Department of Health and Human Services, Office of Adolescent Health, "Promoting Health Among Teens! Abstinence-Only Intervention; Reviewed Studies," https://tppevidencereview.aspe.hhs.gov/document.aspx?rid=3&sid=166&mid=5 (accessed February 22, 2017).

107. Jemmott et al., "Efficacy of a Theory-Based Abstinence-Only Intervention over 24 Months," p. 153.
108. Ibid., p. 154.
109. Ibid.
110. Ibid., p. 156.
111. Ibid., p. 157, Table 3.
112. Ibid.
113. Ibid.
114. Ibid., p. 158.
115. Ibid.
116. Elaine M. Walker, Rafael Inoa, and Nanci Coppola, *Evaluation of Promoting Health Among Teens Abstinence-Only Intervention in Yonkers, NY, Final Impact Report* for Program Reach, Inc., March 9, 2016.
117. Ibid., p. 15, Table 2.
118. Ibid., p. 23, Table 5; p. 24, Table 6; and p. 24, Table 7.
119. Ibid., p. 23, Table 3.
120. Ibid., pp. 25 and 26.
121. Ibid., p. 26.
122. Rick S. Zimmerman et al., "Effects of a School-Based, Theory-Driven HIV and Pregnancy Prevention Curriculum," *Perspectives on Sexual and Reproductive Health*, Vol. 40, No. 1 (March 2008), pp. 42–51; Valerie F. Reyna and Brian A. Mills, "Theoretically Motivated Interventions for Reducing Sexual Risk Taking in Adolescence: A Randomized Controlled Experiment Applying Fuzzy-Trace Theory," *Journal of Experimental Psychology*, Vol. 143, No. 4 (2014), pp. 1627–1648; Betty M. Hubbard, Mark L. Giese, and Jacquie Rainey, "A Replication Study of Reducing the Risk, a Theory-Based Sexuality Curriculum for Adolescents," *Journal of School Health*, Vol. 68, No. 6 (August 1998), pp. 243–247; Angela Ebrero et al., "Effects of Peer Education on the Peer Educators in a School-Based HIV Prevention Education Research Program: Where Should Peer Education Research Go from Here?" *Health Education & Behavior*, Vol. 29, No. 4 (August 2002), pp. 411–423; Douglas Kirby et al., "Reducing the Risk: Impact of a New Curriculum on Sexual Risk-Taking," *Family Planning Perspectives*, Vol. 23, No. 6 (November/December 1991), pp. 253–263; Richard P. Barth et al., "Preventing Adolescent Pregnancy with Social and Cognitive Skills," *Journal of Adolescent Research*, Vol. 7, No. 2 (April 1992), pp. 208–232; and U.S. Department of Health and Human Services, Office of Adolescent Health, "Reducing the Risk: Reviewed Studies," https://tppevidencereview.aspe.hhs.gov/document.aspx?rid=3&sid=182&mid=5 (accessed February 22, 2017).
123. Hubbard, Giese, and Rainey, "A Replication Study of Reducing the Risk, a Theory-Based Sexuality Curriculum for Adolescents," and Ebrero et al., "Effects of Peer Education on the Peer Educators in a School-Based HIV Prevention Education Research Program: Where Should Peer Education Research Go from Here?"
124. Rick S. Zimmerman et al., "Effects of a School-Based, Theory-Driven HIV and Pregnancy Prevention Curriculum," *Perspectives on Sexual and Reproductive Health*, Vol. 40, No. 1 (March 2008), pp. 42–51, and Reyna and Mills, "Theoretically Motivated Interventions for Reducing Sexual Risk Taking in Adolescence: A Randomized Controlled Experiment Applying Fuzzy-Trace Theory."
125. Zimmerman et al., "Effects of a School-Based, Theory-Driven HIV and Pregnancy Prevention Curriculum," and U.S. Department of Health and Human Services, Office of Adolescent Health, "Study Details."
126. Ibid., p. 43.
127. Ibid.
128. Ibid., p. 47.
129. Ibid., p. 43, Table 3.
130. Ibid., p. 47, Table 5.
131. Kirby et al., "Reducing the Risk: Impact of a New Curriculum on Sexual Risk-Taking"; Barth et al., "Preventing Adolescent Pregnancy with Social and Cognitive Skills"; and U.S. Department of Health and Human Services, Office of Adolescent Health, "Study Details."
132. Kirby et al., "Reducing the Risk: Impact of a New Curriculum on Sexual Risk-Taking."
133. Ibid., p. 258, Table 3. The study used the simple χ^2 tests and t-tests to determine impact.
134. Ibid., p. 259.
135. Ibid.
136. Ibid., p. 260.
137. Barth et al., "Preventing Adolescent Pregnancy with Social and Cognitive Skills."
138. Ibid., p. 223.
139. Reyna and Mills, "Theoretically Motivated Interventions for Reducing Sexual Risk Taking in Adolescence."
140. Ibid., p. 1635.

141. U.S. Department of Health and Human Services, Office of Adolescent Health, "Study Details."
142. Reyna and Mills, "Theoretically Motivated Interventions for Reducing Sexual Risk Taking in Adolescence."
143. Ibid., p. 1638, Table 3.
144. Ibid., p. 1639, Table 5.
145. Anita P. Barbee et al., "Impact of Two Adolescent Pregnancy Prevention Interventions on Risky Sexual Behavior: A Three-Arm Cluster Randomized Control Trial," *American Journal of Public Health*, Vol. 1, No. S1 (2016), pp. S85–S90, and Meridith Kelsey et al., "Replicating Reducing the Risk: 12-Month Impacts of a Cluster Randomized Controlled Trial," *American Journal of Public Health*, Vol. 1, No. S1 (2016), pp. S45–S52.
146. Barbee et al., "Impact of Two Adolescent Pregnancy Prevention Interventions on Risky Sexual Behavior."
147. Ibid., p. S85.
148. Ibid., p. S87.
149. Ibid.
150. Ibid., p. S86.
151. Ibid. Assignment was not truly random because five students were not randomly assigned in order to ensure gender balance among clusters and that members of the same household were within the same cluster to avoid cross-contamination.
152. Ibid., p. S87, Table 1.
153. Ibid.
154. Ibid., p. S88, Table 2.
155. Ibid., p. S89, Table 3.
156. Ibid., p. S88, Table 2.
157. Ibid.
158. Ibid., pp. S89–S90.
159. Kelsey et al., "Replicating Reducing the Risk."
160. Ibid., p. S46.
161. Ibid., p. S49.
162. Ibid., p. S49, Table 2.
163. Ibid., p. S50.
164. Ibid., p. S51.
165. Kelsey et al., "Replicating Reducing the Risk," p. S45.
166. Lydia A. Shrier et al., "Randomized Controlled Trial of a Safer Sex Intervention for High-Risk Adolescent Girls," *Archives of Pediatrics and Adolescent Medicine*, Vol. 155 (January 2001), pp. 73–79, and U.S. Department of Health and Human Services, Office of Adolescent Health, "Study Details."
167. Shrier et al, "Randomized Controlled Trial of a Safer Sex Intervention for High-Risk Adolescent Girls."
168. Ibid.
169. Ibid., p. 77, Table 2.
170. Eric Jenner et al., "Evaluation of Safer Sex Intervention in New Orleans, LA," *Final Impact Report* for Louisiana Public Health Institute, January 22, 2016, and Meredith Kelsey, Jessica T. Walker, Jean Layzer, Crisofer Price, and Randall Juras, "Replicating the Safer Sex Intervention: 9-Month Impact Findings of a Randomized Controlled Trial," *American Journal of Public Health*, Vol. 1, No. S1 (2016), pp. S53–S59.
171. Jenner et al., "Evaluation of Safer Sex Intervention in New Orleans, LA."
172. Ibid., p. 25.
173. Ibid.
174. Ibid., p. 25, Table IV.1.
175. Ibid., p. 26, Table IV.2.
176. Kelsey et al., "Replicating the Safer Sex Intervention."
177. Ibid., p. S54.
178. Ibid., p. S56.
179. Ibid., p. S58, Table 2.
180. Ibid., p. S56.

181. Julie S. Downs et al., "Interactive Video Behavioral Intervention to Reduce Adolescent Females' STD Risk: A Randomized Controlled Trail," *Social Science & Medicine*, Vol. 59 (2004), pp. 1561–1572, and U.S. Department of Health and Human Services, Office of Adolescent Health, "Study Details."
182. Ibid., p. 1565.
183. Ibid., p. 1566.
184. Ibid., p. 1567.
185. Ibid.
186. Joan Eichner et al., "Evaluation of Seventeen Days in Ohio, Pennsylvania, and West Virginia," *Final Impact Report* for University of Pittsburgh, Office of Child Development, August 31, 2015.
187. Ibid., p. 15.
188. Ibid., p. 16, Table 3.3; pp. 16 and 17, Table 3.4; and pp. 17 and 18, Table 3.5.
189. Ibid., p. 21.
190. Ibid., p. 23, Table 4.3.
191. Ibid., p. 22, Table 4.2, and p. 23, Table 4.3.
192. Ibid., p. 23, Table 4.4.
193. Joseph P. Allen et al., "Preventing Teen Pregnancy and Academic Failure: Experimental Evaluation of a Developmentally Based Approach," *Child Development*, Vol. 68, No. 4 (1997), pp. 729–742; Joseph P. Allen and Susan Philliber, "Who Benefits Most from a Broadly Targeted Prevention Program? Differential Efficacy Across Populations in the Teen Outreach Program," *Journal of Community Psychology*, Vol. 29, No. 6 (2001), pp. 637–655; and U.S. Department of Health and Human Services, Office of Adolescent Health, "Reviewed Studies," https://tppevidencereview.aspe.hhs.gov/document.aspx?rid=3&sid=237&mid=5 (accessed February 15, 2017).
194. Allen and Philliber, "Who Benefits Most from a Broadly Targeted Prevention Program?" and U.S. Department of Health and Human Services, Office of Adolescent Health, "Study Details."
195. Allen et al., "Preventing Teen Pregnancy and Academic Failure."
196. Ibid., p. 730.
197. Ibid., p. 731.
198. Ibid., p. 734.
199. Ibid., p. 735, Table 2.
200. Ibid.
201. Ibid., p. 735.
202. William T. Robinson et al., "Randomized Trials of the Teen Outreach Program in Louisiana and Rochester, New York," *American Journal of Public Health*, Vol. 1, No. S1 (2016), pp. S39–S44, and Kimberly Francis et al., "Scalability of an Evidence-Based Adolescent Pregnancy Prevention Program: New Evidence from 5 Cluster-Randomized Evaluations of the Teen Outreach Program," *American Journal of Public Health*, Vol. 1, No. S1 (2016), pp. S32–S38.
203. Francis et al., "Scalability of an Evidence-Based Adolescent Pregnancy Prevention Program," p. S33, Table 1.
204. Ibid., p. S33, Table 1.
205. Ibid., p. S34.
206. Ibid., S33, Table 1.
207. Ibid., p. S36, Table 3, and p. S37, Table 4.
208. Ibid., p. S37.
209. Robinson et al., "Randomized Trials of the Teen Outreach Program in Louisiana and Rochester, New York."
210. Ibid., p. S42, Table 2, and p. S43, Table 3.
211. Ibid., p. S42.
212. Barbee et al., "Impact of Two Adolescent Pregnancy Prevention Interventions on Risky Sexual Behavior."
213. Stuart M. Butler and David D. Muhlhausen, "Can Government Replicate Success?" *National Affairs* (Spring 2014), pp. 25–39, http://www.nationalaffairs.com/publications/detail/can-government-replicate-success (accessed March 2, 2017).
214. David B. Muhlhausen, "Evidence-Based Policymaking: A Primer," Heritage Foundation *Backgrounder* No. 3063, October 15, 2015, http://www.heritage.org/budget-and-spending/report/evidence-based-policymaking-primer.
215. Flay, "Efficacy and Effectiveness Trials (and Other Phases of Research) in the Development of Health Promotion Programs."
216. Ibid.

217. Muhlhausen, *Do Federal Social Programs Work?,* pp. 311–313, and David B. Muhlhausen, "Evaluating Federal Social Programs: Finding Out What Works and What Does Not," testimony before the Subcommittee on Human Resources, Committee on Ways and Means, U.S. House of Representatives, July 17, 2013, https://waysandmeans.house.gov/UploadedFiles/David_Muhlhausen_Testimony_071713.pdf (accessed March 1, 2017).
218. Philliber et al., "Preventing Pregnancy and Improving Health Care Access Among Teenagers," p. 249, Table 3.
219. Markham et al., "Sexual Risk Avoidance and Sexual Risk Reduction Interventions for Middle School Youth: A Randomized Controlled Trial," p. 284, Table 3.
220. Ibid., p. 156, Table 2.
221. Potter et al., "It's Your Game…Keep It Real in South Carolina," p. 566, Table 3.
222. Kelsey et al., "Replicating ¡Cuídate!: 6-Month Impact Findings of a Randomized Controlled Trial," p. S75.