# The Teen Pregnancy Prevention Program (2010-2015): Synthesis of Impact Findings

The US Department of Health and Human Services (HHS) Office of Adolescent Health (OAH) Teen Pregnancy Prevention (TPP) Program is a national, tiered, evidence-based program that funds diverse organizations nationwide working to prevent adolescent pregnancy. The OAH TPP Program invests a larger share of its grant funds in the implementation of evidence-based programs—those programs proven through rigorous evaluation to reduce adolescent pregnancy and risky behavior associated with adolescent pregnancy (tier 1). These diverse programs include sex education, youth development, abstinence education, clinic-based programs, and programs specifically designed for vulnerable populations, including parenting adolescents and youths in juvenile detention (Table 1). The OAH TPP Program also invests a smaller portion of its grant funds in research and demonstration projects to develop and test new models and innovative strategies to address gaps in what is known about how to prevent adolescent pregnancy (tier 2).

## FIRST FIVE-YEAR COHORT

The first five-year cohort of the OAH TPP Program began in 2010 as one of six new federal evidence-based initiatives. Beginning in 2010, OAH provided $100 million annually to 75 organizations to replicate evidence-based TPP programs (tier 1) and to 27 organizations to develop and evaluate new and innovative approaches to prevent adolescent pregnancy (tier 2). The tier 1 grantees implemented one or more programs from a list of 28 models identified by the HHS Pregnancy Prevention Evidence Review (from here on referred to as the HHS Evidence Review) through an independent, systematic, and comprehensive review of the literature to reduce adolescent pregnancy, births, sexually transmitted infections (STIs), and associated risk behaviors through rigorous evaluation. The research quality standards used to identify these 28 programs became the criteria for what constituted an evidence-based TPP program and set the standards for research quality for future TPP evaluations. The tier 2 grantees focused on developing programs or approaches to address gaps in the existing evidence base, with the goal that programs found to be effective would feed into the menu of evidence-based TPP programs and become eligible for replication in future cohorts of the TPP Program.

Across the TPP Program, OAH funded a total of 41 program evaluations, including 19 evaluations of evidence-based programs and 22 evaluations of new or innovative approaches (Figure 1). The evaluations assessed the effectiveness of each program in at least one of the following areas: reducing adolescent pregnancy and births, delaying sexual initiation, improving contraceptive use, and reducing STIs (Table A, available as a supplement to the online version of this article at http://www.ajph.org, has descriptions of each evaluation). The evaluations represent a mix of independent grantee-led evaluations conducted through cooperative agreements with OAH and OAH-led evaluations conducted through contracts with research firms. Findings from 21 of these evaluations are featured in this *AJPH* themed supplement issue. The remaining evaluation findings are contained in reports held in a collection at the National Library of Medicine, available through a link on the OAH Web site (http://www.hhs.gov/ash/oah/oah-initiatives/evaluation/grantee-led-evaluation/grantees-2010-2014.html).

## REQUIREMENTS FOR EVALUATION STUDIES

All projects 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. Implementations 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. A standard set of performance measurement data related to fidelity, dosage, reach and retention, partnerships, training, and dissemination were collected and reported to OAH every six months and reviewed to monitor progress of the project. Grantees had to adhere to evaluation expectations, primarily meeting the standards for research quality as established by the HHS Evidence Review. Furthermore, tier 2 programs had to be packaged and implementation ready by the end of the five-year grant period. All evaluations were required to collect three time points of data, including a baseline prior to program implementation as well as short- and long-term follow-ups, and evaluations had to collect at least one behavioral outcome measure from the HHS Evidence Review.

## SUPPORT AND TECHNICAL ASSISTANCE

OAH provided a strong system of support to all evaluation grants in the TPP Program. OAH project officers are assigned to grants for the life of the grant period, hold at least monthly structured calls with each grantee during the five-year grant, and provide as-needed support and guidance. An evaluation training and technical assistance contract provided all evaluation grantees with intensive technical assistance, support, and monitoring to ensure that the design, implementation, analyses, and reporting met the HHS Evidence Review standards

**ABOUT THE AUTHORS**

*Amy Feldman Farb and Amy L. Margolis are with the US Department of Health and Human Services, Office of the Assistant Secretary for Health, Office of Adolescent Health, Rockville, MD.*

*Correspondence should be sent to Amy Feldman Farb, Senior Evaluator, Office of Adolescent Health, Department of Health and Human Services, 1101 Wootton Parkway, Suite 700, Rockville, MD, 20852 (e-mail: amy.farb@hhs.gov). Reprints can be ordered at http://www.ajph.org by clicking the "Reprints" link.*

*This editorial was accepted June 24, 2016.*

*doi: 10.2105/AJPH.2016.303367*

**TABLE 1—Diversity of Programs Included in the Teen Pregnancy Prevention Evaluations: United States, 2010–2015**

| Type of Program | Evaluations, % (No.) |
|---|---|
| Comprehensive sex education | 39 (16) |
| Youth development | 39 (16) |
| Abstinence education | 7 (3) |
| Clinic-based | 7 (3) |
| Special populations | 5 (2) |
| Relationship education | 2 (1) |

for research quality (see Zief et al.[1] in this issue). OAH also employs an evaluation specialist to provide guidance to OAH project officers and grantees, and to manage the evaluation training and technical assistance contract.

## SYNTHESIS GOALS

In this article, we present the results of a synthesis across the 41 TPP cohort 1 evaluations. The goal of the synthesis is to address two research questions aligned with the broader goals of the TPP Program: (1) Do programs previously identified as effective by the HHS Evidence Review continue to reduce births, pregnancies, STIs, and associated sexual risk behaviors when replicated at a large scale in new settings and with new populations? and (2) Can new TPP programs or significant adaptations to evidence-based TPP programs be identified that reduce births, pregnancies, STIs, and associated sexual risk behaviors?

## PROGRAM EFFECTIVENESS

The first research question was important to address in light of a newly established evidence base. The evidence base established in 2010, in preparation for the TPP Program, was the first time a systematic review was conducted on the topic of adolescent pregnancy prevention. Individual studies were separately reviewed regardless of content or philosophical approach against rigorous standards for research quality.[2] All but one of the program models meeting the standards of research quality demonstrated evidence of effectiveness through a single study, often conducted by the developer of the program. The review team noted the lack of replication studies as a gap in the evidence base and called for subsequent, independent evaluations to determine the effectiveness of the programs with broader populations and in real-world conditions.

Research 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.[2–6] The number has been estimated to be as low as 10% to 20% of randomized controlled trials (RCTs) that result in the same significant positive impacts as found in the original study.[6,7]

## NEW PROGRAM IDENTIFICATION

OAH's goals with the TPP program evaluations were to build a body of evidence about where, when, and with whom individual evidence-based programs are effective, and to contribute new programs having some evidence of effectiveness to be replicated and further evaluated in the future. The tier 1 evaluations implemented evidence-based programs as intended and with quality, but in new settings and with new populations, therefore adding evidence about where, when, and with whom the program is effective. In addition, the research questions differed across individual evaluations. Some evaluations focused on a test of the program itself, others tested the program as a replacement of usual activities, and others tested the program in additional to usual activities. These "replication" evaluations replicated the program implementation, not the evaluation for the purposes of verifying the original study results.

## METHODS

To conduct this synthesis of the TPP Program evaluations, we used (1) the final impact reports from the 41 federal and grantee program evaluations funded during cohort 1 of the TPP Program and (2) the grantee performance measure data. We summarized the data and themes that emerged from the evaluation reports across the projects, counted behavioral outcomes and assessed the implementation quality of each project. A formal meta-analysis is currently being conducted, and a final report will be available in 2017.

### Final Evaluation Reports

All 41 cohort 1 TPP Program final evaluation reports were included in this synthesis. All evaluation reports consisted of five key elements:

1. the research questions—primary and secondary;
2. detailed descriptions of the treatment and control conditions;
3. the evaluation design including recruitment, data collection, outcomes, sample, baseline equivalence, and methods;
4. impact findings for both implementation and impact analyses; and
5. a brief conclusion.

OAH's Evaluation Training and Technical Assistance contractor reviewed all of the final evaluation reports to ensure they adhered to the approved design and analysis plan, that descriptions of the treatment and control conditions were accurate, that analyses were correct and would meet the HHS Evidence Review standards, and that conclusions drawn were justified by the data and accurate. A detailed description of the evidence review can be found at http://tppevidencereview.aspe.hhs.gov/Default.aspx. OAH used these reports to assess each evaluation's implementation quality and outcome findings for this synthesis.

The reports consisted of

- 19 tier 1 replication evaluations (Figure 1). The 19 tier 1 replication evaluations tested 10 of the evidence-based TPP programs from the original list of 28 programs identified as effective by the HHS Evidence Review.
- 22 tier 2 evaluations of new and innovative approaches (Figure 1). The 22 tier 2 evaluations tested the effectiveness of 22 different programs intended to reduce adolescent pregnancies, births, STIs, and associated sexual risk behaviors that were previously untested through rigorous evaluation. The 22 evaluations consisted of 10 new TPP programs, seven significant adaptations to tier 1 evidence-based programs, and five existing but previously untested programs.

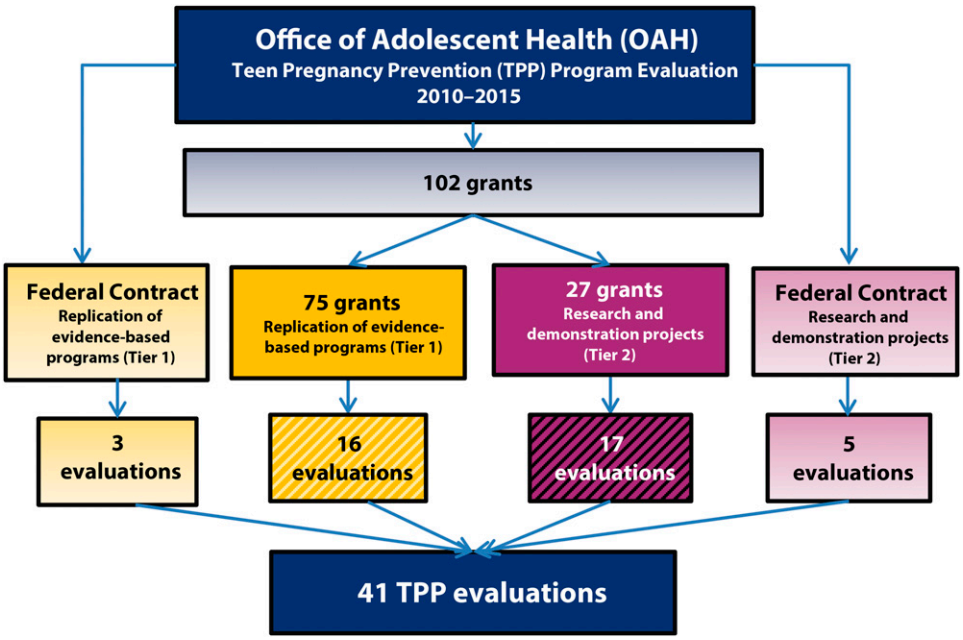



FIGURE 1—Flow of Teen Pregnancy Prevention Program Grants and Contracts Into Evaluations

## Performance Measures

OAH required all grantees to collect performance measure data and report on them twice a year. These data provided OAH with regular updates about the performance of individual grantees and the TPP program overall, including the number and types of people served, the quality of program implementation, and the dissemination of program results. Program implementation data included fidelity to the model and quality of implementation, both monitored and rated by an independent observer using program fidelity logs and an OAH-provided rating scale for implementation quality. The data also included mean attendance and data on the percentage of youths who attended at least 75% of the program sessions. All performance measure data were collected for individual youths by individual program sessions, for the duration of the program implementation, and reported to OAH at the individual level into a Web-based data repository system. These data allowed OAH to judge the quality of implementation in regard to sample size, fidelity to the program model, and participant dosage.

## Review and Synthesis Method

We reviewed and synthesized information from the final evaluation reports, including the study quality assessments of OAH's evaluation training and technical assistance contractor and the HHS evidence review, along with the performance measure data reported by the evaluation grantees. Two individuals coded each evaluation into one of three categories, and disagreements were discussed and resolved between coders. Coders disagreed on the classification of three evaluations, but the disagreements were resolved through discussion. Using "evaluation" as the unit of analysis, data were used to place evaluations into one of the following three categories:

- **Implementation and evaluation were strong; statistically significant positive behavioral impacts were found.** Evaluations placed in this category are characterized by high fidelity and quality of implementation, and good attendance. The evaluations meet the standards for research quality of the HHS Evidence Review including low or corrected-for attrition, baseline equivalence, strong contrast, and no confounding factors. The evaluations were appropriately powered to detect the intended impacts. And finally, a statistically significant positive impact was demonstrated on at least one of the Evidence Review behavioral outcomes. Studies in this category contribute to the body of evidence in TPP by identifying where, when, and for whom programs are effective.
- **Implementation and evaluation were strong; statistically significant positive behavioral impacts were *not* found.** Evaluations placed in this category have similar high quality implementation and rigorous evaluation as defined in the previous category. However, null or negative impacts were demonstrated on the Evidence Review behavioral outcomes. Studies in this category also contribute to the body of evidence by identifying where and with whom programs were not found to be effective.
- **Inconclusive; implementation and evaluation experienced challenges.** Evaluations placed in this category experienced challenges with either program implementation, quality of the evaluation, or both. Evaluations in this category do not provide confidence in their findings (or lack of findings) because of these challenges and should not be interpreted as a true test of the program model. In this category, challenges with implementation include poor fidelity, contrast, quality of implementation, and attendance, making it almost impossible to detect an effect of the intervention. Evaluation challenges in this category include very small sample sizes (not sufficiently powered to reasonably detect impacts), lack of sexual activity among the sample at follow-up, and analyses or reporting that do not meet the standards for research quality defined by the

HHS Evidence Review (did not correct for attrition, did not have baseline equivalence, matching procedure was not strong). Studies in this category do not contribute to the body of evidence other than to indicate further research is needed.

## RESULTS

This synthesis summarizes findings across the TPP evaluations according to the two research questions presented previously.

### Evaluation Designs

All of the program evaluations were rigorous designs; 37 (90%) were RCTs, and four were rigorous quasi-experimental designs (QEDs; Table 2). Twenty-two evaluations used cluster-level random assignment, and 15 used individual-level random assignment. Tier 1 consisted of seven individual-level RCTs, 10 cluster RCTs, and 2 QEDs. Tier 2 consisted of eight individual-level RCTs, 12 cluster RCTs, and two QEDs. Forty-nine percent of the evaluations were conducted in a school setting (during or after school), 20% in community-based organizations, 7% in clinics, and 5% online (Table 2). Fewer than half of the evaluations provided a program to the control group, examples include health and nutrition classes, college or career training, safe driving, and mentoring. Most (53%) of the evaluations compared their program to "business as usual." Business as usual ranged from no other sexual or reproductive health education, to fairly generous sexual or reproductive health education. Evaluations were conducted with a fairly even split of participants in middle school (29%), high school (29%), and high school and older (24%), and a smaller proportion spanning both middle- and high-school (17%; Table 2). The majority of evaluations examined abstinence or sexual activity (73%) and condom or contraceptive use (80%). Pregnancy (22%) and frequency of sex (20%) were also common behavioral outcomes and a small number of evaluations examined STI rates and number of sexual partners (Table 2).

### Program Features

The TPP Program served over 500 000 youths aged 10 to 19 years between 2010 and 2015 in 39 states and the District of Columbia. The majority of participants were aged 14 years or younger (74%), 18% were aged 15 to 16 years, and 8% were aged 17 years or older. The participants were diverse: 37% were Latino, 33% were Black, and 30% were White or other. During cohort 1, more than 6100 new facilitators were trained and more than 3800 new community partnerships were established. Ninety-five percent of all TPP sessions were implemented with high fidelity (as intended), and 92% of all sessions observed by an independent facilitator (n = 30 010) were rated as either very high or high quality. On average, youths attended 86% of all sessions, and 83% of youths attended at least 75% of all program sessions.

### Additional Evidence on the Evidence-Based Models

Of the 19 replication evaluations, four evaluations were strong implementations and evaluations, and found statistically significant positive behavioral impacts measured by the HHS evidence review (see the "replicated" box in Figure 2). All met the HHS evidence review standards for a moderate or high rating for research quality, and our review of their implementation data indicated a strong contrast between treatment and control groups, high fidelity, high quality, and high dosage. The four studies included one evaluation each of the Carrera Program, Reducing the Risk, Safer Sex Intervention, and Teen Outreach Program (TOP) and demonstrated reductions in sexual initiation, sexual intercourse in the last 90 days, sexual intercourse without birth control, and pregnancies.

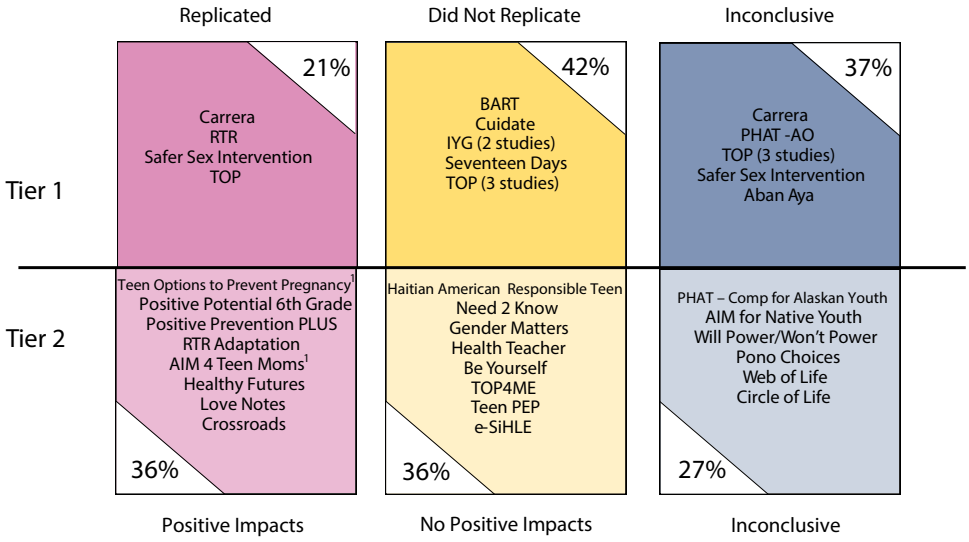
Eight of the 19 evaluations were strong implementations and evaluations but found no positive impacts on the behavioral outcomes measured (see the "did not replicate" box in Figure 2). All eight met the HHS evidence review standards for a moderate or high rating for research quality, and our review of their implementation data indicated a strong contrast between treatment and control groups, high fidelity, high quality, and high dosage. They included evaluations of the evidence-based TPP programs Becoming a Responsible Teen, Cuidate!, Seventeen Days, two evaluations of It's Your Game, and three evaluations of TOP. Seven of these evaluations demonstrated null findings. One of these evaluations demonstrated a null impact at the end of program implementation and a negative impact one year after the program. Further investigation indicated that the control group received another evidence-based pregnancy prevention program that the treatment group did not during the year after the program.

Seven of the 19 evaluations were coded into the inconclusive category (see the "inconclusive" box in Figure 2). While none of these evaluations demonstrated positive behavioral impacts, all seven had challenges that resulted in an invalid test of the program. These included evaluations of Carrera, Promoting Health Among Teens! Abstinence-Only, three studies of TOP, Safer Sex Intervention and Aban Aya. These evaluations suffered from low program attendance, lack of contrast between treatment and control groups, low rates of sexual activity in the sample,

**TABLE 2—Number and Proportion of Evaluations of Office of Adolescent Health Teen Pregnancy Prevention Programs: United States, 2010–2015**

| Characteristic | Evaluations, % (No.) |
|---|---|
| Evaluation design | |
| Randomized cluster | 54 (22) |
| Randomized youths | 37 (15) |
| Quasi-experimental | 10 (4) |
| Setting | |
| During school | 39 (16) |
| Community-based | 20 (8) |
| Multiple settings | 20 (8) |
| After school | 10 (4) |
| Clinic | 7 (3) |
| Online | 5 (2) |
| Target population | |
| High school only | 29 (12) |
| Middle school only | 29 (12) |
| High school or older | 24 (10) |
| Middle and high school | 17 (7) |
| Outcomes measured | |
| Condom/contraceptive use | 39 (33) |
| Abstinence/sexual activity | 36 (30) |
| Frequency of sex | 10 (8) |
| Pregnancy | 10 (8) |
| Number of sexual partners | 4 (3) |
| STI rates | 2 (2) |

*Note.* STI = sexually transmitted infection.




*Note.* AIM = Adult Identity Monitoring; BART = Becoming a Responsible Teen; IYG = It's Your Game; PEP = Prevention Education Program; PHAT-AO = Promoting Health Among Teens! Abstinence-Only; RTR = Reducing the Risk; SiHLE = Sisters, Informing, Healing, Living, Empowering; TOP = Teen Outreach Program; TOP4ME = TOP Plus Text Message Enhancement.

[a]ACF-funded PREIS Grants that are sites in the Office of Adolescent Health –funded pregnancy prevention approach evaluation.

**FIGURE 2—Evaluations of Teen Pregnancy Prevention Program Models by Tier, Program Model, and Ability to Demonstrate Positive Behavioral Impacts**

and in one case, the analyses were not conducted in a manner consistent with the HHS Evidence Review criteria for research quality.

## Demonstrations of New Programs

Out of the 22 tier 2 research and demonstration evaluations, eight were considered strong implementations, strong evaluations, and demonstrated statistically significant positive impacts on behavioral outcomes measured by the HHS evidence review: AIM 4 Teen Moms, Crossroads, Healthy Futures, Love Notes, Positive Prevention PLUS, Positive Potential 6th grade, an adaptation of Reducing the Risk, and Teen Options to Prevent Pregnancy (see the "positive impacts" box in Figure 2). These programs demonstrated impacts on sexual initiation, recent sex, sex without condoms or birth control, and pregnancies. The programs in this category represent a diverse range including relationship education, youth development, a program for youths at risk for dropping out of school, two programs for adolescent mothers, and three sexuality education programs.

Eight of the evaluations were coded as strong implementations and strong evaluations, yet found no statistically significant positive impacts on the measured behaviors (see the "no positive impacts" box in Figure 2). These programs represent a mix of types, including general health education, peer education, sexuality education, youth development, programs designed for Haitian-American and Native American youths, and a program to address gender norms.

Six of the tier 2 program evaluations were coded as inconclusive (see the "inconclusive" box in Figure 2). These evaluations suffered from high attrition, samples with very low rates of sexual activity at follow-up (1.3%, 2%, 10%), weak contrasts between treatment and control groups, and in one case, the analyses were not conducted in a manner consistent with the HHS Evidence Review criteria for research quality. Unfortunately, most of the evaluations identified as inconclusive from tier 2 were conducted with special populations—Alaska natives, Hawaiian youths, and tribal youths.

## DISCUSSION

The results from OAH's TPP Program evaluations have made a significant contribution to the field by helping to build a body of evidence for individual evidence-based programs and greatly expanding information about program effectiveness (see Goesling,[8] in this issue, for discussion of how the TPP impact findings contribute to the evidence base). In a review of the evidence for the 20 years prior to the first cohort of the TPP Program, the HHS Evidence Review identified 31 programs as effective in reducing pregnancies, STIs, and associated sexual risk behaviors.[2] In 2010 through 2015, the OAH TPP Program provided an additional 41 rigorous, independent studies that yielded eight new TPP programs with evidence of effectiveness, identified additional settings and populations where four evidence-based programs are effective, and contributed further data on 10 previously identified evidence-based TPP programs. The number of evaluations demonstrating statistically significant positive impacts on behavioral outcomes represents a larger proportion than found in large evaluation efforts from other fields.[2–7]

Most importantly, the results from these evaluations provide information about where, when, and with whom programs are effective, which is critical for communities to make informed decisions about which programs are the best fit for them. We should not expect any one program to be a magic bullet, effective with anyone, anywhere. The TPP "replications" were simply defined as delivering the program model as it was originally designed and evaluated. The majority of the TPP evaluations were conducted in new settings, with new populations and new evaluation designs compared with the original studies. The TPP Program sought to provide more information about the effectiveness of these programs to aide communities in choosing the most effective program for their population's needs. As Valentine et al. first noted,[9] with more

high-quality replications and advanced methods for reviewing the multiple studies, prevention science will be in a better position to positively impact public health.

## Failure to Replicate Does Not Mean Original Study Was Wrong

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. The landscape of adolescence is constantly changing. An evidence base needs to be dynamic and flexible to continue to meet the needs of its target population. Furthermore, we need to continue to evaluate our target populations to ensure that programs resonate, are engaging, and continue to be effective.

## Fit, Implementation, and Evaluation Are Essential

From our synthesis of the TPP Program evaluations, we conclude that communities need to spend more time selecting programs that are the best fit and ensuring quality implementation. It is important that decisions about which programs to implement be driven by community needs, organizational capacity, and intended outcomes to ensure that the programs selected are a good fit, thereby increasing the likelihood that the program will be effective. Ensuring high-quality implementation, that participants receive the full dosage of the program, and that staff have the comfort, capacity, and skills to implement the program well are also critical to enhancing program effectiveness.

## Existing Models and Measures Should Be Revisited

As expected, the results of the evaluations across the TPP Program are mixed. As our synthesis demonstrates, evaluations of the same program model sometimes produced both positive and null impacts. In attempting to understand these findings, moving beyond impacts on behavioral outcomes to the how and the why these programs produced findings, or did not, may require revisiting their logic models and underlying theories. In other cases, the measures or instruments being used to evaluate the interventions may have led to mixed findings. Many of the TPP evaluations saw positive impacts on measures such as knowledge and attitudes; however, these findings did not translate into positive behavioral changes. As we try to interpret the meaning of the findings across the program, we will need a better understanding of these issues. See Jenner et al.,[10] in this issue, for a deeper discussion.

## New Strategies to Evaluate Hard-to-Reach Populations

Five of the six tier 2 programs put in the inconclusive category were program evaluations conducted in special populations. Sample sizes, attendance, mobility, and navigating tribal or institutional review boards proved challenging for these evaluations. We believe that individual studies are not enough to move the field forward with respect to special populations. Instead, a group of coordinated studies that pool data for analysis and take advantage of the benefits that recent technology and methods afford us may be more efficient when working with special populations. See Kaufman et al.,[11] in this issue, for a discussion of the challenges of working with tribal youth populations.

## Alternative Behavioral Measures for Younger Youths

Lack of sexual activity was a common issue encountered by the TPP Program evaluations. Many of the evaluations were of younger youths with low if not nonexistent rates of sexual activity, making it almost impossible to detect an effect of the intervention on that outcome. While short-term follow-up surveys are beneficial for measuring program impacts on key mediating outcomes such as skills, attitudes, and intentions, longer-term follow-ups are often better for measuring program impacts on behaviors that can take longer to emerge.[2] The behavior outcomes measured by the HHS evidence review—condom or contraceptive use—may be too far down the road for today's adolescents to be detected by an 18- to 24-month follow-up. To continue to build an evidence base for younger, nonsexually active youths, we need to identify behaviors that occur prior to the initiation of sexual activity that are predictive of sexual initiation and risk-taking behavior. See Coyle and Glassman,[12] in this issue, for a discussion of strategies for measuring adolescent sexual activity.

## Cultural Shift Needed in Reporting and Publication Biases

Null and negative findings, most importantly in replication studies, are extremely important to report and publish.[9] A single research study cannot provide confidence of a program model's effectiveness when taken to scale in different settings and populations. A body of evidence is needed to determine when, where, and for whom programs are most effective (as well as which programs should be revisited and re-evaluated, or even walked away from). Recent articles have cautioned against ignoring null or inconsistent findings,[3,13] and instead argue to consider the results of replication studies and all of the trials before them[9] to determine whether the evidence has grown weaker or stronger. Positive findings from a single study does not indicate "it works," just as null findings from a single study should not be interpreted as "it doesn't work." Each study is another critical piece of evidence adding to the body of evidence to be considered when making decisions about which programs to implement where. See the editorial by Cole,[14] in this issue, for discussion of the importance of nonsignificant impact findings. AJPH

*Amy Feldman Farb, PhD*
*Amy L. Margolis, MPH, CHES*

**CONTRIBUTORS**
Both authors contributed equally to this editorial.

**REFERENCES**

1. Zief S, Knab J, Cole RP. A framework for evaluation technical assistance. *Am J Public Health*. 2016;106(suppl 1): S24–S26.

2. Goesling B, Colman S, Trenholm C, Terzian M, Moore K. Programs to reduce teen pregnancy, sexually transmitted infections, and associated sexual risk behaviors: a systematic review. *J Adolesc Health*. 2014;54(5): 499–507.

3. Tseng V. Evidence at the crossroads pt 11: the next generation of evidence-based policy. The William T. Grant Foundation. Available at: http://wtgrantfoundation.org/evidence-crossroads-pt-11-next-generation-evidence-based-policy. Accessed September 19, 2016.

4. Buck S. Important lessons about the reproducibility in science. The Laura and John Arnold Foundation. Available

at: http://www.arnoldfoundation.org/important-lessons-about-reproducibility-in-science. Accessed September 19, 2016.

5. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*. 2015;349(6251):943–951.

6. Coalition for Evidence-Based Policy. Randomized controlled trials commissioned by the Institute of Education Sciences since 2002: how many found positive versus weak or no effects. Available at: http://coalition4evidence.org/wp-content/uploads/2013/06/IES-Commissioned-RCTs-positive-vs-weak-or-null-findings-7-2013.pdf. Accessed September 19, 2016.

7. Manzi J. *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society*. New York, NY: Basic Books; 2012.

8. Goesling B. Informing The evidence base on adolescent pregnancy and sexually transmitted infections: important lessons. *Am J Public Health*. 2016;106(suppl 1):S7–S8.

9. Valentine JC, Biglan A, Boruch RF, et al. Replication in prevention science. *Prev Sci*. 2011;12(2):103–117.

10. Jenner E, Jenner LW, Walsh S, Demby H, Gregory A, Davis E. Adolescent pregnancy prevention programs and research: a time to revisit theory. *Am J Public Health*. 2016;106(suppl 1):S28–S29.

11. Kaufman CE, Schwinn TM, Black K, Keane EM, Big Crow CK. The promise of technology to advance rigorous evaluation of adolescent pregnancy prevention programs in American Indian and Alaska Native tribal communities. *Am J Public Health*. 2016;106(suppl 1):S18–S20.

12. Coyle KK, Glassman JR. Exploring alternative outcome measures to improve pregnancy prevention programming in younger adolescents. *Am J Public Health*. 2016;106(suppl 1):S20–S22.

13. Aschwanden C. *Failure Is Moving Science Forward*. FiveThirtyEight. March 24, 2016. Available at: http://fivethirtyeight.com/features/failure-is-moving-science-forward. Accessed September 19, 2016.

14. Cole RP. Comprehensive reporting of adolescent pregnancy prevention programs. *Am J Public Health*. 2016;106(suppl 1):S15–S16.

# Comprehensive Reporting of Adolescent Pregnancy Prevention Programs

What is the takeaway of a special issue of a journal that contains a number of small, nonsignificant impacts of adolescent pregnancy prevention programs on behavioral outcomes? The Office of Adolescent Health (OAH) funded a large number of evaluations to improve the evidence in the field—and presenting the entire body of results, including the nonsignificant findings, is good science. This is a collection of high-quality evaluations, with analyses and results that have been guarded against identifying spurious findings (*P*-hacking) as a result of prespecified analysis plans and multiple rounds of independent review.[1] Therefore, we can trust these impact results as credible estimates of program effectiveness, and they should become a part of the knowledge base for adolescent pregnancy prevention research.

Above and beyond providing funds to generate evidence of the effect of new programs and replication evidence of existing programs, OAH also funded comprehensive evaluation technical assistance support to these grantee-led studies to increase the likelihood of high quality, credible impact evaluations that showed statistically significant effects of the programs on behavioral outcomes.[2] The evaluation designs were strengthened through an initial review process, the analytic approaches were prespecified during an analysis plan review, the impact analyses were conducted with multiple sensitivity and robustness assessments to guard against potential error, and the final analyses and reporting underwent several rounds of independent review. Because of this evaluation technical assistance effort, these studies produced credible impact estimates of the effect of the programs, as implemented in each study setting.

## INGREDIENTS TO A STATISTICALLY SIGNIFICANT IMPACT

Despite these investments, many of the evaluations did not show favorable, statistically significant results on behavioral outcomes. One common interpretation of nonsignificant impacts for a study is that the program did not actually change participant behavior in a given setting. Another common interpretation is that the study did not have adequate power to state that the impact was significantly different from zero. It is important to remember that the statistical significance of an impact estimate can be operationalized as whether the difference in mean outcomes across conditions is approximately twice the standard error of the difference. In general, when this ratio is larger than a threshold (for most studies, a *t*-statistic of 1.96 is used to define a type I error rate of $\alpha = 0.05$), we state that the impact estimate is statistically significant. This means there are at least two reasons why an impact estimate might be nonsignificant:

1. The numerator (difference in mean outcomes across conditions) was too small. Several factors could contribute to this, including a weak contrast in experience across conditions because of a strong counterfactual, poor program attendance, implementation of the intervention with inadequate fidelity, or using outcomes that are not well aligned with the theory of change of the intervention (or outcomes unlikely to change, given the young age of the sample). Of course, another explanation is that the program is not effective in changing participant outcomes.
2. The denominator (standard error of the difference) was too large. Again, several factors could contribute to this, including smaller-than-expected sample sizes resulting from difficulties during recruitment or enrollment, or low response rates.

The evaluation technical assistance that grantees received guided them in improving aspects contributing to both the numerator and the denominator

**ABOUT THE AUTHOR**

*Russell P. Cole is a Senior Researcher at Mathematica Policy Research, Princeton, NJ. Correspondence should be sent to Russell P. Cole, Senior Researcher, Mathematica Policy Research. PO Box 2393, Princeton, NJ, 08543-2393 (e-mail: rcole@mathematica-mpr.com).*
*Reprints can be ordered at http://www.ajph.org by clicking the "Reprints" link.*

*This editorial was accepted June 19, 2016.*

*doi: 10.2105/AJPH.2016.303332*