ASPE RESEARCH BRIEF
OFFICE OF THE ASSISTANT SECRETARY FOR PLANNING AND EVALUATION
OFFICE OF HUMAN SERVICES POLICY - U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

## MAKING SENSE OF REPLICATION STUDIES

# Guidance for Teen Pregnancy Prevention Researchers

*To date, most teen pregnancy prevention programs have been evaluated only once, often in small-scale efficacy trials involving the program developer. In recent years, however, a growing number of studies have sought to test how these programs perform when implemented on a broader scale, in different settings, or with different populations. These replication studies have the potential to greatly advance the field of teen pregnancy prevention research and help sustain the recent drop in teen birth rates in the United States. Achieving these goals, however, will require a careful interpretation and synthesis of study findings that avoids overly simplistic notions of replication "failure" and "success."*

In the past 30 years, teen pregnancy prevention researchers have achieved notable success in identifying programs that can be effective in reducing teen pregnancy, sexually transmitted infections (STIs), and associated sexual risk behaviors. On the basis of findings from an ongoing systematic review of the teen pregnancy prevention research literature, the U.S. Department of Health and Human Services (HHS) currently recognizes more than 30 teen pregnancy prevention programs as having demonstrated evidence of favorable effects (Goesling et al., 2014). These programs range from short one-time clinical or counseling sessions to broad multi-year school curricula and youth development programs. Much of the supporting research evidence comes from rigorous randomized controlled trials, considered the "gold standard" in evaluation research.

But how do these programs perform when taken beyond their original research studies and implemented on a broader scale, in new settings, or with different populations? To date, most teen pregnancy prevention programs have been evaluated

## ABOUT THIS RESEARCH BRIEF

*This ASPE Research Brief was written by Brian Goesling of Mathematica Policy Research as part of an ongoing systematic review of the teen pregnancy prevention literature. Since 2009, Mathematica and its partner, Child Trends, have conducted the Teen Pregnancy Prevention Evidence Review under contract with HHS. The review aims to identify, assess, and rate the rigor of program impact studies on teen pregnancy and STI prevention programs. The review findings are available online at tppevidencereview.aspe.hhs.gov*

Office of the Assistant Secretary for Planning and Evaluation

Office of Human Services Policy

US Department of Health and Human Services

Washington, DC 20201

only once. Many of the studies are designed as small-scale "efficacy trials" (Flay et al., 2005) conducted in closely managed settings, often by the program developer. These efficacy trials are important for establishing initial evidence of a program's success. However, it is equally important to know how programs perform when implemented on a broader scale, with different populations, or in new settings (Valentine et al., 2011).

A growing number of studies have sought to address this gap. As early as 1998, the National Institutes of Health (NIH), recognizing a clear need for more information on program replication and dissemination, launched a broad initiative to identify and test replicable community-based HIV prevention programs for youth (Bell et al., 2007). The initiative funded a set of six studies designed to implement and test established HIV prevention programs with new populations or in new settings. More recently, the Office of Adolescent Health within HHS has funded more than a dozen ongoing replication studies as part of the federal Teen Pregnancy Prevention (TPP) grant program (Kappeler and Farb, 2014). These studies represent the single largest group of coordinated program replication studies ever undertaken in teen pregnancy prevention research. Study findings will be released on a rolling basis beginning in 2015.

This research brief provides practical guidance for making sense of this growing body of research. It is intended primarily for researchers, policymakers, and practitioners working in the field of teen pregnancy prevention research, to help them navigate and make best use of findings from the growing number of replication studies. The brief starts by discussing different definitions of the term "replication" and distinguishing different types of replication studies. The brief then identifies a series of practical steps researchers and other stakeholders can take to accurately interpret findings from replication studies. The guidance presented in this brief draws on input received from an HHS-sponsored expert meeting on replication held in December, 2013. The meeting convened a group of over 40 federal officials and research experts to discuss the issue of replication and methods for interpreting findings from replication studies. Most of the meeting participants have substantive and methodological expertise in fields outside teen pregnancy prevention, such as health services research, education, and criminal justice. The meeting sought to draw lessons from these neighboring fields to help inform the emerging issue of replication in teen pregnancy prevention research.

## Defining Replication

The term "replication" is used commonly across many fields of scientific research but often in different ways. Understanding what people mean by replication is an important but often overlooked first step in making sense of findings from replication studies. In psychology and medicine, researchers often use the term replication in the sense of reproducing findings from a formal lab experiment. In recent years, the issue of replication has become a major source of controversy in these fields, as researchers have often struggled to replicate findings from even simple lab experiments (Ioannidis, 2005; Pashler and Harris, 2012). Possible reasons for these struggles range from statistical chance to publication bias (i.e., a bias toward publishing results that are positive and statistically significant) or subtle differences in the methods used to carry out the experiments. In economics, researchers use the term replication more commonly in the context of secondary data analysis. Economists will try to re-create, or replicate, findings from published journal articles by conducting their own analyses of the same or different secondary

datasets (Hamermesh, 2007). In this context, replication failure often stems from subtle differences in analysis methods or computer programming code.

For teen pregnancy prevention research, a more directly relevant perspective on replication comes from prevention science research. Prevention scientists have defined replication as "an ongoing process of assembling a body of empirical evidence that speaks to the authenticity, robustness, and size of an effect" (Valentine et al., 2011, p. 104). By this definition, replication is an ongoing, dynamic process that involves compiling and synthesizing evidence across multiple studies. Unlike in psychology or medicine, these studies are not necessarily conducted in the controlled environment of a lab experiment. They are more often conducted in "real world" settings by different groups of researchers, in different settings, and with different populations. The studies share the goal of estimating a common effect. No one single study provides definitive evidence of the "true" effect. Rather, assessments of the effect build and change over time as new evidence emerges.

This definition provides a useful way to think about replication studies in teen pregnancy prevention research. The body of evidence for any one teen pregnancy prevention program begins with the findings from an initial supporting impact study. From this starting point, the process of replication involves expanding the available evidence to include findings from additional supporting studies. At minimum, these additional studies will involve testing program effects among new research samples. They may also involve larger samples, different settings, or different populations. Through this process, the body of evidence for any one program expands from the findings of a single supporting impact study to a synthesis of evidence from multiple studies.

### REPLICATING PROGRAMS OR REPLICATING EFFECTS?

In interpreting findings from replication studies, it is critical to distinguish (1) the replication of programs from (2) the replication of program *effects*. Replication of a program is often viewed in terms of fidelity to the program model. Was the program implemented in the same way as in the original research study or as intended by the program developer? If yes, the replication may be dubbed a success. However, another key question of interest is about the replication of program *effects*. Did the program produce similar *effects* to those reported in the original research study? To answer this question, it is necessary to have some level of consistency in program services or fidelity to the program model. However, it is also necessary to have consistency in study design and research methods. Two studies may examine the impacts of the exact same program implemented under very similar conditions, but if the study design or research methods differ in fundamental ways, there may be little basis for drawing firm conclusions about the replication of program effects. Consistency in study design and research methods can have great practical importance when interpreting findings from replication studies, as discussed in other sections of this brief.

## Types of Replication Studies

Researchers may conduct a replication study for many different reasons, and the motivation for the study can play an important role in interpreting the results. Valentine et al. (2011) usefully

distinguish between five types of replication studies, all falling within the common definition of replication provided in the previous section of this brief:

1. **Statistical replication.** The purpose of a statistical replication is to mirror the original study as closely as possible, in attempt to help rule out the possibility that any observed effects were found by statistical chance. A statistical replication may still require drawing a new sample of study participants. However, other features of the study should be designed to match the original as closely as possible. To date, this type of replication study has been uncommon in teen pregnancy prevention research.
2. **Generalizability replication.** With a generalizability replication, the purpose is to determine whether program effects found in the original study generalize to other target populations or settings. For example, in teen pregnancy prevention research, Stanton et al. (2005) conducted a type of generalizability replication in evaluating whether the *Focus on Kids* program, which was originally developed for youth in urban areas, could have similar impacts among rural youth. In a generalizability replication, the replication study should ideally match the original study in all respects except the one key variable of interest—for example, a new target population. In practice, however, it may be hard to change the target population without also making at least some minimal revision to the program or implementation characteristics.
3. **Implementation replication.** An implementation replication is designed to assess the sensitivity of program effects to some change in the way the program is implemented. For example, researchers may be interested in learning whether school teachers are more or less effective in delivering a classroom-based teen pregnancy prevention program than trained outside facilitators or health professionals. This question could be tested with an implementation replication by changing the mode of program delivery. As when conducting a generalizability replication, researchers should try to match the original study as closely as possible except for the one key variable of interest.
4. **Theory development replication.** The purpose of a theory development replication is to unpack the "black box" of program effects and test the pathways or mechanisms through which a program may work. For example, Coyle et al. (2013) conducted a type of theory development replication in an effort to unpack the effects of classroom-based instruction versus outside service learning activities that together make up the *All4You!* teen pregnancy, STI, and HIV prevention program. To isolate the theoretical question of interest, a theory development replication should ideally match the original study on all other key program and study features, such as target population and implementation characteristics.
5. **Ad hoc replication.** Unlike other types of replication studies, which involve a specific planned or intentional departure from the original study, ad hoc replications involve a mix of planned and *unplanned* deviations. For example, researchers may intentionally plan to evaluate a program with a new target population or in a new setting. However, during the course of the study, the researchers find that the quality of program implementation in the replication study also deviates from the original—for example, because the change in study setting also brought changes in program

staff. Ad hoc replications may deviate from the original study in any number of ways, some intentional and others by chance.

In teen pregnancy prevention research, it is likely that most existing or ongoing replication studies are either ad hoc replications or a mixture of the other types of replication. For example, the recent study of the *All4You!* program by Coyle et al. (2013) may be cited as an example of a theory development replication, because it intends to explore the pathways or mechanisms through which the program works. However, the study also involves elements of a generalizability replication (because the study sample and setting were slightly different) and an implementation replication (because of possible differences in the quality of program implementation). Such combinations of difference are not surprising when implementing programs outside the controlled environment of a formal lab experiment. However, these differences can also make it difficult to compare study findings or isolate possible reasons for any differences in results.

## Guidance for Interpreting Evidence

Findings from replication studies can be difficult to interpret. They have little substantive meaning when interpreted on their own, in isolation from other research. Making sense of the findings invariably requires a comparison or synthesis with results from other studies, and this type of comparison or synthesis is not always straightforward. This section of the brief aims to address this challenge by suggesting five practical guidelines or steps researchers and other stakeholders can take to accurately interpret findings from replication studies.

### 1. Set Realistic Expectations

In teen pregnancy prevention research, replication studies are not typically designed as exact statistical replications of the original study. They often involve different implementing organizations working with different program staff, with different study participants, and in different settings. In most cases, the studies are led by different groups of researchers using at least slightly different study designs, measures, and analysis methods.

If for only these reasons, no one should expect the findings of replication studies to exactly match those of the original research studies. The replication studies might show program impacts for a different set of outcomes, different analytic samples, or at different follow-up periods. Even if the general pattern of findings is similar, the exact magnitude of program impacts will likely differ. If researchers in the fields of psychology and medicine have trouble reproducing findings from controlled lab experiments, researchers operating in the much-less-well-controlled environment of teen pregnancy prevention research should not expect to meet the standard of a perfect match.

### 2. Compare Study Design and Program Characteristics

In setting realistic expectations for replication studies, it is useful to begin by comparing study design and program characteristics. Compared to the original study, did the replication study focus on a different target population? Was the study conducted in a different setting? Did the study involve any changes to the program or mode of implementation? Did the study use a similar research design and assess comparable outcome measures? Were the data analyzed in

comparable ways? In answering these questions, it may be useful to classify the replication study into one of the five different types of replication discussed previously in this brief (statistical, generalizability, implementation, theory development, or ad hoc). However, the comparison should also account for features of the study design and research methods, which are not specific to any one particular type of replication.

As an example, consider two ongoing replication studies of the *Children's Aid Society (CAS) Carrera Adolescent Pregnancy Prevention Program*. The *CAS Carrera* program is a broad, multi-year youth development program designed to reduce teen pregnancy and associated sexual risk behaviors. Findings from the original evaluation of the *CAS Carrera* program, conducted in New York City in the late 1990s, showed that, three years after the program started, female adolescents participating in the intervention were significantly less likely to report having been pregnant or being sexually active (Philliber et al., 2002). With funding from the Office of Adolescent Health TPP grant program, two replication studies of the *CAS Carrera* program are now underway—one in Chicago and one in Atlanta and surrounding areas of Georgia. Both studies are designed to examine program impacts on rates of youth sexual activity. However, because of the relatively young ages and risk profiles of the study samples, the studies may have limited ability to measure long-term program impacts on teen pregnancy rates. Provided that other features of the program implementation and evaluation design are comparable with the original study, these replication studies will provide a basis for comparing or synthesizing evidence of the effects of *CAS Carrera* on rates of youth sexual activity. However, they may provide a more limited basis for assessing the replication of program effects on the longer-term outcome of teen pregnancy rates.

### 3. Look Beyond Reported Statistical Significance Levels

The statistical significance levels (*p*-values) reported in any one particular study have little relevance for the issue of replication. In most cases, the statistical significance levels reported in a study reflect a test of the estimated program impact against a null hypothesis of no program effect. This significance level is determined by the magnitude of the reported impact estimate, sample size, and degree of variability in outcomes within a single study. The issue of replication, however, implicitly or explicitly involves comparing or synthesizing program effects *across* different studies. The statistical significance levels reported in any single study are generally not designed to address such cross-study comparisons, and there is little value or meaning in comparing reported *p*-values from one study to another (Gelman & Stern, 2006).

A simple example illustrates the limitations of using reported statistical significance levels to address the issue of replication. Suppose a teen pregnancy prevention program has been evaluated in three independent studies (see Table 1 on next page). The first study found that the program reduced rates of sexual activity by 10 percentage points. With a sample size of 500 participants, this reported impact is statistically significant at the standard 5-percent level (*p*-value = .0147). The second study also found a program impact on rates of sexual activity of 10 percentage points. However, because the sample size for this study was smaller (300 participants), the reported impact estimate is *not* statistically significant (*p*-value = .0588). The third study found a smaller program impact of 5 percentage points. The sample size for this study was much larger than for the others (1,500 participants) so the reported impact estimate reaches statistical significance at the standard 5-percent level (*p*-value = .0387).

**Table 1. Avoid Using Statistical Significance Levels to Study Replication**

| Study | Control Group | Treatment Group | Program Impact | Sample Size | *p*-value | Statistically significant? |
|---|---|---|---|---|---|---|
| Study 1 | 35% | 25% | 10% | 500 | .0147 | Yes |
| Study 2 | 35% | 25% | 10% | 300 | .0588 | No |
| Study 3 | 35% | 30% | 5% | 1,500 | .0387 | Yes |

In this example, comparing statistical significance levels would lead to the counterintuitive finding that Study 3 had successfully replicated the original study (Study 1) but that Study 2 had not. The example is constructed so that the counterintuitive findings result from differences in sample sizes across the three studies. However, even in the context of studies with equal sample sizes, the statistical significance levels reported in each individual study are not meaningfully compared. As Gelman and Stern (2006) put it, when comparing different effect estimates, the difference between "significant" and "non-significant" is not itself significant.

## 4. Align Analysis Methods with Context

If comparing statistical significance levels is misleading, then what is a better approach? The answer depends both on the research questions of interest and the number of studies under review. In many cases, the research question of interest is whether the estimated program effect reported in a replication study is similar to or different from the effect reported in prior research, or whether the combined body of evidence across all relevant studies suggests a positive program effect. The methods of answering these questions may be different in the context of a single replication study than when multiple replication studies are available.

With only a single replication study available, it may be enough to conduct a simple descriptive comparison of the reported magnitude of program effects. As discussed earlier in this brief, making such comparisons requires some level of consistency across studies. The estimates of program effects must be measured on a comparable metric, pertain to similar outcome measures, and derive from comparable analysis methods. To determine whether reported effect sizes (i.e., the magnitude of a program effect for a particular outcome) are similar across studies, one simple approach is to assess whether the effect size for the replication study lies within the confidence interval of the program effect reported in the original study (Valentine et al., 2011). As an alternative, effect sizes from both studies could be compared to some external benchmark or minimum effect size deemed clinically or practically important. Effect sizes provide useful information beyond whether group differences are statistically significant, since statistical significance does not indicate the practical size of the reported difference and it is relatively sensitive to sample size.

With multiple replication studies available, it becomes most sensible to statistically combine or average effect size estimates across studies. Depending on the context, the synthesis methods could range from taking a simple average across studies (What Works Clearinghouse, 2014) to conducting a more formal meta-analysis (Valentine et al., 2011). For any synthesis method, the effect size estimates must be comparable across studies. The synthesis should also account for any differential weighting or prioritization of studies (discussed below). The results will provide a summary estimate of program effectiveness accounting for the full body of evidence available for the particular program in question.

### 5. Prioritize Studies or Assign Weights

For practical purposes, the findings of replication studies are often used in part to identify programs that should be considered for broader dissemination. If the weight of the combined evidence from the original study and replication studies suggests that the program has favorable effects, policymakers and program funders may be interested in increasing the scale of the program or delivering it in new settings. If, however, the weight of the combined evidence suggests little or no favorable effects, there may be less support for broader dissemination.

In making these determinations, it is not necessary or even preferable to give every study of the program equal weight. Put another way, the combined body of evidence for a program should *not* be determined by simply counting up the total number of studies with positive or negative effects, or by taking a simple average of estimated effects without regard to other features of the program implementation or study design and quality.

Instead, before synthesizing the evidence, it is important to first determine whether the studies under review all have equal relevance to the practical question or decision at hand. For example, in some cases, the research quality of the replication study may be higher (or lower) than that of the original study. One study may involve a rigorous randomized controlled trial whereas the other is based on a quasi-experimental design. Other things equal, findings from the randomized controlled trial may be given more weight. In other cases, the replication study may more closely reflect how the program would be implemented if later disseminated on a broad scale. In these cases, findings from the replication study may weigh most heavily in synthesizing the body of evidence and making decisions about broader dissemination.

There is no one single rule in determining whether and how different studies should be prioritized or weighted. Rather, the decision must be made on a case-specific basis accounting for the features of each study and ultimate use of the resulting evidence. For example, if the ultimate use of the evidence is to make decisions about broader dissemination, there may be reason to give most or all weight to those studies that most closely approximate the conditions of future dissemination efforts. However, this perspective assumes the studies are of equal quality and that the conditions of future dissemination efforts are known. In practice, there will likely be at least some minimum difference in study quality and uncertainty about the conditions of future dissemination efforts.

## Summary and Conclusion

The growing focus on replication is clearly a positive development for the field of teen pregnancy prevention research. It will help strengthen the body of evidence in support of individual programs. It will help advance current understanding of the conditions under which programs have positive effects. More broadly, it will support ongoing efforts to identify and support evidence-based approaches to teen pregnancy prevention, which in turn can help sustain the recent drop in teen birth rates in the United States.

Achieving these goals, however, will require a careful interpretation and synthesis of replication study findings that avoids overly simplistic notions of replication “failure” and “success.” Researchers, policymakers, and practitioners must have a clear sense of what they mean by

“replication” and their motivations for conducting replication studies. They must be realistic in what to expect from replication studies and avoid setting too high of standards for replication success. Interpretations of study findings must look beyond reported statistical significance levels and account for differences in study design and program characteristics. Most importantly, replication must be viewed as an ongoing, dynamic process in which no one single study provides definitive evidence of “true” program effects. From this perspective, the hard work of replication in teen pregnancy prevention research has just begun.

## References


Bell, Stephanie, Susan F. Newcomer, Christine Bachrach, et al. "Challenges in Replicating Interventions." *Journal of Adolescent Health*, vol. 40, no. 6, 2007, pp. 514-520.

Coyle, K.K., J.R. Glassman, H.M. Franks, S.M. Campe, J. Denner and G.M. Lepore. "Interventions to Reduce Sexual Risk Behaviors Among Youth in Alternative Schools: A Randomized Controlled Trial." *Journal of Adolescent Health*, vol. 53, no. 1, 2013, pp. 68-78.

Flay, Brian, A. Biglan, R.R. Boruch, et al. "Standards of Evidence: Criteria for Efficacy, Effectiveness, and Dissemination." *Prevention Science*, vol. 6, 2005, pp. 151-175.

Gelman, Andrew and Hal Stern. "The Difference Between 'Significant' and 'Not Significant' is Itself Not Statistically Significant." *The American Statistician*, vol. 60, no. 4, 2006, pp. 328-331.

Goesling, Brian, Silvie Colman, Christopher Trenholm, Mary Terzian, and Kristin Moore. "Programs to Reduce Teen Pregnancy, Sexually Transmitted Infections, and Associated Sexual Risk Behaviors: A Systematic Review." *Journal of Adolescent Health*, vol. 54, no. 5, 2014, pp. 499-507.

Hamermesh, Daniel S. "Replication in Economics." *Canadian Journal of Economics*, vol. 40, no. 3, 2007, pp. 715-733.

Ioannidis, John P.A. "Why Most Published Research Findings are False." *PLoS Medicine*, vol. 2, no. 8, 2005, p. e124.

Kappeler, Evelyn M. and Amy Feldman Farb. "Historical Context for the Creation of the Office of Adolescent Health and the Teen Pregnancy Prevention Program." *Journal of Adolescent Health*, vol. 54, 2014, pp. S3-S9.

Pashler, Harold and Christine R. Harris. "Is the Replicability Crisis Overblown? Three Arguments Examined." *Perspectives on Psychological Science*, vol. 7, no. 6, 2012, pp. 531-536.

Philliber, S., J. W. Kaye, S. Herrling and E West. "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.

Stanton, B., J. Guo, L. Cottrell, J. Galbraith, X. Li, C. Gibson, et al. "The Complex Business of Adapting Effective Interventions to New Populations: An Urban to Rural Transfer." *Journal of Adolescent Health*, vol. 37, no. 2, 2005, pp. 163e17–163e26.

Valentine, Jeffrey C., Anthony Biglan, Robert F. Boruch, Felipe González Castro, Linda M. Collins, Brian R. Flay, Sheppard Kellam, Eve K. Mościcki, and Steven P. Schinke. "Replication in Prevention Science." *Prevention Science*, vol. 12, 2011, pp. 103-117.

What Works Clearinghouse. "Procedures and Standards Handbook, Version 3." Washington, DC: Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, U.S. Department of Education, March 2014.