MAY 2015

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

IMPROVING THE RIGOR OF QUASI-EXPERIMENTAL IMPACT EVALUATIONS

Lessons for Teen Pregnancy Prevention Researchers

Quasi-experimental evaluations of social and public health interventions often face a risk of selection bias—the chance that observed program impacts reflect underlying differences between study participants in the intervention and comparison groups, rather than a true effect of the program. This research brief highlights three ways to reduce the risk of selection bias and thereby improve the rigor of quasi-experimental impact evaluations, focusing specifically on evaluations of teen pregnancy prevention programs.

Randomized controlled trials are often considered the "gold standard" for evaluating the impacts of social and public health interventions. In these studies, individuals or groups of individuals are randomly assigned to a group that receives the intervention ("intervention group") or a group that does not ("control group"). Random assignment helps ensure that, were it not for the intervention, the intervention and control groups would have similar outcomes on average except by chance. In well-executed randomized controlled trials, the differences in outcomes for the intervention and control groups can be attributed with confidence to the intervention.

When randomized controlled trials are not feasible, researchers sometimes consider quasi-experimental comparison group designs as a next best approach for estimating intervention effects. Studies using these designs compare outcomes for an intervention group and a non-randomly selected comparison group that is used to represent the counterfactual or control condition. Because the groups are formed through a non-random process, the differences in the outcomes of the groups may be due to the intervention, or they may reflect differences in other observed or unobserved characteristics between the two groups. Consequently, quasi-

ABOUT THIS RESEARCH BRIEF

This ASPE Research Brief was written by Brian Goesling and Joanne Lee 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.

Office of the Assistant Secretary
for Planning and Evaluation

Office of Human
Services Policy

US Department of Health
and Human Services

Washington, DC 20201

experimental studies generally have a greater risk of selection bias than randomized controlled trials.

This brief highlights three ways to reduce the risk of selection bias and thereby improve the rigor of quasi-experimental impact evaluations, focusing specifically on evaluations of teen pregnancy prevention programs. A growing body of methodological research from outside the field of teen pregnancy prevention suggests that, under certain circumstances, quasi-experimental comparison group designs can effectively approximate the impact estimates produced by rigorous randomized controlled trials (Cook et al., 2008; Shadish et al., 2008; Steiner et al., 2010). However, this measure of success does not come easily, and teen pregnancy prevention research has yet to fully benefit from the latest methodological insights about quasi-experimental comparison group designs. This brief suggests several practical steps teen pregnancy prevention researchers can take to improve the rigor of quasi-experimental studies of teen pregnancy prevention programs.

Lesson 1: Be Strategic in Choosing a Comparison Group

The overarching goal of a quasi-experimental comparison group design is to approximate the findings of a true experiment, usually by selecting intervention and comparison groups that are as similar as possible. By minimizing any differences between the intervention and comparison groups, researchers can have more confidence that any observed differences in outcomes reflect true impacts of the intervention. Selecting a well-matched comparison group is thus one of the best ways to improve the overall quality and rigor of a quasi-experimental comparison group design.

To date, teen pregnancy prevention researchers have often selected comparison groups on the basis of data availability or logistical constraints. For example, some studies have formed comparison groups using publically available national- or state-level data from surveys such as the Youth Risk Behavioral Surveillance System or the National Survey of Family Growth. Studies have also made comparisons to national- and state-level teen birth rates reported in the National Vital Statistics System. Other studies have collected primary data from a “self-selected” or volunteer comparison group—for example, a group of individual youth or schools that declined to participate in the intervention but agreed to provide data for the purpose of comparison.

Recent methodological research suggests that it pays to be more selective in choosing a comparison group. For example, studies show that researchers can usually reduce the risk of selection bias by collecting primary data for both the intervention and comparison groups from the same local community, rather than relying on existing state- or national-level estimates to provide the comparison data. Drawing primary data from the same local community can help minimize potential differences between the intervention and comparison groups. Studies also suggest that it helps to form the intervention and comparison groups on the basis of observed, measurable characteristics. For example, rather than allowing individuals or groups of youth to self-select or volunteer for placement in the comparison group, researchers could form the groups on the basis of known risk factors or other observed characteristics. Forming the groups on the basis of observed, measureable characteristics helps to quantify the risk of selection bias and avoid the potential for other, unmeasured characteristics to bias the study results.

Lesson 2: Collect More and Better Data

The relative strength of a quasi-experimental comparison group design also depends critically on the amount and quality of data collected from the intervention and comparison groups. For example, in education research on interventions to improve student test scores, studies show that measuring and controlling for prior student test scores can greatly reduce the risk of selection bias in quasi-experimental comparison group designs, because prior test scores explain much of the variation in future test scores (Cook et al., 2009; Fortson et al., 2012). By contrast, measuring only student demographic characteristics generally does little to reduce the risk of selection bias in quasi-experimental studies of education interventions, because demographic characteristics are relatively weaker predictors of future student achievement than prior student test scores (Cook et al. 2008). Alternatively, studies show that measuring and controlling for the mechanism used to form the intervention and comparison groups can also greatly reduce the risk of selection bias in quasi-experimental comparison group designs.

For quasi-experimental studies of teen pregnancy prevention programs, these findings suggest the need to collect more and better data from both the intervention and comparison groups. Teen pregnancy prevention researchers commonly collect data on basic demographic characteristics such as age, race, and gender. In some cases, they may also collect information on pre-intervention health status or associated risk behaviors, if possible. However, these commonly measured characteristics typically explain only a limited amount of variation in the targeted outcome measures. For example, using longitudinal data from the National Longitudinal Survey of Youth 1997 (NLSY97), we found that demographic characteristics such as age, gender, and race typically explain less than 10 percent of the variation in the types of outcomes commonly targeted by teen pregnancy prevention interventions (see the text box on the following page). To effectively reduce the risk of selection bias, teen pregnancy prevention researchers must expand their efforts and collect data for a broader range of sample characteristics. The collected measures should extend beyond basic demographic characteristics to include at least a minimum number of baseline risk characteristics, as well as any other measures that capture the selection process used to form the research groups or that may be strongly predictive of the targeted outcome measures.

Lesson 3: Recognize the Strengths and Limits of Analytic Methods

In quasi-experimental studies of teen pregnancy prevention programs, researchers typically use a regression framework to estimate program impacts on youth outcomes. This framework allows researchers to statistically adjust for baseline covariates and any observed differences in characteristics between the intervention and comparison groups. A regression framework can also accommodate many different types of outcome measures—for example, continuous, binary, or count outcomes—through the selection of an appropriate model. In some cases, researchers may also use propensity score matching or other matching techniques to improve the similarity of the intervention and comparison groups.

HOW PREDICTIVE ARE COMMONLY MEASURED COVARIATES?

To determine the potential for measures of demographic characteristics and other commonly measured covariates to reduce the risk of selection bias in quasi-experimental studies of teen pregnancy prevention interventions, we used data from the National Longitudinal Survey of Youth 1997 (NLSY97) to examine correlations between commonly measured covariates and commonly targeted outcome measures. The results showed that commonly measured covariates explain only a limited amount of variation in the targeted outcomes, which limits their ability to reduce the risk of selection bias.

The following table shows variance explained statistics from a series of regression models that assess how the predictive power of the models changes with the addition of more covariates.

	Outcomes				
Covariates	Ever had sex	Had sex without birth control in last year	Ever pregnant (females)	Number of times had sex in last year	Number of partners
Age	0%	0%	0%	2%	2%
+ Gender	0%	1%	n/a	3%	4%
+ Race/ethnicity	1%	1%	2%	3%	8%
+ Baseline measure of outcome or sexual activity	n/a	7%	14%	22%	69%
+ Frequency of cigarette use	4%	10%	15%	24%	70%
+ Frequency of marijuana use	5%	11%	15%	24%	71%
+ Frequency of drinking alcohol	6%	11%	15%	24%	71%
+ Peer substance use	6%	12%	18%	25%	71%
+ Socioeconomic status	6%	12%	19%	25%	71%
+ Percent of peers who have had sex and expectations for pregnancy in next year	7%	13%	22%	26%	72%
+ Expectations for attaining a college degree	7%	14%	24%	26%	72%

Source: Authors' calculations from Rounds 1 and 2 of the NLSY97.

These methods can play an important role in helping establish the rigor and credibility of program impact estimates from a quasi-experimental comparison group design. In particular, they can help assure readers that the reported program impact estimates do not reflect observed differences between the intervention and comparison groups. They can also help establish the credibility of the selected comparison group as an appropriate match for the members of the intervention group.

However, in addition to acknowledging these strengths, researchers must also recognize that common analytic methods such as regression adjustment and matching cannot overcome limitations in the underlying evaluation design or data collection methods. For example, regression adjustment or matching will not fully account for risk of bias introduced when drawing comparison group data from publically available national- or state-level source. In this case, the risk of bias is best addressed by identifying a different source for comparison group data, not through the application of a particular analytic method. Similarly, common analytic methods cannot account for the risk of bias introduced by using a self-selected or volunteer comparison group, unless the researchers also happened to collect data on an especially rich set of sample characteristics. Put another way, although the analytic methods used to estimate program impacts are important, they are usually a secondary concern to the overall evaluation design and amount and quality of data collected.

Summary and Conclusions

Not all quasi-experimental comparison group designs are created equal. A growing body of methodological research suggests that these designs have the potential to approximate the impact estimates produced by rigorous randomized controlled trials. However, this potential exists only in certain circumstances. To effectively minimize the risk of bias inherent in these designs, researchers must be strategic in choosing a comparison group that provides a good match for the intervention group. Researchers must have rich data on the characteristics of the intervention and comparison groups, and they must recognize that even the most sophisticated analytic methods cannot overcome underlying limitations in the evaluation data or design. By following these steps, researchers can improve the rigor of quasi-experimental impact evaluations of teen pregnancy prevention programs and maximize the contribution of these studies to the broader field.

References

Cook, T. D., W. Shadish, and V. C. Wong. “Three Conditions Under Which Observational Studies Produce the Same Results as Experiments.” *Journal of Policy Analysis and Management*, vol. 27, 2008, pp. 724–750.

Cook, T. D., P. M. Steiner, and S. Pohl. “How Bias Reduction is Affected by Covariate Choice, Unreliability and Mode of Data Analysis: Results From Two Types Of Within-Study Comparisons.” *Multivariate Behavioral Research*, vol. 44, no. 6, 2009, pp. 828-847.

Fortson, K., N. Verbitsky-Savitz, M. Kopa, and P. Gleason. “Using an Experimental Evaluation of Charter Schools to Test Whether Nonexperimental Comparison Group Methods Can Replicate Experimental Impact Estimates.” NCEE Technical Methods Report 2012-4019. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2012.

Shadish, W. R*.*, M. H. Clark*,* and P. M. Steiner*.* “Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random To Nonrandom Assignment*.*” *Journal of the American Statistical Association*, vol. 103, 2008, pp. 1334-1356.

Steiner, P. M., T. D. Cook, W. R. Shadish, and M. H. Clark. “The Importance of Covariate Selection in Controlling For Selection Bias in Observational Studies.” *Psychological Methods*, vol. 15, 2010, pp. 250-267.