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

of the statistical significance calculation. However, because of constraints such as fixed budgets, specialized populations of interest, and service-rich environments, among others, it was not always possible to make the necessary changes to the evaluations to have the ratio reach the necessary threshold used to determine statistical significance. That said, each study offers useful information, regardless of the constraints faced.

MAGNITUDE OF THE EFFECT

Although some studies in this supplement may have been underpowered because they did not hit recruitment and retention targets (and thus had a large standard error), it is also important to look carefully at the magnitude of the difference in means, as Goesling's editorial in this volume points out.[3] For most behavioral outcomes discussed in this supplement, the impact estimate (difference in prevalence rates or means) is relatively small—in most cases, the difference in prevalence rates for behavioral outcomes was less than five percentage points across conditions, and often, was markedly closer to zero.

Therefore, the issue of statistical power is less critical—the programs were not substantively changing participant behavior in these settings. That is, in the particular settings where these interventions occurred, there was only a small difference in the behavior across treatment and control groups, a finding that is independent of sample size and statistical power. The small differences in behavioral outcomes is likely a function of multiple issues including, but not limited to, having a strong counterfactual condition, poor attendance, inadequate implementation of the intervention with intended fidelity, or examining outcomes unlikely to differ across conditions (e.g., because of sample age or lack of alignment with the logic model of the intervention). Even if statistical power could have been improved, given the small impacts observed, the study would have required a massive increase in sample size for the results to be classified as statistically significant.[4] And thus, using the lens of statistical significance as a means to understand the substantive effect of these interventions is less informative than focusing on the observed magnitude of the difference between groups.

THE IMPORTANCE OF REPORTING AND TRANSPARENCY

The American Statistical Association recently released a policy statement on statistical significance. It stated recommendations for researchers and policymakers, with suggestions including: (1) authors should present effect sizes along with *P* values, and (2) that policy decisions should be made on information above and beyond whether a *P* value is below a given threshold. However, one other recommendation seems especially pertinent for this supplement: researchers should report on all tests conducted instead of just selectively reporting the statistically significant findings.[5] In the context of the OAH grant funding effort, this suggestion is analogous to ensuring that all of the findings from the funded impact evaluations are made available. That is, it is better science to disseminate the findings from all of the evaluations rather than cherry picking and highlighting the subset of evaluations that produced statistically significant findings.

This illustrates that the compendium of results in *AJPH* represents good science and an important contribution to the field: these are high-quality evaluations with full transparency in reporting. In particular, this type of journal issue is necessary to help overcome problems of publication bias. Studies with nonsignificant findings are less likely to be published than those with statistically significant results, leading to the file drawer problem.[6] By making the results of these studies with nonsignificant findings available, future meta-analyses can incorporate these results and ensure a more comprehensive understanding of the effects of adolescent pregnancy programs across settings. In addition, understanding both the successes and the failures observed in these evaluations may contribute lessons learned for developing and refining interventions for adolescent pregnancy prevention and youth development. AJPH

Russell P. Cole, PhD

ACKNOWLEDGMENTS

This work was conducted under a contract (HHSP233201300416G) with the Office of Adolescent Health within the Department of Health and Human Services.

REFERENCES

1. Gelman A, Loken E. The statistical crisis in science. *Am Sci*. 2014;102(6):460.

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

3. 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.

4. Moreno L, Cole RP. *Calculating Minimum Detectable Impacts in Teen Pregnancy Prevention Impact Evaluations*. Washington, DC: US Department of Health and Human Services, Administration on Children, Youth and Families, Office of Adolescent Health; 2014.

5. Wasserstein R, Lazar N. The ASA's statement on p-values: context, process, and purpose. *Am Stat*. 2016;70(2): 129–133.

6. Rosenthal R. The "file drawer problem" and the tolerance for null results. *Psychol Bull*. 1979;86(3):638–641.