Evaluation Technical Assistance Brief

SEPTEMBER 2020 • NUMBER 5

Russell Cole

Small samples due to lower-than-planned enrollment in impact evaluations: what to do?

Impact evaluations in child welfare and other fields often struggle because of smaller-than-planned sample sizes. Multiple factors might contribute to the problem: The program's target population might be smaller than was projected, or recruiting and enrolling eligible participants into the study might have proven unexpectedly difficult.

Small sample sizes can create difficulties and limitations when estimating the impacts of programs—especially when you had not planned for them during the evaluation design phase. This brief, presented as a series of questions and answers, addresses this specific problem and offers guidance for analyzing data and reporting findings when it occurs. It *does not* discuss approaches for addressing small samples due to missing data, nor does it discuss ways to maximize sample enrollment and retain participants to maintain focus on a single topic.[1]

Who should read this brief?

The Children's Bureau funded this brief for groups that receive Regional Partnership Grants (RPG) or other grants and want to evaluate the impacts of their programs. The brief summarizes key challenges that occur when studies have markedly smaller sample sizes than planned for estimating program impacts, resulting from lower-than-expected sample enrollment.

The remainder of this brief is structured as a series of questions and answers. Program staff will find the short answer to each question useful to inform discussions of these issues with their evaluators. The detailed answers and footnotes provide more technical information.

Why is small sample size an important consideration for an impact evaluation?

Short answer:

All else equal, an appropriately powered study leads to greater confidence that an impact estimate represents a genuine program effect, as opposed to a random chance difference between the treatment and comparison groups. As a result, achieving an appropriate sample size is critical to interpreting a study's findings.

More details:

A typical impact evaluation compares the average outcome of people assigned to receive the program (the program group) with the average outcome of people not assigned to the program (the comparison group).[2] The difference in the average outcomes is the estimated effect of the program. Impact estimates can be influenced by random error—that is, chance differences between the treatment and comparison groups. The primary measure of the potential magnitude of random error is the standard error. The *standard error* is used to calculate confidence intervals and *p*-values (see text box on the next page for additional

[1] For guidance on evaluation design principles to prospectively maximize sample enrollment and minimize the likelihood of small sample sizes, see Avellar et al. (2017a). Alternately, for guidance on how sample loss at follow-up can lead to bias in impact evaluations of child welfare programs, see Chapter 5 of Wilson et al. (2019). See Deke and Puma (2013) for analytic approaches for dealing with nonresponse in impact analyses.

[2] See Avellar et al.(2017b) for a description of comparison group evaluations—particularly evaluations using randomized controlled trials and quasi-experimental designs—and tips for choosing and successfully conducting the evaluations.

Mathematica
Progress Together

Children's Bureau

RPG
Regional Partnership Grants and Cross-Site Evaluation

Understanding what a *p*-value means

Understanding what a *p*-value represents helps researchers communicate and decision makers interpret impact evaluation findings correctly. The *p*-value is the probability of estimating a program impact of the magnitude observed in the study (or larger) if the true impact is zero (that is, if the program does not actually change participants' outcomes relative to the comparison group).

A *p*-value and the resulting statistically significant label are not sufficient by themselves to fully understand the effect of a program. An impact estimate with a *p*-value less than 0.05 might not be substantively important. A study with a large sample size could detect a difference in participants' outcomes that is so small it is not substantively meaningful yet has a *p*-value less than 0.05. Conversely, an impact estimate with a *p*-value greater than or equal to 0.05 might be substantively important. For example, a study with a small sample might detect large differences in participants' outcomes, but with a *p*-value greater than 0.05, is not labeled statistically significant.

information on *p*-values). All else equal, a larger sample size will lead to a smaller standard error, a narrower confidence interval, and a smaller *p*-value. Furthermore, a smaller standard error means that an impact estimate is more likely to be statistically significant in cases when a program truly has a favorable effect. Although emerging guidance strongly cautions against using *p*-values alone as the means for determining whether a program is effective, some audiences interpret a *p*-value less than 0.05 (often referred to as a *statistically significant* finding) as a signal that the program is effective.[3]

When designing impact evaluations, researchers commonly estimate the statistical power for their prospective study, using information about expected sample sizes, expected magnitude of program impacts, and the goal of observing *p*-values of a certain size (typically, $p < 0.05$). The statistical power of a study is the probability that the null hypothesis will be rejected, if the program is truly effective at moving participants' outcomes a given amount, for a given sample size and *p*-value threshold used for the hypothesis test. Therefore, the size of a study sample plays an important role in the study's statistical power, or stated differently, the likelihood that the study will produce a statistically significant impact on participant outcomes.

Why is small sample size, stemming from lower-than-expected enrollment, potentially a problem for an impact evaluation?

Short answer:

Smaller-than-expected sample sizes make it harder to detect genuine program effects, especially if those effects are small. Findings from a study with a sample that is smaller than the sample required for an adequately powered study must be presented carefully to ensure readers interpret the findings appropriately. This is particularly important because some researchers and policymakers use statistical significance as a criterion for judging a program as evidence-based.

More details:

Small sample size resulting from lower-than-expected enrollment does not adversely affect the study's ability to produce an unbiased estimate of the program's effects. On average, a well-implemented impact evaluation with a small sample size will produce the correct (unbiased) estimate of the effect of the program (Holland 1986).

However, as noted previously, the standard error of the impact estimate is in part determined by the size of the study sample. Therefore, when testing the effect of a program, having smaller-than-expected

[3] The text box on *p*-values briefly discusses common misconceptions about the meaning of a *p*-value. See Wasserstein and Lazar (2016) for a fuller discussion of the proper interpretation of *p*-values. Making decisions based on cutoffs involving *p*-values represents what is called the classical or frequentist approach to assessing the probability that the estimated impacts reflect true program impacts. An alternative, called the Bayesian approach, makes such decisions by combining estimates from one evaluation with prior information from related evaluations or educated guesses. For discussions of the Bayesian approach, see Gelman and Weakliem (2009) and Finucane and Deke (2019).

sample sizes will produce larger-than-expected standard errors, and thus higher *p*-values than the *p*-values the study was likely designed to obtain. When a study has markedly lower sample sizes than originally anticipated at the design phase (and everything else about the study remains constant, notably, the expected size of the difference in outcomes between the program and comparison groups), the study will have reduced power and therefore will be less likely to produce *p*-values less than 0.05.

Researchers and policymakers often use the statistical significance of an impact estimate as a criterion for labeling a program as evidence based. Some decision makers consider a program to be evidence based if a peer-reviewed journal published the study demonstrating program effectiveness. However, many peer-reviewed journals tend to prioritize publication of studies that show statistically significant findings (Franco et al. 2014). In addition, for a child-welfare program to be labeled as "promising," "supported," or "well supported" (that is, evidence based) according to the new Title IV-E Prevention Services Clearinghouse, it must have evidence that (1) one or more outcomes in the program group is more favorable than the outcome in the comparison group and (2) this difference in outcomes is statistically significant (see Administration for Children and Families [2018] and Wilson et al. [2019] for details). Given this information, the statistical significance of a program's impact is being used as a key metric for the level of evidence by which programs like RPG will potentially be judged. Therefore, smaller-than-expected sample sizes, and the resulting larger-than-expected standard errors and *p*-values, might cause some audiences to conclude that the program being evaluated is not effective or evidence based.

However, consensus is emerging in the statistical field that the phrase and implication of a "statistically significant" result is misleading. More specifically, the American Statistical Association has stated that relying solely on the *p*-value of an impact estimate is an inappropriate decision rule as a means to label whether a program is effective (Betensky 2019; Wasserstein et al. 2016, 2019).

Therefore, although having precise impact estimates with small *p*-values is certainly important for impact evaluations, this emerging guidance suggests that additional information should be used to inform decision-making. The concluding section on "What to report in an impact evaluation with a smaller-than-expected sample size" provides guidance on the additional pieces of information that impact studies should report to enable better decision- making.

Is it possible to obtain a statistically significant impact with smaller-than-expected sample sizes?

Short answer:

It is possible to obtain statistically significant impact estimates from a study with markedly smaller-than-expected sample sizes. This can happen for two reasons: (1) the program produces an effect markedly larger than what was expected at the design stage (when the power analysis was conducted) and/or (2) the impact estimate is influenced by a large random error. However, in such instances, researchers must examine and discuss how the observed impacts are not due to chance alone (that is, a Type I error).

More details:

If a program has a large impact estimate, then the study can produce small *p*-values of 0.05 or less, even with smaller-than-expected sample sizes. However, the few instances in which studies with smaller-than-expected samples have large enough impacts to produce small *p*-values are often judged with skepticism by research audiences.

With smaller-than-expected samples, either the program or comparison group is more likely to have received a disproportionate number of participants with background traits that predisposed them to certain types of outcomes (either good or bad). Therefore, when a study with a smaller-than-expected sample finds a statistically significant impact estimate, researchers often worry that the size of the program effect has been exaggerated (called a magnitude error), due to this type of unequal allocation of sample members to condition.

In such situations, the onus is on the researcher to appropriately interpret the (large) impact estimate in the context of the smaller-than-expected sample size. To better understand and interpret the impact estimate, researchers can examine and discuss the strength of the contrast in services across the treatment and comparison conditions, drawing on available implementation data. To further assist the readers in interpreting the study findings as part of the discussion sections in reports or articles, researchers might be able to draw on the literature of comparable interventions—if large impacts are commonly found in the literature this may provide additional support for the validity of the (large) impact estimate.[4]

Should an evaluation with a smaller-than-expected sample size proceed to analysis and reporting?

Short answer:

Although an impact evaluation with a smaller-than-expected sample size might not produce *p*-values that are small enough to be labeled statistically significant, the current guidance in the field strongly suggests that all credible studies present their findings, to ensure openness and transparency (Wasserstein et al. 2019). With less focus on *p*-values alone, such studies can make a contribution based on other key information they present. Comprehensively reporting on all impact findings (regardless of statistical significance) is the expectation for all government reporting, including the final reports required from all funded RPG grantees.

More details:

Publication bias in the scientific literature results when authors are more likely to submit, or editors more likely to accept, study results with favorable, statistically significant findings—rather than null or nonsignificant results. A consequence of such publication bias is that the current literature provides an inaccurate portrayal of a program's true effectiveness. To combat the "file drawer problem," in which evidence failing to reach statistical significance is left unpublished, experts recommend that all impact evaluations proceed to analysis and reporting, even when they have small sample sizes (Rosenthal 1977). That is, it is better to learn something from the originally designed study, and take advantage of opportunities to share the results, even if sample recruitment and retention did not go as expected and resulted in smaller-than-anticipated sample sizes.[5]

An additional benefit of reporting results can be realized in future meta-analyses. Although a single under-powered evaluation of a program might produce nonsignificant results, the findings from multiple small studies can be pooled together to produce a more powerful test of a program in a meta-analysis. To ensure that the appropriate information is shared, see the final section on guidance on information to report in an impact study with a smaller-than-expected sample size.

How do smaller-than-expected sample sizes affect impact analysis

Short answer:

A smaller-than-expected sample size does not change the basic principles for estimating impacts; however, the analytic approach might vary for studies with extremely small samples.

More details:

In a comparison group study, researchers typically establish the set of people used to estimate the effect of the program (that is, the analytic sample); check whether the program and comparison groups in the analytic sample are similar on key characteristics before the program begins (that is, assess baseline equivalence); and estimate impacts by comparing outcomes at follow-up for the program and comparison

[4] In addition, researchers can calculate a variety of additional interpretive statistics to help determine whether a statistically significant impact from a study with a small sample is likely due to the program truly having a large effect, sampling variability producing a large amount of statistical noise, or a combination of a moderate program effect and moderate statistical noise. These statistics include Type M (magnitude) errors; Type S (sign) errors, that is, when the actual effect is the reverse sign of the finding; the false positive rate (or false discovery rate); and the false negative rate (or false non-discovery rate). Papers discussing these statistics and ways to estimate them include Benjamini and Hochberg (1995), Colquhoun (2014), Gelman and Carlin (2014), Genovese and Wasserman (2002), and Storey (2003).

[5] Possible publication avenues include the Journal of Articles in Support of the Null Hypothesis, which publishes nonsignificant results twice a year, and the Public Library of Science, which publishes nonsignificant findings in its Missing Pieces supplements.

groups.[6] A small sample size can affect how researchers conduct hypothesis test analyses and estimate the *p*-values used to draw conclusions about baseline differences and program impacts. For example, the traditional t-test used to calculate a *p*-value for the difference between the program and comparison groups assumes the sample is sufficiently large and the outcomes are normally distributed. Sometimes researchers use regression approaches to estimate program impacts after adjusting for underlying differences—small sample sizes might limit the number of baseline covariates that can be adjusted for in the analysis. Also, if the sample size is very small (for example, smaller than 30 observations, where normality assumptions are likely violated), other nonparametric tests that relax these assumptions might be necessary to obtain correct *p*-values for hypothesis testing.[7]

Can post hoc matching approaches to address sample imbalance in baseline characteristics improve statistical power?

Short answer:

In many cases, conducting a post hoc matching analysis to address sample imbalance will exacerbate the problems with statistical power associated with smaller-than-expected sample sizes. It does so by eliminating poorly matched observations to better enable program effectiveness to be credibly estimated.

More details:

Using post hoc matching approaches (for example, propensity score matching) will typically exacerbate problems with smaller-than-expected sample sizes in impact evaluations. In the context of an impact evaluation, researchers most commonly use matching approaches to ameliorate problems with baseline equivalence among the sample members used to estimate the effect of a program.[8] Although restricting the analytic sample to program and comparison group members who match well on baseline characteristics enables an evaluation to more credibly show the effectiveness of the program, it does so by eliminating people who do not match well, which reduces the sample size and thus decreases statistical power. Therefore, matching is not a panacea for solving the problem of small sample sizes.

Are there any approaches researchers can use to increase power to detect significant impacts (if they truly exist) in studies with smaller-than-expected sample sizes?

Short answer:

There are analytical strategies researchers can use to attempt to mitigate smaller-than-expected sample sizes and increase the power of an impact analysis (assuming that the program is truly effective). By incorporating these strategies into the analyses, producing credible and statistically significant program impact estimates might be possible.[9]

More details:

The *p*-value of an impact estimate is a function of two components: (1) the magnitude of the difference in the outcomes being compared across the program and comparison groups and (2) the standard error of the difference. By statistically adjusting for variables that are good predictors of a given outcome, researchers can reduce error variance in the outcome. This reduction will reduce the standard error and lower the resulting *p*-value of the impact estimate.

[6] See Kautz and Cole (2017) for guidance on various benchmark and sensitivity approaches to consider when estimating program impacts.

[7] An impact evaluation with a very small sample (for example, $n < 30$) might need to use a nonparametric inference test. For example, the Mann-Whitney U test is the nonparametric analog to the traditional t-test for testing the statistical significance of differences between program and comparison groups. Textbooks such as Hollander et al. (2014) explain this issue in technical terms.

[8] Researchers commonly use matching to obviate large or statistically significant differences in key baseline characteristics that are likely to influence the outcome of interest—for example, when the baseline measure of an outcome of interest is more than 0.25 standard deviations different across conditions and regression adjustment will be insufficient to reduce bias.

[9] Best practice in impact analysis requires pre-specifying analytic approaches before conducting analyses and disclosing any changes or alterations in the eventual analytic approach. Such pre-specification and transparency in reporting helps guard against any possible perceptions of *p*-hacking. In the event of smaller-than-expected sample sizes, the approaches described below can be included as part of a pre-specified analysis plan or included as part of a disclosure as face-valid ways of estimating impacts

In general, the best variable to adjust for in an impact analysis is a baseline assessment of the outcome of interest. These variables often have a strong relationship (that is, a high correlation coefficient) with the outcome of interest, and thus, including them in an analysis will greatly reduce the standard error of an impact estimate. By adjusting for any chance differences in baseline measures of the outcome, this approach might also improve the face validity of the impact estimate.[10] Adjusting for additional variables (in particular, those the literature shows to be predictive of the outcome) can further improve the statistical precision of the impact estimate.[11]

A second analytic approach might be possible when a given outcome domain has multiple outcome measures. For example, in the RPG cross-site evaluation, grantees collect multiple outcomes and assessments that all measure constructs within the child well-being domain. In such a situation, researchers can conduct impact analyses that pool information across several outcomes within a domain. By pooling information across multiple outcomes that each tap into a common or underlying latent domain (such as different assessments of child well-being, which grantees collect for the RPG cross-site evaluation), the study will have a more reliable assessment of the underlying domain—that is, one with less error variance. In such situations, composite inferential tests such as composite t-tests or a multivariate regression approach will have greater statistical power than individual tests on each outcome. Therefore, although a study might not be able to have statistically significant impacts on any one individual measure (of child well-being, in the RPG example), the study might show that the program had a statistically significant impact on the domain as a whole.[12]

Conclusion

What to report in an impact evaluation with a smaller-than-expected sample size

Small sample sizes in an impact evaluation stemming from lower-than-planned enrollment rates decrease a study's likelihood of showing a statistically significant effect (assuming the program is truly effective). For now, many audiences still use statistical significance to determine whether programs are evidence based. As such, it may be that smaller-than-expected sample sizes can result in more Type II errors (situations in which programs may actually be effective, but the *p*-values are too large to label the program as effective or evidence based).

Despite this, transparently reporting findings from an underpowered evaluation has great value. As noted previously, the estimate of program effectiveness from a well-executed impact evaluation can be unbiased and thus can and should be reported, even if it is imprecise (due to a large standard error). Also as previously mentioned, full reporting of impact findings can be extremely useful for future meta-analyses, in which several under-powered studies of an intervention can be pooled together to produce a more powerful test of that program.

The statistical field is evolving and is suggesting new practice for researchers in terms of communicating results from statistical analyses—notably, greater transparency in reporting to better inform decision making. To enable findings to be useful for a broad audience, we recommend reporting a variety of information from RPG impact analyses, especially (but not only) from impact evaluations with smaller-than-expected sample sizes:

(1) Report the impact estimate and its standard error. It is critical to state the estimated difference in outcomes between the program and comparison groups, as

[10] In a valid experiment, randomization should balance all observed and unobserved characteristics between the program and comparison groups. Therefore, the only reason that any variable differs at baseline across the two groups is due to random sampling error.

[11] See Kautz and Cole (2017) for additional considerations with respect to covariate adjustment. This brief is particularly relevant regarding articulating a face-valid process for selecting covariates that does not create criticisms about data mining.

[12] See Appendix C in Schochet (2008) for information on composite analytic approaches for impact evaluations.

this might be an unbiased estimate of the true effect of the program, assuming the evaluation was conducted well. The observed impact estimate, along with an interpretation of the size of the effect, will help readers understand the potential policy relevance of the difference, regardless of the *p*-value. Describe the effect in terms of the magnitude of the impact, which will help readers interpret whether the program made a large or small difference in moving outcomes. Supplementing the impact estimate with the standard error will provide researchers with additional information necessary for future meta analyses, calculation of alternate confidence intervals, or Bayesian interpretation.

(2) Provide a confidence interval around the estimate. A confidence interval provides a visual representation of an impact estimate's precision. A confidence interval is centered at the impact estimate discussed in Point 1. The width of a confidence interval is a function of the standard error of the impact estimate. Because small sample sizes produce large standard errors, small sample sizes also produce wide confidence intervals. The confidence interval indicates a range of values that are likely to contain the true difference in participants' outcomes. When a study discusses the interpretation of both the high and low bounds of the interval, readers will have a more complete description of the possible effect of the program.

(3) Present the *p*-value of the impact estimate, with appropriate interpretation. Studies commonly provide a categorical summary of the *p*-value, such as $p < 0.05$ (or an asterisk indicating that a *p*-value is less than 0.05 or 0.01). Instead, researchers should present the actual *p*-value estimated (such as $p = 0.04$). In addition, as in the findings section, researchers should present the appropriate interpretation of the *p*-value of the impact estimate: The *p*-value is the probability of estimating a program impact of the magnitude observed in the study (or larger) if the true impact is zero (that is, if the program does not actually change participant outcomes relative to the comparison group).

In addition to reporting those specifics on the impact estimates, researchers should describe the limitations in the observed data, relative to what was intended. The study should report what the original sample size target was, along with power calculations or minimum detectable impact calculations to demonstrate what the study was originally designed to detect. This information can provide context for whether the observed impacts are in line with what was expected or intended at the design phase.

In addition to describing the initial sample size target, the study should describe the reasons why the observed sample was smaller than expected. This information will illustrate potential pitfalls for future researchers to be aware of and for which researchers can prospectively plan as means to mitigate comparable limitations in future studies.

Finally, transparently discussing limitations of the enrolled sample will provide information on generalizability. If the enrolled study sample is markedly smaller than the intended target population by virtue of missing key subsets of the population, the observed impact estimates not only suffer from imprecision but also generalizability. The observed impact will not necessarily represent a contrast of interest if only a certain subset of the target population was enrolled into the study and included in the analysis.

REFERENCES

Administration for Children and Families, U.S. Department of Health and Human Services. "Decisions Related to the Development of a Clearinghouse of Evidence-Based Practices in Accordance with the Family First Prevention Services Act of 2018." 83 *Federal Register* 29122, pp. 29122–29124.

Avellar, S., K. Borradaile, and D. Strong. "Tips for Enrolling and Retaining Evaluation Participants." Evaluation Technical Assistance Brief, No. 4. Washington, DC: Mathematica Policy Research, 2017a.

Avellar, S., R. Santillano, and D. Strong. "Tips for Planning an Impact Evaluation." Evaluation Technical Assistance Brief, No. 3. Washington, DC: Mathematica Policy Research, 2017b.

Benjamini, Y., and Y. Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society, Series B,* vol. 57, no. 1, 1995, pp. 289–300.

Betensky, R. A. "The *p*-Value Requires Context, Not a Threshold." The American Statistician, vol. 73, suppl. 1, 2019, pp. 115–117. https://doi.org/10.1080/00031305.2018.1529624

Colquhoun, David. "An Investigation of the False Discovery Rate and the Misinterpretation of *p*-Values." *Royal Society Open Science*, vol. 1, no. 3, 2014, p. 140216.

Deke, J., and M. Puma. "Coping with Missing Data in Randomized Controlled Trials." Evaluation Technical Assistance Brief 3. Washington, DC: U.S. Department of Health and Human Services, Office of Adolescent Health, 2013.

Finucane, M., and J. Deke. "Moving Beyond Statistical Significance: The BASIE (BAyeSian Interpretation of Estimates) Framework for Interpreting Findings from Impact Evaluations." Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2019.

Franco, A., N. Malhotra, and G. Simonovits. "Publication bias in the Social Sciences: Unlocking the File Drawer." *Science*, vol 345, no. 6203, 2014, pp. 1502–1505.

Gelman, A., and J. Carlin. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science*, vol. 9, no. 6, 2014, pp. 641–651.

Gelman, A., and D. Weakliem. "Of Beauty, Sex and Power: Too Little Attention Has Been Paid to the Statistical Challenges in Estimating Small Effects." *American Scientist*, vol. 97, no. 4, 2009, pp. 310–316.

Genovese, C., and L. Wasserman. "Operating Characteristics and Extensions of the False Discovery Rate Procedure." *Journal of the Royal Statistical Society*, Series B, vol. 64, no. 3, 2002, pp. 499–517.

Holland, P. "Statistics and Causal Inference." *Journal of the American Statistical Association*, vol. 81, 1986, pp. 945–960.

Hollander, M., D. A. Wolfe, and E. Chicken. *Nonparametric Statistical Methods* (3rd edition). Hoboken, NJ: John Wiley & Sons, Inc., 2014.

Kautz, T., and R. Cole. "Selecting Benchmark and Sensitivity Analyses." Evaluation Technical Assistance Brief. Washington, DC: Office of Adolescent Health, U.S. Department of Health and Human Services, 2017. Available at https://www.hhs.gov/ash/oah/sites/default/files/selecting-benchmark-and-sensitivity-analyses.pdf. Accessed July 27, 2018.

Rosenthal, R. "File Drawer Problem and Tolerance For Null Results." *Psychological Bulletin*, vol. 86, no. 3, 1979, pp. 638–641. doi:10.1037/0033-2909.86.3.638.

Schochet, P. "Technical Methods Report: Guidelines for Multiple Testing in Impact Evaluations." NCEE 2008-4018. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2008.

Storey, John D. "The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value." *The Annals of Statistics*, vol. 31, no. 6, 2003, pp. 2013–2035.

Wasserstein, R. L., and N. A. Lazar. "The ASA's Statement on *p*-Values: Context, Process, and Purpose." *The American Statistician*, vol. 70, no. 2, 2016, pp. 129–133. doi:10.1080/00031305.2016.1154108.

Wasserstein, R. L., A. L. Schirm, and N. A. Lazar. "Moving to a World Beyond 'p < 0.05.'" *The American Statistician*, vol. 73, suppl. 1, 2019, pp. 1–19, doi:10.1080/00031305.2019.1583913.

Wilson, S. J., C. S. Price, S. E. U. Kerns, S. D. Dastrup, and S. R. Brown. "Title IV-E Prevention Services Clearinghouse Handbook of Standards and Procedures, Version 1.0." OPRE Report No. 2019-56. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2019.