Tool 4.3 – Program Evaluation Resources

This tool includes selected references to relatively accessible information about various program evaluation topics. The list is neither exhaustive nor intended to include all foundational references to particular topics. Instead, it aims to provide additional "how-to" information and beginning points for further exploration.

A resource that aims to strengthen program managers' understanding of and readiness for program evaluation is "The Program Manager's Guide to Evaluation," published by the Office of Planning, Research, and Evaluation (OPRE) within the Administration for Children and Families. It explains what program evaluation is, its importance, and different steps in the evaluation process, including how to engage an evaluation team, prepare for and design an evaluation, gather credible evidence and analyze data, and share lessons learned.

In addition, OPRE periodically organizes meetings to convene scientists and research experts to advance critical topics in social science research methodology. The meetings provide an opportunity to discuss how innovative methodologies can be applied to policy-relevant questions and help to ensure that government-supported research represents the most scientifically rigorous approaches available. Additional resources on some of the topics listed below can be found under Past Meetings on the OPRE site.

Topics

- Difference in Differences
- Effect Sizes
- Evaluation Design—General
- Interrupted Time Series (ITS) and Comparative Interrupted Time Series (CITS)
- Matching
  - General
  - Propensity Score Methods
  - Synthetic Comparison Methods
- Multiple Hypothesis Testing
- Null Results
- Regression Discontinuity (RD)
- Subgroups
- Theory of Change and Logic Models

Difference in Differences

Somers, Marie-Andrée, Pei Zhu, Robin Jacob, and Howard Bloom. 2013. "The Validity and Precision of the Comparative Interrupted Time Series Design and the Difference-in-Difference Design in Educational Evaluation," MDRC Working Paper on Research Methodology. New York: MDRC.

Wing Cody, Kosali Simon, and Ricardo A. Bello-Gomez. 2018. "Designing Difference in Difference Studies: Best Practices for Public Health Policy Research." *Annual Review of Public Health* 39: 453-469.

Bloom, Howard S., Charles Michalopoulos, Carolyn J. Hill, and Ying Lei. 2002. "Can Nonexperimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?" MDRC Working Paper on Research Methodology. New York: MDRC.

Effect Sizes

Bloom, Howard S. 1995. "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs." *Evaluation Review* 19, 5: 547-556.

Bloom, Howard S., Carolyn J. Hill, Alison Rebeck Black, and Mark W. Lipsey. 2008. "Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions." New York: MDRC.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.

Durlak, Joseph A. 2009. "How to Select, Calculate, and Interpret Effect Sizes." *Journal of Pediatric Psychology* 34, 9: 917-928.

Hill, Carolyn J., Howard S. Bloom, Alison Rebeck Black, and Mark W. Lipsey. 2008. "Empirical Benchmarks for Interpreting Effect Sizes in Research." *Child Development Perspectives* 2, 3: 172-177.

Lipsey, Mark W., Kelly Puzio, Cathy Yun, Michael A. Herbert, Kasia Steinka-Fry, Mikel W. Cole, Megan Roberts, Karen S. Anthony, and Matthew D. Busick. 2012. *Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms. (NCSER 2013-3000).* Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2015. *Mastering Metrics: The Path from Cause to Effect*. Princeton, NJ: Princeton University Press.

Evaluation Design—General

El Mallah, S., Gutuskey, L., Hyra, A., Hare, A., Holzwart, R., & Steigelman, C. 2022. "The Program Manager's Guide to Evaluation" (OPRE Report 2022-208). U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation.

Murnane, Richard J., and John J. Willett. 2010. *Methods Matter Improving Causal, Inference in Educational and Social Science Research*. Oxford, England: Oxford University.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton, Mifflin and Company.

Bloom, Dan. 2018. "How Is Random Assignment Like a Frying Pan?" *MDRC Reflections on Methodology* (blog), November. Website: https://www.mdrc.org/work/publications/how-random-assignment-frying-pan.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2008. "Using Randomization in Development Economics Research: A Toolkit." Pages 3895-3962 in *Handbook of Development Economics*. Amsterdam: Elsevier.

Interrupted Time Series (ITS) and Comparative Interrupted Time Series (CITS)

Bloom, Howard S. 2003. "Using 'Short' Interrupted Time-Series Analysis To Measure The Impacts Of Whole- School Reforms: With Applications to a Study of Accelerated Schools." *Evaluation Review* 27, 1: 3-49.

Somers, Marie-Andrée, Pei Zhu, Robin Tepper Jacob, and Howard Bloom. 2013. "The Validity and Precision of the Comparative Interrupted Time Series Design and the Difference-in-Difference Design in Educational Evaluation." New York: MDRC.

Wing, Coady, Kosali Simon, and Ricardo A. Bello-Gomez. 2018. "Designing Difference in Difference Studies: Best Practices for Public Health Policy Research." *Annual Review of Public Health* 39: 453-469.

Shadish, William R., Thomas D. Cook, Donald T. Campbell, and C. S. Reichardt. 2002. "Experimental and Quasi-Experimental Designs for Generalized Causal Inference," Book review. *Social Service Review* 70, 3: 510–514.

Tuttle, Christina Clark, Brian Gill, Philip Gleason, Virgina Knechtel, Ira Nichols-Barrer, and Alexandra Resch. 2013. "KIPP Middle Schools: Impacts on Achievement and Other Outcomes, Final Report." Washington, DC: Mathematica Policy Research, Inc.

Wong, Manyee, Thomas D. Cook, and Peter M. Steiner. 2009. "No Child Left Behind: An Interim Evaluation of Its Effects on Learning Using Two Interrupted Time Series Each With Its Own Non-Equivalent Comparison Series." Working Paper 09-11, 18 (7). Evanston, IL: Institute for Policy Research, Northwestern.

Matching

General

Stuart, Elizabeth A. 2010. "Matching Methods for Causal Inference: A Review and Look Forward." *Statistical Science* 25, 1: 1-21.

Propensity Score Methods

Guo, Shenyang. and Mark W. Fraser. 2014. *Propensity Score Analysis: Statistical Methods and Applications*. Sage Publications: Thousand Oaks, CA.

Stuart, Elizabeth A. 2011. "The Why, When, and How of Propensity Score Methods for Estimating Causal Effects." Unpublished Paper. Fairfax, VA: Society for Prevention Research.

Randolph, Justus J., Kristina Falbe, Austin Kureethara Manuel, and Joseph L. Balloun. 2019. "A Step-by-Step Guide to Propensity Score Matching in R." *Practical Assessment, Research, and Evaluation* 19, 18.

Heinrich, Carolyn, Alessandro Maffioli, and Gonzalo Vazquez. 2010. *A Primer for Applying Propensity-Score Matching*. Washington, DC: Office of Strategic Planning and Development Effectiveness, Inter-American Development Bank.

Caliendo, Marco, and Sabine Kopeinig. 2008. "Some Practical Guidance for the Implementation of Propensity Score Matching." *Journal of Economic Surveys* 22, 1: 31-72.

Synthetic Comparison Methods

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105: 490, 493-505.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2014. "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science* 59, 5: 495-510.

Cook, Thomas D., William R. Shadish, and Vivian Wong. 2008. "Three Conditions Under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons." *Journal of Policy Analysis and Management* 27, 4: 724-750.

Multiple Hypothesis Testing

Schochet, Peter M. 2008. *Technical Methods Report: Guidelines for Multiple Testing in Impact Evaluations*. Washington, DC: Institute of Education Sciences, U.S. Department of Education.

Null Results

Herrington, Carolyn D., Rebecca Maynard. 2019. "Editors' Introduction: Randomized Controlled Trials Meet the Real World: The Nature and Consequences of Null Findings." *Educational Researcher* 48, 9: 577-579.

Jacob, Robin T., Fred Doolittle, James Kemple, and Maire-Andree Somers. 2019. "A Framework for Learning From Null Results." *Educational Researcher* 48, 9: 580-589.

Landis, Ronald S., Lawrence R. James, Charles E. Lance, Charles A. Pierce, and Steven G. Rogelberg. 2014. "When is Nothing Something?" Editorial, Null Results Special Issue, *Journal of Business and Psychology* 29, 2: 163-167.

Regression Discontinuity (RD)

Bloom, Howard S. 2012. "Modern Regression Discontinuity Analysis." *Journal of Research on Educational Effectiveness* 5, 1: 43-82.

Imbens, Guido W. and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142, 2:615–635, 2008.

Jacob, Robin, Pei Zhu, Marie-Andree Somers, Howard Bloom. 2012. "A Practical Guide to Regression Discontinuity." New York: MDRC.

Howell, Sabrina T. 2016. "Financing Innovation: Evidence from R&D Grants." Unpublished Paper. New York: New York University.

Subgroups

Bloom, Howard S., Charles Michalopoulos. 2013. "When is the Story in the Subgroups?" *Prevention Science* 14, 179–188.

Theory Of Change and Logic Models

Bangser, Michael. 2014. "A Funder's Guide to Using Evidence of Effectiveness in Scale-Up Decisions." New York: MDRC.

Epstein, Diana and Jacob Alex Klerman. 2012. "When is a Program Ready for Rigorous Impact Evaluation? The Role of a Falsifiable Logic Model." *Evaluation Review* 36, 5: 375-401.

W.K. Kellogg Foundation. 2004. "Logic Model Development Guide: Using Logic Models to Bring Together Planning, Evaluation, and Action." Battle Creek, MI: W.K. Kellogg Foundation.