Anglin, K. L., Krishnamachari, A., & Wong, V. (2020). *Methodological Approaches for Impact Evaluation in Educational Settings*. Oxford Bibliographies. https://doi.org/10.1093/obo/9780199756810-0244

Methodological Approaches for Impact Evaluation in Educational Settings

INTRODUCTION

Since the start of the War on Poverty in the 1960s, social scientists have developed and refined experimental and quasi-experimental methods for evaluating and understanding the ways in which public policies, programs, and interventions affect people's lives. The overarching mission of many social scientists is to understand "what works" in education and social policy. These are causal questions about whether an intervention, practice, program, or policy affects some outcome of interest. Although causal questions are not the only relevant questions in program evaluation, they are assumed by many in the fields of public health, economics, social policy, and now education to be the scientific foundation for evidence-based decision making. Fortunately, over the last half-century, two methodological advances have improved the rigor of social science approaches for making causal inferences. The first was acknowledging the primacy of research designs over statistical adjustment procedures. Donald Campbell and colleagues showed how research designs could be used to address many plausible threats to validity. The second methodological advancement was the use of potential outcomes to specify exact causal quantities of interest. This allowed researchers to think systematically about research design assumptions and to develop diagnostic measures for assessing when these assumptions are met. This article reviews important statistical methods for estimating the impact of interventions on outcomes in education settings, particularly programs that are implemented in field, rather than laboratory, settings. We begin by describing the causal inference challenge for evaluating program effects. Then four research designs are discussed that may be used for estimating program impacts. The article highlights what the Campbell tradition identifies as the

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305B140026 to the Rectors and Visitors of the University of Virginia. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

strongest causal research designs: the randomized experiment and the regression-discontinuity designs. These approaches have the advantage of transparent assumptions for yielding causal effects. The article then discusses weaker but more commonly used approaches estimating effects, including the interrupted time series and the non-equivalent comparison group designs. For the interrupted time series design, differences-in-differences are discussed as a more generalized approach to time series methods; for non-equivalent comparison group designs, the article highlights propensity score matching as a method for creating statistically equivalent groups on the basis of observed covariates. For each research design, references are included that discuss the underlying theory and logic of the method, exemplars of the approach in field settings, and recent methodological extensions to the design. The article concludes with a discussion of practical considerations for evaluating interventions in field settings, including the external validity of estimated effects from impact studies.

GENERAL OVERVIEWS

The fundamental problem of causal inference is that we cannot observe both what happens to a student when they receive an intervention and what would have occurred in an alternate reality in which the same student did not receive an intervention. For example, researchers can observe what happens to children in a preschool program but cannot observe what would have happened to the same children had they not entered preschool. To study the causal effect of a program or intervention, one needs a counterfactual, or something that is contrary to fact. Given that researchers never observe the counterfactual, we look for approximations (e.g., older siblings, neighborhood children, children in a nationally representative survey, or randomly assigned control children not exposed to the treatment). The Rubin Causal Model introduced in Rubin 1974 formalizes this reasoning mathematically. It is based on the idea that every unit has a

potential outcome based on its "assignment" to a treatment or control condition. Using a potential outcomes framework, researchers are able to define a causal estimand of interest for a well-defined treatment and inference population, as well as assumptions required for a research design to yield a valid effect. Campbell and Stanley 1963 demonstrates how these assumptions may be violated in field settings through their list of "validity threats." Cook and Campbell 1979 and Shadish et al. 2002 extend this idea by introducing four types of validity threats, including threats to internal, external, statistical conclusion, and construct validity. Angrist and Pischke 2009 provides an up-to-date overview of common methodological approaches from an econometric perspective and discusses estimation procedures for producing causal estimates. Angrist and Pischke 2015 offers a more approachable overview of the same material intended for an undergraduate audience. Imbens and Rubin 2015 and Morgan and Winship 2007 straddle the econometric and statistics literature and offer additional insights about causal inference from a potential outcomes perspective and a causal graph theory perspective, respectively. For an overview of key experimental and quasi-experimental designs specific to the field of education, see Murnane and Willett 2011 and Stuart 2007.

Angrist, J., and J.-S. Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion.* Princeton, NJ: Princeton Univ. Press. [ISBN: 9780691120348][class:book]

This book is a reference on methods of causal inference using a potential outcomes framework. It covers randomized experiments, statistical matching, instrumental variables, difference-in-differences, and regression discontinuity. The book describes each design and its assumptions formally through a series of proofs and informally through applied examples. Though written for a graduate student audience, it is a useful resource for any evaluator with training in probability and statistics.

Angrist, J., and J. -S. Pischke. 2015. *Mastering 'metrics: The path from cause to effect*. Princeton, NJ: Princeton Univ. Press. [ISBN: 9780691152837][class:book]

Angrist and Pischke 2015 provides a more approachable and conversational companion to Angrist and Pischke 2009. While both books describe the same methods of causal inference (randomized control trials, statistical matching, instrumental variables, regression discontinuity, and differences-in-differences designs), this book focuses more on conceptual understanding than on formal proofs—though brief proofs are provided. The book is written as an introduction to causal inference for undergraduate economics students.

Campbell, D. T., and J. C. Stanley. 1963. *Experimental and quasi-experimental design for research*. Boston, MA: Houghton Mifflin.[class:book]

This seminal book outlines the major threats to internal validity (*Did the intervention cause the observed effect?*) and external validity (*To what population, settings, treatments, and outcomes can this effect be generalized?*) and provides an overview of how design features can address these threats. While the book discusses quasi-experimental designs, it is best suited for an overview of conceptual challenges related to causal inference rather than for guidance in statistical methods in estimating effects.

Cook, T. D., and D. T. Campbell. 1979. *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin. [ISBN: 9780395307908][class:book]

Similar to Campbell and Stanley 1963, the first chapters of this book introduce the challenge of causal inference and threats to validity. The book updates Campbell and Stanley 1963 by also addressing analytical approaches. Helpfully, the book concludes with a section outlining major obstacles to conducting randomized experiments and describing situations that are particularly conducive to experimental evaluation.

Imbens, G., and D. Rubin. 2015. *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge, UK: Cambridge Univ. Press. [ISBN: 9780521885881][class:book]

This textbook provides a rigorous introduction to the potential outcomes framework. Because the book relies on formal mathematical derivations, it is most appropriate for those with a solid understanding of probability and statistics. The book discusses randomized experiments (including instrumental variables for non-compliance) and matching methods but does not provide an overview of quasi-experimental designs. Applied examples from education, social science, and biomedical science are used to illustrate concepts.

Morgan, S., and C. Winship. 2007. *Counterfactuals and causal inference: Methods and principles for social research.* Cambridge, UK: Cambridge Univ. Press. [ISBN: 9780521856157][class:book]

This textbook discusses how to answer causal questions using observational data rather than data where researchers have the opportunity to manipulate the treatment assignment. The book discusses randomized experiments primarily as a starting point to further understanding on non-experimental research designs, but several concepts, including the potential outcomes framework, are explained in detail with the help of causal diagrams, structural models, and examples from the social sciences.

Murnane, R., and J. Willett. 2011. *Methods matter: Improving causal inference in educational and social science research*. New York: Oxford Univ. Press. [ISBN: 9780199753864][class:book]

This book is a broadly accessible reference to causal inference in education research. It illustrates important concepts in the design and analysis of randomized experiments, quasi-

experiments (including the difference-in-difference, regression discontinuity, and instrumental variables approaches), and observational studies. High-quality causal studies in the field of education are used to demonstrate and evaluate the decisions researchers make in the design and analysis of a study.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66.5: 688–701. [doi:10.1037/h0037350]

Provides the fundamental building blocks for modern program evaluation. Rubin conceptualizes the fundamental challenge of causal inference using a series of potential outcomes—individual outcomes in the presence of treatment and in the absence of treatment. This conceptualization allows for the formalization of both experimental and non-experimental design assumptions and is often referred to as the Rubin causal model.

Shadish, W. R., T. D. Cook, and D. T. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin. [ISBN: 9780395615560][class:book]

This book is a successor to Campbell and Stanley 1963 and Cook and Campbell 1979. Provides a comprehensive discussion of the design elements a researcher may include to improve internal validity and provides the conceptual theory for research design choices. The latter part of the book proposes a theoretical framework for generalized causal inference.

Stuart, E. A. 2007. Estimating causal effects using school-level data sets. *Educational Researcher* 36.4: 187–198. [doi:10.3102/0013189X07303396][class:journalArticle]

Stuart provides a survey of evaluation approaches with school-level data, including randomized experiments, regression discontinuity, interrupted time series, and non-equivalent comparison group designs. The article provides an overview of the National Longitudinal School-Level

State Assessment School Database (NLSLASD) and key considerations to keep in mind when using the NLSLASD or other school-level datasets to answer causal questions.

RANDOMIZED CONTROL TRIALS

The most credible evaluations use random assignment to determine access to an intervention. The modern design of randomized experiments can be attributed to Fisher 1935. In a randomized experiment, researchers assign participants to a "treatment" or "control" group using a deliberately random procedure such as a coin toss. The treatment group participates in some program or intervention while the control group does not. Assuming a large enough sample, the random assignment procedure creates two or more groups that are equivalent on average for all baseline characteristics and potential outcomes. When this happens, the evaluator may estimate program impacts by comparing the average outcomes in the treatment and control groups and interpreting the difference in the two means as the average treatment effect (ATE) in the study population. The random assignment procedure helps to ensure that differences in outcomes between the treatment and control groups are due to the treatment or policy under investigation, and not some unobserved factors related to both treatment assignment and the outcome. Over the last twenty years, there have been increasing calls for experimental evaluations of treatments, programs, and policies in education settings. Cook 2002 and Mosteller and Boruch 2002 offer arguments for conducting randomized control trials in field settings and provide advice for addressing the political and moral considerations that may arise. Beyond even political and moral concerns, randomized control trials can be challenging to implement in field settings. Problems include randomization failure, interference between units, attrition, treatment noncompliance, and missing data. For comprehensive guides on addressing (and preventing) these challenges, we recommend Gerber and Green 2012 and Duflo, et al. 2005. Barnard, et al. 2003 also offers

additional insights into addressing treatment non-compliance and missing data. Finally, Angrist, et al. 1996 discusses the use of instrumental variables (IV) to answer causal research questions when there is treatment non-compliance and Gennetian 2002 uses an IV approach to identify effects of intervening variables (or mediators) with the aim of improving experimental design and informing policy decisions.

Angrist, J. D., G. W. Imbens, and D. B. Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American statistical Association* 91.434: 444–455.[class:journalArticle]

This paper outlines the use of IV to estimate the treatment effect on treated individuals in the case of treatment non-compliance. The use of IV is formulated using the Rubin Causal model, and the authors outline the identifying assumptions required to identify treatment-on-the-treated effects.

Barnard, J., C. E. Frangakis, J. L. Hill, and D. B. Rubin. 2003. Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *Journal of the American Statistical Association* 98.462: 299–323.[class:journalArticle]

This article discusses the benefits of implementing a randomized experiment and outlines potential complications in experiments that involve human subjects. These include missing background and outcome data and noncompliance with randomly assigned treatment. The article details and addresses these complications using a principal stratification framework.

Cook, T. D. 2002. Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis* 24.3: 175–199. [doi:10.3102/01623737024003175][class:journalArticle]

Despite the widespread belief that experiments provide the best warrant for causal claims, experiments have only recently started making their way into schools and classrooms. In this article, Cook discusses five common critiques of experiments and provides concrete examples of how experiments may be designed to counteract these concerns.

Duflo, E., R. Glennerster, and M. Kremer. 2007. *Using randomization in development economics research: A toolkit *Handbook of development economics*, *4*, 3895-3962.[doi:10.1016/s1573-4471(07)04061-2][class:journalArticle]

Duflo and colleagues provide an in-depth "toolkit" for practitioners and researchers who are interested in implementing randomized field experiments. The paper explains why randomized experiments are considered the best design to answer causal research questions, examines the conditions under which random assignments yields such causal claims, and discusses implementation procedures for successful studies.

Fisher, R. A. 1935. *The design of experiments.* Oxford, UK: Oliver & Boyd.[class:book]

As the first introduction to null hypothesis testing, Fisher's *Design of Experiments* is considered a foundational work in experimental design. The book discusses several types of experimental designs and shows how conclusions can be drawn from such designs by formulating and disproving null hypotheses.

Gennetian, L. A., J. M. Bos, and P. A. Morris. 2002. *Using instrumental variables analysis to learn more from social policy experiments[https://www.mdrc.org/sites/default/files/full_599.pdf]*. *MDRC Working Papers on Research Methodology.* New York, NY: MDRC.[class:report]

This report discusses the use of IV for examining causal claims. In their report, the authors explore the feasibility of applying IV strategies to data from experimental designs, review

policy questions that can be answered, and examine necessary conditions for estimating mediating effects. Provides guidance on the use of IV to design more effective interventions and inform broader policy decisions.

Gerber, A. S., and D. P. Green. 2012. *Field experiments: Design, analysis, and interpretation*. New York: W. W. Norton. [ISBN: 9780393979954][class:book]

This is an introductory textbook on field experiments in the social sciences and covers major topics in the design, implementation, and analysis of experiments in field settings. Readers also learn how to handle common implementation challenges that arise in field experiments, including treatment non-compliance, violations to participant non-interference assumptions, and missing data. Overall, this is a great resource for new researchers to familiarize themselves with the "how-to" of experiments.

Mosteller, F., and R. F. Boruch. 2002. *Evidence matters: Randomized trials in education research*. Washington, DC: Brookings Institute Press. [ISBN: 9780815702054][class:book]

In this edited volume, authors discuss the necessity of experiments, theorize reasons for their relative absence in education compared to other fields, and offer advice in addressing the political and moral challenges of conducting randomized experiments in education. Cook 2002 and this volume together provide a comprehensive overview of the status of experiments in education and the reasons they are sparsely implemented in educational settings.

REGRESSION DISCONTINUITY DESIGN

In a regression-discontinuity design, participants are assigned to treatment and comparison groups on the basis of a cutoff score from a quantitative assignment variable (also called a "running variable"). Here, individuals who score above the cutoff are assigned to the treatment (or control), while individuals who score below the cutoff are assigned to the control (or

treatment). Treatment effects are estimated by examining the conditional mean difference in the outcome measure for individuals who score just above and below the cutoff along the assignment variable. The regression discontinuity (RD) design is particularly useful in education settings where scarce resources are allocated on the basis of a standardized test score (e.g., remediation and retention, curriculum interventions, gifted education). This section provides a methodological overview of RD designs, as well as examples of how the design has been implemented in real world settings. Cook 2008 and Cook and Wong 2008 describe the history of RD designs and conduct empirical tests of the validity of such designs. Lee and Lemieux 2010 discusses the strengths and weaknesses of RD and the conditions necessary for such designs to succeed in producing reliable estimates of the impact of a program. Similarly, Schochet, et al. 2010 reviews the conditions necessary for RD designs to meet What Works Clearinghouse (WWC) standards, while Schochet 2009 discusses the sample size requirements required for RD designs to yield precise effect estimates in educational settings. The seminal paper on RD is Thistlethwaite and Campbell 1960, broadly recognized as the first paper to introduce an RD design. Building on Thistlethwaite and Campbell 1960, McCrary 2008 is the first paper to offer a formal test for manipulation of the treatment assignment variable, a violation of the continuity assumption valid RD estimates hinge upon. Angrist and Rokkanen 2015 offers a method of testing causal claims away from the cutoff through a case study of exam school admissions. Finally, Jacob, et al. 2012 provides practitioners with a guide to understanding and implementing RD designs in simple language aimed at a broader audience.

Angrist, J. D., and M. Rokkanen. 2015. Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association* 110.512: 1331–1344. [doi:10.1080/01621459.2015.1012259][class:journalArticle]

This article addresses a central challenge of RD designs; causal claims are most credible for the population near the point of discontinuity, but we often want to know the impact on populations who are farther away from the small window surrounding the cutoff. Using data from Boston's exam schools, the article offers a test of the validity of causal claims away from the cutoff.

Cook, T. D., and V. C. Wong. 2008. Empirical tests of the validity of the regression discontinuity design. *Annals of Economics and Statistics/Annales d'Économie et de Statistique* 91.92: 127–150. [doi:10.2307/27917242].[class:journalArticle]

Provides a detailed overview of the extent to which RD design estimates can reliably reproduce estimates from a randomized experiment. Traces the history of the RD design and discusses the conditions necessary for RD estimates to correspond to estimates from RCTs.

Cook, T. D. 2008. 'Waiting for life to arrive': A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics* 142.2: 636–654. [doi:10.1016/j.jeconom.2007.05.002].[class:journalArticle]

Provides readers with an in-depth historical account of RD designs. Aimed at a wide audience, the article describes the basic RD design introduced by Thistlethwaite and Campbell 1960 and highlights contributions by different researchers that built upon the basic RD framework.

Jacob, R. T., P. Zhu, M. -A. Somers, and H. S. Bloom. 2012. *A practical guide to regression discontinuity[https://www.mdrc.org/sites/default/files/RDD%20Guide_Full%20rev%202016_0.pdf]*. *MDRC Working Papers.* New York, NY: MDRC.[class:report]

Written as an easy-to-read guide for practitioners looking to implement regression discontinuity designs, this paper discusses various techniques available and illustrates their strengths and weakness using a simulated dataset. In addition, the paper concludes with a

helpful glossary of widely used terms and a checklist for researchers and practitioners to follow when implementing an RD.

Lee, D. S., and T. Lemieux. 2010. Regression discontinuity designs in economics. *Journal of Economic Literature* 48.1: 281–355. [doi:10.1257/jel.48.2.281].[class:journalArticle]

This paper summarizes the "nuts and bolts" of implementing RD designs in field settings. The paper stresses the close relationship between randomized experiments and RD designs and emphasizes that RD designs should be conceptualized as a particular data generating process (like random assignment) rather than a method of data analysis.

McCrary, J. 2008. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics* 142.2: 698–714.[class:journalArticle]

One of the necessary conditions for RD designs to yield valid effect estimates is that participants cannot manipulate their assignment variable to gain (or avoid) access to treatment. In this article, McCrary proposes a formal test for such sorting (known widely as the McCrary test) that is based on an estimate of the discontinuity occurring at the cutoff point along the running variable.

Schochet, P. Z. 2009. Statistical power for regression discontinuity designs in education evaluations. *Journal of Educational and Behavioral Statistics* 34.2: 238–266.[class:journalArticle]

Schochet discusses the comparatively large sample sizes required for precise RD estimates. Useful as an empirical guide when designing and implementing RD designs for educational evaluation, the paper examines statistical power of RD designs in clustered settings (such as schools and classrooms) as well as the cost of implementing RD designs in such settings.

Schochet, P. Z., T. D. Cook, and J. Deke. 2010. *_Standards for regression discontinuity designs_[https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_rd.pdf]*. Washington, DC: Institute of Education Sciences.[class:report]

This panel report presents the conditions under which RD designs meet What Works Clearinghouse (WWC) evidence standards. The article describes the criteria used to assess whether a study qualifies as an RD under WWC guidelines and details the applicable standards.

Thistlethwaite, D. L., and D. T. Campbell. 1960. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology* 51.6: 309.[class:journalArticle]

Recognized as the first paper to introduce an RD design, this seminal article by Thistlethwaite and Campbell presents the method as an alternative to an ex post facto experiment. Using data from a national scholarship competition, the authors first use an aptitude cutoff score to estimate the effects of increased public recognition on each student's chances of winning the scholarship and then compare these estimates to those resulting from a matching design.

INTERRUPTED TIME SERIES DESIGNS

For cases in which measures of the same outcome are available both prior to and after the introduction of a policy, evaluators may use repeated measures designs to estimate impacts. With a difference-in-difference (DID) design, intervention impacts are determined by whether the treatment group deviates from its baseline mean by a greater amount than the comparison group. Thus, the design controls for baseline outcome differences between groups. When a researcher has outcome measures for multiple time points prior to the introduction of treatment, they may use a comparative interrupted time series design (CITS). With CITS, interventions are evaluated

by looking at whether the treatment group deviates from its baseline *trend* by a greater amount than the comparison group after implementation, thus controlling for baseline differences in outcomes and in growth. This section provides references for two comprehensive guides to CITS in education research. Somers, et al. 2013 is a comprehensive report on the validity of comparative repeated measures designs and acts as a reference in the design of CITS and DID studies. Hallberg, et al. 2018 focuses on the challenges of CITS with only a few data points before and after treatment implementation, as is commonly the case with school-level outcomes. We also borrow from the health literature: Wing, et al. 2018 offers an approachable overview of DID design assumptions. For a more formal exposition of assumptions, see Lechner 2010, Abadie 2005, and Imbens and Wooldridge 2009. Finally, Bloom 2003 offers guidance in the case that a researcher does not have ready access to a comparison group and so must rely on an interrupted time series design (ITS).

Abadie, A. 2005. Semiparametric difference-in-differences estimators. *The Review of Economic Studies* 72.1: 1–19. [doi:10.1111/0034-6527.00321].[class:journalArticle]

Begins with a technical overview of the DID design and its assumptions within the potential outcomes framework. The authors then propose weighting schemes to be used in cases where treatment and comparison group trends are not expected to be parallel due to covariate imbalance.

Bloom, H. S. 2003. Using ‘short’ interrupted time-series analysis to measure the impacts of whole-school reforms. *Evaluation Review* 27.1: 3–49. [doi:10.1177/0193841X02239017][class:journalArticle]

This article covers the ITS design in educational settings. Bloom lays out the assumptions of the design and model specifications (which may be adapted to the comparative case) and

describes extensions such as controlling for changes in student characteristics and combining impact estimates. The article also describes how researchers may conduct analyses on the impact of an intervention on overall score distributions.

Hallberg, K., R. Williams, A. Swanlund, and J. Eno. 2018. Short comparative interrupted time series using aggregate school-level data in education research. *Educational Researcher* 47.5: 295–306. [doi:10.3102/0013189X18769302][class:journalArticle]

This paper acts as a practical guide for researchers implementing CITS designs using school-level data. It provides an overview of design assumptions, threats to validity, and analytic decisions that researchers face. The paper also includes a summary table listing CITS design considerations in education evaluations.

Imbens, G. W., and J. M. Wooldridge. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47.1: 5–86. [doi:10.1257/jel.47.1.5][class:journalArticle]

Imbens and Wooldridge review recent methodological innovations in DID and CITS, including recent non-parametric approaches, the use of artificial control groups, and appropriate standard error corrections.

Lechner, M. 2010. The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics* 4.3: 165–224. [doi:10.1561/0800000014][class:journalArticle]

Lechner surveys the history of the DID design, provides an approachable formalization of the model and its assumptions, and discusses key analysis considerations. The article also includes a discussion of DID extensions, including non-linear DID and DID with matching.

Somers, M. -A., P. Zhu, R. Jacob, and H. Bloom. 2013. *The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational evaluation[https://www.mdrc.org/sites/default/files/validity_precision_comparative_interrupted_time_series_design.pdf]*. *MDRC Working Papers on Research Methodology.* New York, NY: MDRC.[class:report]

Examines the conditions under which CITS designs are able to estimate internally valid treatment effects. The study also outlines design choices available to researchers implementing CITS and/or DID studies in education, including model specification, comparison group choice, and optional design features such as matching.

Wing, C., K. I. Simon, and R. A. Bello-Gomez. 2018. Designing difference in difference studies: Best practices for public health policy research. *Annual Review of Public Health* 39:453–469. [doi:10.1146/annurev-publhealth-040617-013507][class:journalArticle]

Though written for a health policy audience, this article provides an approachable overview to the assumptions and design decisions researchers face with DID studies. The article discusses the construction of comparison groups and the design of sensitivity and robustness checks that may be used to help validate DID assumptions.

NON-EQUIVALENT COMPARISON GROUP DESIGNS

Non-equivalent comparison group designs are perhaps the most widely used approach for examining program effects. Like the randomized experiment, this design includes both a treatment and comparison group. However, unlike the randomized experiment, individuals self-select into treatment conditions or are selected into treatment by a third party. The researcher often does not know all factors that are related to individuals' selection into treatment and their

outcome. As a consequence, the simple mean difference between the treated and untreated participants is very likely biased. When confounders are known and measured reliably by the researcher, covariate adjustment (e.g., regression), and matching methods may be used to estimate causal treatment effects. In this section, we include references of papers that use non-equivalent comparison group designs such as matching via propensity score methods as well as articles examining the importance of pre-test covariates and measures in reducing selection bias. Rosenbaum and Rubin 1983 introduces the concept of a propensity score and explains how propensity score matching can generate unbiased treatment effects if several stringent conditions are met. While the 1984 article is meant for a technically advanced audience, Rosenbaum 1999 can be used as a beginner's textbook: it discusses the principles of statistical inference and the use of observational designs to make such inferences. For a comprehensive review on the use of matching methods to make causal inferences, we recommend Imbens and Wooldridge 2009 and Stuart 2010. We also include references to papers that discuss methods to reduce selection bias in observation designs, including case matching in Cook and Steiner 2010 and reliable covariate measurement in Steiner, et al. 2011 and Wong, et al. 2017.

Cook, T. D., and P. M. Steiner. 2010. Case matching and the reduction of selection bias in quasi-experiments: The relative importance of pretest measures of outcome, of unreliable measurement, and of mode of data analysis. *Psychological Methods* 15.1: 56–68. [doi:10.1037/a0018536][class:journalArticle]

Reanalyzes experimental data in order to compare experimental estimates to estimates from non-equivalent comparison group designs. Similar to the conclusions of Wong, et al. 2017, the authors find that covariate choice can reduce selection bias, particularly when models include measures of pretreatment outcomes.

Imbens, G. W., and J. M. Wooldridge. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47.1: 5–86. [doi:10.1257/jel.47.1.5][class:journalArticle]

Reviews recent methodological innovations in non-equivalent control group designs and provides a comprehensive discussion of related assumptions and design choices.

Rosenbaum, P. R. 1999. *Design of observational studies*. New York: Springer Series in Statistics. [ISBN: 9781441912138][class:book]

This is an introductory textbook on the principles of statistical inference and the use of observational designs to make such inferences. While it briefly discusses causal inference in the context of randomized experiments, the primary focus is on the use of matching techniques (including propensity scores). Threats to internal and external validity and the role of design sensitivity checks are also highlighted.

Rosenbaum, P. R., and D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70.1: 41–55. [doi:10.1093/biomet/70.1.41][class:journalArticle]

Rosenbaum and Rubin introduce the concept of a propensity score—an estimate of the probability that participants are members of either the treatment group or the control group. The article also provides an overview of the applications of propensity scores in non-equivalent comparison group designs and discuss extensions to simple matching, including multivariate adjustment using subclassification and the visual representation of multivariate covariance adjustment using plots.

Steiner, P. M., Cook, T. D., and W. R. Shadish. 2011. On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics* 36.2: 213–236.[class:journalArticle]

Researchers are often unaware of the extent to which covariate measurement impacts the internal validity of non-equivalent comparison group designs. This article compares estimates from a randomized experiment to those from propensity score matching and systematically examines the impact of measurement error on estimate bias.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science* 25.1: 1–21. [doi:10.1214/09-STS313][class:journalArticle]

Written as a guide for researchers interested in using matching methods, this paper summarizes relevant literature on statistical matching from a variety of disciplines. Stuart describes settings in which matching designs are commonly used and provides a comprehensive review of the history and development of matching methods. Key implementation decisions, like covariate selection and matching methods, are also considered.

Wong, V. C., J. Valentine, and K. Miller-Bains. 2017. Empirical performance of covariates in education observational studies. *Journal of Research on Educational Effectiveness* 10.1: 207–236. [doi:10.1080/19345747.2016.1164781][class:journalArticle]

Wong and colleagues compare estimates from twelve observational studies to their experimental benchmarks in order to identify the role of covariates in reducing selection bias. Demonstrates that matching units on pretest measures of the outcome substantially reduces (but does not entirely eliminate) bias in observational studies in education settings.

PRACTICAL CONSIDERATIONS FOR PLANNING IMPACT EVALUATIONS

So far this article has described and provided references for experiments and quasi-experimental designs that yield valid causal estimates under stringent identification strategies. However, even well-designed causal studies can be challenging to implement in field settings. In this section, we highlight practical considerations that researchers must be mindful of when implementing educational evaluations. In no particular order, the considerations address issues of statistical power, heterogeneous or subgroup effects, treatment fidelity, dependence between units in education settings, questions of practical significance, and generalizability of effects from impact evaluations.

Consideration 1: Statistical Power

For evaluations to yield valid impact estimates, researchers must ensure sufficient statistical power for detecting significant effects. Hedges and Rhoads 2010 and Raudenbush, et al. 2007 both provide guidance toward estimating power in education settings. Bloom, et al. 2005 provides estimates of how covariates increase precision while Hedges and Hedberg 2007 provides estimates of interclass correlations in education settings.

Bloom, H. S., Richburg-Hayes, L., and A. R. Black. 2005. Using covariates to improve precision: Empirical guidance for studies that randomize schools to measure the impacts of educational interventions. *Educational Evaluation and Policy Analysis* 29.1: 30-59. [doi: 10.3102/0162373707299550 [class:journalArticle]

Examines how controlling for baseline covariates (especially pretreatment outcomes) improves the precision of educational studies. Bloom shows that baseline covariates can reduce the number of randomized schools needed for a given level of precision by one-tenth to one-half of what would be otherwise needed.

Hedges, L. V., and E. C. Hedberg. 2007. Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis* 29.1: 60–87.[class:journalArticle]

Provides guidance on considerations of intraclass correlations when designing group-randomized experiments. According to the authors, knowledge of intraclass correlation structure to compute statistical power and sample sizes is particularly important to inform experiments that measure the effects of interventions by randomizing schools or classrooms. To fill this gap, the article compiles a range of values of academic achievement and covariate effects that can be used for such calculations.

Hedges, L., and C. Rhoads. 2010. *Statistical power analysis in education research (NCSER 2010–3006)*. Washington, DC: U.S. Department of Education.[class:report]

This report provides a guide for calculating statistical power in multilevel designs that are commonly required in education research. The report focuses on hierarchical and blocked randomized experiments, showing how statistical power depends on intraclass correlations, sample sizes at different levels, the expected effect size, correlation between covariates and the outcome, and the heterogeneity of treatment effects.

Raudenbush, S. W., A. Martinez, and J. Spybrook. 2007. Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis* 29.1: 5–29.[class:journalArticle]

Educational evaluations commonly require that treatment implementation and randomization occur at the school rather than individual level, a design commonly referred to as a group-randomized experiment. This article presents a non-technical guide on the extent to which pre-

treatment blocking and statistical covariate adjustment can increase the statistical power of such group-randomized experiments.

Consideration 2: Subgroup and Heterogeneous Treatment Effects

Researchers are rarely content with estimating only an average treatment effect; they also want to know about treatment effects for particular subgroups and settings. Gerber and Green 2012 provides a comprehensive chapter on designing and analyzing field experiments for heterogeneous effects.

Gerber, A. S., and D. P. Green. 2012. *Field experiments: Design, analysis, and interpretation*. New York: W. W. Norton. [ISBN: 9780393979954][class:book]

Chapter 9 of this book covers methods of detecting heterogeneous treatment effects while also describing the limits to what experimental data can tell us about subgroup effects. The authors argue for a design approach to estimating heterogeneity when researchers are interested in studying heterogeneity in treatment effects by condition characteristics.

Consideration 3: Treatment Fidelity

Interventions are rarely implemented in ways that conform entirely with their original program theory and so careful attention must be paid to ensure that treatment and control conditions are implemented as intended. O'Donnell 2008 and Smith, et al. 2007 provide comprehensive reviews of the treatment fidelity literature, covering how researchers may plan for, collect, and report fidelity data.

O'Donnell, C. L. 2008. Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research* 78.1: 33–84. [doi:10.3102/0034654307313793][class:journalArticle]

This article acts as a guide for researchers hoping to understand how fidelity of implementation impacts the interpretation of treatment effects. Through a review of K–12 curriculum evaluations, O'Donnell clarifies the definition, conceptualization, and measurement of fidelity of implementation.

Smith, S. W., A. P. Daunic, and G. G. Taylor. 2007. Treatment fidelity in applied educational research: Expanding the adoption and application of measures to ensure evidence-based practice. *Education and Treatment of Children* 30.4: 121–134.[class:journalArticle]

This article argues the importance of planning for, collecting, and reporting fidelity data so that evaluators may ensure that the intervention is delivered with accuracy and conformity. The authors also review the five areas of treatment fidelity proposed by the Health Behavior Change Consortium: study design, training, treatment delivery, treatment receipt, and treatment enactment.

Consideration 4: Dependence Among Units

In addition to implementation challenges, educational evaluations suffer from lack of independence between units: children are nested within classrooms, classrooms are nested within schools, and so on. In such cases, researchers are advised to use hierarchical models to address the dependence between units. Raudenbush and Byrk 2001 is a seminal text for using hierarchical linear models to address nested data structures in educational settings, while Hox 2010 introduces readers to more advanced forms of linear modeling along with example datasets for practical understanding.

Hox, J. J. 2010. *Multilevel analysis: Techniques and applications*. 2d ed. Routledge. [ISBN: 9781848728455][class:book]

This book is useful for graduate students and researchers who are interested in using multilevel models. It extends the introductory coverage of multilevel models in Raudenbush and Bryk 2001 by including advanced extensions and by providing data sets that readers can use to sharpen their understanding of the estimation and interpretation of such models.

Raudenbush, S. W., and A. S. Byrk. 2001. *Hierarchical linear models: Applications and data analysis methods.* 2d ed. Los Angeles: SAGE. [ISBN: 9780803946279][class:book]

Used widely as a textbook for graduate students, this book comprehensively covers the applications and analysis of hierarchical linear models and provides extensions. The book provides illustrative examples and easy-to-understand explanations of the theory and applications of hierarchical linear models and is particularly useful in education since most education research involves hierarchical data structures.

Consideration 5: Practical Significance of Effect Estimates

Meaningful effect sizes are needed for interpreting the practical importance of results. Cohen 1988 provides ad hoc guidelines for judging "small," "medium," and "large" effects while Hill, et al. 2007 uses results from education interventions to generate empirical distributions of education effect sizes for comparing results.

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences*. Mahweh, NJ: Lawrence Erlbaum Associates. [ISBN: 9780805802832][class:book]

Offers a nontechnical guide to power analysis for research planning and analysis, covering t-tests for means, differences between correlation coefficients, differences between proportions, chi-square tests for goodness of fit, F-tests of means, and F-tests of variance proportions. The publication assumes the reader has a working knowledge of each of the statistical analyses

covered. Importantly, the book also introduces guidelines for interpreting the magnitude of effect sizes.

Hill, C. J., H. S. Bloom, and M. W. Lipsey. 2007. *Empirical benchmarks for interpreting effect sizes in research[https://www.mdrc.org/sites/default/files/full_84.pdf]*. *MDRC Working Papers on Research Methodology.* New York, NY: MDRC.[class:report]

Offers three types of empirical benchmarks for interpreting effect sizes in education settings: expectation for student growth in achievement over time; policy relevant gaps in student achievement; and effect size results from previous research on similar interventions. The article is broadly accessible for researchers and research consumers.

Consideration 6: Generalizability

Finally, though experiments are the gold standard for internal validity, their samples and settings rarely match the conditions of interest. Educational evaluations do not only need to contend with the internal validity of their study but also the external validity of the study or the extent to which study findings may be generalized to populations, settings, and treatments of interest. Bracht and Glass 1968 and Cronbach and Shapiro 1982 provide frameworks for defining and assessing threats to external validity. Stuart, et al. 2011 and Tipton 2014 propose metrics for assessing the generalizability of experimental samples to a population of interest.

Bracht, G. H., and G. V. Glass. 1968. The external validity of experiments. *American Educational Research Journal* 5.4: 437–474.[class:journalArticle]

This article builds upon Campbell and Stanley 1963 (cited under *General Overviews*) by refining, expanding, and elaborating on the book's list of potential threats to external validity. The article defines two types of external validity—population validity and ecological

validity—and describes in-depth the potential threats to each. The article is broadly relevant for researchers and research consumers.

Cronbach, L. J., and K. Shapiro. 1982. Designing evaluations of educational and social programs. San Francisco: Jossey-Bass. [ISBN: 9780875895253][class:book]

In this seminal book, Cronbach and Shapiro suggest that conversations of external validity should be guided by defining UTOS: units (the population on which a conclusion is sought), treatments (the program as planned and implemented), operations (the data collected), and settings (the social context of the program). The book argues that when researchers clearly define UTOS, questions of generalizability may be more clearly assessed.

Stuart, E. A., S. R. Cole, C. P. Bradshaw, and P. J. Leaf. 2011. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society* 174.2: 369–386. [doi:10.1111/j.1467–985x.2010.00673][class:journalArticle]

Proposes the use of propensity-score metrics to quantify the similarity of the sample of a randomized experiment to the target population for generalization. Instead of using propensity score methods to estimate the probability of treatment, authors use propensity score methods to estimate the probability of belonging to the experimental sample. The propensity score is used both to quantify the similarity of the two groups and to match the control group outcomes to the population of interest.

Tipton, E. 2014. How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics* 39.6: 478–501. [doi:10.3102/1076998614558486][class:journalArticle]

This article proposes an index which may be used to assess the generalizability of a randomized experiment to a target population of interest. Like Stuart, et al. 2011, the metric

relies on propensity scores, but the article proposes that the metric can be bounded between zero and one using the Bhattacharyya coefficient. The article is appropriate for readers with an advance understanding of statistics.