Evaluating Program Impacts:
BIBLIOGRAPHY

Agodini, 2004, 'Are Experiments the Only Option? A Look at Dropout Prevention Programs,' The Review of Economics and Statistics, Vol. 86, No. 1, Pages 180-194. (the paper argues that unobserved factors often exert powerful influences on outcomes and these factors are often difficult to control for using statistical methods, I totally agree)

Aid Leap, 2017, Contribution vs Attribution – A Pointless Debate. (well-argued and cuts through the confusion).
See: Contribution vs Attribution – A Pointless Debate – AID LEAP

Almutairi, A., Gardner, G. and McCarthy, A. 2014, 'Practical Guidance for the Use of a Pattern-Matching Technique in Case-study Research: A Case Presentation", Nursing and Health Sciences, 16(2), pp. 239-244. (useful example of how to undertake pattern-matching in case studies)

Anderson, 2010, Proven Programs are the exception, not the rule, blog post: http://blog.givewell.org/2008/12/18/guest-post-proven-programs-are-the-exception-not-the-rule/ (The author argues that examples of proven programs are rare. Their scarcity stems from two main factors: 1) the vast majority of social programs and services have not yet been rigorously evaluated, and 2) of those that have been rigorously evaluated, most, including those backed by expert opinion and less-rigorous studies, turn out to produce small or no effects, and, in some cases negative effects)

Angrist J. & Pischke J. 2015, Mastering Metrics: The paths from cause to effect. (easy to read guidance on what the authors see as the five most valuable econometric methods - random assignment, regression, instrumental variables, regression discontinuity designs, and differences in differences)

Antonakis, J. et al 2010, 'On Making Causal Claims: A Review and Recommendations', The Leadership Quarterly, Volume 21, Issue 6, 1086-1120. (this paper assessed 110 published impact studies in the field of leadership and found that most were significantly flawed, this paper illustrates the challenges of undertaking causal analysis using econometric methods with passive observational data and also includes better practice guidelines, highly recommended)

Arnold Ventures, 2016, Key Items to Get Right When Conducting Randomized Controlled Trials of Social Programs.
(brief but helpful guidance)

Asher, 1983, Causal Modeling, Sage. (an introduction to causal/statistical modeling)

Aston. T. 2021 Complexity; context, checklists? (blog) (an interesting discussion of evaluation, complexity and interactions between the program, beneficiaries, and context in international development). Available at: Complexity; context, checklists?. Michael Bamberger recently offered a… | by Thomas Aston | Medium

Asian Development Bank, 2006, Impact Evaluation: Methodological and Operational Issues. (a brief introduction, interesting discussion of common myths about impact studies)

AusAID, 2012, Impact Evaluation: A discussion paper for AusAID practitioners. (a brief introduction to the topic)

Bamberger, et al 2006, RealWorld Evaluation, Sage. (an excellent overview of how to undertake evaluations of development programs while facing various types of constraints, also includes a discussion of the most commonly used designs for evaluating the impact of development programs)

Bamberger, 2012, Introduction to Mixed Methods in Impact Evaluation, InterAction.

Banerjee, A. et al, 2021, 'Selecting the Most Effective Nudge: Evidence from a Large-Scale Experiment on Immunization', National Bureau of Economic Research, Working Paper 28726, DOI 10.3386/w28726. (an excellent example of undertaking a factorial experiment to identify which combination of 3 interventions with what level of dosage is the most effective and the least expensive).

Barlow & Hersen, 1989, Single Case Experimental Designs, Pergamon. (an under-utilized approach in my opinion)

Barnett and Munslow 2014, Process Tracing The Potential and Pitfalls for Impact Evaluation in International Development, Institute of Development Studies. (a useful and balanced overview of the method, one of the better papers that I have come across on this topic)

Barnow, B. 2010, 'Setting up social experiments: the good, the bad, and the ugly', Journal for Labour Market Research, 43, 91-105. (this paper explores common problems and offers suggestions on ways to deal with them)

Barnow & Greenberg, 2020, 'Conducting Evaluations Using Multiple Trials', American Journal of Evaluation, 41(4), 564-580. (discusses the rationale for multiple trials and why programs vary across sites, recommended)

Beach, D. 2019, 'Multi-Method Research in the Social Sciences: A Review of Recent Frameworks and a Way Forward', Government and Opposition, 0, 1–20. (a good discussion and comparison of variable-based vs case-based research strategies)

Becker, 2000, Discussion Notes: Causality, http:/web.uccs.edu/lbecker/Psy590/cause.htm (a brief summary of different philosophical perspectives on causality)

Bédécarrats, Guérin, and Roubaud, (eds) 2020, Randomized Control Trials in the Field of Development: A Critical Perspective, OUP Oxford. (what is the exact scope of the experimental method? Which sorts of questions are RCTs able to address and which do they fail to answer? This book provides answers to these questions, explaining how RCTs work, what they can achieve, why they sometimes fail, how they can be improved and why other methods are both useful and necessary, this book gives more weight to the anti RTC camp)

Befani, B. 2012, Models of Causality and Causal Inference – Annex to Stern et al DFID Working Paper 38, UK Department for International Development. (a helpful summary)

Befani, B. 2016, Choosing Appropriate Evaluation Methods: A Tool for Assessment and Selection, Bond, UK. (a review of various approaches)

Befani, B. 2020, 'Diagnostic evaluation and Bayesian Updating: Practical solutions to common problems', Evaluation, vol 26, no 4. (discusses practical issues relating to the application of diagnostic principles to theory based evaluation, useful but be careful confusing theory building and refinement with causal inference, also has some insightful observations about the potential for confirmation bias in process tracing approaches)

Bennet, A. and George, A. 1997, Process Tracing in Case Study Research, MacArthur Foundation Workshop on Case Study Methods. (a comprehensive discussion of the strengths and weaknesses of this method, a challenging method to apply rigorously, high recommended)

Bernal, J. Cummins, S. and Gasparrini, A. 2018, 'The use of controls in interrupted time series studies of public health interventions', International Journal of Epidemiology, Volume 47, Issue 6. (a user friendly summary of the approach, recommended)

BetterEvaluation, 2013, Understand Causes of Outcomes and Impacts. (a general overview of methods)

BetterEvaluation, 2018, Impact Evaluation. (resources for impact evaluation). Available on the internet: https://www.betterevaluation.org/en/themes/impact_evaluation

Bigelow, J. et al 2021, A Guide for using Administrative Data to Examine Long Term Outcomes in Program Evaluation, OPRE Report 2021-145. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Blankshain & Stigler, 2020, 'Applying Method to Madness: A User's Guide to Causal Inference in Policy Analysis', Policy and Academia, vol 3, issue 3. (a useful summary for the policy community)

Blatter & Blume, 2008, 'In Search of Co-variance, Causal Mechanisms or Congruence? Towards a Plural Understanding of Case Studies', Swiss Political Science Review, 14(2): 315–56. (an excellent discussion of the difference between variable based and case based research, plus the difference between process tracing vs congruence/pattern marching approaches, highly recommended)

Block, 1999, Flawless consulting, Pfeiffer. (highly recommended, excellent chapter on working with resistant clients)

Bloom, Hill & Riccio, 2003, 'Linking Program Implementation and Effectiveness: Lessons from a Pooled Sample of Welfare-to-Work Experiments', Journal of Policy Analysis and Management, Vol. 22, No. 4, pp. 551-575. (implementation can have a major effect on outcomes)

Bloom, Michalopoulos, Hill, & Lei, 2002, Can Nonexperimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs? (the authors say no) Available free on the net at: http://aspe.hhs.gov/pic/reports/acf/7541.pdf

Blossfeld, Golsch & Rohwer, 2007, Event History Analysis with Stata, LEA. (Excellent discussion of causal modelling with longitudinal passive observational data and the limitations of using cross-sectional surveys for causal inference, recommended)

Bohte, J. and Meier, K. 2000, 'Goal Displacement: Assessing the Motivation for Organizational Cheating', Public Administration Review, 60, 173-182. (Discusses the motivations, incentives, and methods used by agencies to present an overly positive view of their performance. The lesson here is that accountability and incentive systems need to be designed with great care.)

Bonell, et al 2015, 'Dark Logic: Theorising the Harmful Consequences of Public Health Interventions', Journal Epidemiol Community Health, 69, 95-98. (an interesting paper on identifying and preventing the adverse effects / outcomes of public health interventions and the need to understand the underlying causal mechanisms)

Boruch, 2005, Randomized Experiments for planning and evaluation: A practical guide, Sage. (a good introduction to the topic)

Brady, 2002, Models of Causal Inference: Going Beyond the Neyman-Rubin-Holland Theory, Paper Presented at the Annual Meetings of the Political Methodology Group, University of Washington, Seattle, Washington. (paper reviews four of the more common theories of causality)

Brinkerhoff, 1991, Improving Development Program Performance: Guidelines for Managers, Lynne Rienner. (includes a discussion of the most common causes of performance problems in development programs, recommended)

Brodeur, Cook & Heyes, 2020, 'Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics', American Economic Review, Vol. 110, No. 11. (the credibility revolution in economics has promoted causal identification using randomized control trials (RCT), difference-in-differences, instrumental variables and regression discontinuity design. Applying multiple approaches to over 21,000 hypothesis tests published in 25 leading economics journals, the authors found that the extent of p-hacking and publication bias varied greatly by method. IV and to a lesser extent DID are particularly problematic)

Brown, E. and Tanner, J. 2019, Integrating Value for Money and Impact Evaluations, World Bank. (overview of the issues and opportunities)

Budhwani, S. and McDavid, J. 2017, 'Contribution Analysis: Theoretical and Practical Challenges and Prospects for Evaluators', Canadian Journal of Program Evaluation, 32.1, 1-24. (a good summary of CA's strengths and weaknesses and how it has evolved over time, recommended).

Campbell, 1969, 'Reforms as Experiments', American Psychologist, 24, p.409-429. (classic article on the politics of reforms and practical research designs)

Campbell & Stanley, 1963, Experimental and Quasi-experimental Designs for Research, Rand McNally. (the all-time classic text, discusses the strengths and weaknesses of various research designs for assessing program impacts, highly recommended)

Carroll, C. et al 2007, 'A conceptual framework for implementation fidelity', Implementation Science, 2:40 doi:10.1186/1748-5908-2-40. (only by understanding and measuring whether an intervention has been implemented with fidelity can evaluators and practitioners gain a better understanding of how and why an intervention works, and the extent to which outcomes can be improved)

Carter, Klein and Day, 1992, How organisations measure success: the use of performance indicators in government, Routledge. (offers a useful typology of different types of performance indicators and how they can be used/misused)

Christie and Alkin, 2019, 'Theorists' Models in Action: A Second Look, New Directions for Evaluation, no 163. (a useful illustration of how different theorists would approach the same evaluation task)

Coalition for Evidence-Based Policy, 2014, Which comparison-group (quasi-experimental) study designs are most likely to produce valid estimates of a program's impact? (excellent brief summary of the research evidence) Available free on the net.

Coly, A., & Parry, G. 2017, Evaluating Complex Health Interventions: A Guide to Rigorous Research Designs, AcademyHealth. (a useful overview of research designs for undertaking impact evaluations). Available free on the net at: https://www.academyhealth.org/sites/default/files/AH_Evaluation_Guide_FINAL.pdf

Cook, 2000, 'The false choice between theory-based evaluation and experimentation', in Rogers et al (eds) Program Theory in Evaluation: Challenges and Opportunities, Jossey-Bass. (excellent discussion of the main limitations of theory based methods for impact assessment)

Cook, 2018, Twenty-six assumptions that have to be met if single random assignment experiments are to warrant "gold standard" status: A commentary on Deaton and Cartwright, Social Science & Medicine 210 37-40. (a useful clarification of key issues, strengths and weaknesses)

Cook et al, 2010, 'Contemporary Thinking About Causation in Evaluation: A Dialogue With Tom Cook and Michael Scriven', American Journal of Evaluation, 31(1) 105-117; (highly recommended, a debate between two experts with different views, I find Cook's position convincing)

Cook & Campbell, 1979, Quasi-experimentation, Houghton Mifflin. (excellent, contains a useful review of different theories of causality and how to test for causal relationships as well as the application of quasi-experiments for impact evaluations, highly recommended)

Cook, Shadish & 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. (this reference recommends regression-discontinuity designs, matching geographically local groups on pre-treatment outcome measures, and modeling a known selection process)

Corlazzoli and White, 2013, Measuring the Un-Measurable, DFID. (a useful resource).

Cox & Wermuth, 2001, 'Causal Inference and Statistical Fallacies', International Encyclopedia of the Social & Behavioral Sciences, Pages 1554-1561. (paper outlines different definitions of causality and some common errors)

Cracknell, B. E. 2000. Evaluating Development Aid: Issue, Problems and Solutions. Sage Publications, New Delhi. (interesting discussions of evaluating for learning vs for accountability plus the politics of evaluation)

Cracknell, B. E. 2001, 'Knowing is all: Or is it? Some reflections on why the acquisition of knowledge focusing particularly on evaluation activities, does not always lead to action'. Public Administration and Development, 31, 371-379.

Craig, et al 2018 (draft), Developing and Evaluating Complex Interventions, Medical Research Council. (some useful guidance for health programs)

Cragun, D. et al, 2016, 'Qualitative Comparative Analysis: A Hybrid Method for Identifying Factors Associated with Program Effectiveness', Journal of Mixed Methods Research, 10(3): 251–272. (a useful introduction to the method, good for identifying patterns of factors associated with positive outcomes)

Creswell, 2020, Research Design: Qualitative, Quantitative, and Mixed Methods Approaches, Sage. (an introduction to the tropic)

Cronbach, et al 1980, Toward Reform of Program Evaluation, Jossey-Bass. London. (a great discussion of evaluation's potential purposes and operating in a political environment, highly recommended

Davidson, E. J. 2000. 'Ascertaining causation in theory-based evaluation'. In P.J. Rogers, T.A. Hacsi, A. Petrosino, and T.A. Huebner (eds.), "Program theory in evaluation: challenges and opportunities". New Directions in Evaluation, Number 87:17-26, San Francisco, CA. (provides an overview of various methods)

Davidson, E. J. 2004, Evaluation methodology basics: The nuts and bolts of sound evaluation, Sage. (suggests 8 techniques for causal inference)

Davis, 1985, The Logic of Causal Order, Sage. (worth a quick read)

Dawid, A. 2007, Fundamentals of Statistical Causality. (a good intro and description of the challenges of using passive observational research designs). https://pdfs.semanticscholar.org/c4bc/ad0bb58091ecf9204ddb5db7dce749b0d461.pdf

Deaton, A. 2010, What Can We Learn From Randomized Control Trials? Chicago: Milton Friedman Institute http://mfi.uchicago.edu/events/20100225_randomizedtrials/index.shtml

Deaton, A. & Cartwright, N. 2018, 'Understanding and misunderstanding randomized controlled trials', Social Science & Medicine 210 2–21. (a summary of key issues, RTC strengths and weaknesses, Cook 2018 refutes some of their comments)

Delahais, T and Toulemonde. J. 2017, 'Making Rigorous Causal Claims in a Real-life Context: Has Research Contributed to Sustainable Forest Management?' Evaluation, 23, 4, 370-388. (The authors conclude that substantiating and quantifying casual claims is not something that contribution analysis can deal with on its own. However by combining CA with other analytical approaches such as process tracing and realist evaluation it is possible to formulate rigorous contribution claims)

Denniss, R. 2021, Econobabble – How to decode political spin and economic nonsense, Black Inc, Australia. (some sharp criticisms of how economic modeling is sometimes misused)

Dept of Finance, 1987, Evaluating Government Programs, Australian Government Publishing Service. (introductory, includes a useful table comparing different types of research designs)

Dept of Finance and Administration, 2006, Handbook of Cost-Benefit Analysis, Australian Government Publishing Service. (an easy to read introduction). This book is available free on the net at: http://www.finance.gov.au/FinFramework/fc_2006_01.html

Dimova, R. 2019, 'A Debate that Fatigues…: To Randomise or Not to Randomise; What's the Real Question?', The European Journal of Development Research, Volume 31, Issue 2, pp 163–168. (argues that the RTC debate is increasingly fatiguing and tends to overemphasize methodological peculiarities at the expense of conceptual issues, the resolution of which is crucial for successful policy making)

Dixon, V. and Bamberger, M. 2022, Incorporating process evaluation into impact evaluation, 3IE. (a useful discussion). This document is available free on the net at: Incorporating process evaluation into impact evaluation: what, why and how | 3ie (3ieimpact.org)

Dominici, F. et al, 2021, "From Controlled to Undisciplined Data: Estimating Causal Effects in the Era of Data Science Using a Potential Outcome Framework", Harvard Data Science Review, 10.1162/99608f92.8102afed. (a discussion of the principles of causal inference, the challenges of observational research designs, and the use of big data).

Donaldson, Christie & Mark, 2009, What Counts as Credible Evidence in Applied Research and Evaluation Practice?, Sage. (reviews the debates on this topic, offers a range of perspectives)

Dong, N. et al 2017, "Can Propensity Score Analysis Approximate Randomized Experiments Using Pretest and Demographic Information in Pre-K Intervention Research?" Evaluation Review, XX(X). (The answer is no; propensity score analysis can sufficiently remove bias only if certain key assumptions are satisfied. In practice these assumptions are usually not tested for and when they are tested we generally find that the assumptions do not hold)

Ebneyamini & Moghadam, 2018, "Toward Developing a Framework for Conducting Case Study Research", International Journal of Qualitative Methods, Volume 17: 1–11. (helpful introduction)

Elze, M. C. 2017, "Comparison of Propensity Score Methods and Covariate Adjustment, Evaluation in 4 Cardiovascular Studies", Journal of the American College of Cardiology, vol 69, no 3. (PS methods are not necessarily superior to conventional covariate adjustment, and care should be taken to select the most suitable method)

Epstein & Klerman, 2012, "When is a Program Ready for Rigorous Impact Evaluation?" Evaluation Review, vol 36, pp. 373-399. (when it has a plausible theory of change in place and its implementation has been assessed as sound)

European Evaluation Society, 2007, Statement: The Importance of a Methodologically Diverse Approach to Impact Evaluation.

Evidence in Governance and Politics, 2020, Methods Guides. (a very useful website with technical guidance for conducting impact evaluations). See: https://egap.org/list-methods-guides

Evidence in Governance and Politics, 2020, 10 Things to Know about Causal Inference. (highly recommended). See: https://egap.org/methods-guides/10-things-you-need-know-about-causal-inference

Evidence in Governance and Politics, 2020, 10 Types of Treatment Effects You Should Know About. (highly recommended). See: http://egap.org/methods-guides/10-types-treatment-effect-you-should-know-about

Faddar, J. et al, 2018, 'School self-evaluation: self-perception or self-deception? The impact of motivation and socially desirable responding on self-evaluation results', School Effectiveness and School Improvement, 29:4, 660-678, DOI:10.1080/09243453.2018.1504802 (motivation and socially desirable responding strongly affect schools' self-evaluations)

Fredriksson et al, 2019,' Impact evaluation using difference in differences', RAUST Management Journal, 54, 4, 519-532. (a good discussion of the assumptions and limitations of this method)

Freedman and Collier, 2009, Statistical Models and Causal Inference: A Dialogue with the Social Sciences, Cambridge University Press. (argues that statistical techniques are seldom an adequate substitute for substantive knowledge of the topic, having a good research design, relevant data and undertaking empirical testing in diverse settings; I totally agree)

Fretheim A., Tomic O. 2015, 'Statistical process control and interrupted time series: a golden opportunity for impact evaluation in quality improvement', BMJ Qual Saf; 24:748–752. (helpful introduction to this topic)

Gaarder, M. 2019, A Commentary to 'Bridging to Action Requires Mixed Methods, Not Only Randomised Control Trials', The European Journal of Development Research, Volume 31, Issue 2, pp 169–173. (this paper puts the debates about RTCs into context, highly recommended)

Gates & Dyson, 2017, "Implications of the Changing Conversation About Causality for Evaluators", American Journal of Evaluation, vol. 38, no. 1 (an introductory overview of issues for consideration plus six guidelines for evaluators seeking to make causal claims: (1) being responsive to the situation and intervention, (2) building relevant and defensible causal arguments, (3) being literate in multiple ways of thinking about causality, (4) being familiar with a range of causal designs and methods, (5) layering theories to explain causality at multiple levels; and (6) justifying the causal approach taken to multiple audiences)

Gao, X. et al 2019, 'Evaluating Program Effects: Conceptualizing and Demonstrating a Typology, Evaluation and Program Planning, 72, 88-96. (some interesting comments about collective impact and different types of program effects)

Gelman, A. and Aki Vehtari, A. 2021, "What are the most important statistical ideas of the past 50 years?", arXiv:2012.00174 [stat.ME], Cornell University. (quite interesting, for the enthusiast)

Gerring, J. 2009, Causal Mechanisms: Yes, But… (well worth reading, points out the challenges in examining causal mechanisms, in my view most evaluation discussions of mechanisms confuse building and refining theories with making causal inferences, if you are interested in process tracing or pattern matching you should read this paper)
http://people.bu.edu/jgerring/documents/CausalMechanisms_Extended.pdf

Gertler, P. et al 2010, Impact Evaluation in Practice, World Bank. It is available free on the net at:
http://documents.worldbank.org/curated/en/2011/01/13871146/impact-evaluation-practice

Glazerman, Levy & Myers, 2002, Nonexperimental Replications of Social Experiments: A Systematic Review, Mathmatica Policy Research Inc. (this research paper concludes that more often than not, statistical models do a poor job of estimating program impacts, highly recommended). This report is available free on the net at: http://www.mathematica-mpr.com/publications/PDFs/nonexperimentalreps.pdf

Gleason, et al 2018, "RD or Not RD: Using Experimental Studies to Assess the Performance of the Regression Discontinuity Approach", Evaluation Review, 42,(1), 3-33. (paper concludes that RDs often provide accurate estimates of impacts although the results are sensitive to the manipulation of the assignment variable)

Glennerster & Takavarasha, 2013, Running Randomized Evaluations: A Practical Guide, Princeton University Press.
(a comprehensive and handy guide to undertaking randomized impact evaluations of social programs)

Glewwe & Todd, 2022, Impact Evaluation in International Development: Theory, Methods and Practice, World Bank Group. (reasonably comprehensive at 400 pages but written from the standard economists' perspective)

Gilovich, 1991, How we know what isn't so: The fallibility of human reason in everyday life, Free Press, New York. (demonstrates how cognitive, social and motivational processes distort our thoughts, beliefs, judgments and decisions)

Goodman, L. et al, 2018, "Beyond the RCT: Integrating Rigor and Relevance to Evaluate the Outcomes of Domestic Violence Programs", American Journal of Evaluation, vol. 39, no. 1 (argues that it is important to match the evaluation design to the nature of the intervention, in addition a more inclusive conceptualization of credible evidence is required)

Gouleta, M. et al, 2020, 'Understanding the dynamic interinfluences of implementation processes: An illustration by multiple case studies', Evaluation and Program Planning, https://doi.org/10.1016/j.evalprogplan.2020.101798. (a clear illustration of how implementation fidelity can affect the achievement of desired outcomes)

Gray, K. 2019, Making Causal Inferences from Observational Data, LinkedIn. (a helpful short summary of the challenges plus some useful references).

Gray, K. 2020, How to Tell Good Science from Bad Science, LinkedIn. (a particularly good summary).

gsocialchange, 2017, How do you know whether your intervention had an effect (a website with a range of resources) https://sites.google.com/site/gsocialchange/cause

Guba and Lincoln, 1989, Fourth Generation Evaluation, Sage. (the authors argue that 'cause and effect' do not exist except by imputation, a constructivist perspective; I don't agree with this approach but many 'qualitative' researchers do)

Gugerty & Summer 2018, "Ten Reasons Not to Measure Impact and What to do Instead", Stanford Social Innovation Review, (argues that impact evaluations are only a good investment in the right circumstances, i.e. when matched to the evaluation question, it's feasible to undertake an impact evaluation, the program is stable, etc. Recommended)

Handan-Nader, Ho and Elias, 2020, 'Feasible Policy Evaluation by Design: A Randomized Synthetic Stepped Wedge Trial of Mandated Disclosure in King County', Evaluation Review, Vol. 44(1) 3-50. (a good example of using a non-traditional RTC design to overcome practical constraints)

Handley, M. et al, 2018, 'Selecting and improving quasi experimental designs in effectiveness and implementation research', Annual Review Public Health; 39:5–25. (a good summary of options and the merits of different approaches)

Hall J. 2020, 'Assessing the effectiveness of development co-operation: Method matters', Dev Policy Rev; 00:1–17. https://doi.org/10.1111/dpr.12486 (we are asking the wrong questions and using the wrong methods for our questions, well argued)

Hatry et al 1981, Practical Program Evaluation for State and Local Governments, The Urban Institute Press (a classic introductory text with an excellent discussion of when experiments are feasible and appropriate)

Hawkins, A. 2014, 'The Case for Experimental Design in Realist Evaluation', Learning Communities: International Journal of Learning in Social Contexts, 14, 46-59. DOI: http://doi.org/10.18793/LCJ2014.14.04. (a well-argued paper that helps to correct certain misunderstandings)

Hay, C. 2016, 'Process tracing: a laudable aim or a high-tariff methodology?' New Political Economy, 21 (5). pp. 500-504. ISSN 1356-3467 (explains the significant limitations of process tracing, I think the method has been over sold)

Hedström, P. 2009, 'Studying Mechanisms To Strengthen Causal Inferences In Quantitative Research', in Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier (eds) The Oxford Handbook of Political Methodology. (Benefits of studying mechanisms: an understanding of the mechanisms at work can improve statistical inference by guiding the specification of the statistical models to be estimated; mechanism-based models can strengthen causal inferences by showing why, acting as they do, individuals bring about the social outcomes they do)

Hedström & Ylikoski, 2010, 'Causal Mechanisms in the Social Sciences', Annual Review of Sociology. (a review of common theories and key issues in this approach)

Hernán M. and Robins, J. 2020, Causal Inference: What If, CRC Press. (a book that aims to help health and social scientists generate and analyze data to make causal inferences that are explicit about both the causal question and the assumptions underlying the data analysis, comprehensive and very technical, useful discussions of interactions and causal graphs)

HM Treasury, 2020, Magenta Book Annex A - Analytical methods for use within an evaluation, (excellent summary of the strengths and limitations of various methods, recommended). Available free on the internet at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/877202/Magenta_Book_Annex_A._Analytical_methods_for_use_within_an_evaluation.pdf

Holland, P. 1986, 'Statistics and Causal Inference'. Journal of the American Statistical Association. Vol. 81 pp. 945-960.

Hopkins, A. et al, 2020, The Experimenter's Inventory: A Catalogue of Experiments for Decision-Makers and Professionals, Alliance for Useful Evidence, Nesta, London. (a useful plain English summary of the pros and cons of different research designs for impact evaluation, addresses common criticisms of RTCs, uses nonstandard terminology at times which can be a bit confusing)

Hopkins, L. 2021, Tools and tips for implementing contribution analysis: A quick guide for practitioners, Itad Ltd. (discussion of the theory vs the practice of CA).

Hughs & Hutchings, 2011, Can We Obtain the Required Rigour Without Randomisation? Oxfam GB's Non-Experimental Global Performance Framework, 3IE. (suggests alternative approaches for NGOs)

Illari, P. 2014, Causality: Philosophical Theory meets Scientific Practice, OUP Oxford. (an introduction to the philosophy of causality, useful for evaluators unacquainted with philosophy)

Impact and Innovation Unit, Government of Canada, 2019, Measuring Impact by Design: A guide to methods for impact measurement: (a new reference guide for those involved in the design, delivery, procurement or appraisal of impact measurement strategies in Canada, a helpful summary of different methods).

Informed Choices Network, 2019, A framework for thinking critically about claims, evidence, and choices. (useful resources). Available free on the internet at: https://thatsaclaim.org/

Intrac 2017, Most significant Change, (a description of MSC plus a summary of strengths and weaknesses). Available free on the internet at: https://www.intrac.org/wpcms/wp-content/uploads/2017/01/Most-significant-change.pdf

Intrac, 2017, Impact Assessment, (a useful summary of issues to consider). Available free on the internet at: https://www.intrac.org/wpcms/wp-content/uploads/2017/01/Impact-Assessment.pdf

Intrac, 2018, Reporting Change, (a useful easy to read overview, recommended). Available free on the internet at: https://www.intrac.org/wpcms/wp-content/uploads/2017/01/Reporting-change.pdf

Intrac, 2021, Qualitative Impact Protocol. (I don't find this method very persuasive, similar limitations as the most significant change approach) https://www.intrac.org/wpcms/wp-content/uploads/2019/05/QUIP.pdf

Ioannidis, J. 2018, 'Randomized controlled trials: Often flawed, mostly useless, clearly indispensable: A commentary on Deaton and Cartwright', Social Science & Medicine, Volume 210, Pages 53-56. (insightful, reviews problems with RTCs and the common limitations of observational studies)

J-PAL, 2017, Impact Evaluation Toolkit. (a guide to impact evaluation methods)

J-PAL, 2020, Library of Research Resources. (with an emphasis on experiments) https://www.povertyactionlab.org/research-resources?utm_source=newsletter&utm_medium=email&utm_campaign=aug20&view=toc

Jabeen, S. 2018, 'Unintended Outcomes Evaluation Approach: A Plausible Way to Evaluate Unintended Outcomes of Social Development Programmes', Evaluation and Program Planning, 68, 262-274. (a good paper addressing one of evaluation's weak areas, highly recommended)

Jacob, R. et al 2012, A Practical Guide to Regression Discontinuity, MDRC. (the title says it all, available free on the internet)

Jimenez. E. 2019, Be careful what you wish for: cautionary tales on using single studies to inform policymaking, 3IE Blog, (explains why we need to be careful about generalising from singe studies) https://www.3ieimpact.org/blogs/be-careful-what-you-wish-cautionary-tales-using-single-studies-inform-policymaking

Johns, M. et al 2012, 'Evaluating the NYC Smoke-Free Parks and Beaches Law: A Critical Multiplist Approach', American Journal of Community Psychology, DOI: 10.1007/s10464-012-9519-5. (a helpful explanation of this approach, recommended)

Johnson & Ahn, 2017, 'Causal Mechanisms' in The Oxford Handbook of Causal Reasoning, Waldmann (ed), (this chapter reviews empirical and theoretical results concerning knowledge of causal mechanisms—beliefs about how and why events are causally linked)

Judd & Kenny, 1981, Estimating the Effects of Social Interventions, Cambridge. (heavy emphasis on statistical applications, for the enthusiast)

Kahneman, D. 2013, Thinking, Fast and Slow; Farrar, Straus and Giroux. (discusses why cognitive biases are common across all aspects of our lives and why human beings are generally unable to accurately perceive causal relationships, highly recommended)

Kazdin, A. 2011, Single-Case Research Designs, Oxford University Press. (an under-utilized approach in my opinion)

Kenny, 2004, Correlation and causality, (a very technical book about analysing causal impacts using statistical models). This book is available free on the net at: http://davidakenny.net/cm/cc.htm

King, J. 2018, OPM's Approach to Assessing Value for Money, Oxford Policy Management. (useful and practical guidance)

Knight, C. 2015, 'Mechanism-Based Causal Analysis', International Encyclopedia of the Social & Behavioral Sciences. (this article investigates five major approaches to causal mechanisms toward the goal of identifying major points of consensus and contention. It further suggests that there are two distinct approaches to causal mechanisms: 'top-down' approaches that seek to generalize empirical events under widely instantiated causal patterns, and 'bottom-up' approaches that seek to disaggregate 'average causal effects' by opening up the 'black-boxes')

Kotvojs, F. and Carolina Lasambouw, C. nd, MSC: Misconceptions, Strengths and Challenges. (a good summary). Available free on the internet at: https://www.aes.asn.au/images/stories/files/conferences/2009/Papers/Kotvojs,%20Fiona%20-%20MSC.pdf

Krämer et al. (2021), Rigorous Impact Evaluation: Evidence generation and take-up in German development cooperation, German Institute for Development Evaluation (DEval), Bonn. (useful insights into organizational issues: What are existing barriers to (a) the initiation of RIE and (b) the take-up of RIE evidence?)

Kraft MA. 2019, "Interpreting Effect Sizes of Education Interventions", Educational Researcher; 49 (4) :241-253. (this tends to be a weak area in most published evaluations, the author presents guidelines for interpreting effect sizes that are applicable across the social sciences, recommended)

Krauss, A. 2018, "Why All Randomized Controlled Trials Produce Biased Results", Annals of Medicine, 50:4. (highly recommended, RTCs are never conducted without some degree of bias and this paper explains why. The idea of a single study that provides the ultimate definitive answer is flawed. We slowly and steadily build up our knowledge base over time. Undertaking research is similar to an Easter egg hunt combined with a jigsaw puzzle with no instructions; we are putting a mosaic together piece by piece over time)

LaLonde, 1986, "Evaluating the Econometric Evaluations of Training Programs with Experimental Data", American Economic Review, vol 76, pp. 604-620. (the classic article explaining why the econometric analysis of correlational research designs usually fail to achieve accurate estimates of program impact)

Lam, S. 2020, "Toward Learning from Change Pathways: Reviewing Theory of Change and Its Discontents", Canadian Journal of Program Evaluation doi: 10.3138/cjpe.69535 (this paper identifies 7 common problems with ToC based evaluations, makes some good points)

Lam & Valencia, 2019, 'Retrospective Pretest and Counterfactual Self-Report: Different or Same?', Journal of MultiDisciplinary Evaluation, Volume 15, Issue 33. (the evidence shows that people are very inaccurate judges of change over time)

Lance, P. et al 2014, How Do we Know If a Program Made a Difference? A Guide to Statistical Methods for Impact Evaluation. Chapel Hill, North Carolina: MEASURE Evaluation. (a useful overview of methods)

Landsittel, Douglas; Srivastava, Avantika; Kropf, Kristin. 2020, 'A Narrative Review of Methods for Causal Inference and Associated Educational Resources', Quality Management in Health Care: October/December 2020 - Volume 29 - Issue 4 - p 260-269 (the available literature is vast and difficult to summarise, RTCs are often not feasible while using observational data for causal inference is quite challenging))

Langbein, 1980, Discovering Whether Programs Work, Goodyear. (good but technical)

Larson, 1980, Why government programs fail, Praeger. (The reasons: a faulty theory of change/strategy; poor implementation; a changing external environment; or the evaluation itself is faulty)

Ledford, J. 2018, "No Randomization? No Problem: Experimental Control and Random Assignment in Single Case Research", American Journal of Evaluation, vol. 39, no. 1. (an overview of the use of single subject designs for impact evaluation to assess changes in level, trend and variability)

Lee et al, 2019, 'Investigating causal mechanisms in randomised controlled trials', Trials, (there are various methodological issues and assumptions that should be considered when mediation analyses of randomised trials are used to inform clinical practice and policy decisions) https://trialsjournal.biomedcentral.com/articles/10.1186/s13063-019-3593-z

Leeuw, F. and Vaessen, J. 2009, Impact Evaluations and Development: Nonie Guidance on Impact Evaluation, The Network of Networks on Impact Evaluation. (a useful overview from a World Bank economic perspective)

Li, S. and Liu, Y. 2020. Using big data to evaluate the impacts of transportation infrastructure investment: the case of subway systems in Beijing, 3ie Impact Evaluation Report 115, New Delhi: International Initiative for Impact Evaluation (3ie). Available at: https://doi.org/10.23846/DPW1IE115. (this report uses interrupted time series analysis combined with comparison groups and pattern matching, a good example of using multiple methods to arrive at a defensible conclusion a la Reynolds & West)

Light, P. 2014, A Cascade of Failures: Why Government Fails and How to Stop It, Centre for Effective Public Management at Brookings. (analysis of government failures in the USA, insightful for evaluators. The reasons: poor policy; inadequate resources; culture; structure; lack of leadership)

Liao & Deviatko 2015, 'History of Causal Analysis', International Encyclopedia of the Social & Behavioral Sciences, 2nd edition, Volume 3. (recommended)

Loannidis, J. 2005, 'Why Most Published Research Findings Are False', PLOS Medicine. (discusses a range of potential biases). Available free on the internet at: https://journals.plos.org/plosmedicine/article/file?id=10.1371/journal.pmed.0020124&type=printable

Lwamba, E. et al 2021, Protocol: Strengthening women's empowerment and gender equality in fragile contexts towards peaceful and inclusive societies: a systematic review and meta-analysis, 3ie. (an excellent example of a protocol for an evaluation synthesis study)

Lynn, J. et al 2021, 'Lost Causal: Debunking Myths About Causal Analysis in Philanthropy', The Foundation Review, vol 13, Issue 3. (a useful rebuttal of common myths about causal analysis)

McClintock, C. 1990, "Evaluators as Applied Theorists", American Journal of Evaluation 11(1):1-12 DOI:10.1177/109821409001100102. (highlights the value of applying program and implementation theory when undertaking impact evaluations)

Macintyre, S. 2011, 'Good Intentions and Received Wisdom are Not Good Enough', Journal of Epidemiology and Community Health, 65(7), 564-567. (reviews the arguments for and against RTCs)

McMillan, 2007, Randomized Field Trials and Internal Validity: Not So Fast My Friend, (good overview of the limitations). Available free on the net at: http://pareonline.net/pdf/v12n15.pdf

McMurry et al, 2015, "Propensity scores: Methods, considerations, and applications", Journal of Thoracic and Cardiovascular Surgery, 150:14-9. (the authors conclude that results from most of the examples of PS that they examined were not convincing due to methodological problems)

Manning, R., Ian Goldman, I. and Licona, G. 2020, The Impact of Impact Evaluation, WIDER Working Paper 2020/20, United Nations University. (a discussion of various political aspects of impact evaluations)

Mark & Reichardt, 2004, 'Quasi-experimental and correlational designs: Methods for the real world when random assignment isn't feasible'. In Sansone, Morf and Panter, (eds), Handbook of methods in social psychology, (pp. 265-286), Sage. (useful introductory overview, recommended)

Markman, A. 2015, 'Two Ways to Keep Your Data from Tricking You', Harvard Business Review. (the author argues for developing and testing predictions/hypotheses and examining statistical interactions, I think this is great advice)

Masset, et al, 2019. Successful Impact Evaluations: Lessons from DFID and 3ie, Centre of Excellence for Development Impact and Learning. (achieving the goals of credibility, relevance and policy impact is very challenging)

Masset, E., Shrestha, S. and Juden, M. (2021). 'Evaluating Complex Interventions in International Development'. CEDIL Methods Working Paper 6. Centre of Excellence for Development Impact and Learning (CEDIL), London and Oxford. Available from: https://doi.org/10.51744/CMWP6 (this paper has been strongly criticized by Aston 2021 who disagrees with how Masset defines complexity. I think Aston has a good point but he himself misunderstands epistemology and causal analysis)

Maxwell, J. 'Using Qualitative Methods for Causal Explanation', Field Methods, vol. 16, No. 3, pp 243–264. (an easy to read introduction to the topic of developing and testing causal explanations)

Mayne, J. 2008, Contribution analysis: An approach to exploring cause and effect, ILAC Brief 16. (a popular approach based on using program theory and performance indicators - shares similar strengths and weaknesses)

Mayne, J. 2019, A Brief on Contribution Analysis: Principals and Concepts, available via LinkedIn. (explains what CA can and cannot do, although I disagree with his approach to assessing causality).

Mayne, J. 2019, 'Revisiting Contribution Analysis', Canadian Journal of Program Evaluation, 34.2, 171-191. (an update on the method which continues to evolve over time becoming more rigorous and also more complex/resource intensive)

Mayne, J. 2019, Assessing the Relative Importance of Causal Factors, Centre for Development Impact, Practice Paper. (written from the perspective of CA and hence this approach shares the same advantages and disadvantages, in my opinion path analysis is a much better methodology for comparing the strength of causal pathways)

Mayo, D. 2018, Statistical Inference as Severe Testing: How To Get Beyond the Statistics Wars, Cambridge: Cambridge University Press. (essential reading for those interested in Bayesian vs frequentist methods, and what constitutes credible evidence)

McIntyre, L. 2019, The Scientific Attitude, MIT Press. (this text places scientific methods into their broader context which I found helpful when thinking about different approaches to impact evaluation and what counts as a credible finding and the role of sound scientific processes and more importantly the scientific 'attitude' and 'community')

Michalopoulos, 2004, 'Can Propensity-Score Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?' The Review of Economics and Statistics, Vol. 86, No. 1, Pages 156-179. (the answer: occasionally, but not consistently)

Miles and Huberman, 1994, Qualitative data analysis, Sage. (contains examples of undertaking causal analysis with qualitative data, recommended)

Mill, J. S. 1843, A System of Logic Ratiocinative and Inductive, 8th edn. New York: Harper and Brothers. (Mill proposes 3 criteria for testing causal relationships: association, temporal order; and non-spuriousness)

Mohr, 1995, Impact Analysis for Program Evaluation, Sage. (an advanced discussion of research designs and impact analysis)

Mohr, 1999, 'The Qualitative Method of Impact Analysis', American Journal of Evaluation, 20, 1, pp. 69-84. (useful introduction to the topic)

Morgan, Stephen L. and Winship, C. 2014, Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research), New York: Springer. (an easy to read introduction for social researchers)

Muller, J. 2018, The Tyranny of Metrics, Princeton University Press. (a useful discussion of the limits of measurement as a management tool)

Meyer, B. D. 1995, 'Natural and quasi-experiments in economics', Journal of Business & Economics Statistics, 13(2), 151−161. (a useful paper, discusses internal validity from an econometric perspective)

Murnane and Willett, 2010, Methods Matter: Improving Causal Inference in Educational and Social Science Research, Oxford University Press. (a good introduction, not overly technical and includes many examples)

National Institute for Health Research, 2016, Assessing claims about treatments effects: Key concepts that people need to understand. (a useful summary). Available on the net at: http://www.testingtreatments.org/key-concepts-for-assessing-claims-about-treatment-effects/?nabm=1

Network of Networks on Impact Evaluation, 2009, Impact Evaluations and Development – NONIE Guidance on Impact Evaluations. (a review of the methods commonly used by development agencies). Available free on the net at: http://www.worldbank.org/ieg/nonie/guidance.html

Nisbett and Ross, 1985, Human Inference: Strategies and Shortcomings of Social Judgments, Prentice-Hall. (explains why we all struggle to accurately perceive causal relationships, basically people are terrible at this due to various 'involuntary' cognitive biases)

Nobbs, J. 2021, The Problem with Observational Studies (Epidemiology), (makes some good points). Blog at: https://www.jeffnobbs.com/posts/the-problem-with-observational-studies-epidemiology

Noble, J. et al, 2020, Understanding Impact – Using your theory of change to develop a measurement and evaluation framework, NPC.

Norad, 2008, The Challenge of Assessing Aid Impact: A Review of Norwegian Evaluation Practice, (provides a number of examples of problematic impact evaluations along with various lessons for better practice). Available free on the net at: http://www.norad.no/default.asp?V_ITEM_ID=12314

Norgbey, E, 2016, 'Debate on the Appropriate Methods for Conducting Impact Evaluation of Programs within the Development Context', Journal of Multidisciplinary Evaluation, vol 12, issue 27. (argues that method choices should respond to contextual and situational aspects of the program, I agree)

Nutley, et al, 2013, What counts as Good Evidence? Alliance For Useful Evidence. (useful overview of the topic)

Nutt, 2002, Why Decisions Fail, Berrett-Koehler Publishers. (interesting review of why strategic decisions often fail, e.g. lack of consultation, poor analysis, faulty implementation,)

Nuttall & Houle, 2008, 'Liars, Damn Liars, and Propensity Scores', Anesthesiology, Vol. 108, 3-4. (argues that PSM has some serious limitations that are common to most passive observational research methods, I agree)

OECD, 2020, Building Capacity for Evidence-Informed Policy-Making: Lessons from Country Experiences. (a useful review of strategies to stimulate demand for and use of evidence, recommended)

Olofsgard, A. 2014, Randomized Control Trials: Strengths, Weaknesses and Policy Relevance, EBA. (a balanced thoughtful summary of the main issues).

Olsen & Orr, 2016, "On The 'Where' of Social Experiments: Selecting More Representative Samples to Inform Policy", New Directions in Evaluation, no 152. (useful suggestions for improving the external validity of experiments through better sampling)

Orr, L. et al 2019, 'Using the Results from Rigorous Multisite Evaluations to Inform Local Policy Decisions', Journal of Policy Analysis and Management. (this is generally a difficult undertaking as the size of the local error is often equal or larger than the estimated effect size).

Patton, 1982, Practical Evaluation, Sage. (of all his books this one is my favorite, great chapters on data analysis and on preparing useful recommendations, highly recommended)

Patton, 1987, How to use qualitative methods in evaluation, Sage. (excellent discussion of combining qualitative and quantitative methods)

Patton, 1990, Qualitative Evaluation and Research Methods, Sage. (good all round reference, helpful description of different types of purposeful sampling)

Pawson, R. 2002, "Evidence based policy: The promise of realist synthesis'. Evaluation, 8(3), 340-358.

Pawson. R. 2008, Causality for Beginners, unpublished. (unpublished but you can find this on the web. This paper compares three longstanding modes of causal explanation: 'succesionist', 'configurational' and 'generative')

Pearl, 2000, Causality: Models, Reasoning, and Inference, Cambridge University Press. (advanced technical examination of causal/statistical modeling)

Pearl, J. 2009, Myth, confusion, and science in causal analysis (Technical Report No. R-348). Retrieved from University of California, Los Angeles website: http://www.cs.ucla.edu/~kaoru/r348.pdf

Pearl, J. and Mackenzie, D. 2019, The Book of Why: The New Science of Cause and Effect, Penguin Books. (an advanced text for the enthusiast, challenging to read)

Peck, (ed) 2016, "Social Experiments in Practice: The What, Why, When, Where and How of Experimental Design and Analysis", New Directions in Evaluation, no 152. (a good overview of various issues from an econometric perspective)

Peck, L. 2017, When is Randomization Right for Evaluation? (offers principles for when experiments are appropriate). See: http://abtassociates.com/Perspectives/March/When-Is-Randomization-Right-for-Evaluation

Peck, L. 2020, Experimental evaluation design for program improvement, Sage. (this book explains how experiments can be used to unpack the block box of interventions, a good reference for evaluators)

Peck, L 2022, 'Section editor's note: Insights into the generalizability of findings from experimental evaluations', American Journal of Evaluation, vol 43,, pp. 66-69. (a good summary of the issues, also see the related articles in the same issue of the journal)

Perrin, Burt. 1998, 'Effective Use and Misuse of Performance Measurement'. American Journal of Evaluation. Vol. 19 (1):367-379. (a classic article, highly recommended)

Perrin, Burt. 1999, 'Performance Measurement: Does the Reality Match the Rhetoric? A Rejoinder to Bernstein and Winston'. American Journal of Evaluation. Vol. 20(1).

Perrin, Burt. et al (eds) 2015, Evaluations that Make a Difference, (examples of evaluations from around the world that have had a positive impact on the public). Available on the net at: https://evaluationstories.wordpress.com/evaluation-story-publications/

Peters, B. 2020, Qualitative Methods in Monitoring and Evaluation: Causal Mechanisms: Let's Consider Golf Balls. (interesting perspective). See: https://programs.online.american.edu/msme/masters-in-measurement-and-evaluation/resources/qualitative-methods-and-causal-mechanisms

Peters, B. 2020, Qualitative Methods in Monitoring and Evaluation: The Philosophy of Science and Qualitative Methods. (useful introduction). See: https://programs.online.american.edu/msme/masters-in-measurement-and-evaluation/resources/science-and-qualitative-methods

Posavac & Carey, 2002, Program Evaluation: Methods and Case Studies, Prentice Hall. (good all round text, includes a summary of the types of evaluation questions that can be answered by particular types of research designs)

Posthumus, H. and Wanitphon, P. 2015, Measuring Attribution: a practical framework to select appropriate attribution methods, with cases from ALCP in Georgia, MDF in East Timor, Propcom Mai-Karfi in Nigeria and Samarth-NMDP in Nepal, Hans Posthumus Consultancy. (a useful introduction from the DCED perspective with some helpful graphical illustrations).

Powell, K. and Prasad, V. 'Where are randomized trials necessary: Are smoking and parachutes good counterexamples?', 2021, European Journal Clinical Investigation, https://doi.org/10.1111/eci.13730. (an interesting paper that dispels and some common misunderstandings)

Pritchett, L. 2021, Lets Take the Con Out of Randomized Experiments, CID Faculty Working Paper No. 399, Harvard University. (this paper argues that experiments have major problems with external validity, while the author has a point he takes a very simplistic view and ignores the work of people such as Peck who have shown how experiments can be quite effectively used to unpack the black box and enhance external validity)

Pritchett & Sandefur 2013, Context Maters for Size: Why External Validity Claims and Development Practice Don't Mix, Working paper 336, Center for Global Development. (this paper argues that impact evaluation findings are context dependent and hence we need to be very careful when seeking to generalize/apply findings from one context to another, even when using RCTs; I agree with this view)

Ramalingham B. 2011, Learning how to learn: eight lessons for impact evaluations that make a difference, ODI, London.

Ravallion, M. 2001, 'The Mystery of the Vanishing Benefits: An Introduction to Impact Evaluation', The World Bank Economic Review, vol 15, issue 1, pages 115–140. (This article provides an introduction to the concepts and methods of impact evaluation. The article takes the form of a short story about a fictional character's on-the-job training in evaluation)

Reddy, S. G. 2019, 'Economics' Biggest Success Story Is a Cautionary Tale: Field experiments now dominate development economics—often at the expense of the world's poor', Foreign Policy. (a critique of the use of RTCs in development economics, raises the usual arguments and makes some good points but in my opinion also misunderstands some key issues). See: https://foreignpolicy.com/2019/10/22/economics-development-rcts-esther-duflo-abhijit-banerjee-michael-kremer-nobel/

Reichardt, C. 2019, Quasi-Experimentation: A Guide to Design and Analysis - (Methodology in the Social Sciences), The Guilford Press. (a great text on the application of quasi-experiments and why they are sometimes preferable to true experiments, for the enthusiast)

Reichardt, C. 2000, 'A typology of strategies for ruling out threats to validity'. In Bickman (ed) Research Design: Donald Campbell's' legacy, Sage. (very insightful)

Reinertsen, Bjørkdahl, and McNeill, 2017, Confronting the Contradiction- An Exploration into the Dual Purpose of Accountability and Learning in Aid Evaluation, SIDA. (their main conclusion is that, in practice, the dual purposes of accountability and learning leads to difficult trade-offs; there are tensions and sometimes direct contradictions between maximizing accountability vs learning)

Reschovsky, Heeringa and Colby, 2018, Selecting the Best Comparison Group and Evaluation Design: A Guidance Document for State Section 1115 Demonstration Evaluations, Mathematica Policy Research. (an excellent paper with a useful decision making flow chart, highly recommended). Available on the net at: https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/comparison-grp-eval-dsgn.pdf

Reynolds & West 1987, 'A multiplist strategy for strengthening nonequivalent control group designs', Evaluation Review, 11, 6, 691-714. (an excellent example of how to fix up a weak research design by adding additional features thereby improving your overall assessment of the program's impact, a classic article and highly recommended)

Rogers et al 2000, 'Program Theory in Evaluation: Challenges and Opportunities', New Directions for Evaluation, No. 87. (a series of papers on the strengths and weaknesses of using program theory to assist with causal analysis)

Rogers et al, 2015, Choosing Appropriate Designs and Methods for Impact Evaluation, Office of the Chief Economist, Department of Industry Innovation and Science. (helpful introduction)

Rohrer, J. 2018, 'Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data', Advances in Methods and Practices in Psychological Science, Vol. 1(1) 27–42. (this paper argues that it is very difficult to undertake causal analysis with observational data although graphical methods can help, highly recommended)

Roodman, 2008, Through the Looking Glass, and What OLS Found There: On Growth, Foreign Aid, and Reverse Causality, Working Paper 137, Center for Global Development. (discussion of assessing the impact of foreign aid)

Rosenbaum, P. 2019, Observation and Experiment: An Introduction to Causal Inference, Harvard University Press. (contains interesting examples and is quite insightful but it is not an easy read)

Rosling, H. 2018, Factfulness: Ten Reasons We're Wrong About the World--and Why Things Are Better Than You Think, Sceptre. (highly recommended for its real-world examples of outcome trajectories)

Rossi, Lipsey & Freeman 2003, Evaluation – A Systematic Approach, Sage. (recommended, includes an excellent discussion of different types of research designs and when to use each of them)

Rothgang & Lageman, 2021, "The unused potential of process tracing as evaluation approach: The case of cluster policy evaluation", Evaluation, Vol. 27(4) 527–543. (one of the better articles I have seen on this topic).

Rothman & Greenland, 2005, 'Causation and Causal Inference in Epidemiology', American Journal of Public Health, Vol 95, No. S1.

Rubin, 2008, "For Objective Causal Inference, Design Trumps Analysis", Annals of Applied Statistics, vol 2, pp. 808-840. (the title says it all and I agree with this position)

Salkind, N. J. (ed) 2010, Encyclopedia of Research Design, Sage. (very comprehensive at 1800 pages, and expensive)

Sauerbrei, W., Perperoglou, A., Schmid, M. et al. 2020, 'State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues' Diagnostic and Prognostic Research 4, 3. https://doi.org/10.1186/s41512-020-00074-3. (multivariate analysis requires both advanced technical skills and expert subject matter knowledge, a number of subjective analytical choices are required when undertaking observational research and each choice can generate quite different results).

Scher, L. et al 2015, Designing and Conducting Strong Quasi Experiments in Education, Institute of Education Sciences. (a good overview of key issues)

Schmit, J. 2020, 'Causal Mechanisms in Program Evaluation', New Directions for Evaluation, no 167. (explores the topic of causal mechanisms)

Scriven, M. 1976, 'Maximizing the power of causal investigations: The modus operandi method'. In Gene V. Glass (ed.) Evaluation studies review annual, Volume 1, 101-118, Beverly Hills, CA: Sage Publications. (in my view Scriven's argument that undertaking an impact evaluation is similar to a detective investigating a murder is simply incorrect)

Shadish, W. R. 1993, 'Critical multiplism: A research strategy and its attendant tactics', New Directions for Program Evaluation, vol 60, https://doi.org/10.1002/ev.1660. (in my view critical multiplism has a lot to offer the field of impact evaluation, its my preferred paradigm)

Shadish, Clark, & Steiner, 2008, 'Can nonrandomized experiments yield accurate results? A randomized experiment comparing random and nonrandom assignments;, Journal of the American Statistical Association, 103(484), pp. 134-1343. (yes, provided that all key variables are observed and we have good covariates to facilitate adjustment)

Shadish and Cook, 1999, 'Comment-Design Rules: More Steps toward a Complete Theory of Quasi-Experimentation', Statistical Science, Vol. 14, No. 3, pp. 294-300. (describes the design elements used in constructing quasi-experiments, argues that statistics are not effective in resolving basic design problems)

Shadish, Cook and Campbell, 2002, Experimental and Quasi-Experimental Designs for Generalized Causal Inference, Houghton Mifflin. (a classic advanced text, recommended)

Shadish, Cook, and Houts, 1986, 'Quasi-Experimentation in a Critical Multiplist Mode', in W. M. K. Trochim (ed.). 'Advances in Quasi-Experimental Design and Analysis', New Directions for Program Evaluation, no. 31. San Francisco: Jossey-Bass. (in my view critical multiplism has a lot to offer when undertaking impact evaluations)

Shadish, Cook and Leviton, 1991, Foundations of Program Evaluation, Sage. (the final chapter contains an excellent summary of evaluation theory in relation to program design, evaluation practice, and theory of use, highly recommended)

Social Programs that Work, 2020, https://evidencebasedprograms.org/
(This site seeks to identify those social programs shown in rigorous studies to produce sizable, sustained benefits to participants and/or society, so that they can be deployed to help solve social problems)

Somers, M. et al, 2013, The Validity and Precision of the Comparative Interrupted Time Series Design and the Difference-in Difference Design in Educational Evaluation, MDRC. (technical but interesting)

Spector, 1981, Research Designs, Sage. (a basic introduction)

St. Clair, T., Cook, T. and Hallberg, K. 2014, 'Examining the Internal Validity and Statistical Precision of the Comparative Interrupted Time Series Design by Comparison With a Randomized Experiment', American Journal of Evaluation, vol. 35, no 3, pp. 311-327, 2014. (CITS often provide accurate estimates but are sensitive to choices in design parameters)

Stame, N. 2010, 'What Doesn't Work? Three Failures, Many Answers', Evaluation, 16, 4, 371-387. (useful review of debates on impact evaluation methodology)

Staines, G. L. 2007, 'Comparative outcome evaluations of psychotherapies: Guidelines for addressing eight limitations of the gold standard of causal inference', Psychotherapy: Theory, Research, Practice, Training, 44(2), 161–174. https://doi.org/10.1037/0033-3204.44.2.161

Stern, E. 2012, Broadening the Range of Designs and Methods for Impact Evaluations, DFID (useful discussion of key issues).
Available on the net at: http://www.oecd.org/dataoecd/0/16/50399683.pdf

Stern E. 2015, Impact Evaluation: A Guide for Commissioners and Managers, Bond (a useful non technical overview).
Available on the net at: http://www.bond.org.uk/data/files/Impact_Evaluation_Guide_0515.pdf

Straight Talk on Evidence, 2020, Second RCT of Nevada reemployment program has found sizable earnings gains, providing actionable evidence for U.S. policy post-pandemic. (results illustrate that impact size often depends more on the specific program model {in this case the Nevada REA} than on the general approach {reemployment programs more broadly} as other rigorously-tested REA programs with different features have produced small or no impacts)

Straight Talk on Evidence, 2017, How "Official" Evidence Reviews Can Make Ineffective Programs Appear Effective (part one in a series). (insightful). See: http://www.straighttalkonevidence.org/2017/11/27/how-official-evidence-reviews-can-make-ineffective-programs-appear-effective/

Straight Talk on Evidence, 2019, Beware the pitfalls of short-term program effects: They often fade (this paper includes several real world examples). See: https://www.straighttalkonevidence.org/2019/04/03/beware-the-pitfalls-of-short-term-program-effects-they-often-fade/

Straight Talk on Evidence, 2019, Why most non-RCT program evaluation findings are unreliable (and a way to improve them). (useful) See: https://www.straighttalkonevidence.org/2019/12/12/why-most-non-rct-program-evaluation-findings-are-unreliable-and-a-way-to-improve-them/

Streiner & Norman, 2012, 'The Pros and Cons of Propensity Scores', CHEST Journal, Volume 142, Issue 6, Pages 1380–1382, DOI: https://doi.org/10.1378/chest.12-1920 (PSM's apparent simplicity masks a number of rather arbitrary statistical assumptions and tradeoffs)

Stuart, E. A., & Rubin, D. B. 2008, 'Matching with multiple control groups with adjustment for group differences'. Journal of Educational and Behavioral Statistics, 33(3), 279–306. (an illustration of this method)

3IE, 2012, 3ie impact evaluation glossary, International Initiative for Impact Evaluation: New Delhi, India.

3IE, 2018, International Initiative for Impact Evaluation, (resources for impact evaluation from an economist's perspective). Available on the internet: http://www.3ieimpact.org/en/

The International Campbell Collaboration, 2020, https://campbellcollaboration.org/
(this site offers a registry of systematic reviews of evidence on the effects of interventions in the social, behavioral, and educational arenas)

Tipton et al 2017, "Implications of Small Samples for Generalization: Adjustments and Rules of Thumb", Evaluation Review, 41(5), 472-505. (illustrates why it is important to be careful when seeking to generalise from small nonprobability samples)

Ton, G. et al 2019, Contribution Analysis and Estimating the Size of Effects: Can we Reconcile the Possible with the Impossible? Centre for Development Impact Practice Paper 20, Brighton. (an interesting discussion of how contribution analysis might be stretched to give some sense of the importance of a contribution in a quantitative manner. However, I do not agree with the criteria the authors apply to test for causal inference). Available free on the net at: https://opendocs.ids.ac.uk/opendocs/handle/123456789/14235

Treasury Board of Canada, no date, Program Evaluation Methods: Measurement and Attribution of Program Results, (useful overview). It is available free on the net at: http://www.tbs-sct.gc.ca/eval/pubs/meth/pem-mep_e.pdf

Trochim, 1984, Research designs for program evaluation: The regression discontinuity approach, Sage. (excellent method for evaluating impacts where entry into the program depends upon meeting a numerical eligibility criterion, e.g. income less than X, academic grades more than Y)

Trochim, 1989, 'Outcome Pattern Matching and Program Theory'. Evaluation and Program Planning, Vol. 12:355-366. (an interesting alternative for impact evaluations using case studies and pattern matching)

United Nations, 2013, Impact Evaluations in UN Agency Evaluation Systems - Guidance on Selection Planning and Management. Available on the web at: http://www.uneval.org/normsandstandards/index.jsp

United Nations Working Group on Evaluation, 2020, Compendium of Evaluation Methods Reviewed - Volume 1, United Nations Evaluation Group. (a useful short summary although I don't agree with some of their views about the rigor of passive observational studies)

U.S. Department of Education, 2003, Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide. (a useful summary of what counts as good evidence in evaluation, this publication caused some controversy in the USA)

U.S. Department of Education, 2008, Technical Methods Report: Guidelines for Multiple Testing in Impact Evaluations.

USAID 2018, Guide for planning long-term impact evaluations. (useful introductory guidance)

Vanhoof, J. and Van Petegem, P. 2010, 'Evaluating the quality of self-evaluations: The (mis)match between internal and external meta-evaluation', Studies in Educational Evaluation, 36, 20–26. (there is limited correlation between the findings of internal and external meta evaluation at Finnish schools, differences in evaluation goals and approaches of internal and external evaluation – though focusing on the same aspects of self evaluation – however inevitably lead to different procedures and measurement instruments)

Vaessen, Jos, Sebastian Lemire, and Barbara Befani. 2020, Evaluation of International Development Interventions: An Overview of Approaches and Methods, Independent Evaluation Group. Washington, DC: World Bank. (free on the internet, a standard econometric overview that is quite comprehensive, I disagree with their text on quasi-experiments.)

Wauters and Beach, 2018, 'Process Tracing and Congruence Analysis to Support Theory Based Impact Evaluation', Journal of Evaluation. (this article illustrates the difference between process tracing and congruence analysis and their relative advantages).

Weakliem, 2016, Hypothesis Testing and Model Selection in the Social Sciences, The Guilford Press. (some helpful observations about the role of hypothesis testing in the evaluation of theories)

Weisburd, D. 2010, 'Justifying the use of non-experimental methods and disqualifying the use of randomized controlled trials: challenging folklore in evaluation research in crime and justice', Journal of Experimental Criminology, volume 6, pages 209–227. (The key limitation of passive observational impact evaluation methods is that they require an assumption that all confounding factors related to treatment are identified in the statistical model being applied. The author explains why this assumption is so critical and challenges what he describe as "folklores" that are used to justify the use of non-randomized studies despite this being a major problem).

Weiss, C. 1997, "How Can Theory-Based Evaluation Make Greater Headway?" Evaluation Review, 21, 4, 501-525. (this article explores the problems of theory-based evaluation as well as the potential benefits)

Weiss, M. et al, 2013, A Conceptual Framework for Studying the Sources of Variation in Program Effects, MDRC Working Papers on Research Methodology. (The goal of the framework is to enable researchers to offer better guidance to policymakers and program operators on the conditions and practices that are associated with larger and more positive effects)

Weisz, J., Han, S. and Valeri, S. 1997, "More of What? Issues Raised by the Fort Bragg Study", American Psychologist, 52, 5, 541-545. (argues that demonstration projects must be combine both process and sound outcome evaluations, comparison groups chosen for their low scores can improve over time even without any treatment, clients can be highly satisfied even though the intervention is ineffective).

West, S. and Thoemmes, F. 2010, "Campbell's and Rubin's Perspectives on Causal Inference", Psychological Methods, Vol. 15, No. 1, 18–37. (Campbell's approach tends to be used in education and psychology and Rubin's in economics, medicine and statistics, they are actually complementary)

Weyrauch, V. and Langou, G. D. 2011, Sound Expectations: From Impact Evaluations to Policy Change, 3ie Working Paper 12. London: 3iE. It is available free on the net at: http://www.3ieimpact.org/3ie_working_papers.html.

What Works Clearinghouse, 2017, Standards Handbook. (A good but very technical set of standards for impact evaluation. The What Works Clearinghouse identifies existing research on education interventions, assesses the quality of this research, and then summarizes and disseminates the evidence from studies that meet the Clearinghouse's standards)

White, H. 2010, 'A Contribution to Current Debates in Impact Evaluation', Evaluation, 16(2) 153-164. (a useful summary of some of the issues and controversies)

White, H. And Raitzer, D. 2017, Impact Evaluation of Development Interventions – A Practical Guide, Asian Development Bank. (a useful resource with an econometric flavour). It is available free on the net at: https://www.adb.org/publications/impact-evaluation-development-interventions-practical-guide

Williams, M. 2019, "Making the Case for 'Mechanism Mapping': External Validity and Policy Adaptation: From Impact Evaluation to Policy Design", The World Bank Research Observer, Volume 35, Issue 2, pp 158–191, https://doi.org/10.1093/wbro/lky010. (well worth reading, raisies some good points)

Williams, R. 2015, Logic of Scientific Inference/ What is Causality? University of Notre Dame, https://www3.nd.edu/~rwilliam/. (a concise overview, explains Mill's 3 causal criteria)

Winship, C. and Morgan S. 1999. "The estimation of causal effects from observational data", Annual Review of Sociology. 25: 659-706. (a useful summary of the field; the author's overall conclusion: the challenges of estimating causal effects with observational data are often formidable, the authors tend to favour longitudinal methods)

Winston, J. 1993, 'Performance indicators: Do they perform?' Evaluation News and Comments, 2(3), 22-39. (this paper outlines a number of concerns with how KPIs get used)

Wong, Steiner & Anglin, 2018, "What Can Be Learned from Empirical Evaluations of Nonexperimental Methods?", Evaluation Review, 42(2), 147-175. (a lot, this paper provides a summary of the key issues)

Woolcock, M. 2019, Reasons for Using Mixed Methods in the Evaluation of Complex Projects, CID Faculty Working Paper No. 348, Centre for International Development at Harvard University. (a helpful overview of the topic)

Work-Learning Research, 2019, Levels of Evidence for the Learning Profession. (a useful summary as to what counts as better evidence). See: https://www.worklearning.com/2019/11/20/levels-of-evidence-for-the-learning-profession/

World Bank (Independent Evaluation Group) no date, Impact Evaluation- The Experience of the Independent Evaluation Group of the World Bank, author. It is available free on the net at:
http://documents.worldbank.org/curated/en/475491468138595632/Impact-evaluation-the-experience-of-the-independent-evaluation-group-of-the-World-Bank

World Bank (Independent Evaluation Group) 2006, Conducting Quality Impact Evaluations Under Budget, Time and Data Constraints, author. (this text is a highly summarized version of Bamberger's book). It is available free on the net at:
http://www.worldbank.org/ieg/ecd/conduct_qual_impact_eval.html

World Bank, 2016, Impact Evaluation in Practice. It is available free on the net at:
http://www.worldbank.org/en/programs/sief-trust-fund/publication/impact-evaluation-in-practice

World Bank, 2018, DIME Wiki (resources for impact evaluation from an economist's perspective). Available on the internet:
https://dimewiki.worldbank.org/wiki/Main_Page

Yang and Hendra, 2018, "The Importance of Using Multiple Data Sources in Policy Assessments: Lessons from Conditional Cash Transfer Programs in New York City", Evaluation Review, 1-25. (this paper argues that it is important to triangulate data sources in order to reach accurate conclusions about program effects)

Yeaton and Thompson, 2016, “Transforming the Canons of John Stuart Mill from Philiopshy to Replicative, Empirical Research: The Common Cause Design”, Journal of Methods and Measurement in the Social Sciences, Vol. 7, No . 2, p 122-143. (discussion of the strengths and weakensses of this little known approach).

Yeh, R. et al, 2018, “Parachute use to prevent death and major trauma when jumping from aircraft: randomized controlled trial”, British Medical Journal;363:k5094, doi:10.1136/bmj.k5094. (a tongue in cheek/satirical article that implies the mindless application of RTCs is a waste of time and resouces, quite amusing)

Yin, 2000, 'Rival Explanations as an Alternative to Reforms as Experiments', in Bickman (ed) Validity and Social Experimentation, Sage. (good review of how to identify and test rival explanations when evaluating reforms or complex social change)

Yin, 2003, Applications of Case Study Research, Sage. (very good reference, includes advice on undertaking causal analysis using case studies and pattern matching - in my view this is preferable to process tracing)

Young, J. and Mendizabal, E. 2009, Helping researchers become policy entrepreneurs, ODI Briefing Papers 53. London: ODI http://www.odi.org.uk/resources/download/1127.pdf (guidance on the research:policy interface)

Zhao, Y. 2017, “What Works May Hurt: Side Effects in Education, Journal of Educational Change, DOI10.1007/s10833-016-9294-4. (If an intervention can potentially help it can also potentially harm. Unintended side effects are inseparable from intended effects – both are outcomes from the same intervention; highly recommended)

Scott Bayley
5 December 2022