Experimental and Quasi-Experimental Evaluation Methods

Experimental: experiments that have treatments, outcome measures, and experimental conditions AND use random selection and assignment to treatment conditions.

1. Randomized controlled trial:
    a. Participants are randomly selected to participate in the study; participants are randomly assigned to treatment and control groups; pre- and post-tests are administered to treatment and control groups.
        i. Internal validity (IV): threats are very low when well implemented
        ii. External validity (EV): threats are moderate, but can be high depending on the quality of implementation. Low if accompanied by process evaluation or qualitative work to triangulate findings and provide context.

Quasi-Experimental: experiments that have treatments, outcome measures, and experimental conditions but do not use random selection and assignment to treatment conditions.

1. Regression discontinuity design
    a. Participants are assigned to treatment or comparison group based on a "cut score", or score on a covariate located on either side of a fixed threshold (Imbens and Lemieux, 2007; Trochim, 2006).
        i. IV: threats are as low as threats for RCTs when well designed (e.g. proper cut off criteria defined, pre/post distribution is properly specified, program is delivered uniformly to all recipients); possible threats from selection, regression to the mean.
        ii. EV: similar to those with RCTs
        iii. Notes: needs roughly 2.75 times as many participants to achieve the level of power of an RCT.
    b. Sharp and fuzzy designs- sharp designs require that the probability of receiving the treatment change from 0 to 1 at the threshold (cut score). The criteria for falling below or above the cut score is very rigid (e.g. a test score); fuzzy designs allow for a smaller change in probability of receiving the treatment at the threshold (e.g composite score on a health questionnaire).
2. Differences-in-differences
    a. An econometric model where an equation is specified, and the differences in pre/post-tests between treatment and comparison groups is calculated (Bertrand, Duflo, and Mullainathan, 2004; Meyer, 1995).
        i. IV: low if well implemented and all or most major assumptions are met (e.g. model is correctly specified; error term is, on average, 0; the error term is uncorrelated with other variables in the equation[also known as the parallel trend assumption, a solution if this is uncertain can be to obtain more data on time periods before and after treatment to detect pre-existing differences in trends; utilizing multiple comparison groups can help])
        ii. EV: low generalizability is a problem; time scale for some studies might not be sufficient for all independent variables to affect the dependent variable (a

solution is using longer time lags); endogeneity of the changes can be a problem (Remler, Ch. 13, 2014).

iii. Notes: potential for biased estimators if standard errors are auto-correlated or equation is misspecified; needs panel data.

3. Time series: comparative interrupted time series
    a. A series of pre-intervention measures (at least 4) are taken of a treatment and comparison group; an intervention is administered to the treatment group, but not the comparison group; post-intervention measures are taken of both groups (at least 3-4); a statistical model estimates the treatment effect.
        i. IV: if well implemented, low. Selection can be a problem, depending on if/how the comparison group was matched to the treatment group.
        ii. Notes: the level of rigor needed in matching the treatment and comparison groups is currently under debate; groups may not need to be matched as rigorously as in other QEDs (Clair, Cook, and Halberg, 2014; Somers, Zhu, Jacob, and Bloom, 2012).
4. Pre/post-test with matched comparison group
    a. A comparison group is constructed through matching based on the composition of the treatment group. Both groups take a pre-intervention test; the intervention is administered to the treatment group; both groups take a post-intervention test.
        i. IV: if matching is rigorous and design is well implemented, internal validity is high.
        ii. EV: depending on design, low to moderate.
    b. The comparison group can be created several ways:
        i. Propensity score matching
        ii. Case matching
        iii. Instrumental variable
        iv. Determining group equivalence post-hoc based on demographic variables

Works Cited

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates?. The Quarterly Journal of Economics, 119(1), 249-275.

Clair, T. S., Cook, T. D., & 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.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. Journal of Econometrics, 142(2), 615-635.

Remler, D., and Van Ryzin, G. (2014). Research methods in practice: strategies for description and causation. Sage Publications. Link location: http://www.sagepub.com/upm-data/33935_Chapter13.pdf.