WHEN TO USE A RANDOMIZED CONTROLLED TRIAL AND WHEN NOT TO

Bernadette Wright, PhD
Meaningful Evidence, LLC
Last updated May 4, 2013

SUMMARY

A randomized controlled trial (RCT) to measure your outcomes is expensive. You do not want to conduct one unless it is going to be useful. This post introduces ways you can tell what approach might work for your situation, including:

* Issues that can affect whether a RCT is likely to provide valid and useful results

* Examples of some of the many types of designs that can establish strong conclusions about effectiveness and value

* Considerations and techniques that can help to determine what approaches might be suitable for your situation

A randomized controlled trial (RCT) is expensive. Many other types of impact evaluation designs are less expensive, because they rely on existing data or they do not require recruiting research participants who do not receive the treatment.

ALL DESIGNS HAVE STRENGTHS AND LIMITATIONS.

The major strength of RCTs is precision at estimating what happened in the experiment. In a randomized controlled trial (RCT), researchers randomly assign people (or facilities, towns, or other unit) to receive a treatment or to not receive treatment (and/or to different versions of a treatment), then compare results for the different groups. The idea is that, because assignment to the different groups is random, they should be the same except for whether or not they received the treatment. This increases confidence that the effect that is measured is really a result of the treatment and not due to random chance or other causes.

These designs take an approach similar to consumer product tests, like in *Consumer Reports*. When independent researchers test several different brands of the same product in the exact same way, we can be reasonably certain that the results are due to actual differences in the quality of the products.

One limitation of RCTs is the risk that random chance or conditions of the experiment could affect results. The logic behind a RCT is to estimate the difference between actual the program outcomes and what the outcomes would have been without the program. Researchers call this hypothetical situation of "what would have been" the counter-factual. In the current preferred theory behind RCTs, the counter-factual would be the same person receiving the treatment and not receiving the treatment at exactly the same time and in exactly the same situation, which the laws of physics prevent (Cook et al., 2010). Actually, we know from quantum physics that, even if we could look at two different universes where everything was the same except whether or not participants received the treatment, we could not completely control for the effects of random chance. Thus, we cannot claim that randomly assigned groups the same for any particular study. The theory is that RCTs could estimate the true counter-factual if we could make an infinite number of random assignments.

Going back to the *Consumer Reports* example, a product model could be high quality, but a small percentage of them could still randomly malfunction. [1] The testers would then need to conduct more tests to determine whether the malfunction was due to quality or an isolated issue.

Another major risk to validity in RCTs is contamination of results by conditions of the experiment, especially when a placebo is not possible. In a test of consumer products, like different models of refrigerators, experts can control for this by putting all the test refrigerators in the same temperature-controlled room at the same time, using them to freeze the exact same amount of the same type of food, and so on.

In studies with human participants, this is more difficult. This is why the ideal RCT is a double-blind study, where both the person receiving the treatment and the person delivering the treatment are "blind" as to whether the person is getting the treatment or a placebo (Cook et al., 2009). Arguably, a RCT should use triple-blinding, so that neither the person

[1] ConsumerReports.org, "How We Test: Appliances and Home Products," http://www.consumerreports.org/content/cro/en/about-us/whats-behind-the-ratings/testing/appliances-home.html

receiving the treatment, the person administering the treatment, or the person evaluating the results know what treatment any individual received.

In one study that a friend once heard about, the group that received the placebo showed better results than the group that received the treatment. It turned out that this was because the person administering the placebo really sold people on it, while the person administering the treatment did not. Double blinding could have prevented this problem.

In RCTs of real-world programs and services (also called "randomized field experiments"), blinding is rarely possible. In impact evaluations where RCTs are suitable, a "good practice" is to combine RCTs with other methods (European Evaluation Society, 2007).

In addition, RCTs are weak at explaining how and why things happened or what will happen in other places, settings, and times. RCTs measure the effect of a specific intervention under specific conditions. The trade-off is that they are weak at explaining whether the results are due to the intervention design, how it was carried out in a particular situation, or something else unique to the locations and groups that participated in the study.

Combining evidence from multiple designs and studies provides stronger evidence than relying on any one design. Using multiple designs and data sources increases chances for getting well-supported and adequately explained findings. A set of many well-designed studies addressing the same issue is better than one well-designed study.

MULTIPLE DESIGNS CAN PROVIDE STRONG EVIDENCE OF VALUE WITHOUT THE EXPENSE OF A RANDOMIZED EXPERIMENT.

Several alternatives to experimental designs can establish conclusions about program effects just as well as RCTs, often better in particular circumstances (Cook et al., 2009). Standard texts on social science and program evaluation methods tend to present an array of research designs and descriptions of each in a way that resembles a "cookbook" (Trochim, 2006, "Designing Designs for Research"). However, an effective program evaluation is not about choosing from a few standard research designs, but constructing appropriate research strategies to get the evidence that funders and stakeholders need to make decisions. Trochim's website (http://www.socialresearchmethods.net/kb/desdes.php) describes

the basic conditions needed to establish cause-effect relationships, minimize threats to validity, and construct a good research design.

These include both quasi-experimental designs (ways of testing effects on specified outcomes) and ways of demonstrating how and why it works and ruling out rival hypotheses. These categories are overlapping and not mutually exclusive. In impact evaluations for real-world interventions, the best strategy is often to combine many approaches to create a hybrid design for the specific situation.

QUASI-EXPERIMENTAL DESIGNS (WAYS OF TESTING EFFECTS ON SPECIFIC OUTCOMES)

Quasi-experiments are designs that follow a similar logic and purpose as RCTs, but they do not randomly assign participants into treatment and control groups. Examples include:

- **Regression discontinuity.** A regression discontinuity design can work when assignment to treatment is by known criteria, such as reading level for students. Evaluators can estimate program impact by comparing results for participants just above and below the cut-off point for receiving the intervention. This approach may be more ethically acceptable than a RCT, because if resources for services are limited, you can give them to those who need it most. Social researchers have not frequently used this design (Trochim, "The Regression-Discontinuity Design," http://www.socialresearchmethods.net/kb/quasird.php). A new guide by MDRC (Jacob et al., 2012) explains this lesser known design.
- **A geographically local matched comparison group.** This type of design compares a group of people who are participating in the treatment with a nearby group of similar individuals who are not participating. It uses statistical techniques to construct a group that is as similar as possible to the participant group. The major challenge is to find a group that is the same except for participating in the program or to adjust for all differences.
- **Use of a perfectly known process of assignment.** When researchers know and can measure the process of selection into treatment, this can be a way to adjust for differences between the participant and comparison groups. The key is to account for all the important factors that are associated with selection into treatment and outcomes.
- **Interrupted time-series design**. An interrupted time series can provide strong evidence of impact by comparing outcomes at several

points in time before and after the intervention begins. This requires collecting enough data points before and after start of the program.

WAYS OF DEMONSTRATING HOW AND WHY IT WORKS

Although many evaluations focus on the question of what impact did the program have compared to no program, this is not always a key question that decision-makers need answered (Stern et al., 2012). Below are major types of designs that can answer questions like how something works, why it works, and how it fits in with what it takes to solve the problem.

Although some may consider these types of designs to be "less scientific" than RCTs and quasi-experiments, the reality is the quality of evidence from an impact evaluation is not a function of the type of design. It is a matter of how well the designs and methods fit the situation, whether the researchers properly apply the design and methods used, and how well the data support the conclusions (Stern et al., 2012). As Michael Patton (2013) said, "wise evaluators tailor their approach to the complexities of the circumstances they face."

- **Collecting information from participants, program staff, partners, and other stakeholders.** These methods provide a way to learn from experience. Examples include case studies, site visits, surveys, focus groups, and key informant interviews. In non-laboratory situations, an advantage of systematically conducted case studies is that their ability to address plausible rival hypotheses, which is the core of the scientific method (Yin, 1994, forward by Campbell, p. ix; Flyvbjerg, 2006). Engaging participation from stakeholders throughout an evaluation is an overarching technique that can add important insight to any design.
- **Unobtrusive measures.** When you only have data from interviews or surveys, you have inadequate knowledge of the rival hypothesis that results may stem from individuals' awareness of being tested or something related to their interaction with the investigator (Webb et al., p. 175). Adding data from unobtrusive measures like review of documents or observation can increase strength of the overall findings.
- **Statistical designs.** Statistical approaches like statistical models, longitudinal studies, and econometric models provide information about relationships among variables (Stern et al., 2012).
- **Evidence syntheses.** Combining evidence form multiple related studies and evaluations is a cost-effective way to get evidence using existing data and increase strength of the overall findings. Looking at what past studies have found before collecting new data can help you

avoid wasting money re-testing things that past research has already adequately demonstrated.

- **Theory-based approaches.** Any complete impact evaluation study includes explanation and interpretation of what the results mean in relation to our best understanding about how something works to solve a problem. Developments in theory-based approaches provide systematic ways to establish causation by identifying rival explanations and systematically ruling out each one. For example, the general elimination method, or exhaustive alternative causal explanation elimination design, works in cases where an improvement has been noticed, but what caused the improvement is unclear (Cook et al., 2010; Duignan, 2011). Researchers list all other rival explanations and systematically eliminated. The challenge is to identify the right causes and to collect convincing evidence to rule out each one. Contribution analysis and process tracing are some of the many other examples of this type of approach (Stern et al., 2012).

YOU NEED A GOOD HYPOTHESIS BEFORE YOU CAN DECIDE HOW TO TEST IT.

A complete evaluation study involves several stages:

- **Clarifying understanding of how what is being evaluated works**
- **Refining the evaluation questions**
- **Collecting and analyze evidence to answer the questions**
- **Interpreting/reporting what the results mean**

Several questions can help determine what options to assess impact might be appropriate for your situation.

HOW DO STAKEHOLDERS EXPECT THE INTERVENTION WILL WORK?

You probably have some evidence that the services you are offering will work to address the problem. You would not attempt to save the Chesapeake Bay by reciting poetry to it (Ruesga, 2013) or try to help the poor by giving them lupines, like in the Monty Python "Dennis Moore" skit. However, an important first step before developing and testing something is to listen to participants about how they see the problem and what they need. This can save you from wasting time and money carrying out and testing an intervention that will not achieve its outcomes because it is poorly planned (Duignan, 2013). For a good example of this, see Ernesto Sirolli's video on TED.com and heed his warning about, "At least we fed the hippos" ("Want to help someone? Shut up and listen!"

http://www.ted.com/talks/ernesto_sirolli_want_to_help_someone_shut_up_and_listen.html).

WHAT IS THE TIMEFRAME FOR ACHIEVING RESULTS?

Oftentimes, an ultimate goal that funders want to see, like total savings to health care programs, can take several years to achieve. For example, several federal and state studies have found that investing in services to support people with disabilities living at home and in the community helped control Medicaid spending in the long-run (Mollica et al., 2009).

You do not want to measure an outcome before enough time has passed that it has a reasonable chance to happen. When the ultimate desired outcomes are longer-term, you may want to focus on intermediate indicators of progress (e.g., decreases in hospitalizations, increased cooperation across agencies) and/or plan a longer-term study.

IS THE PROJECT ASSOCIATED WITH ONE DIRECT OUTCOME OR MANY OUTCOMES /INDIRECT EFFECTS?

Many services for people with disabilities are linked to multiple direct and indirect, planned and unanticipated, individual-level and system-level, short-term and long-term effects. For example, a community program that aims to prevent injuries from falls might also lead to increased activity and socialization, improved health, reduced hospitalizations, reduced nursing home use, or other inter-related outcomes.

To capture the full outcomes of these services, you need impact evaluation strategies that can assess all the effects across stakeholders.

DOES IT WORK BY ITSELF OR IN COMBINATION WITH OTHER CAUSES?

A growing number of services for people with disabilities and chronic illnesses involve partnerships with multiple providers. For example, a program to help people age in place at home might coordinate services from occupational therapists and other professionals to conduct home assessments, providers of home modifications and repairs, and home care providers. Several new programs under the Patient Protection and Affordable Care Act (ACA), like accountable care organizations, provide new opportunities for providers of services of home and community services for people with disabilities to coordinate with medical care providers.

If an evaluation of this type of program measured only the overall effects, the results would be hard to interpret without more information. In order to know whether the model of services can work, you would need to know

whether results were because of the service model design or because of something unique to the partnerships and conditions of the particular situation.

DOES EVERYONE GET THE SAME THING, OR DO YOU TREAT PEOPLE LIKE INDIVIDUALS?

Research shows that services are most effective when offered as part of a comprehensive system of long-term care services tailored to service recipients' and caregivers' individual needs and preferences (Kassner et al., 2008; Fox-Grage & Walls, 2013; The Lewin Group, 2010). Outstanding questions that funders want to know are what combination of services and strategies are most effective.

FOR FURTHER READING

ConsumerReports.org, "How We Test: Appliances and Home Products," 2006-2013, http://www.consumerreports.org/content/cro/en/about-us/whats-behind-the-ratings/testing/appliances-home.html

Cook, T.D., Scriven, M., Coryn, C.L.S., & Evergreen, S.D.H. "Contemporary Thinking about Causation in Evaluation: A Dialogue with Tom Cook and Michael Scriven." *American Journal of Evaluation*, 2010 31 (1), 105-117.

Duignan, P. Impact/outcome evaluation design types: An Outcomes Theory Knowledge Basic Topic [Internet]. Version 1. Outcomes Theory Knowledge Base. 2011 Mar 10. Available from: http://outcomestheory.wordpress.com/article/impact-outcome-evaluation-design-types-2m7zd68aaz774-10/.

Duignan, P. "Putting the Planning back into M&E – PME or PM&E what's the acronym going to be?" Blog, March 5, 2013, http://outcomesblog.org/2013/03/05/pme/

European Evaluation Society, "EES Statement: The importance of a methodologically diverse approach to impact evaluation—specifically with respect to development aid and development interventions." Available from: http://www.europeanevaluation.org/library.htm

Fox-Grage, W. & Walls, J. *State Studies Find Home and Community-Based Services to Be Cost-Effective."* Washington, DC: AARP Public Policy Institute, Spotlight 2, March, 2013. Available from: http://www.aarp.org/health/medicare-insurance/info-03-2013/state-studies-find-hcbs-to-be-cost-effective-AARP-ppi-ltc.html

Flyvbjerg, B. "Five Misunderstanding about Case-Study Research," *Qualitative Inquiry*, April 2006, 12, 2, 219-245.

Jacob, R., Zhu, P., Somer, M., & Bloom, H. *A Practical Guide to Regression Discontinuity*, MDRC, July 2012. Available from: http://www.mdrc.org/publications/644/full.pdf

Kassner, N., Reinhard, S., Fox-Grage, W., Houser, A., Accius, J., Coleman, B., and Milne, D. A. *Balancing Act: State Long-Term Care Reform*. 2008. Washington, DC, AARP Public Policy Institute.

The Lewin Group. *Projected Economic Impact of Eliminating California's Medi-Cal Adult Day Health Care Program*. Report for the Congress of California Seniors. May 18, 2010.

Mollica, R.L., Kassner, E., Walker, L., Houser, A.N. *Taking the Long View: Investing in Medicaid Home and Community-Based Services Is Cost-Effective*, Public Policy Institute, March 2009.

Patton, Michael Q. "Day 3 Keynote." SEA Change Evaluation Conclave 2013, February 26 - March 1, Kathmandu, Nepal. Available from: http://www.seachangecop.org/node/1668

Ruesga, G.A., "Philanthropy's Albatross: Debunking Theories of Change," in "Gotta Be Cruel to Be Kind (In Philanthropy)," Blog, March 25, 2013, http://postcards.typepad.com/white_telephone/2013/03/philanthropy-in-the-stocks-at-the-bradley-center.html

Stern, E., Stame, N., Mayne, J.; Forss, K., Davies, R., & Befani, B. Department for International Development (DFID) Working Paper 38. *Broadening the range of designs and methods for impact evaluations*. DFID, London, UK (2012) vi + 24 pp. Available from: http://www.dfid.gov.uk/r4d/Output/189575/Default.aspx

Trochim, W. M. The Research Methods Knowledge Base, 2nd Edition. Internet www page, at URL: <http://www.socialresearchmethods.net/kb/> (version current as of October 20, 2006), "Designing designs for research," http://www.socialresearchmethods.net/kb/desdes.php.

Webb, E.J., Campbell, D.T., Schwartz, R.D., and Sechrest, L. *Unobtrusive Measures.* SAGE Publications, 2000 (Revised Edition).

Yin, R.K. *Case Study Research: Design and Methods*. SAGE Publications, 1994 (Second Edition), with Forward by Donald T. Campbell.

Meaningful Evidence, LLC – *Where Research Meets Results.*
1069 W. Broad St., #141, Falls Church, Virginia 22046
www.meaningfulevidence.com
703-348-0061, info@meaningfulevidence.com