COMMENTS ON: "UNDERSTANDING AND MISUNDERSTANDING RANDOMIZED CONTROLLED TRIALS" BY CARTWRIGHT AND DEATON

GUIDO IMBENS - STANFORD UNIVERSITY

Deaton and Cartwright (DC2017 from hereon) view the increasing popularity of randomized experiments in social sciences with some skepticism. They are concerned about the quality of the inferences in practice, and fear that researchers may not fully appreciate the pitfalls and limitations of such experiments. I am more sanguine about the recent developments in empirical practice in economics and other social sciences, and am optimistic about the ongoing research in this area, both empirical and theoretical. I see the surge in use of randomized experiments as part of what Angrist and Pischke [2010] call the credibility revolution, where, starting in the late eighties and early nineties a group of researchers associated with the labor economics group at Princeton University, including Orley Ashenfelter, David Card, Alan Krueger and Joshua Angrist, led empirical researchers to pay more attention to the identification strategies underlying empirical work. This has led to important methodological developments in causal inference, including new approaches to instrumental variables, difference-in-differences, regression discontinuity designs, and, most recently, synthetic control methods (Abadie et al. [2010]). I view the increased focus on randomized experiments in particular in development economics, led by researchers such as Michael Kremer, Abhijit Banerjee, Esther Duflo, and their many coauthors and students, as taking this development even further.[1] Nothwithstanding the limitations of experimentation in answering some questions, and the difficulties in implementation, these developments have greatly improved the credibility of empirical work in economics compared to the standards prior to the mid-eighties, and I view this as a major achievement by these researchers. It would be disappointing if DC2017 takes away from this, and were to move empirical practice away from the attention paid to identification and the use of randomized experiments. In the remainder of this comment I will discuss four specific issues. Some of these elaborate on points I raised in a previous discussion of D2010, Imbens [2010].

In general, in causal studies it is helpful to make distinction between the choice of *design* of a study (e.g., choice of population and sampling design, and methods for allocating treatments) and

[1] The move towards pre-analysis plans and transparency (e.g., Miguel et al. [2014]) is in the same spirit.

the *analysis* (involving choices of methods for estimation and inference). See for a general discussion Rubin [2008]. Given a design involving randomization, simple estimation and inferential methods may be sufficient and optimal at the analysis stage for narrow questions regarding average treatment effects under weak assumptions, whereas more complex structural models may be required for answering more complicated counterfactual questions, relying on stronger assumptions for large sample properties.[2] Given data from an observational (non-randomized) design, the researcher is typically forced at the analysis stage to use non-experimental methods, with often-increased sensitivity to modeling choices. Conditional on the statistical methods a researcher chooses to use at the analysis stage, there are considerable advantages to use randomization at the design stage to assign treatments in terms of robustness. Because DC2017 do not make this distinction between design and analysis explicit, it is unclear to me whether it is the design of randomized experiments they take issue with, or the analysis, or both, and, more specifically, what Cartwright and Deaton see as the alternatives at each stage. Here I agree with Senn [2013], who writes, "I cannot end without expressing my exasperation with some of the critics of randomization. There may be good reasons to doubt the value of randomization, but one should not underestimate what Fisher provided. He not only produced a method of allocating treatments to experimental units but also developed an approach to analyzing the data, the analysis of variance that matched analysis to design. If you do not like this and want to propose something better, then you should make its details clear." (Senn [2013], page 1448).

Second, DC2017 express concern with the lack of sophistication in the understanding of randomized experiments among researchers, including some at organizations that are leaders in the evaluation business. They present quotes to illustrate their concerns that some researchers fail to understand, for example, the difference between covariate balance conditional on a particular draw from the assignment distribution, which randomization cannot deliver, and the balance in expectation that randomization guarantees. I am less pessimistic about the level of understanding in the profession. I think the quotes are unfortunate and poorly phrased, possibly with a non-technical audience in mind, but I think these quotes are not at all reflective of the level of understanding of their authors. Similarly, when DC2017 themselves write that "while we cannot

[2] Incidentally, the combination of experimental designs with observational study analysis methods is one of the exciting areas of ongoing research involving randomized experiments – see the discussion in Imbens [2010].

observe the *individual* [italics in original] treatment effects, we can observe their mean." (DC2017, page 3, column 1), I do not think they really mean that we can *observe* the mean treatment effect – we obviously cannot, and I think Cartwright and Deaton know that perfectly well even though they say differently.[3] DC2017 also raise concerns regarding inference in small samples. I think these are largely overblown. One of the advantages of randomization is that it makes the analyses more robust to changes in specification than they would be in observational studies. As a result, I think the concerns with using refinements to confidence intervals based on the literature on Behrens-Fisher problem, raised here and in D2010, are generally misplaced, especially in the light of the increasing use of large scale experiments – see below. The findings in Young [2015] appear largely to reflect the use of linear regression methods, rather than differences in means, in settings with high leverage observations where the regression methods can be sensitive to the presence of such observations.

Third, I want to briefly mention current research on design and analysis of randomized experiments. Although not discussed in DC2017, there is currently much innovative research in this area. See Athey and Imbens [2017] for a recent survey. Much of this is interdisciplinary, and involves researchers from computer science, statistics, as well as social sciences. Its impetus does not come from the fields where randomized experiments have a long tradition, such as biomedical settings. Instead the researchers are often motivated by the many (i.e., thousands of) large scale randomized experiments run in big tech companies such as Google, Facebook and Amazon, as well as in smaller ones, where high stakes decisions are systematically based on such experiments. There is widespread agreement in these settings regarding the fundamental value of randomization and experimentation for decision making, with a deep suspicion of having decisions driven by what DC2017 charitably call "expert knowledge" (DC2017, p.1) and Kohavi et al. [2007] call, in more colorful language, the HIPPO (Highest Paid Person's Opinion). The research literature focuses on improving the design and analysis of experiments to exploit the benefits from randomization as much as possible. One active area of research focuses on the value of personalized treatment assignments, exploiting heterogeneity in treatment effects by covariates

[3] The issues are in fact quite subtle, and it is not simply a matter of replacing "observe" with "estimate unbiasedly". Suppose we have an experiment with a single unit where we flip a fair coin to determine the treatment $W \in \{0, 1\}$ and then observe the realized outcome $Y = Y(W)$. In that case there is an unbiased estimator for the treatment effect $Y(1)-Y(0)$ for that particular unit, namely $2(2W-1)Y$, which has expectation $E[2(2W-1)Y] = (1/2)(2(2-1)Y(1) + (1/2)(2(-1))Y(0) = Y(1) - Y(0)$. Although this shows there exists an unbiased estimator for individual level treatment effect, inference is difficult relative to that for average effects in a large sample because limit theorems do not apply. See Athey and Imbens [2016].

(e.g., Athey and Imbens [2016], Wager and Athey [2017]). There is also much research analyzing settings where Rubin's SUTVA condition (e.g., Imbens and Rubin [2015]) that treatments for one unit do not affect outcomes for other units is violated. In settings with social interactions, such violations are common. These violations may be simply a nuisance to be taken into account, or the main focus of the analyses. Examples include the analysis of the effect of job training programs with assignment rates varying across labor markets in Cr´epon et al. [2013], exact randomization-based tests for the presence of interactions (Aronow [2012], Athey et al. [2017]). Another literature focuses on combining observational and experimental data, for example to look at long-term outcomes that are not easily measured in experiments (e.g., Athey et al. [2016]). Some of the literature studies the efficiency of sequential designs using multi-armed bandits, where units are assigned to various treatment arms based on initial estimates of efficacy, with these estimates continually updated as new information comes in. Such designs are particularly effective in settings where researchers are interested in comparing multiple (possibly many) treatments (Scott [2010], Li et al. [2010], Dimakopoulou et al. [2017]).

Finally, I want to make a comment on the distinction between internal and external validity. Despite the suggestions in DC2017, internal and external validity are well-understood concepts, and it would be helpful if DC2017 had used them in the standard manner rather than proposing new terms. By the standard usage I mean, for example, Shadish et al. [2002] who define internal validity as "the validity of inferences about whether observed covariation ... reflects a causal relationship," and external validity as "the validity of inferences about whether the cause-effect relationship holds over variation in persons, settings, treatment variables, and measurement variables. Rosenbaum [2002] writes in a similar spirit, "A randomized study is said to have a high level of 'internal validity' in the sense that the randomization provides a strong or 'reasoned' basis for inference about the effects of the treatment ... on the ... individuals in the experiment," and " 'external' validity refers to the effects of the treatment on people not included in the experiment." Cartwright [2007a] appears to have a very different perspective on these concepts, leading her to conclude that "despite the claims of RCTs to be the gold standard, economic models have all the advantages when it comes to internal validity" and "But it seems that RCTs have the advantage over economic models with respect to external validity," (Cartwright [2007b], p. 19). The standard perspective is more helpful for thinking about the increasing attention paid to meta studies, e.g., Meager [2015] and hierarchical modelling strategies (e.g., Gelman and Hill [2006]).

References

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. Journal of the American Statistical Association, 105(490):493–505, 2010.

Joshua D Angrist and J¨orn-Steffen Pischke. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. Journal of economic perspectives, 24(2):3–30, 2010.

Peter Aronow. A general method for detecting interference between units in randomized experiments. Sociological Methods & Research, 41(1):3–16, 2012.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27):7353–7360, 2016.

Susan Athey and Guido W Imbens. The econometrics of randomized experiments. Handbook of Economic Field Experiments, 1:73–140, 2017.

Susan Athey, Raj Chetty, Guido Imbens, and Hyunseung Kang. Estimating treatment effects using multiple surrogates: The role of the surrogate score and the surrogate index. arXiv preprint arXiv:1603.09326, 2016.

Susan Athey, Dean Eckles, and Guido W Imbens. Exact p-values for network interference. Journal of the American Statistical Association, pages 1–11, 2017.

Nancy Cartwright. Hunting causes and using them: Approaches in philosophy and economics. Cambridge University Press, 2007a.

Nancy Cartwright. Are rcts the gold standard? BioSocieties, 2(1):11–20, 2007b.

Bruno Cr´epon, Esther Duflo, M. Gurgand, R. Rathelot, and P. Zamora. Do labor market policies have displacement effects? evidence from a clustered randomized experiment. Quarterly Journal of Economics, 128(2):531–580, 2013.

Maria Dimakopoulou, Susan Athey, and Guido Imbens. Estimation considerations in contextual bandits. arXiv preprint arXiv:1711.07077, 2017.

Andrew Gelman and Jennifer Hill. Data analysis using regression and multilevel/hierarchical models. Cambridge university press, 2006.

Guido Imbens. Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). Journal of Economic Literature, pages 399–423, 2010.

Guido W Imbens and Donald B Rubin. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, 2015.

Ron Kohavi, Randal M Henne, and Dan Sommerfield. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 959–967. ACM, 2007.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In Proceedings of the 19th international conference on World wide web, pages 661–670. ACM, 2010.

Rachael Meager. Understanding the impact of microcredit expansions: A bayesian hierarchical analysis of 7 randomised experiments. arXiv preprint arXiv:1506.06669, 2015.

Edward Miguel, Colin Camerer, Katherine Casey, Joshua Cohen, Kevin M Esterling, Alan Gerber, Rachel Glennerster, Don P Green, Macartan Humphreys, Guido Imbens, et al. Promoting transparency in social science research. Science, 343(6166):30–31, 2014.

Paul R Rosenbaum. Observational studies. In Observational Studies. Springer, 2002.

Donald B Rubin. For objective causal inference, design trumps analysis. The Annals of Applied Statistics, pages 808–840, 2008.

Steven L Scott. A modern bayesian look at the multi-armed bandit. Applied Stochastic Models in Business and Industry, 26(6):639–658, 2010.

Stephen Senn. Seven myths of randomisation in clinical trials. Statistics in medicine, 32(9): 1439–1450, 2013.

William R Shadish, Thomas D Cook, and Donald T Campbell. Experimental and quasi-experimental designs for generalized causal inference. Houghton, Mifflin and Company, 2002.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, (just-accepted), 2017.

Alwyn Young. Channelling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. E, 0:0–0, 2015.