# Comments on Best Quasi-Experimental Practice

Larry V. Hedges

Northwestern University

Presented at the Future of Implementation Evaluation,
National Science Foundation, Arlington, VA, October 28, 2013

# Opening Caveats

This work is ***very*** important, given that there is widespread belief that randomization is often impossible

It gives us reason to be optimistic

This work shows the wide gulf between ***best practice*** and typical practice (or the practice that is sometimes observed)

Given those caveats, my job is to raise issues about this work

# Organization of My Talk

1. Causal inference
2. Theoretical justification
3. Issues of research strategy
4. Practical Issues

# Causal Inference
# (Two Group Comparison)

The Rubin/Neyman explication of causal inference

The $i^{th}$ individual has two potential outcome values:

their outcome if treated $Y_i^T$

their outcome if in control group $Y_i^C$

The causal effect of T versus C on outcome *Y* for individual unit *i is*

$$\tau_i = Y_i^T - Y_i^C$$

As you know, we can't identify an estimate of $\tau_i$ because no person gets both T and C (one of $Y_i^T$ or $Y_i^C$ is missing).

Note that I have made a hidden assumption in the notation: I assume that the potential outcomes of unit *i* don't depend on the assignment of other units

# Randomized Experiments Give Estimates of Average Causal Effects

Randomized experiments can give estimates of average causal effects

To talk about this we need a little more notation to make it clear what is going on with assignment

Let $Y_i^T(\mathrm{T}) = Y_i^T$ be the outcome value of unit *i* when assigned to T and $Y_i^C(\mathrm{C}) = Y_i^C$ be the outcome when unit *i* is assigned to C

Let $Y_j^C(\mathrm{C}) = Y_j^C$ be the outcome value of unit *j* when assigned to C and $Y_j^C(\mathrm{T}) = Y_j^T$ be the outcome when unit *j* is assigned to T

In randomized experiments $\mathrm{E}\{Y_i^C(\mathrm{C})\} = \mathrm{E}\{Y_j^C(\mathrm{C})\}$

# Randomized Experiments Give Estimates of Average Causal Effects

In a randomized experiment we observe an estimate of $E\{Y_i^T(T)\} - E\{Y_j^C(C)\}$

But

$$E\{Y_j^C(C)\} = E\{Y_i^C(T)\}$$

So

$$E\{Y_i^T(T)\} - E\{Y_j^C(C)\} = E\{Y_i^T(T)\} - E\{Y_i^C(T)\}$$

$$= E\{Y_i^T(T) - Y_i^C(T)\} = E\{\tau_i\}$$

Note that this requires no modeling assumptions!

# What Happens Without Randomization?

We do not have a "guaranteed" estimate of $E\{Y_i^C(T)\}$ , so we have to construct one

Such a construction must depend on further modeling assumptions:

- Knowing the variables to include in the model
- Knowing the functional form of the model
- Measuring the variables necessary for the model

Matching ***is*** modeling

It is usually difficult to make *deductive* claims for veracity of models

# Issue 2: Theoretical Justification

The justification of causal claims from randomized experiments is ***deductive*** not inductive

Causal claims based on the (sharp cutoff) RDD also have a deductive justification (but there is often fuzziness)

Justification of causal claims from other quasi-experiments must be ***inductive*** rather than deductive

The program of research Cook and Halberg report here is a great start on building an inductive knowledge base, but must be evaluated like any other set of empirical research

# Modeling Assumptions

One empirical finding that is very important is that small numbers of covariates may be sufficient to remove nearly all bias in some cases

This helps reduce dependence on second order modeling assumptions involved in matching on large numbers of covariates

Multivariate matching is complicated, many strategies (e.g., propensity scores, Mahalanobis distance) reduce dimensionality (to 1 dimension)

However the dimension reduction strategies involve modeling assumptions too

Moreover these specifications can be hard to get right

# Propensity Scores

The propensity score e(**x**) is the probability of a unit with vector of covariate values **x** is in the treatment group

The beauty of propensity scores is that they are balancing scores, units that have the same propensity score have the same distribution of the values on each of the components of **x** that went into that propensity score

In other words, matching on e(**x**) is the metaphorical equivalent of matching on all the covariates in **x**

This is exceptionally useful in multivariate matching

But note that this doesn't apply to just *any* **x**, it has to be the **x** that determines the probability of being in the treatment group

# Propensity Scores

What covariates go into **x**?

What is the functional form of the covariates (e.g., just linear or should extreme scores be over- or under-weighted, are interactions needed)?

Note that this is a slightly different question than which covariates control *bias* in outcome

Theory must provide a covariate set to screen, empirical methods can help rule out certain covariates

The good news: Once you get the covariates right, the *estimated* propensity score actually leads to better balance than the true propensity score!

# Issue 3: Research Strategy

This program of research is generative and is developing a body of results

There is a  mixture of field-based work and “laboratory” work (as there should be)

The ultimate question is how conclusions hold up in field-based applications

It would be useful to differentiate these parts of the work a bit more

The 4 armed WSC (randomization to experimental or  quasi-experimental designs) has been more like laboratory work

# Research Strategy

We need to know how well findings from this laboratory work hold up in the field

It is useful to remember that most experimental laboratory findings do *not* hold up in field experiments

This has implications for the empirical research findings about quasi-experiments

It also has implications for the applications to policy research

Here other examples of how laboratory work holds up in field experiments may give a relevant guideline (an analogy) for what to expect

# Most Interventions Tested in the Field Don't Work

**Education:** In about 120 IES RCTs since 2002, approximately 90% found weak or non-positive effects

**Employment/training:** In 13 US DOL RCTs since 1992, about 75% found weak or non-positive effects

**Medicine:** Reviews find that 50-80% of positive results in phase II studies are overturned in subsequent phase III RCTs

**Business:** Of 13,000 RCTs of new products/strategies conducted by Google and Microsoft, 80-90% have reportedly found no significant effects

Source: Coalition for Evidenced Based Policy

# Most Interventions Tested in the Field Don't Work

These were large, well implemented, adequately powered RCTs

The interventions tested were things that worked on a smaller scale (at least in the "laboratory") or they wouldn't have been tested in large scale RCTs

I am not sure why we should expect the situation to be different in the case of WSC

By definition, laboratory studies provide better control for isolating effects

The situation in the field is messier, selection effects more complex and harder to understand in field situations

However progress in the lab can be made more rapidly so it is an essential first step

# Issue 4: Practicalities

Best practice is not *typical* practice

Typical practice makes QEDs look easy and cheap (easier than experiments)

It is not clear that best practice QEDs are either easier or cheaper than randomized experiments

Certainly you don't have to be very smart to get decent causal inferences from experiments, you do have to be smart to get decent causal inferences from QEDs

Smart, well trained people are in short supply (but I think we can change that)

# How Hard Are Best Practice RDDs?

The case for RDDs may be best articulated both theoretically and empirically

Best practice standards exist (e.g., WWC) call for sophisticated analyses (e.g., nonparametric regression, specification searches, etc.)

In the case of fuzziness, the analyses may involve instrumental variables

These analyses may be beyond the reach of many scientists (this is not just a matter of ‘pushing a button’ on software)

RDDs have wide, but limited application

# How Hard are Best Practice ITSs?

The case for ITSs is less well articulated theoretically and empirically

Best practice standards have not been accepted yet, but are also likely to involve rather sophisticated analyses (e.g., to deal with autocorrelation structures)

It is hard to know whether these analyses will be within the reach of most scientists

However this is crucially important because ITSs have very wide application

# How Hard Are Best Practice NECGs?

NECG designs are the workhorse QED, used more often than any other design

The substantive knowledge required to develop selection theories is quite sophisticated

Generic guidelines may help, but it is hard to believe they will be the end of the story

The statistical methods required are also not simple (e.g., multi-level, multivariate matching, trading off local and focal matches, developing propensity score models)

All of these things require sophisticated knowledge and judgment

# How Hard Are Best Practice NECGs?

Two considerations have implications for expense and feasibility of best practice for NECGs

Good *matching* requires a large eligibility pool, typically several times the size of the treatment group

Good *measurement* covariates implicated by a selection model adds cost and complexity

Measurement of the all the potential matches in the eligibility pool may add considerable expense

The cost and complexity of best practice can easily exceed that of doing a randomized experiment

# How Hard is it to Understand Best Practice for QEDs?

Randomized experiments have the great virtue of simplicity

They are easy to explain to non-experts, including policy makers, making then transparent

Basic QEDs are also reasonably transparent

The more model-based methods become, the less transparent they are

The loss of transparency may disadvantage best practice QEDs

# Conclusions

When will best practice QEDS be a better alternative than randomized experiments?

When experiments cannot be done

When the expertise to carry out best practice QEDs is readily available

For NECGs, both of the and when data collection is cheap

When they are not so complex that they lose transparency