HCEO
Summer School on Socioeconomic Inequality

Professor Jeffrey Smith
Department of Economics
University of Michigan
econjeff@umich.edu

Program Evaluation I: Heterogeneous Treatment Effects

University of Chicago
July 18-22, 2016

My research

Social experiments

Non-experimental methods (especially matching and regression discontinuity)

Active labor market programs

University quality / mismatch and related issues in higher education

Statistical treatment rules

Performance management

Outline of lecture

Notation

Parameters of interest

Experiments require assumptions?

Heterogeneous impacts in experiments

Everything generalizes to observational data after you deal with selection

External validity

Experiments as benchmarks

General equilibrium effects

Conclusions

Notation

The Mill-Frost-Fisher-Neyman-Roy-Quandt-Rubin potential outcomes framework

Let Y_1 denote the outcome in the treated state

Let Y_0 denote the outcome in the untreated state

The fundamental evaluation problem is that we observe at most one of these two outcomes for each person

Let D be an indicator for participation in the program

Let R be an indicator for randomization into the treatment group in an experiment

Usual parameters of interest

The usual parameter of interest is the impact of treatment on the treated given by

$$E(Y_1 - Y_0 \mid D = 1) = E(Y_1 \mid D = 1) - E(Y_0 \mid D = 1).$$

This is sometimes called the ATET or TOT or ATT or just TT

In an experiment, this parameter is estimated by

$$E(Y_1 \mid D = 1, R = 1) - E(Y_0 \mid D = 1, R = 0).$$

In non-experimental evaluations, the unobserved counterfactual is obtained through econometric manipulation of the outcomes of non-participants: persons with $(D = 0)$

Simple model of program participation and outcomes

A simple model is useful to help organize our thinking. This model comes from Heckman and Robb (1985). See also Heckman, LaLonde and Smith (1999) and McCall, Smith and Wunsch (2016).

Outcomes: $Y_{it} = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \beta_{Di} D_{it} + \eta_i + \varepsilon_{it}$

Define $Y_{it} = D_{it} Y_{1it} + (1 - D_{it}) Y_{0it}$

Y_{1it} =treated outcome, with $Y_{1ik} = 0$

Y_{0it} =untreated outcome

Note structure on outcome equation for simplicity.

Note additive, separable, heterogeneous treatment effect (and no interactions, yet)

Note the fixed effect on outcomes.

Assume that treatment is available only in period *k* as in the classic Heckman and Robb (1985) setup

A simple model (continued)

Participation equation:

$$D_i^* = \gamma_0 + \gamma_C C_i + \gamma_Y Y_{0it} + \phi_\beta \beta_{Di} + U_{ik};$$

$D_{it} = 1$ iff $D_i^* > 0$ and $t \geq k$ else $D_{it} = 0$

This is a standard latent index model where D_i^* represents the net utility from participation.

Clarify Y_{0it} in terms of the definition of the outcome equation.

This formulation assumes agents know their treatment effect. It is easy to specialize to the case where they simply have some, possibly ill-informed, beliefs about it.

What agents know about their treatment effects is a wide-open area for research.

Implications of the simple model of program participation

Q: Assuming that the β_i are independent of everything else in the simple model, and that they are known to agents, what are the impacts, if any, of the simple model for the relative magnitudes of ATET, ATE and ATNT?

Put differently, is there selection on impacts and, if so, what is the nature of it?

Q: Is there selection into treatment based on untreated outcome levels? If so, what is the nature of it?

Implications of the simple model (continued)

The simple model also has many implications for thinking about non-experimental evaluation strategies.

When does simply running a regression provide unbiased estimates?

Exogeneity: $E(\eta_i + \varepsilon_{it} \mid X_{1i}, ..., X_{ki}, D_{it}) = 0$.

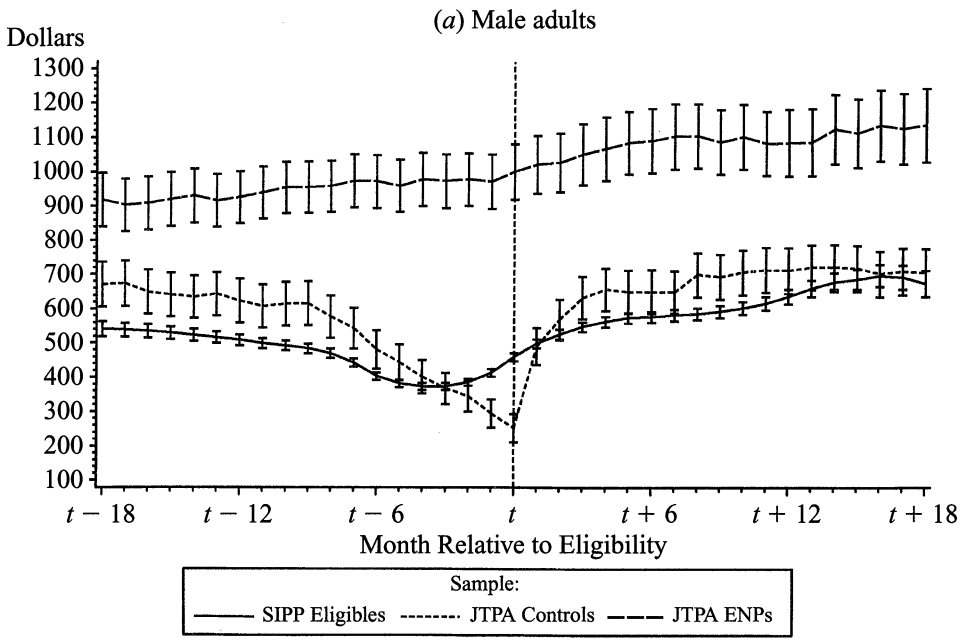
Conditional independence (CIA): $E(\eta_i + \varepsilon_{it} \mid X_{1i}, ..., X_{ki}, D_{it}) = E(\eta_i + \varepsilon_{it} \mid X_{1i}, ..., X_{ki})$

Think about conditioning variables in the context of the model

An experiment (i.e. "randomized control" trial) makes the conditional independence assumption true (in the population) by construction.

Foreshadow the use of costs as an instrumental variable.

Think about the use of difference-in-differences in the context of the model



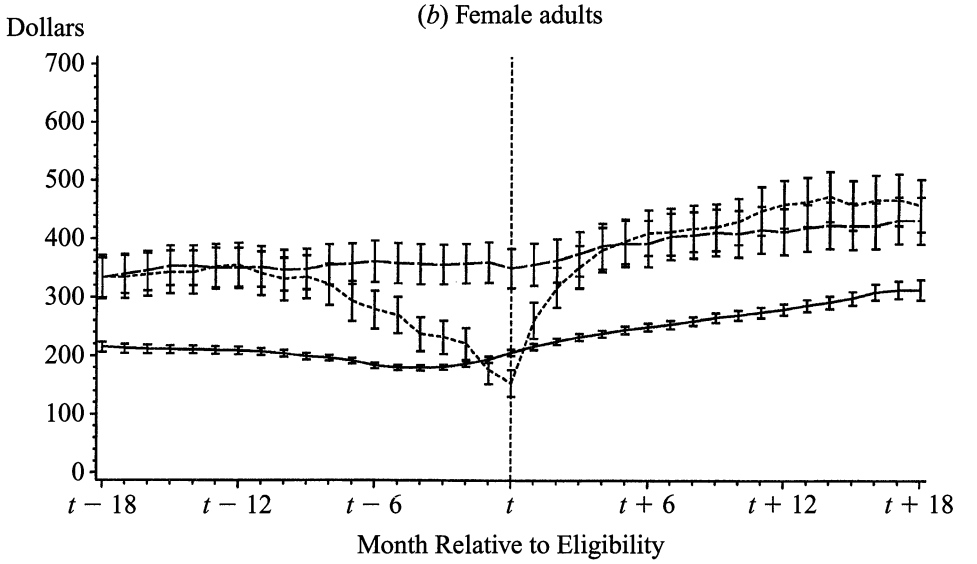


Fig. 1. *Mean Self-Reported Monthly Earnings: SIPP Eligibles and JTPA Controls and ENPs*
Notes: SIPP uses all JTPA–eligible person–month observations of respondents present in both the first and last months of the panel. Controls are randomised–out participants from the National JTPA Study. Observations based on quasi-rectangular sample. ENPS are JTPA–eligible non-participants at the same sites as the controls from the National JTPA Study. Observations based on quasi-rectangular sample. Standard error bars +/− 2 standard errors of the means.

Experimental evaluations require assumptions

No randomization bias (different from no Hawthorne effects or John Henry effects)

We still know very little about the empirical importance of this
See the recent paper using the ERAD data from Barbara Sianesi at IFS

No treatment group dropping out and no control group substitution (or different interpretation of the estimates)
Heckman, Smith and Taber (1998) *ReStat*
Heckman, Hohmann and Smith (2000) *QJE*

No general equilibrium effects

In statistics, this is SUTVA, for Stable Unit Treatment Value Assumption

Key: Experiments require untestable assumptions too!

Variants of random assignment

Random assignment at the time of participation

Random assignment of eligibility; Self-Sufficiency Project example

Random assignment at the margin; see Black, Smith, Berger and Noel (2003)

Multi-stage random assignment (useful for statistical treatment rules)

Random assignment of incentives to participate; creating your own instrument (also called the "randomized encouragement" design)

Group random assignment (e.g. classrooms in the TFA evaluation)

Each variant answers a different question; some may be more politically palatable than others

Other parameters of interest

Parameters requiring the joint distribution of outcomes

Experiments and standard non-experimental methods provide only the marginal outcome distributions $f(y_1 | D=1, R=1)$ and $f(y_0 | D=1, R=0)$

Some parameters require the joint distribution $f(y_1, y_0 | D=1)$

Example: Fraction gaining or losing from the program.

Example: Percentiles (e.g. median) of the impact distribution.

Example: Impact variance.

Example: Outcome correlation.

Other parameters of interest (continued)

Heckman, Smith and Clements (1997) *ReStud* discuss how to use bounds and other methods to obtain information on parameters that depend on the joint distribution

Recent application and survey: Djebbari and Smith (2008) *Journal of Econometrics*

More work to be done here on how the economics can narrow the identified set for these parameters.

Bounds / set identification / partial identification

Simple example: 2 x 2

Rows: employment status in control state: NE: 0.4, E: 0.6
Columns: employment status in treatment state: NE: 0.2, E: 0.8

These are the marginal distributions provided by an experiment

Note that there are two bits of information but three unknowns

Assumptions can provide the initial bit, as with an assumption of no negative effects

Bounds (continued)

The formula for the Frechét-Höffding bounds is:

$$\max[F_1(Y_1) + F_0(Y_0) - 1, 0)] \leq F(Y_1, Y_0) \leq \min[F_1(Y_1), F_0(Y_0)].$$

The intuition for the upper bound is easy: the cell may not exceed either marginal

The intuition for the lower bound can be seen by thinking about (0.6, 0.4) marginals; the min is then 0.2 in the (0,0) cell because if you go lower, the sum exceeds one.

Apply to the case of the (NE, NE) or (0, 0) cell in the example

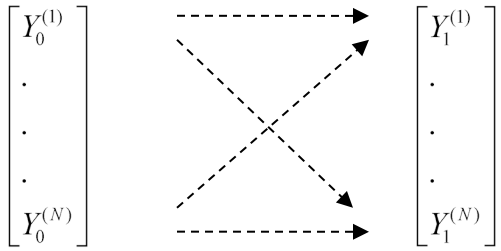
$F_1(0) = 0.2$ and $F_0(0) = 0.4$

$$\max[0.2 + 0.4 - 1, 0)] \leq F(0,0) \leq \min[0.2, 0.4].$$

Thus, the probability of (NE, NE) is between 0.0 and 0.2. Cool!

Bounds (continued)

The bounds correspond to the cases of rank correlations of -1 and 1.



Rank preservation minimizes the impact variance while rank inversion maximizes it.

There are unique distributions of impacts associated with rank correlations of -1 and 1 and sets of possible distributions of impacts for rank correlations in (-1, 1).

Testing the null of a zero impact variance

Heckman, Smith and Clements (1997) *ReStud*, Appendix E show how to use the estimated distribution of impacts in the rank preservation case to test the null of the common effect model, i.e. of a zero impact variance.

The test operates under rank preservation because the variance is minimized there.

To see how to implement this, imagine collapsing the treated and untreated outcome distributions into percentiles and taking differences to obtain impacts at each percentile.

The key is thinking about how to get the distribution of the estimated impact variance under the null. The solution is to resample from the control distribution.

Link to broader literature on boundary (of the parameter space) issues in econometrics. Here $\text{var}(Y_1 - Y_0) \geq 0$ is the relevant boundary.

Note the relationship to bootstrap inference.

Quantile treatment effects

Compare quantiles of the treated and untreated outcome distributions

Interpretation 1: impacts on quantiles of the outcome distribution

Interpretation 2: impacts at quantiles of the outcome distribution

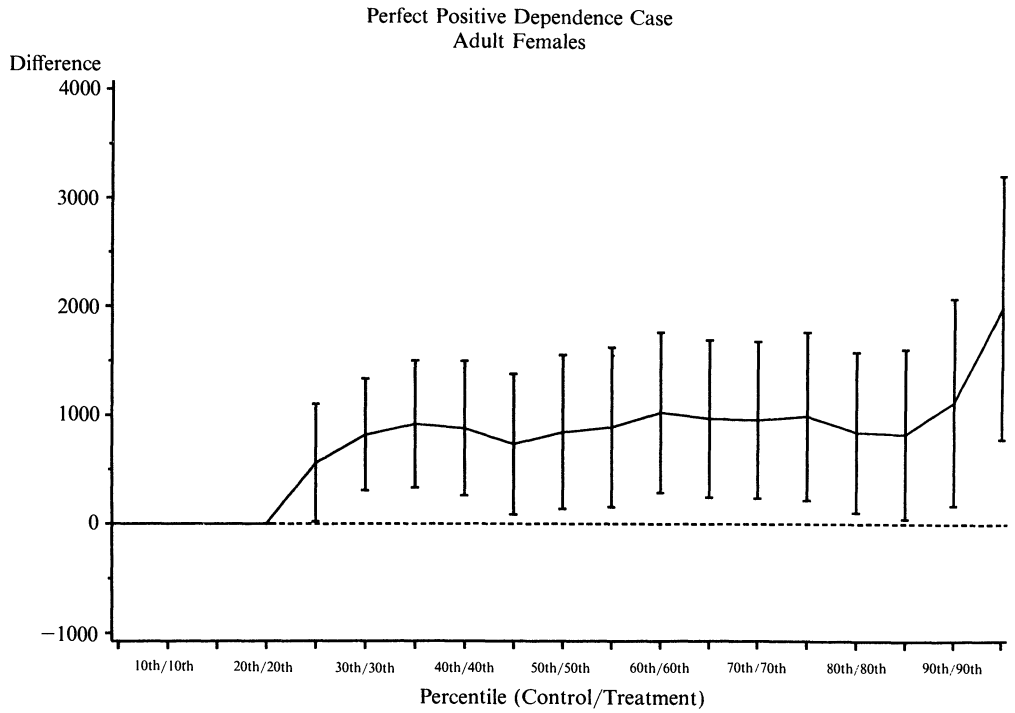
The latter interpretation requires an assumption about the joint distribution of outcomes, namely that it embodies a rank correlation of one. This is sometimes called the "rank preservation" assumption

An implication of rank preservation can be tested: see Bitler, Gelbach and Hoynes (2005) NBER

Intuition: covariate balance at percentiles of the outcome distribution

The QTEs interpreted via rank preservation provide the lower bound on the impact variance; i.e. they correspond to the impacts implied by the F-H lower bound distribution.

Q: Why are QTEs not routinely reported in experimental evaluations?



1 National JTPA Study 18 month impact sample
2. Standard errors for the quantiles are obtained using methods described in Csorgo (1993)

FIGURE 1
Treatment-control differences at percentiles of the 18 month earnings distribution

In an Appendix available on request, we explore the sensitivity of these estimates to measurement error in earnings. Our basic inferences are not altered, including our major inference bounding the variability in programme impacts away from zero.

(*c*) *The discrete case*

The Fréchet–Hoeffding bounds apply to all bivariate outcome distributions.[20] Variables may be discrete, continuous or both discrete and continuous. In this section, we use the bounding distributions to establish the variability in the distribution of impacts on employment status. The latent distribution underlying this situation is multinomial.[21] Let (E, E) denote the event "employed with treatment" and "employed without treatment" and let (E, N) be the event "employed with treatment, not employed without treatment." Similarly, (N, E) and (N, N) refer respectively to cases where a person would not be employed if treated but would be employed if not treated, and where a person would not be employed in either state. The probabilities associated with these events are P_{EE}, P_{EN}, P_{NE} and P_{NN}, respectively. This model can be written in the form of a contingency table. The columns refer to employment and non-employment in the untreated state. The rows refer to employment and non-employment in the treated state.

20. Formulae for multivariate bounds are given in Tchen (1980) and Rüschendorf (1982).
21. The following formulation owes a lot to the missing cell literature in contingency table analysis. See, e.g. Bishop, Fienberg and Holland (1975).

Conditional and unconditional quantile treatment effects

Conditional and unconditional quantiles are quite different

Consider the context of the Michigan Medical School Salary Study

An assistant professor may be in the 20^{th} percentile of the unconditional salary distribution but the 80^{th} percentile of the conditional (on being an assistant professor) salary distribution.

This study looked (in an explicitly non-causal way) at the "treatment effect" of being female on conditional (on a small number of variables) quantiles of earnings

Describe the findings and link them to theories of differential male/female outcomes

Note that the heterogeneity in the impacts matters substantively.

Random coefficient models

An alternative to the rank preservation assumption for identifying the joint distribution of outcomes (and thereby the distribution of impacts)

Assumes impacts uncorrelated with untreated outcome. In notation, assumes

$$(Y_1 - Y_0) \perp Y_0 \mid D$$

In a simple setup without covariates, the variance in the treatment effect equals the difference between the variance of the treated outcome and the variance of the untreated outcome:

$$\text{var}(Y_1) = \text{var}(Y_0 + \Delta) = \text{var}(Y_0) + \text{var}(\Delta)$$

Rearranging yields

$$\text{var}(\Delta) = \text{var}(Y_1) - \text{var}(Y_0)$$

Note the implicit test here; if the difference in variances is negative, then the random coefficient model is clearly false. This test is informative for one of the groups in the US National JTPA Study.

Random coefficient models (continued)

Can assume a normal distribution (the classic case, often invoked in HLM and in IO)

Go through HLM a bit.

Can assume a flexible parametric form

Can estimate non-parametrically via deconvolution as in HSC (1997)

When is this model economically plausible for mandatory or voluntary programs?

Subgroup effects motivation:

What works for whom?

Adjust program operation to target groups with higher impacts in an informal way

Example: Budgetary changes in the relative funding in the youth and adult components of the US Job Training Partnership Act program following publication of the experimental results showing that the program had possibly negative effects on male youth and at best very modest effects on female youth.

Subgroup effects motivation: statistical treatment rules

Example: US Worker Profiling and Reemployment Services System (profiles on levels)
Example: Canada SOMS
Example: Response to Intervention (RtI)
Example: Susan Murphy SMART (Sequential, Multiple Assignment, Randomized Trial) designs
Example: "Selective incapacitation" (profiles on levels); Bushway and Smith (2008)

Basic idea: use a statistical model to assign individuals to treatment who are expected to benefit the most from it.

See Smith and Staghøj (2008) for a survey and various Manski papers, e.g. Manski (2004), for the conceptual framework.

Subgroup effects motivation: analysis of variation in treatment effects

Distinguishing systemic versus idiosyncratic variation in treatment effects

See e.g. Djebbari and Smith (2008), Bitler, Gelbach and Hoynes (2014)

Simple way: take out as much systematic variation as you can and then bound the remaining variation or apply other approaches to it as above.

Are subgroup effects common?

This is often (implicitly) assumed in the literature

What if effects are heterogeneous within subgroups? Consider an example:

Half of men have impact 10 and half have impact 4
Half of women have impact 12 and half have impact 1

Assume that the cost of participation is five, so top half of both groups participate if agents know their impacts

Evaluation finds program "works better for women" so gender-specific subsidies are provided to induce the remaining women to participate ….

Conditional mean impacts on treated in general do not equal impact on marginal untreated person!

Are subgroup effects structural?

Structural = policy invariant

Subgroups effects may be common, or structural, or both, or neither

Confounders matter here (more on this below)

The estimated subgroup effect may change when the policy changes even if the distribution of treatment effects within groups is structural if the policy changes the program participation process.

Going deeper: is structural always a binary notion?

Models of heterogeneous treatment effects

May or may not be models of subgroup effects

Relates to the question of whether subgroup effects are "structural"

Important for understanding mechanisms
    Ex: Rosenzweig on male / female differences in the impact of education

Can provide testable predictions
    Ex: Bitler, Gelbach and Hoynes (2006) *AER* on Connecticut Jobs First

Can provide restrictions on the joint distribution of outcomes
    Ex: Kline and Tartari (2016) *AER* also on Connecticut Jobs First

Huge opportunities for research here

Site / context effects and external validity

Suppose that individual impacts depend on both unit and site characteristics, as in:

$\Delta_{ic} = g(X_i, X_s)$, where not all characteristics may be observed at either level

Examples of external context characteristics: local labor market conditions, school characteristics

Examples of internal (to the program) context characteristics: nature and quality of implementation, contractor choice, program management style / characteristics

How to generalize the results of an experiment implemented in a small number of sites, with a subset of the possible values of the site characteristics, to the population of sites, which embodies the entire distribution?

This is a problem of extrapolation. It requires, implicitly or explicitly, a model.

External validity (continued)

Key issues identified in Hotz, Imbens and Mortimer (2005):

1. Selection into the study from the population of possible sites

2. Common support

See also Muller (2014).

External validity has important implications for the design of experiments (i.e. for initial site selection) and of non-experimental evaluations. Deaton is too narrow here.

Same problem arises for individual characteristics but typically $n_i >> n_s$

There is a broader debate about the relative importance of internal and external validity when relying on research to inform policy. See e.g. the discussion in the Imbens (2013) review of the Manski (2013) book.

How does "structure" address the external validity issue? Is it more than aspiration?

An aside on Gechter (2014)

Bounding impacts based on information about the joint distribution in one location.

Rank preservation

Assumptions about the degree of dependence

Subgroups / sites / contexts and meta-analysis

Meta-analysis in medicine

Meta regressions in economics: accomplishing a different task

Measurement issues for the dependent variable

Measurement issues for the independent variables
    Inevitably a bit reductionist

Confounding matters here too

Card, Kluve and Weber (2010) *Economic Journal*

Vivalt (2015) job market paper

Site / context effects and learning about program operation

Site selection – convenience or random?

How many sites to include traded off against the number included per site?

Examples of designs:

1. NJS: tried for a random sample of 20, got a convenience sample of 16, randomized all participants at the sites for about 18 months with 1/3 control, 2/3 treatment. Randomized to service-eligible versus not service-eligible.

2. NJCS: ran the evaluation at every site (not as many as JTPA / WIA but still expensive) but only assigned five percent of eligible applicants to the control group at each site.

3. WGSE: tried for a random sample of 30, got 26 plus two replacement sites for a total of 28; randomized all participants for a set period of time, but randomized to training-eligible versus not training-eligible rather than service-eligible versus not service-eligible.

Example of what can be learned: NJCS and performance measures

Example of what cannot be learned: Riverside “miracle”; see Dehejia (2003)

Subgroup effects and fishing

Problem: if you estimate and report enough subgroup impacts, some of them will be statistically significant, even if they all equal zero in the population

Special case of what the literature calls the multiple comparisons problem, and a good illustration of the occasional oddness of classical statistical inference

Responses to fishing:

1. Pre-commitment: confirmatory versus exploratory subgroup analyses

In US DOE evaluations, confirmatory outcomes / subgroups subject to adjustment for multiple comparisons, while exploratory analyses are not.

2. Adjustment of p-values for multiple comparisons; see Schochet (2008)

3. Dimension reduction via domain-specific indices
    Example: MTO studied by Kling, Liebman and Katz (2007)
    Example: Anne Fitzpatrick job market paper

See e.g. Anderson (2008) for an application that worries about these issues.

An aside on classical statistics

One can frame the fishing remedies as solutions to a problem that exists only because researchers (and policymakers and students and pretty much everyone else) takes classical significance testing too seriously.

Issues:

Why privilege zero as a null?

Where does the 0.05 cutoff for statistical significance come from?

Why binary accept / reject rather than p-values as a continuous measure?

Key: substantive versus statistical significance?

Ziliak and McCloskey (2007) *The Cult of Statistical Significance*

Aside on classical statistics (continued)

More generally, what fraction of the overall uncertainty in an estimate typically is due to sampling variation, which is the only sort of uncertainty captured in the standard errors?

Other sources include functional form choice, measurement error / choices, other design issues such as temporal alignment

Example: CETA studies

Example: footnote in Lise, Seitz and Smith (2004)

An aside on hierarchies of evidence

What is the hierarchy of evidence?

RCT
RD
IV
DID
CIA / selection on observed variables
Case studies
Theory

Key empirical question: within identification strategy quality variation versus between identification strategy variation

Studies vary on many dimensions other than their identification strategy

Experiments as benchmarks

Evaluating structural models

Todd and Wolpin (2005) *AER*
Lise, Seitz and Smith (2015) *IZA JOLE*

Evaluating non-experimental identification strategies

LaLonde (1986) *AER*
Dehejia and Wahba (1999, 2002) *ReStat*, *JASA*
Smith and Todd (2005a, 2005b) *Journal of Econometrics*
Many other papers

Much of this literature is confused about what LaLonde showed

Much of this literature is confused about what it is testing

An aside on replication

A very loosely used term in economics!

Example usages:

Re-run the authors' programs on their analysis file

Recreate the analysis file and/or the programs from "scratch"

Repeat the authors' analysis on different data on the same population at the same time

Ex: PSID versus NLSY-79

Repeat the authors' analysis for a different population or a different time period or a different geographic location

Ex: repeating a US study using Canadian data

Ex: repeating an NLSY-79 study using NLSY-97 (e.g. college quality)

More on replication aside

Paper I am writing now (with Sebastian Calonico)

LaLonde (1986)

Heckman and Hotz (1989)

Different data source for outcomes => sample size and composition differs

Dehejia and Wahba (1999, 2002)

Cannot recreate LaLonde subgroups using his data

Smith and Todd (2005a,b)

Urban residence as an exclusion restriction

Calonico and Smith (2015)

Not enough documentation for exact replication

PSID versions

Mysterious PSID replicates

Different types of replication answer different questions and catch different errors

General equilibrium evaluation

What if SUTVA fails? SUTVA likely fails in practice in many contexts

1. Compare across (relatively) isolated markets with variation in treatment intensity
Dahlberg and Forslund (2005) *SJE*
Angelucci and de Giorgi (2009) *AER*
Crepón, Duflo, Gurgand, Rathelot and Zamora (2013) *QJE*

2. General equilibrium models
Johnson (1979, 1980)
Davidson and Woodbury (1993) *JoLE*, UI bonus experiments
Heckman, Lochner and Taber (1998) *AER*, college tuition subsidy
Plesca (2010) JHC, US employment service
Lise, Seitz and Smith (2004) NBER

This is a large fraction of all of the papers. Much more remains to be done here.

Summary and conclusions

Experiments are not a substitute for thinking

There are more parameters of interest than just ATET and ATE

Many methods exist for estimating those parameters, but more research remains to be done, especially to integrate the economics and the econometrics

Subgroup effects and site effects are not as simple as they might seem

General equilibrium effects of programs are substantively important and understudied