# Learning About External Validity from Multiple Literatures

Jeffrey Smith

June 11, 2023

# Epigram

"I know it when I see it"
        - U.S. Supreme Court Justice Potter Stewart

# Motivation: general

- When can we generalize evaluative knowledge?
- Many literatures grappling with this in different ways
- Knowledge is expensive!
- Faulty generalization can be expensive too!

# Motivation: "Structural parameters"

- The U.S. Negative Income Tax (NIT) experiments
- Ways we talk about behavioral parameters ...
  - Labor supply elasticity
  - Elasticity of taxable income
- Changes over time, but uneven ones
- Aside on McCloskey's *The Rhetoric of Economics*
- More of this will appear in the paper

# Motivation: heterogeneous treatment effects

- A second revolution in evaluative empirical work

- The (slow, partial) demise of "the effect"

- Site effects in the National JTPA Study

- TDL: Treatment effect heterogeneity makes external validity interesting

# Dimensions of treatment effect heterogeneity

- Characteristics of treated units
  - ▶ Demographics, earnings histories, etc.
- Characteristics of the treatment
  - ▶ Duration or intensity
  - ▶ Type of training or job search assistance
- Characteristics of the program
  - ▶ Program manager characteristics
  - ▶ Operating organization characteristics
- Characteristics of the context
  - ▶ Local economic conditions
  - ▶ Available local service providers

# Other dimensions of TE heterogeneity

- Scale-up
  - ▶ Buhl-Wiggers et al. scale-up paper
- Relative time (i.e. short run vs. long run)
  - ▶ Minimum wage
  - ▶ Elasticity of taxable income
- Estimand (i.e. LATE versus ATE)
  - ▶ Black, Joo, LaLonde, Smith, and Taylor (2023)

# Empirical importance of treatment effect heterogeneity

- Evidence from the Fréchet-Höffding bounds
  - Heckman, Smith and Clements (1997)
  - Djebbari and Smith (2008)
  - Buhl-Wiggers et al. (2023)
- Evidence from existing evaluations
  - National JTPA Study and demographics
  - National Job Corps Study
- Jacob's revenge: statistical power and subgroups

# Systematizing treatment effect heterogeneity

- Djebbari and Smith (2008) define systematic and idiosyncratic treatment effect heterogeneity.
- Need to systematize the heterogeneity by linking it to observed moderators to make it useful.
  - ▶ Statistical treatment rules
  - ▶ Causal moderators?
- Most relevant for the first four dimensions above.

# What to do with moderators when you have them

- Estimate parametric models of treatment effect heterogeneity
  - The meta-analyses of active labor market program impact estimates by Card, Kluve and Weber (2010, 2018) provide an example of this.
- Reweight data from existing evaluations to look like the target population in terms of some or all of the first four dimensions of treatment effect heterogeneity.
  - See e.g. Tipton and Olsen (2018)
- If many candidate moderators, apply machine learning methods to reduce the dimensionality of the set or discover unanticipated interactions, or both.
  - But "garbage in, garbage out"

# Practical issues with moderators: support issue

- Also called "coverage" or "overlap"
- No way to non-parametrically predict impacts for cases that are not a weighted average of existing evidence.
  - Example: want to predict for women but all existing evaluations are for men
  - Example: current issues in clinical trials
- Disciplinary differences in comfort with using functional form assumptions to project outside the support provided by the existing evaluations.

# Evaluating moderators

- Out-of-sample predictive success
- Aside: prediction versus understanding (or causal moderators versus proxies)
- Reductions in the Fréchet-Höffding lower bound of the impact variance as in Buhl-Wiggers et al. (2023) SCLB.

# An aside on non-sampling variation

- How much of the variation across studies in impact estimates results from non-sampling variation rather than sampling variation?
- Heckman and Smith (2000) on the National JTPA Study
- Dillon, Miller and Smith (2023) metaverse study.

# Mom, where do moderators come from?

- Usual suspect moderators - what is the motivation?
- Arguably, most of the usual suspects are proxies rather than causal moderators.
- Dearth of formal or even verbal models of effect moderation.
- Dearth of advances in the design and measurement of moderators.
- Have we over-invested in applied statistics and econometrics relative to behavioral substance?
- See the discussion in SCLB paper.

## Summary and conclusions

- Treatment effect heterogeneity generates external validity issues.
- We have learned a lot about the applied econometrics of treatment effect heterogeneity.
- Plenty of opportunities remain in terms of advances in moderator theory and measurement.
- We can, and should, do better than Justice Potter.

# References

Black, Dan, Joonhwi Joo, Robert LaLonde, Jeffrey Smith and Evan Taylor. 2022. "Simple Tests for Selection Bias: Learning More from Instrumental Variables." *Labour Economics* 79: 102237.

Buhl-Wiggers, Julie, Jason Kerwin, Juan Sebastián Muñoz, Jeffrey Smith, and Rebecca Thornton. Forthcoming. "Some Children Left Behind: Variation in the Effects of an Educational Intervention." *Journal of Econometrics*. Also available as NBER Working Paper No. 29459.

Card, David, Jochen Kluve and Andrea Weber. 2010. "Active Labour Market Policy Evaluations: A Meta-Analysis." *Economic Journal* 120(548): F452-F477.

Card, David, Jochen Kluve, and Andrea Weber. 2018. "What Works? A Meta-Analysis of Recent Active Labor Market Program Evaluations." *Journal of the European Economic Association* 16(3): 894-931.

## References (continued)

Djebbari, Habiba and Jeffrey Smith. 2008. "Heterogeneous Impacts in PROGRESA." *Journal of Econometrics* 145(1-2): 64-80.

Heckman, James, Jeffrey Smith, with the assistance of Nancy Clements. 1997. "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts." *Review of Economic Studies* 64(4): 487-535.

Heckman, James, and Jeffrey Smith. 2000. "The Sensitivity of Experimental Impact Estimates: Evidence from the National JTPA Study" in David Blanchflower and Richard Freeman (eds.), *Youth Employment and Joblessness in Advanced Countries*, Chicago: University of Chicago Press for NBER, 331-356.

Tipton, Elizabeth, and Robert Olsen. 2018. "A Review of Statistical Methods for Generalizing From Evaluations of Educational Interventions." *Educational Researcher* 47(8): 416-524.