Policy Evaluations for Benefit-Cost Analyses

Jens Ludwig
University of Chicago & NBER

(builds on work done with Philip Cook, Duke and Jonathan Guryan, Northwestern)

Core argument of my talk

Core argument of my talk

- We live in a fallen world

Core argument of my talk

- We live in a fallen world
  - Hard (impossible?) to explain difference between failing to reject vs. accepting the null hypothesis
  - Hard (impossible?) to explain what is a good from bad matching / diff-in-diff / etc. study
  - Plus politics are very, very political

Core argument of my talk

- We live in a fallen world
  - Hard (impossible?) to explain difference between failing to reject vs. accepting the null hypothesis
  - Hard (impossible?) to explain what is a good from bad matching / diff-in-diff / etc. study
  - Plus politics are very, very political

- We need what engineers call “human-error-tolerant design” for program evaluations / BCA
  - Adequately powered up (avoid need for explanation of ns results)
  - Bright lines for study design (avoid need for policymaking process to vet study quality on study-by-study basis)

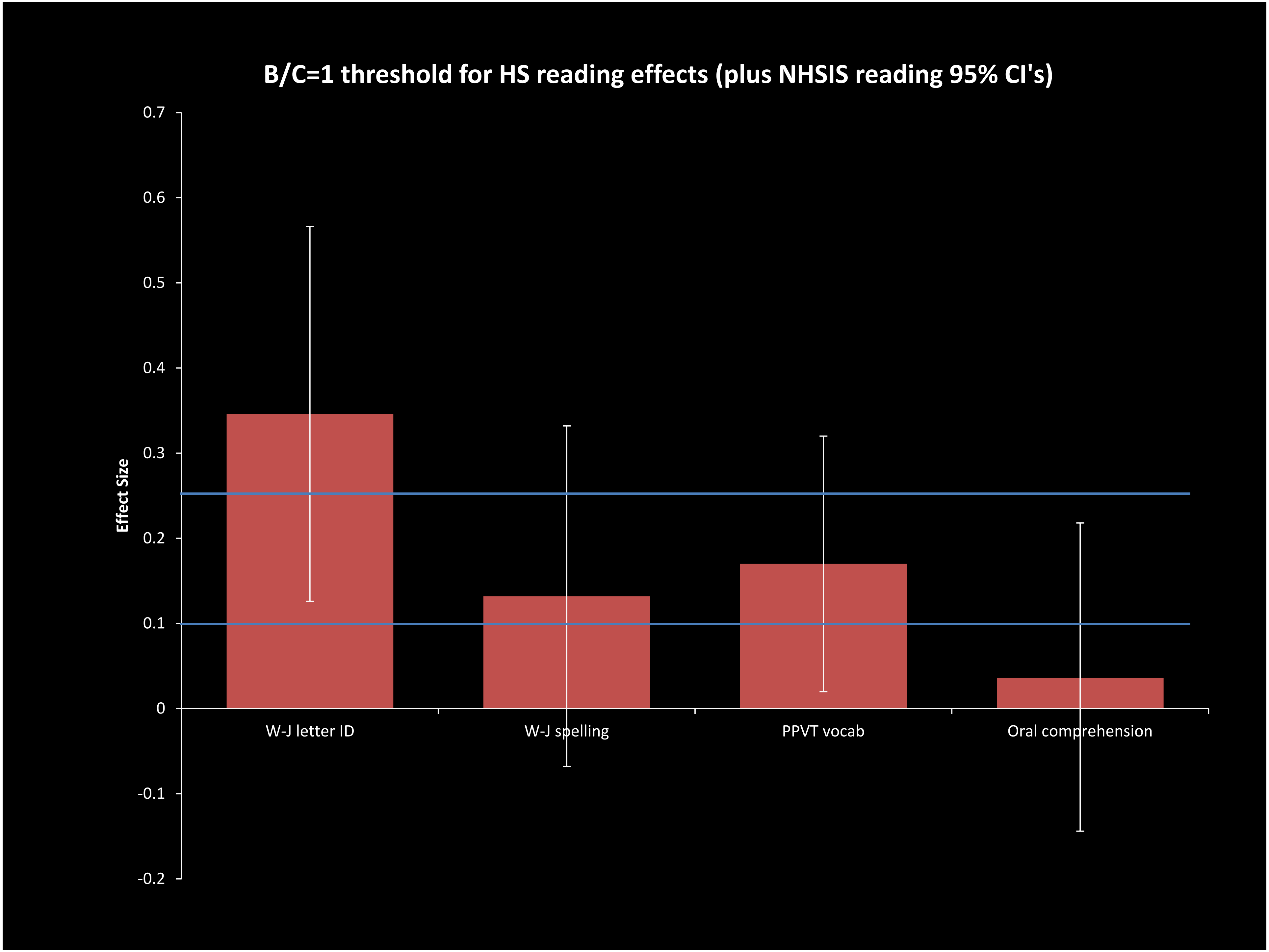
Example 1: Null results & statistical power

- National Head Start Impact Study
  - Evidence is “indisputable” that “Head Start simply does not work”… continued funding is “criminal, every bit as outrageous as tax breaks for oil companies” (Joe Klein, *Time Magazine*, 2011)
  - “Taxpayers get little for their annual investment of $8 billion in Head Start” (Ron Haskins, Brookings, 2010)
  - “Head Start’s broken promise” (Doug Besharov, AEI, 2005)

Example 1: Null results & statistical power

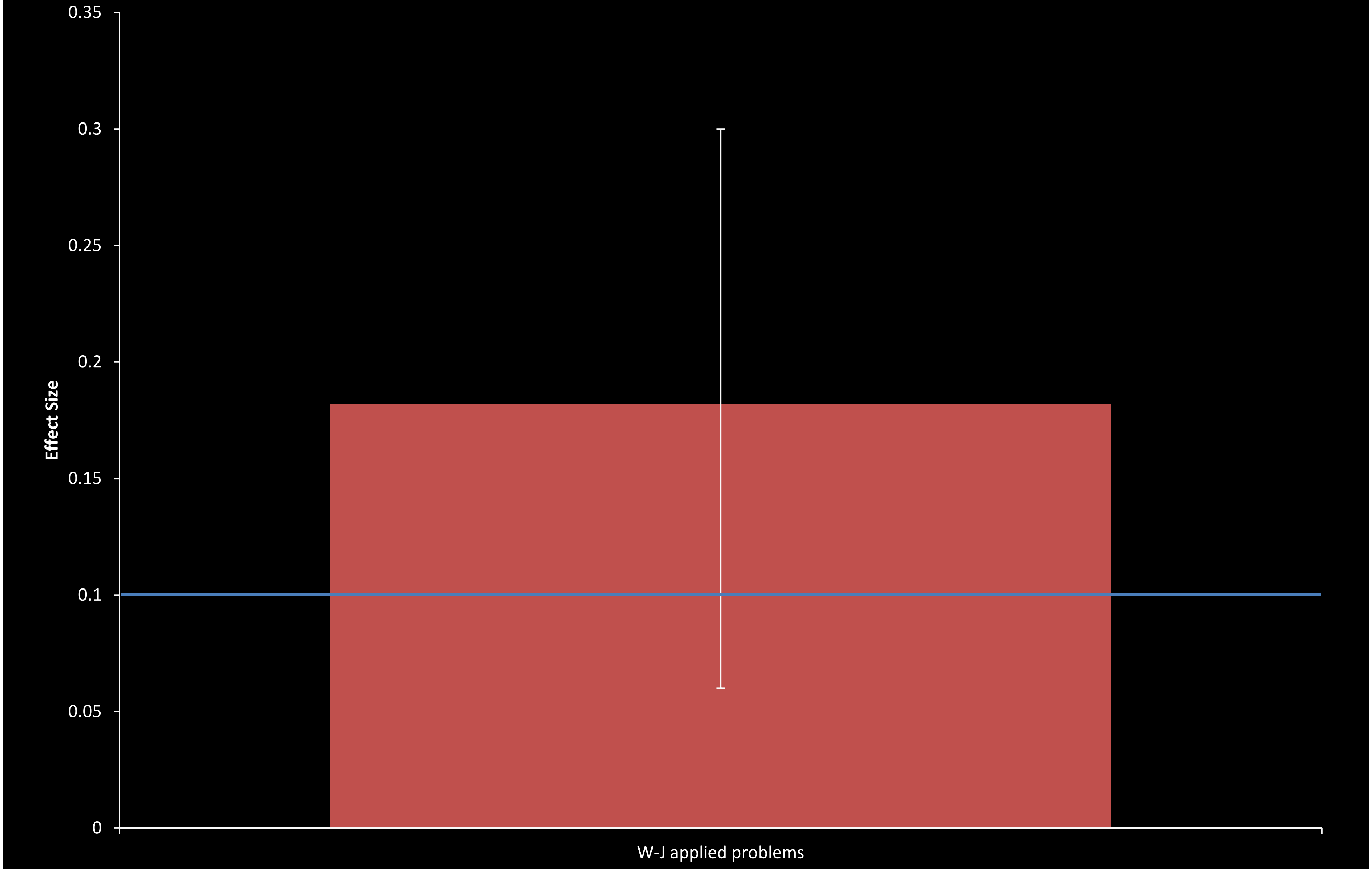
- National Head Start Impact Study
  - Evidence is "indisputable" that "Head Start simply does not work"… continued funding is "criminal, every bit as outrageous as tax breaks for oil companies" (Joe Klein, *Time Magazine*, 2011)
  - "Taxpayers get little for their annual investment of $8 billion in Head Start" (Ron Haskins, Brookings, 2010)
  - "Head Start's broken promise" (Doug Besharov, AEI, 2005)

- Wow! Surely with this sort of piling on we must have a very precise estimate of Head Start B/C ratio w/ 95% CI that rules out possibility that B/C>1, right?



B/C=1 threshold for HS reading effects (plus NHSIS reading 95% CI's)

B/C=1 threshold for HS math effects (plus NHSIS math 95% CI's)



Potential solutions for this problem

- Educate the consumers of our benefit-cost analyses
  - Concern: If even the Washington, DC think-tank commentariat don’t get it, is it realistic to believe we can do large-scale consumer education?

Potential solutions for this problem

- Educate the consumers of our benefit-cost analyses
  - Concern: If even the Washington, DC think-tank commentariat don't get it, is it realistic to believe we can do large-scale consumer education?
- Rebecca Maynard proposal: Ban asterisks
  - I am sympathetic (see Cook and Ludwig 2006) but still requires consumer education

Potential solutions for this problem

- Educate the consumers of our benefit-cost analyses
  - Concern: If even the Washington, DC think-tank commentariat don’t get it, is it realistic to believe we can do large-scale consumer education?
- Rebecca Maynard proposal: Ban asterisks
  - I am sympathetic (see Cook and Ludwig 2006) but still requires consumer education
- Anticipate user error, and build that into design
  - Here’s a potential new standard for power calculations for program evaluations:
  - Minimum detectable effect (MDE) should be small enough to detect B/C ratio =1

Example 2: Cherry-picking results

- Part 1 of problem: Research consumers cannot adjudicate what is a good vs. bad study within a given research design class
  - I have a crisp $100 bill for anyone who can make sense of the John Lott vs. John Donohue debate about whether more guns leads to less (or more) crime

Example 2: Cherry-picking results

- Part 1 of problem: Research consumers cannot adjudicate what is a good vs. bad study within a given research design class
  - I have a crisp $100 bill for anyone who can make sense of the John Lott vs. John Donohue debate about whether more guns leads to less (or more) crime
  - Stamp-of-approval mechanisms also don't seem to work
    - Compare what you would conclude "works" if you looked at WSIPP vs. Blueprints for Healthy Youth Development vs. Coalition for Evidence Based Policy vs. What Works Clearinghouse

Example 2: Cherry-picking results

- Part 1 of problem: Research consumers cannot adjudicate what is a good vs. bad study within a given research design class
  - I have a crisp $100 bill for anyone who can make sense of the John Lott vs. John Donohue debate about whether more guns leads to less (or more) crime
  - Stamp-of-approval mechanisms also don’t seem to work
    - Compare what you would conclude “works” if you looked at WSIPP vs. Blueprints for Healthy Youth Development vs. Coalition for Evidence Based Policy vs. What Works Clearinghouse
- Part 2 of problem: Research consumers *like* having mixed results within design class (cherry picking)

Potential solutions for this problem

- Educate consumers

Potential solutions for this problem

- Educate consumers
- Standardize / strengthen stamp-of-approval mechanisms
  - But still need consumer education to make people realize they should only believe WWC-approved studies etc.

Potential solutions for this problem

- Educate consumers
- Standardize / strengthen stamp-of-approval mechanisms
  - But still need consumer education to make people realize they should only believe WWC-approved studies etc.
- Anticipate user error, and build that into design
  - Do we need bright lines for research design quality?
    - For example: “Believe only RCTs and RD studies. Period.”
  - Note the tradeoff: Throwing away good information from the good non-RCT/non-RD studies, vs. current cherry-picked free-for-all that we have now

Design the car recognizing drivers are (very) fallible

- How we design program evaluations for BCA if, say, Tom Cook were our (mostly) benevolent dictator is different from what we should do in real world
- Build user error into program evaluations
  - Stop under-powering policy experiments (in a BCA sense)
  - Bright lines for study design that raise avg. study quality and make it harder for political cherry-picking of results?
  - More generally would be useful to do after-action report of how BCA is used (and abused) in real world to identify other problems we could try to design out of system