Outcome Reporting Bias in Educational Research

Jeffrey C. Valentine
University of Louisville
December 2, 2013
Acknowledgments

• This presentation is based on two projects:
  – One is with Terri Pigott and Josh Polanin (Loyola-Chicago), Ryan Williams (Memphis), and Dericka Canada (Boston College)
    • Hot off the presses at *Educational Researcher*
  – The other is with Doug Altman (Oxford), Mohammed Ansari (Ottawa), An-wen Chen (Toronto), Jamie Kirkham (Liverpool), and Barney Reeves (Bristol)
    • Manuscript in progress
Skepticism Regarding the Credibility of Scientific Results

• “Consequences of prejudice against the null hypothesis” (Greenwald, 1975)
• “Why most published research findings are false” (Ioannidis, 2005)
• “Why current publication practices may distort science” (Young et al., 2008)
• “Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results” (Wicherts et al., 2011)
• “Restructuring incentives and practices to promote truth over publishability” (Nosek et al., 2012)
Pleas for More Reproduction and Replication

• “On ensuring the availability of evaluation data for secondary analysis” (Hendrick et al., 1978)
• “Shall we really do it again? The powerful concept of replication is neglected in psychological research” (Schmidt, 2009)
• “Replication in prevention science” (Valentine et al. 2011)
• “Psychology’s woes and a partial cure: Replication” (Roediger, 2012)
Decision Points in Research Design and Analysis

• Researchers designing and analyzing a study face many decision points along the way. Examples:
  – Unbalanced randomization
  – Statistical assumptions, outliers, etc.
  – Model specification

• These decisions have some things in common:
  – Judgments are ambiguous
  – No right answers, no strong norms
Absence of Strong Norms: Unbalanced Randomization

- Random assignment equates groups on expectation
  - Groups differ only as a function of chance
- What if part way into randomly assigning participants to conditions you realize that the groups are qualitatively different along some important dimension?
- There are no clear norms governing what researchers should do in this situation
  - Start over, start over and stratify, ignore, use a covariate, etc.
The Biasing Effects of Ambiguous Judgments

• Many of the judgments involved in the research process are ambiguous

• Ambiguous judgments are where judge preferences get revealed
  – foul calls in basketball
  – One team’s goon is another team’s enforcer (hockey)

• Because there are few norms that govern lots of ambiguous judgments, there exists a lot of room for the expression of researcher preference
  – We probably engage in motivated reasoning
So For Example...

• I randomly assign 3rd graders either to get my reading intervention or not
• IQ is related to my post-test (reading comprehension)
• I realize that average IQ is higher in the control group than in the treatment group
• But the difference is not statistically significant
So What Should I Do?
What Am I Likely to Do?

• Do I …
  – Use ANCOVA, ignore, re-randomly assign?
  – To what extent is my choice dependent on what I would like to see happen?
    • Would I likely have made the same choice if the treatment group IQ had been higher?
    • Strongly suspect that tendency is for different rules to be applied depending on the implications for results
Related Lines of Research

• Allegiance effects
  – Clinical research (e.g., Luborsky)
  – Prevention (e.g., Petrosino & Soydan)
  – Drug companies, etc.

• Confirmation bias (i.e., tendency to distill complex findings in a way that supports the story one wants to tell; Littell)

• The “decline” effect
  – Often, effect sizes appear to decline over time
  – Largest effects tend to be observed earlier in the empirical history of a question

• Selective outcome reporting (i.e., reporting only outcomes that tell the story one wants to tell)
More Research

• John, Loewenstein, & Prelec (2012)
  – Surveyed over 2,000 psychologists at U.S. research universities regarding their data analysis and reporting practices
    • Data collected anonymously
    • Directly asked about acceptability of the practices
      – Also asked participants to estimate the extent to which others engaged in these practices
    • One condition had inducements for truth-telling
John et al. (2012): Results

- Failing to report all of a study’s DVs (63%)
- Stopping data collection after checking for statistical significance (56%)
- Selectively reporting studies that “worked” (46%)
- Deciding whether to exclude data after looking at the impact of doing so (38%)

- Truth telling inducements resulted in greater reporting of undesirable behavior

“Some questionable practices may constitute the prevailing research norm”
More Good News

• Francis (2012) How easily can omission of patients, or selection amongst poorly-reproducible measurements, create artificial correlations? Methods for detection and implications for observational research design in cardiology
  – Answer: Very easily

• Simmons, Nelson, & Simonsohn (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant
  – You can guess what these authors found!

• Undisclosed flexibility + motivated reasoning = big problems
Implications

• Many of the things John et al. asked about are not evil per se
• Cumulatively may “stack the deck” resulting in a bias against the null hypothesis
• Francis (2012) and Simmons et al. (2011) illustrate how this can play out
• Suggests serious problems in our ability to use evidence
Outcome Reporting Bias

• Publication bias is a well-understood phenomenon (e.g., Greenwald, 1975)
  – Researchers are less likely to submit, and journal editors and reviewers are less likely to accept, papers that lack statistically significant results on their primary outcomes
  – All else equal, “less statistical significance” = smaller effect sizes

• Outcome reporting bias – the selective reporting of measured outcomes – is a less understood problem
Selective Reporting of Measured Outcomes

• 63% of researchers in the control group of John et al.’s study said they had done this

• Why might researchers selectively report outcomes?
  – Space
    • Occasionally journal editors will say things like “Drop all of these non-significant outcomes”
    • Authors worried about space constraints prior to submission
  – Financial considerations
    • Failing to report on adverse effects (Chan & Altman, 2005)
  – Desire to get published
    • A mix of statistically significant and non-statistically significant effects can be difficult to understand and explain (especially given common statistical misconceptions)

• Worry a lot about significance chasing
Example of Outcome Reporting Bias

- Reported in Vedula et al. (2009) *NEJM*
- Studies of off-label use of gabapentin
  - Used to help control seizures; studies were designed to see if effective for migraines, bipolar disorder, et al.
- Compared outcomes described in *published* reports with those described in protocols and *internal research documents* from the industry sponsors
  - Some of which were obtained as a result of litigation
A Bit of Background

• All drug trials in the U.S. and other western countries are highly regulated
  – The U.S. government (through the FDA) requires that, prior to starting a study, drug companies complete detailed protocols that operationally define the methods and analyses that will be used in the study
    • Sampling plan, outcome measures, data analytic techniques, etc.
  – Also specifies the primary (most important) outcomes and those that are secondary
• Back to Vedula et al.

• According to the protocols and internal documents, the trials started with 21 primary outcomes. In the published documents
  – 6 went unreported
    • None of these were statistically significant
  – 4 were demoted to secondary outcomes
    • These had smaller than expected effects and/or were not statistically significant

• And ended up with 28 primary outcomes
  – 12 were newly introduced
  – 5 were promoted from secondary outcomes
Implications for Inferences

• Selective outcome reporting → potential bias

• Although this has been shown extensively in medicine, there is little empirical evidence to support this phenomenon in education
Our Study of Outcome Reporting Bias

• Educational researchers don’t use protocols
• Similarities between the goals of a protocol and the goals of a methods section of a dissertation
  – Both are meant to be highly operational
  – Plus, relatively strong norms regarding dissertations
• We therefore assumed that the dissertations represent a complete record of the study plan
• We focused on dissertations from all 96 Carnegie designated Research Universities/Very High Research Activity (RU/VH)
Methods

• Using PDDT, we searched for dissertations completed between 2001 and 2005 at each RU/VH institution with keyword “Education”

• Using the abstracts, we identified studies that reported on an intervention for student outcomes in PreK – 12
  – We focused on interventions for ease of identification of primary outcomes
  – We eliminated single-case designs because the analysis usually does not use inferential statistics
Methods (continued)

• After identifying the set of dissertations that studied an intervention with PreK-12 students, we searched for a published version of the dissertation (Google Scholar, others)
• In the matched dissertations, we coded each treatment outcome, the statistical test used and its associated $p$-value, sample size, outcome type (academic vs. socio-emotional), time to publication, etc.
  – We did not use baseline tests, correlation matrices, etc.
• In the matched publication, we coded whether each outcome was reported
Example Study
(Real, but not used in exactly this way)

421 Statistical Tests

212 Statistically Significant

29 Published

209 Not Statistically Significant

3 Published

Publication rate was about 9X greater for statistically significant (13.5%) vs. ns outcomes (1.5%)
Results

• Over 9,500 potentially relevant dissertations were screened
• 621 (6.5%) were on an intervention for preK-12 students
• 79 (13%) were subsequently published

• The 79 dissertations contained 1,559 treatment outcomes
  – That’s about 20 treatment outcomes per dissertation (!)
  – Range was 2-173
• The 79 publications contained 803 outcomes
  – Still high at over 10 per published study!
  – Range was 0-96
• Statistically significant outcomes in the dissertations were 30% more likely to be reported in the follow-up publication (95% CI = +/- 15%; OR =2.3)
  – The probability of being published, conditional on statistical significance, was .79
  – The probability of being published, conditional on non-significance, was .21
Typical Study

20 Statistical Tests

10 Statistically Significant
- 6 Published

10 Not Statistically Significant
- 4 Published

Note: This example slightly overstates the effect we observed.
Other Interesting Results

• ORB risk did not vary as a function of outcome type (academic achievement vs. socio-emotional), time to publication (< 2 years vs. > 2 years), sample size, number of statistical tests conducted

• About 46% of the statistical tests in the dissertations were statistically significant
  – This is in line with other estimates of typical statistical power in the social sciences (e.g., Cohen 1963)
How Much of a Problem is ORB?

• We don’t know

• Correlation between reported and censored outcomes matters
  – If strong (e.g., two highly correlated measures of a construct), then problem is likely minimal
  – If weak (e.g., two measures of two different constructs) then problem will lead to biased inferences

• Our early simulations suggest that typical effect size might be inflated by about 25% given this level of ORB
Lots of Problems, A Few Solutions

What can be done about shoddy data practices?

• Protocols should be routine, detailed, and publically available (e.g., stored online)
• Strong professional norms should promote the use of good data analysis practices and reporting
• Researchers could be required to state explicitly whether or not they (a) excluded or modified data, (b) tested assumptions, (c) addressed failed assumptions, (d) attempted multiple model specifications, etc.
  – Sensitivity analyses ought to be routinely carried out and reported
• Data should be warehoused in an accessible and usable way (so others can attempt to reproduce the results)
  – Strong fair use guidelines
• Full correlation matrices should be reported for every study
Outcome Reporting Bias in Educational Research

Jeffrey C. Valentine
University of Louisville
December 2, 2013
What Works Clearinghouse: Informal Overview

• Department of Education’s Institute of Education Sciences
• Aims to provide independent reviews of research on effects attributable to educational policies, processes, and products
• Produces
  – Quick reviews / single study reviews
  – Intervention reports
  – Practice guides
WWC Activities

• Products
  – Quick reviews / single study reviews
  – Intervention reports
  – Practice guides
  – Dept. of Ed grants competition reviews

• Quality control

• Quality improvement
WWC Review Processes

• WWC reviews are governed by protocols
  – Protocols are generic with some tailoring to the review question (substantive experts do this)
  – Specify the outcomes and subgroups of interest, provide and operational definition of the intervention, etc.
  – Study quality standards are applied uniformly across studies (again, with some tailoring)

• Process is highly similar to that used by the Cochrane and Campbell collaborations
  – Study quality standards based on Valentine & Cooper (2008)
WWC Study “Quality” Standards

• The great majority of studies reviewed by the WWC do not meet standards (≈ 94%)
• In part this is self-inflicted
• In part this is a reflection of reality

• The quality standards are very basic
  – For RCTs the main issue is attrition (post-assignment)
    • If attrition is “too high” treat like a QED
  – For QEDs the main issue is comparability of the analytic sample
WWC Quality Improvement

• Part of the reason that most studies fail to meet WWC standards is “self-inflicted”
• Here’s an example:
• QEDs must allow for an assessment of group balance on protocol-specified covariates
  – Usually just a pretest of the outcome, but sometimes more (postsecondary usually a measure of prior achievement and a measure of SES)
Group Balance in QEDs

- Balance: $d < .05$
- Moderate imbalance: $.05 < d < .25$
- Severe imbalance: $d \geq .25$

- If groups are balanced on covariate(s), nothing needs to be done
- If groups exhibit severe imbalance, study does not meet standards
- If groups exhibit moderate imbalance, study authors must have controlled for imbalance using a WWC approved method
  - Not approved are student fixed effects, matched t-test, and mixed ANOVA
  - The WWC does not ask for new analyses, so if the study author failed to use an approved method the study will not meet standards
Current Quality Improvement Efforts

• Re-examine “approved” adjustment methods
• Re-examine balance standards
  – .05 is clearly too low in my opinion (If IQ is 100.0 and 100.76, groups are moderately imbalanced)
  – .25 is probably too low too

• Re-examine definition of “substantively important effect size”
• Update RD standards
Broader Use of the WWC’s Work by Educational Researchers

• Practice guides and intervention reports have associated literature searches
  – The searches tend to be very well done

• The searches could be the base of a systematic review using different methods, different specification of the review question, etc.

• Intervention Reports and Practice Guides point out areas in need of rigorous research
Broader Use: Example

• WWC Postsecondary Topic Area: Interventions for Students in Developmental Education
• Literature search identified 6,000 + hits
• These were screened (about 5200 screened out)
• Then placed into buckets (about 25 different categories)
• One category was “placement into developmental education”
Placement into Developmental Education

• This is a “policy”, so in theory could be a WWC product
• But, only one (very small) randomized experiment (using an outcome that is not very interesting)
• Most of the rest attempt to use OLS to estimate effects
• The WWC is not set up to review these types of studies
• Valentine, Goldrick-Rab, and Konstantopoulos undertaking this review
Effects of Developmental Education Review

• About 25 between groups studies
• These studies tend have significant problems
  – Compare dev ed students to non-dev ed students
  – Control baseline achievement less than half the time
• They suggest large, negative effects associated with placement into dev ed
  – At least part of this is an illusion
    • Studies that control baseline achievement suggest smaller, but still negative effects
Effects of Developmental Education Review

• We also uncovered 12 RD studies
  – One of the first meta-analyses of RD

• Fairly different cut points for placement
  – Suggests something interesting about the lack of agreement on what constitutes “college ready”

• RD studies certainly do not suggest that placement into developmental education is helpful
  – On balance probably harmful

• Stay tuned for details (submit around the end of 2013)