Paper #2 of 3

Bayesian interpretation of impact estimates: Conceptual Issues, Myth Busting, and Frequently Asked Questions

We have proposed a conceptual framework for interpreting findings from impact evaluations that we call BASIE (BAyeSian Interpretation of Estimates; Deke and Finucane 2019). We developed this framework specifically for the context of impact evaluations of social policy (including education) interventions, where both objectivity and policy-relevance are highly valued. We believe policymakers want an objective, evidence-based assessment of the probability that an intervention worked. Our aim is to provide such an assessment using a "best of both worlds" approach that avoids the pitfalls and controversies of both the purely *frequentist* and *Bayesian* approaches.

BASIE uses Bayes Rule to combine prior evidence with new evidence to assess the probability that an intervention has positive (or meaningful) effects. There is nothing in this framework that is mathematically novel. Instead, the novelty of BASIE is in the combination of ideas that we reject and accept. With BASIE we **reject** both null hypothesis significance testing (often associated with the word *frequentist*) and a subjective definition of probability (often associated with the word *Bayesian*). We **accept** an objective definition of probability based on relative frequencies (often associated with the word *frequentist*) and the use of prior evidence and Bayes Rule (often associated with the word *Bayesian*).

In this talk we will discuss some key conceptual issues that are helpful to understand when applying BASIE, dispel common misconceptions, and answer frequently asked questions. We hope this will prove useful to education researchers who share our goal of providing policymakers with an objective answer to the question – *what's the probability this intervention really worked?*

Core Issues. We will discuss two core issues: (1) the definition of probability and (2) exchangeability.

(1) The definition of probability may seem like an esoteric, philosophical issue of little interest to applied researchers. However, this issue has important implications for how research is conducted and findings interpreted. We define probability in terms of relative frequency because (1) it forces us to be specific and concrete in explaining what a probability does and does not mean, (2) it greatly limits the extent to which personal beliefs (and biases) can influence the findings from an evaluation, and (3) psychologists have shown that people have a more accurate understanding of probability when communicated in terms of frequencies. We will discuss these issues in our presentation

(2) Attempting to eclipse probability in apparent philosophical esotericism is the concept of exchangeability. This concept (which is similar to the concept of independent and identically distributed random variables) comes into play when we are selecting the prior evidence that we will use to help interpret findings from a new evaluation. Findings from the new evaluation will be *exchangeable* with the prior evidence base if we have no evidence that findings from the new evaluation would be anything other than randomly different from findings in the prior evidence base. In our presentation, we will provide

examples of things that are and are not *exchangeable*. We will also explain how defining probability in terms of evidence, rather than belief, can make it easier for researchers to assess exchangeability. Finally, we will suggest what researchers can do when exchangeability is either not satisfied or technically satisfied in an intuitively unsatisfying way.

Misconceptions. We will discuss three misconceptions: (1) some subjectivity cannot be avoided so all subjectivity must be ok, (2) priors can only be based on evidence that is "highly relevant," and (3) Bayesian interpretation is costly.

(1) A subjective definition of probability is sometimes justified by pointing to the many subjective design and analysis decisions that researchers make when conducting evaluations. Given all of that subjectivity, why not also define probability and priors subjectively? Our answer is that a priori subjectivity with respect to research methodology generally does not favor one policy or program over another, but injecting subjectivity into priors can. We provide examples.

(2) Researchers new to Bayesian interpretation sometimes believe that prior evidence must be highly relevant to a new study. An extreme version is that only replication studies can benefit from prior evidence – that is, prior evidence can only be used if it's from studies of the effect of the same intervention, on the same outcome, for the same population as the new study. We argue that prior evidence need not be so highly relevant so long as we are clear about the correct interpretation of probability statements, given that prior evidence. We also point out that using less relevant prior evidence will almost always be preferable to the alternative of misinterpreting *p*-values and statistical significance.

(3) After synthesizing the prior evidence into a normal prior probability distribution function (known as a conjugate prior), we can provide a Bayesian interpretation of estimates using a simple formula. This calculation can be implemented using a formula in a spreadsheet. Though more complex approaches exist, this simple approach adds considerable value relative to the status quo of misinterpreting p-values.

Frequently asked questions. We will discuss 3 frequently asked questions: (1) What do I do when studying a totally new, unique intervention with no prior evidence? (2) How do you account for the fact that the prior evidence is noisy or biased? (3) How is BASIE different from meta-analysis? Short answers are provided below, more detail will be provided in our presentation.

(1) Conduct a larger study and examine sensitivity to a range of plausible priors. With a larger study, findings will be less sensitive since the study's data will receive relatively more weight, compared to the prior, than in a smaller study.

(2) When synthesizing prior evidence into a probability distribution, we recommend employing meta-analysis to give greater weight to more precise estimates. We also suggest an adjustment to account for bias arising from the so-called "file drawer problem."

(3) BASIE is a complement to meta-analysis, not a replacement. A meta-analysis can be used to synthesize the prior evidence into a prior distribution that can be combined, using Bayes Rule, with findings from a new study.