IES Institute of Education Sciences

Toolkit

The BASIE (BAyeSian Interpretation of Estimates) framework for interpreting findings from impact evaluations: A practical guide for education researchers

NCEE 2022-005
U.S. DEPARTMENT OF EDUCATION

A Publication of the National Center for Education Evaluation and Regional Assistance

U.S. Department of Education
Miguel Cardona
Secretary

Institute of Education Sciences
Mark Schneider
Director

National Center for Education Evaluation and Regional Assistance
Matthew Soldner
Commissioner

Thomas Wei
Amy Johnson
Project Officers

The Institute of Education Sciences (IES) is the independent, non-partisan statistics, research, and evaluation arm of the U.S. Department of Education. The IES mission is to provide scientific evidence on which to ground education practice and policy and to share this information in formats that are useful and accessible to educators, parents, policymakers, researchers, and the public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other IES product or report, we would like to hear from you. Please direct your comments to ncee.feedback@ed.gov.

This report was prepared for the Institute of Education Sciences (IES) under Contract 91990020F0052 by Mathematica. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

April 2022

Deke, J., Finucane, M. & Thal, D. (2022). The BASIE (BAyeSian Interpretation of Estimates) framework for interpreting findings from impact evaluations: A practical guide for education researchers. (NCEE 2022-005). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee.

This report is available on the Institute of Education Sciences website at http://ies.ed.gov/ncee.

The BASIE (BAyeSian Interpretation of Estimates) framework for interpreting findings from impact evaluations: A practical guide for education researchers

John Deke
Mariel Finucane
Daniel Thal
Mathematica

NCEE 2022-005
U.S. DEPARTMENT OF EDUCATION

Contents

The BASIE (BAyeSian Interpretation of Estimates) framework for interpreting findings from impact evaluations: A practical guide for education researchers

BASIE is a framework for interpreting impact estimates from evaluations. It is an alternative to null hypothesis significance testing. This guide walks researchers through the key steps of applying BASIE, including selecting prior evidence, reporting impact estimates, interpreting impact estimates, and conducting sensitivity analyses. The guide also provides conceptual and technical details for evaluation methodologists.

Introduction

Readers of impact evaluation reports want to know whether the evaluated intervention improved outcomes—that is, did the thing work? Researchers cannot provide a simple 'yes/no' answer because all impact estimates are subject to statistical errors, leading to uncertainty about whether an intervention worked. Researchers have often used statistical significance and p-values to assess uncertainty resulting from these statistical errors. However, statistical significance and p-values are often misinterpreted (Greenland et al., 2016; Wasserstein & Lazar, 2016). [1] In particular, people often misinterpret statistical significance and p-values to be measures of the probability that an intervention had a meaningful effect, given an impact estimate.

We developed the BAyeSian Interpretation of Estimates (BASIE) as a way to calculate the probability an intervention had a meaningful effect, given the impact estimate and prior evidence regarding the effects of broadly similar interventions (see Deke & Finucane, 2019 for more detail on the rationale for, and conceptual underpinning of, BASIE). BASIE is designed for use in the field of program evaluation, where concepts like transparency and impartiality are highly valued (BASIE is heavily influenced by Gelman, 2011, 2015, 2016; Gelman & Hennig, 2007; and Gelman & Shalizi, 2013). The key steps to apply BASIE are summarized in Exhibit 1.

[1] A report about citing significance in the context of National Center for Education Statistics statistical data reporting is available at https://www.niss.org/research/citing-significance-nces-data-reporting.

Exhibit 1. Summary of key steps to applying BASIE

Key step		Summary
1.	Select prior evidence	Select a distribution of prior intervention effects. The prior distribution is used to calculate a shrunken impact estimate (step 2) and for interpretation (steps 3 and 4). We provide in this guide 108 prior distributions[2] that researchers can use without conducting their own meta-regressions. We also provide a spreadsheet tool for using these prior distributions.
2.	Report impact estimates	Report either the traditional impact estimate (based only on study data) or the shrunken impact estimate (based on both study data and prior evidence) as the study's main impact estimate. Prespecify which of these two estimates will be presented as the study's primary estimate. The shrunken estimate can be calculated using the spreadsheet tool.
3.	Interpret the impact estimates	Interpret impact estimates using Bayesian posterior probabilities (or credible intervals). Posterior probabilities can be calculated using the spreadsheet tool.
4.	Sensitivity analysis	Report sensitivity of shrunken impact estimates and posterior probabilities to which prior distribution is selected in step 1. These sensitivity analyses can be conducted using the spreadsheet tool. Also report whichever impact estimate (traditional or shrunken) was not reported in step 2.

BASIE= BAyeSian Interpretation of Estimates

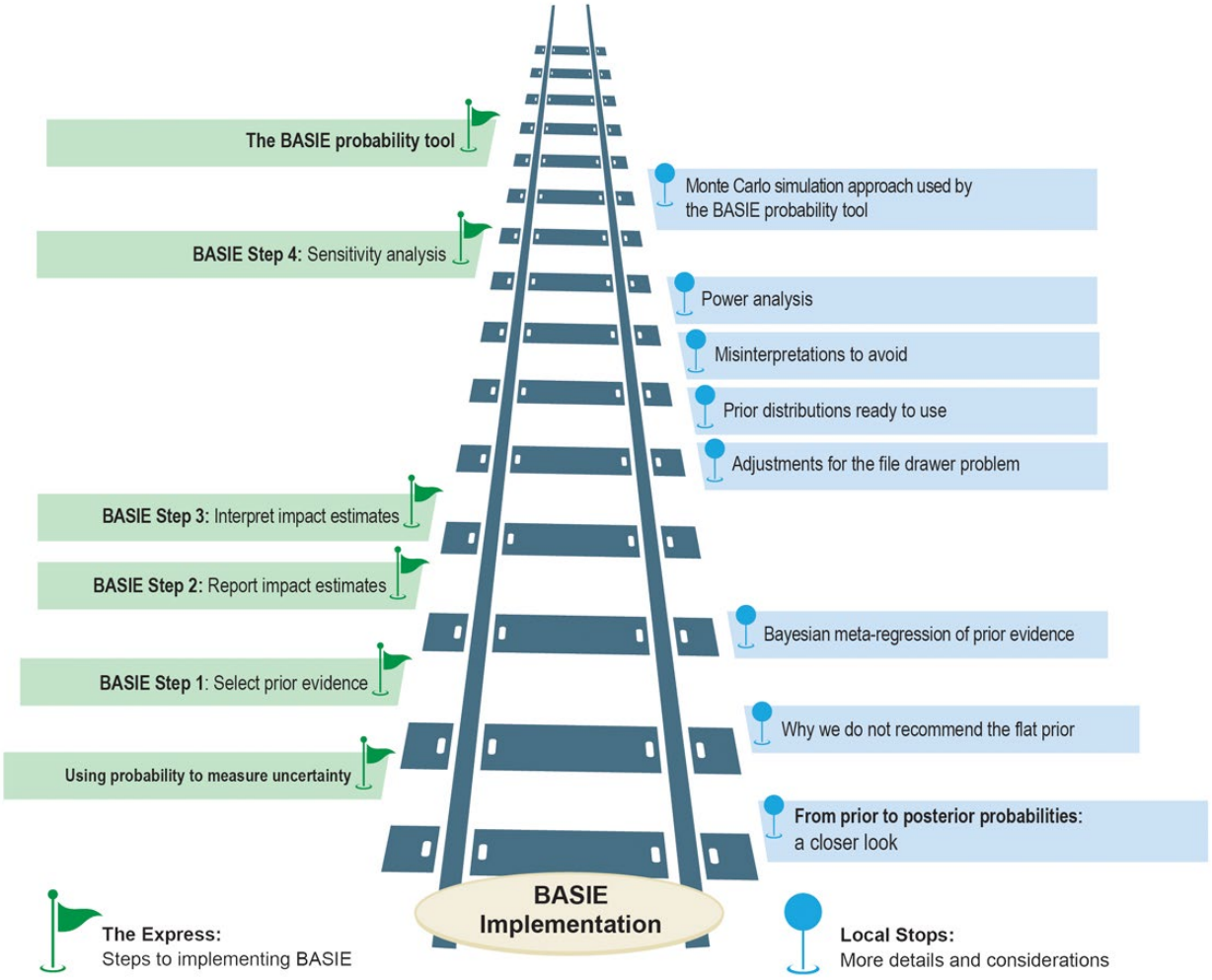
This guide walks through each of these steps. It explains, for a broad audience of applied researchers conducting evaluations of education programs (or interpreting findings from those evaluations), how to use BASIE to better interpret impact estimates. The guide also provides conceptual and technical details for evaluation methodologists. Accompanying this guide on the IES website are supplementary files including computer programs and a spreadsheet tool that can be used to calculate posterior probabilities.

[2] These prior distributions are based on the posterior distributions from a Bayesian meta-analysis described later in this guide.

Choose your own adventure

We offer readers two pathways—the "Express" and "Local Stops" (Exhibit 2). The Express describes the steps to implementing a BASIE analysis for readers who want to start using BASIE right away. Local Stops provides more in-depth information about our tools and methods, including recommendations for how to conduct your own analyses of prior evidence. This guide includes internal links that will enable the reader to move between the Express and Local Stops. We also include an appendix with even more technical details.

Exhibit 2. Pathways for readers of this guide



Using probability to measure uncertainty

Probability is the key tool we need to assess uncertainty. By looking across multiple events, we can calculate what fraction of events had different types of outcomes and use that information to make better decisions. This fraction is an estimate of probability called a relative frequency. For example, a school superintendent seeking to improve math test scores can examine what fraction of different intervention types (tutoring programs, supplementary curricula, education technology, etc.) have improved math test scores in the past. Those probability estimates can be used to guide decisions about which types of interventions to invest in, based on which have shown higher rates of success.

In this section, we describe different probability distributions that we can calculate to help interpret education research findings. A probability distribution is an equation or table that describes the probability (or relative frequency) that a range of different events occur.

Three probability distributions that can help us interpret research evidence

Three distributions that can help us interpret evidence are the prior distribution, the distribution of the impact estimate (a.k.a., the likelihood), and the posterior distribution. The posterior distribution, which can only be calculated if we know the other two distributions, is the one that is most useful for interpreting evidence.

The prior distribution. In general, the prior distribution represents all previously available information regarding a parameter of interest. In the present context, the prior distribution describes how common it is for education interventions to have true (not estimated) effects of varying sizes. Under BASIE, this distribution is based on findings from previously completed studies. If the prior evidence shows that it is unusual for interventions to have large favorable effects, then we would infer that a very large impact from a new study is less likely. By contrast, the more common large effects have been in the past, the more probable it is that a sizeable impact estimate using data from a new study is the result of a true effect rather than random chance. We can calculate this type of probability distribution by analyzing prior evidence of education effects from the What Works Clearinghouse (WWC). We draw on the WWC because it is, to the best of our knowledge, the largest standards-based systematic review of evaluation evidence in the field of education. With the prior distribution in hand, we can calculate the proportion of intervention effects that are positive, negative, greater/less than a specified value, or that fall within a specified range of interest.

The distribution of the impact estimate. We can use this distribution, often referred to as *the likelihood*[3], to calculate the probability of observing large but random differences between a treatment and control group. In a randomized experiment, we do not expect there to be any systematic differences in the characteristics of the treatment and control groups. However, there can be chance differences in the characteristics of study participants between the treatment and control groups arising from the random assignment mechanism, and those characteristics might be related to outcomes. These chance differences mean that even when the true effect of an intervention is zero, the impact estimate will almost never be literally zero. If we were to repeat the experiment over and over, we would see a different impact estimate each time—sometimes positive, sometimes negative, generally small, but occasionally large.

To calculate this probability, we need to know how much the impact estimate would vary across replications of the experiment. For a study of a particular intervention, the main thing that affects

[3] In general, the likelihood describes the probability of observing the data as a function of model parameters.

how much an impact estimate varies is the sample size of the study—bigger studies have more precise (less variable) impact estimates.

The posterior distribution. If we combine the information from the previous two probability distributions, then we can learn something more useful. Specifically, we can use Bayes' Rule to calculate the probability that the intervention we are interested in really had a positive effect for study participants given what we observe in the data (our impact estimate and standard error), and how often interventions have had positive effects in the past.

This is called a posterior probability, and it is what we use to interpret impact estimates under BASIE. It is also what people often misinterpret a p-value to be.

For more details on these distributions, including a numerical example of how to calculate a posterior distribution using the prior distribution and the distribution of the impact estimate, visit Local Stop: From prior to posterior probabilities: a closer look.

BASIE Step 1: Select prior evidence

We have prepared a total of 108 prior distributions for use by education researchers based on impact estimates in the WWC database that meet evidence standards. These prior distributions are described in Local Stop: Prior distributions ready to use. We provide considerations for selecting a prior distribution below, but for readers looking for a single 'safe default' prior distribution, we suggest the one that is (1) based on all estimates in the WWC that meet standards (because it is broadly applicable), (2) adjusted for small-study effects (because small-study effects may be due to selective reporting bias, described below under '**Whether we adjust for small-study effects**'; also see Marks-Anglin and Chen, 2020 for an overview), and (3) centered at zero (because many studies involve a contrast between conditions that can all be viewed as 'treatments'). For readers interested in learning more about why we recommend using a prior based on evidence from the WWC (an evidence-based prior) instead of assuming that effect sizes of all magnitudes are equally likely (a prior distribution known as the *flat prior*), visit Local Stop: Why we do not recommend the flat prior. For readers interested in learning more about the Bayesian meta-regression we used to create these distributions (which advanced users could also use to construct their own custom prior distributions), visit Local Stop: Bayesian Meta-regression of Prior Evidence. For readers accustomed to assuming that prior distributions are normal (Gaussian), we note that we use a more flexible functional form – the skewed generalized t-distribution, as described in Local Stop: Prior distributions ready to use.

The 108 distributions we prepared vary along three dimensions:

1. **The populations of effects for which the distribution is estimated.** The populations vary by outcome domain and school level. There are seven outcome domains and five school levels which combine to form 35 populations of intervention effects. Adding a single overall population creates a total of 36 populations. The outcome domains are math achievement, ELA achievement, science achievement, other achievement, behavioral, attainment, and miscellaneous (for more detail about what is included in each outcome domain, visit Local Stop: Prior distributions ready to use). The school levels are pre-kindergarten, elementary, middle, high, and post-secondary. In cases where a study has impact estimates that span multiple school-levels or outcome domains, we recommend conducting analyses to assess sensitivity to which of the applicable prior distributions are used.
2. **Whether we adjust for small-study effects.** The practical consequence of this adjustment is that it reduces the prior probability of large effects. *Small-study effects* refers to the oft-observed phenomenon in meta-analysis that studies with larger standard errors tend to have more favorable impact estimates (see Marks-Anglin and Chen, 2020 for an overview). One possible explanation for this correlation between impact estimate and standard errors is that some researchers may calculate multiple impact estimates but only report the most favorable estimate—a form of reporting bias. Because smaller studies (which have larger standard errors) tend to have more variability across multiple estimates, the magnitude of reporting bias in small studies tends to be larger. In our meta-regression we make an adjustment for small-study effects under the assumption that these effects are due to reporting bias (described in this Local Stop: Adjustments for small-study effects). An alternative explanation for small-study effects is that researchers follow standard guidance (e.g., Cohen, 1969) to conduct larger studies when they anticipate smaller intervention effects. Thus, the prior distributions with an adjustment for small-study effects corresponds to the pessimistic assumption that a correlation between effect size and standard error is due to reporting bias while the prior distributions without an adjustment corresponds to the optimistic assumption that such a correlation is due to researchers following recommended practice in designing their studies. Because we cannot know for sure whether a correlation between effect sizes and standard errors is due to biased

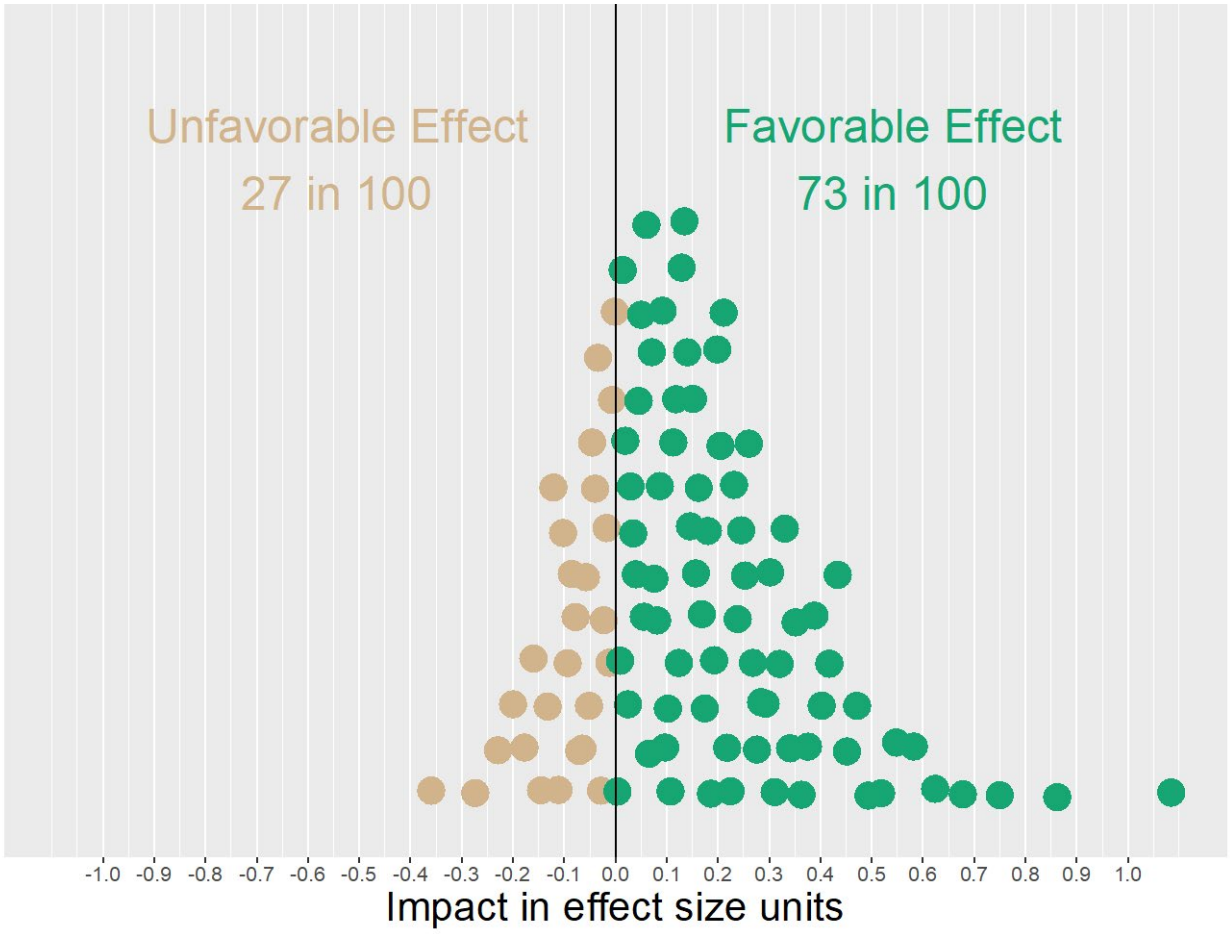
reporting or researchers faithfully following Cohen's guidance (this lack of knowing is an example of *model uncertainty*), we provide prior distributions with and without this adjustment.

3. **Whether the distribution is centered on zero.** The practical consequence of this adjustment is an equal prior probability of positive and negative effects. There are two good reasons to use a zero-centered prior.[4] First, in studies comparing two interventions (for example, two math curricula), a zero-centered prior may be more appropriate, especially if there is no consensus among recognized experts that one intervention is better than the other, for example if experts disagree regarding the best approach to teaching math (known as equipoise in the ethics of clinical research, Freedman, 1987). A zero-centered prior might be less appropriate in cases where the treatment adds substantial resources (either time or money) beyond what is available in a business-as-usual control condition. Second, although our Bayesian meta-regression adjusts for reporting bias associated with small-study effects, this adjustment may not address all possible forms of reporting bias. That is because our adjustment method assumes that researchers calculate multiple estimates and report the one that looks best, even if the best-looking estimate is still unfavorable to the intervention. Instead, researchers whose best-looking estimate is still unfavorable might choose to report nothing at all. Our adjustment procedure cannot account for that type of selection bias. A zero-centered prior provides stronger protection against that type of bias, though this protection comes with a cost—the zero-centered prior may result in an overly pessimistic assessment that the median intervention effect is zero (meaning that half of all interventions have unfavorable effects).

The distribution of effects for the overall population in the WWC is illustrated in Exhibit 3, where green bubbles represent favorable intervention effects and tan bubbles represent unfavorable effects. The overall average effect is 0.16, the median is 0.11, and the standard deviation is 0.29. In Local Stop: Prior distributions ready for use, we provide more detailed information about all 108 prior distributions. The 108 distributions include 36 that are not zero-centered and have an adjustment for small-study effects, 36 that are not zero-centered and do not have an adjustment for small-study effects, and 36 that have an adjustment for small-study effects and are zero-centered (we do not include distributions that are zero-centered but lack an adjustment for small-study effects).

[4] If the intervention is being compared to a control condition that represents the status quo, and the decision informed by the evaluation is to either adopt the intervention or maintain the status quo, then using a zero-centered prior has the effect of erring on the side of conserving the status quo. Whether this is a good reason to use a zero-centered prior depends on the perceived desirability of the status quo, which may vary across contexts and perspectives.

Exhibit 3. The overall distribution of WWC intervention effects



Mean	Standard Deviation	Percentiles 10th	25th	50th	75th	90th
0.16	0.29	-0.11	-0.01	0.12	0.28	0.47

Note: This figure is a discretized representation of the distribution of intervention effects. The distribution is represented by 100 dots corresponding to 100 quantiles (specifically, the 0.5th through 99.5th quantiles) of the prior distribution based on a meta-regression of all WWC studies that met standards. The dots are randomly scattered vertically so that they are visible. The density of the distribution is represented by the number of dots within a range. For example, there are 27 dots to the left of zero on the x-axis, meaning that 27 percent of the distribution is negative.

When using the BASIE probability tool researchers can select which prior distribution to use by specifying the outcome domain, school level, whether to adjust for small-study effects, and whether to use a zero-centered prior.

BASIE Step 2: Report impact estimates

There are two point estimates that should be reported—the traditional estimate and the shrunken estimate. Researchers must choose which option to use as the primary point estimate and should report the other as a sensitivity analysis.

The traditional estimate is based only on study data, not prior evidence. This traditional estimate is familiar to most researchers, representing, for example, the difference in outcomes between the treatment and control groups. We recommend always reporting this estimate to be transparent about what is observed in the study data and to facilitate future meta-analysis (Gelman, 2005).

The shrunken estimate incorporates the prior distribution (the same one selected under Step 1) into the point estimate for the new study. This estimate is essentially a weighted average of the traditional estimate and the mean of the prior distribution. Though the word "shrunken" might make it sound like this will be a smaller estimate, what it actually means is that the impact estimate based on study data is pulled towards the mean of the prior distribution. So, if the mean of the prior distribution is larger than the impact estimate based on study data, then the shrunken estimate will actually be larger than the estimate based on study data. We recommend always reporting this estimate because it is less susceptible to statistical noise, especially in small studies.

We do not provide a recommendation for which impact estimate to report as the primary representation of an intervention's effect. Instead, we recommend researchers pre-specify either the traditional or the shrunken estimate as their primary point estimate and report the other as a sensitivity analysis. Two considerations in selecting which point estimate to lead with are:

1. **The needs of your audience.** Consumers of research may differ in what they need or expect from an evaluation report. For example, an evaluation that randomizes schools within school districts might report district-specific estimates. Should those estimates be shrunken or based only on data from that district? If the primary audience for those estimates are decision makers in those districts, and if the districts are very different from one another, then the traditional estimate based only on data from the district may be preferable. On the other hand, if the primary audience is a state or federal level decision maker, then shrunken estimates may be preferable.
2. **Alignment with the prior distribution.** If the impact estimate at hand cannot be regarded as a member of the population of evidence represented by the prior distribution, then the estimate based on study data is preferable. For example, the WWC uses review protocols to specify the criteria for which evidence will be reviewed in a topic area. Not all evidence produced in the field of education will necessarily fall under the purview of a WWC topic area; some estimated impacts of interest in education may not, strictly speaking, be members of the same population of evidence contained in the WWC. Otherwise, the shrunken estimate may be preferable.

BASIE Step 3: Interpret impact estimates

The core of BASIE is interpretation of impact estimates using posterior probabilities instead of statistical significance and credible intervals rather than confidence intervals. These types of probabilities and intervals can be calculated using the spreadsheet-based BASIE probability tool. This tool calculates probabilities to support statements like:

- We estimate a 90 percent probability that the intervention had a positive effect on reading test scores.
- We estimate a 75 percent probability that the intervention increased reading test scores by at least 0.15 standard deviations.

So that readers fully understand the meaning of these probability statements, we recommend explaining the correct interpretation somewhere in a report or journal article (for example a methods section or an appendix). In particular, researchers should explain to readers that these statements can (1) only be interpreted relative to the selected prior distribution and (2) are not predictive statements about the effects in the future, but instead of retrospective statements about the effect of an intervention in the evaluation context (see Local Stop: Misinterpretations to avoid). To help clarify the meaning, a longer version of the simpler probability statements could be provided that references the prior distribution and uses past tense to emphasize that the probability is not predictive. For example:

> *We estimate a 75 percent probability that the intervention increased reading test scores by at least 0.15 standard deviations,* ***given our estimates and*** **prior evidence on the impacts of reading programs for elementary school students.**

This longer form of the probability statement need not be used throughout the report. It is sufficient to provide it once for clarification.

We also recommend reporting several of these probability statements spanning the range of plausible intervention effects so that readers have a complete understanding of the posterior distribution. For example, report the probability the effect size of the impact is less than -0.20, less than -0.10, less than -0.05, less than 0, greater than 0.05, greater than 0.10, and greater than 0.20.

Without a complete reporting of the posterior distribution, readers of evaluation reports may misinterpret impact estimates with different standard errors. That is because with larger standard errors, there is more uncertainty regarding the true effect of the intervention. By focusing on a select few probability statements, some important implications of that greater uncertainty could be obscured.

For example, consider the difference in interpretation between a standard error of 0.05 (calculated in an evaluation of intervention A) and a standard error of 0.20 (calculated in an evaluation of intervention B) when both evaluations report an impact estimate of 0.10 standard deviations (Exhibit 4). The probability that the true effect of intervention A is greater than 0.15 standard deviations is 0.14. The probability that the true effect of intervention B is greater than 0.15 standard deviations is 0.28. If that is all we report, readers might think intervention B is the better bet. (These probabilities are calculated using the 'safe default' prior distribution we recommended in Step 1.) But now let us look at the probability that the intervention effects are negative. For intervention A that probability is just 0.03, but for intervention B it is ten times larger – it is 0.34. Clearly, Intervention A is a much safer bet than Intervention B (i.e., less likely to cause harm). This is why it is important to provide readers of evaluation reports with a complete understanding of the posterior distribution—we can come to different conclusions about which intervention is preferable depending on our cutoff of interest, and cannot necessarily use performance for one cutoff to determine performance at another.

Exhibit 4. Report enough posterior probabilities to avoid misleading readers

	Intervention A	Intervention B
Impact estimate (based only on study data)	0.10	0.10
Standard error	0.05	0.20
Probability that the true effect size is:		
Greater than 0.15 standard deviations	0.14	0.28
Less than 0	0.03	0.34

Note: This is a hypothetical example involving fictitious interventions.

Finally, we recommend pre-specifying the smallest impact that would be regarded as meaningful and always reporting the probability that the intervention's effect exceeds that minimum meaningful effect size. Selecting the minimum meaningful effect size involves the same considerations used when conducting statistical power analysis. Hill et al. (2008) and Lipsey et al. (2012) suggest multiple substantive benchmarks for assessing what a meaningful impact would be for a given intervention and context.

Cutoffs and characterizations

An advantage of reporting posterior probabilities is that they provide a more nuanced understanding of an intervention's likely effects than statistical significance. Yet in some contexts, cutoffs on those probabilities may be needed to provide a narrative characterization or to choose which findings to highlight. If probability cutoffs or characterizations will be used to interpret findings, we recommend researchers pre-specify those cutoffs or characterizations. Researchers should choose the cutoffs and characterizations that best fit their context, but for illustrative purposes we provide an example in Exhibit 5 of how probabilities are characterized in the field of risk management (Garvey, 2001).

Exhibit 5. Characterization of probabilities

Probability Range	Characterization
0 to 5 percent	Extremely sure not to have occurred
5 to 15 percent	Almost sure not to have occurred
15 to 25 percent	Not likely to have occurred
25 to 35 percent	Not very likely to have occurred
35 to 45 percent	Somewhat less than an even chance to have occurred
45 to 55 percent	An even chance to have occurred
55 to 65 percent	Somewhat greater than an even chance to have occurred
65 to 75 percent	Likely to have occurred
75 to 85 percent	Very likely to have occurred
85 to 95 percent	Almost sure to have occurred
95 to 100 percent	Extremely sure to have occurred

Source: Garvey (2001).

Credible intervals

Closely related to posterior probabilities, credible intervals are another way to understand the likely effect of an intervention. A credible interval tells us the probability that an intervention effect falls within some specified interval of interest. One practical application of credible intervals is to assess the likelihood that there is no meaningful difference between two groups (in other words, the probability that the groups are practically equivalent). With credible intervals, we can make statements like:

- We estimate a 70 percent probability that the effect of the intervention is within +/- 0.05 standard deviations of 0.
- We estimate an 80 percent probability that the effects of interventions A and B were within 0.05 standard deviations of each other.

BASIE Step 4: Sensitivity analysis

In all statistical analyses, we can never be entirely sure about the best model to use—this is the problem of *model uncertainty*.[5] Pre-specified sensitivity analyses are often used in evaluations to assess how findings might change if alternative (but valid) modeling choices were made to calculate and interpret impact estimates. For example, researchers might report how p-values change depending on which covariates are included in a regression model or the approach used to adjust standard errors for clustering.

For BASIE, we recommend that researchers pre-specify sensitivity analysis to determine how posterior probabilities vary across a range of prior distributions. Findings that are consistent across this range can be used with increased confidence to guide decision making. The 108 prior distributions provided in this report provide a wide range of evidence-based priors that can be used for sensitivity analysis. For example, a researcher interpreting an impact estimate on math test scores that was calculated for a combined sample of 5th, 6th, and 7th graders has to decide whether to use a prior distribution for the elementary or middle school level. They could use a prior distribution for middle school as their benchmark (since two of the three grades in the sample are middle school) and then use the prior distribution for elementary as a sensitivity analysis.

While different contexts may call for different sensitivity analyses, we recommend all researchers conduct the following three sensitivity analyses:

1. **Use the zero-centered prior, with a correction for small-study effects, using the overall population of intervention effects in the WWC.** That distribution is broadly applicable and sets a high bar for concluding that an intervention is likely to have meaningful effects. It can also provide a common interpretive benchmark for readers comparing probabilities across different studies. The only exception to using this prior distribution as a sensitivity analysis is if it was the distribution selected in step 1 discussed above.
2. **Assess sensitivity to making an adjustment for small-study effects.** Since we can never know for sure whether a correlation between effect sizes and standard errors is due to biased reporting or due to researchers conducting larger studies when they expect small effects, it will almost always be useful to examine sensitivity to this adjustment.
3. **Report whichever point estimate was not reported in step 2.** If the traditional impact point estimate was reported in step 2, report the shrunken as a sensitivity analysis.

A challenge facing all sensitivity analyses is interpretation: what constitutes high or low sensitivity? We recommend that researchers consider this question in the context of their study and the different ways in which findings from their study might be used. Three factors to consider:

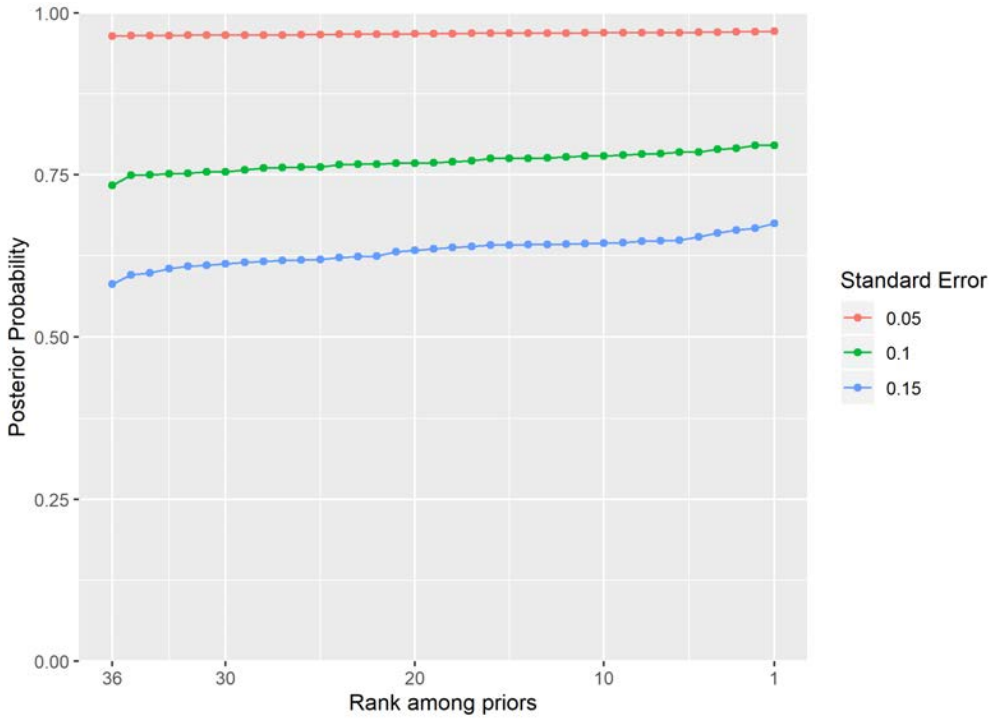
1. **The stakes.** The higher the stakes—for example, the greater the cost of the intervention, the more vulnerable the population being studied, the more people who may be affected by decisions made using the study findings—the lower our likely tolerance for sensitivity to the prior distribution.
2. **Decision cutoffs.** Sometimes decisions depend on key cutoffs. The closer we are to a key cutoff, the lower our tolerance for sensitivity. For example, the decision to replace an existing tutoring program with a new tutoring program might depend on whether an evaluation found at least a 50 percent chance that the new program boosted test scores by at least 0.05 standard deviations. If the posterior probability that the new program boosted test scores by at least 0.05 standard deviations is very close to 50 percent, then our tolerance for sensitivity to the prior distribution is likely to be low.

[5] See Kaplan (2021) for a discussion of quantifying model uncertainty in Bayesian analyses.

3. **The takeaways.** Would main takeaways from your study be different if a different prior distribution were used? If so, then sensitivity can be considered high.

When designing a new study, we suggest researchers calculate sensitivity of possible posterior probability statements to prior distributions as part of their power analysis (see the end of the next section and also Local Stop: Power analysis). If sensitivity appears too high, it can be reduced by increasing the size of the study. In Exhibit 6 we illustrate how sensitivity to the choice of prior distribution varies with the size of the study (as represented by the anticipated standard error). In this figure we plot the probability that an intervention had an impact of at least 0.15 standard deviations, given an impact estimate of 0.25 and three different standard errors (0.05, 0.10, and 0.15). We use 36 different priors, varying with respect to outcome domain and school level. These prior distributions are not centered at zero, but they are adjusted for small-study effects. The figure shows that studies with more precise estimates (a standard error of 0.05) will experience less sensitivity to the prior distribution.

Exhibit 6. Sensitivity of a posterior probability to different prior distributions



Note: The posterior probability that the true effect is greater than 0.15 standard deviations is calculated assuming an impact estimate of effect size 0.25, a standard error of 0.05, 0.10, or 0.15, and each of the 36 prior distributions calculated with an adjustment for small-study effects. These prior distributions are not centered at zero. The priors are ordered by the rank of the posterior probability calculated using each prior.

The BASIE probability tool

In this section, we describe a spreadsheet-based tool with prior distributions based on a meta-regression of findings from the WWC. This tool can be used to calculate probabilities of interest, such as the probability an intervention had an effect greater than a level considered meaningful. For methodological details on the probability tool, see Local Stop: Monte Carlo simulation approach used by the BASIE probability tool. We illustrate how the tool can be used with a hypothetical example.

Example setup

For this example, we start with traditional (that is, not shrunken) impact estimates and standard errors from a hypothetical evaluation of an intervention called Crank It Out!, which is focused on improving homework completion among middle school students. The program is hypothesized to ultimately improve both math and reading test scores. The fictitious evaluation compared outcomes for students randomly assigned to participate in Crank It Out! to outcomes for students assigned to a business-as-usual (BAU) control group. The traditional impact estimates and standard errors are reported in Exhibit 7.

Exhibit 7. Estimated impacts of Crank It Out! on math and reading test scores

Outcome	Impact Estimate (Effect Size)	Standard Error
Math test scores	0.18	0.10
Reading test scores	0.08	0.11

Note: This is a hypothetical example involving a fictitious intervention.

Specify prior distribution in the BASIE probability tool

The BASIE probability tool has three drop-down menus that jointly specify the prior distribution (Exhibit 8).

- Under 'Model', the user can select 'Main,' 'Mirrored,' or 'No small-study effects adjustment.' The 'Main' model includes an adjustment for small-study effects and is not centered at zero. The 'Mirrored' model is centered at zero and includes an adjustment for small-study effects. The 'No small-study effects adjustment' model is not centered at zero and does not include an adjustment for small-study effects.
- Under School level, the user can select 'All,' 'PK,' 'Elementary,' 'Middle,' 'High,' or 'Post-secondary.'
- Under Outcome domain, the user can select 'All,' 'Math achievement,' 'ELA achievement,' 'Science achievement,' 'Other achievement,' 'Behavioral,' 'Attainment,' or 'Misc.'

Exhibit 8. Screenshot of prior selection drop-down menus

	A	B	C
1	**Prior distribution selection drop-down menus**		
2	**Model**	**School level**	**Outcome Domain**
3	Main	Middle	Math achievement

For this example, we will select 'Main,' 'Middle,' and 'Math achievement' to interpret the estimated impact on math test scores and 'Main,' 'Middle," and 'ELA achievement' to interpret the estimated impact on reading test scores. Automatically calculated probabilities from the prior distribution are displayed in the spreadsheet (Exhibit 9). This represents our best estimate of the distribution of effects in the absence of an impact estimate for the intervention we are studying (that is, what we knew about the distribution of intervention effects before we did the study).

Exhibit 9. Screenshot of prior probabilities

14	Prior probabilities		
15	**q**	**P(impact > q)**	**P(impact < q)**
16	**-0.5**	1.00	0.00
17	**-0.4**	0.99	0.01
18	**-0.3**	0.98	0.02
19	**-0.2**	0.96	0.04
20	**-0.15**	0.93	0.07
21	**-0.1**	0.89	0.11
22	**-0.05**	0.82	0.18
23	**0**	0.72	0.28
24	**0.05**	0.64	0.36
25	**0.1**	0.55	0.45
26	**0.15**	0.45	0.55
27	**0.2**	0.37	0.63
28	**0.3**	0.24	0.76
29	**0.4**	0.15	0.85
30	**0.5**	0.09	0.91

Input traditional impact estimate

To calculate posterior probabilities, the user needs to enter into the BASIE probability tool the traditional impact estimate (not the shrunken estimate), standard error, and degrees of freedom (typically the number of randomized units minus the number of parameters estimated in a regression model) (Exhibit 10).

Exhibit 10. Screenshot of estimates entered by the user

5	Input traditional impact estimates		
6	**impact estimate**	**stndard error**	**degrees of freedom**
7	0.18	0.1	30

Specify simulation precision

The BASIE probability tool uses a simulation to calculate posterior probabilities. The user can choose the tradeoff between precision and computational cost (Exhibit 11). There are three options:

- 'minimize file size'—this option minimizes the size of the spreadsheet but the reported probabilities are not reliable (that is, the precision of the reported probabilities is *very* low)
- 'low precision'—this option yields fast calculation but probabilities that will tend to 'bounce around' upon recalculation (usually within +/- 3 percentage points)
- 'high precision'—this option will take more time for calculations to finish, but probabilities will 'bounce around' less upon recalculation (usually within +/- 1 percentage point)

Exhibit 11. Screenshot of simulation control parameters

9	**Simulation precision**	**Press F9 to re-run simulation**
10	low precision	

Shrunken estimate, credible intervals, and posterior probabilities

Given all the inputs described above, the BASIE probability tool will report the mean and standard deviation of the posterior distribution along with several predefined credible intervals and posterior probabilities (Exhibit 12). The mean of the posterior distribution is also known as the shrunken impact estimate. The results in Exhibit 12 correspond to the estimated impact on math test scores from Exhibit 7. In this case, the shrunken estimate is 0.16 (compared to the traditional estimate of 0.18), the probability that the true effect of the intervention is positive is 96 percent, and the probability that the true effect is at least 0.10 is 75 percent (these probabilities are highlighted in yellow). Users can change values of 'q' (q represents an effect size quantity of interest) in the spreadsheet to calculate different probabilities.

When reporting the impact estimate and probabilities, researchers could write:

> We estimate the impact of Crank It Out! on math test scores is 0.18 standard deviations. Given this estimate, there is a 96 percent probability that Crank It Out! really did improve math test scores. Furthermore, there is a 75 percent probability that it improved test scores by at least 0.10 standard deviations.

If researchers prefer to emphasize the shrunken estimate, they could replace '0.18 standard deviations' with '0.16 standard deviations.'

Exhibit 12. Screenshot of shrunken estimate, credible intervals, and posterior probabilities

Credible interval -- asessing equivalence of treatment and control groups				
q1	**q2**	**P(q1 < impact < q2)**		
-0.05	0.05	0.11		
-0.1	0.1	0.25		
Credible interval centered at the posterior mean				
Interval width	**q1**	**q2**	**P(q1 < impact < q2)**	
0.05	0.11	0.21	0.41	
0.1	0.06	0.26	0.71	
Posterior distribution mean and standard deviati				
mean	0.16			
standard deviation	0.09			
Posterior probabilities				
q	**P(impact > q)**	**P(impact < q)**		
-0.5	1.00	0.00		
-0.4	1.00	0.00		
-0.3	1.00	0.00		
-0.2	1.00	0.00		
-0.15	1.00	0.00		
-0.1	1.00	0.00		
-0.05	0.99	0.01		
0	0.96	0.04		
0.05	0.88	0.12		
0.1	0.75	0.25		
0.15	0.55	0.45		
0.2	0.34	0.66		
0.3	0.07	0.93		
0.4	0.01	0.99		
0.5	0.00	1.00		

Findings and sensitivity analyses

Using the BASIE probability tool, we can report the main findings and results of sensitivity analyses for the example evaluation. In this case, we choose to report the traditional impact estimate (not the shrunken) along with four posterior probabilities for our main findings (Exhibit 13).

Exhibit 13. Impacts of Crank It Out! on math and reading test scores

Outcome	Impact Estimate (Effect Size)	Standard Error	Probability that the true effect is: Less than -0.05	Less than 0	Greater than 0.05	Greater than 0.10
Math test scores	0.18	0.10	0.02	0.05	0.87	0.73
Reading test scores	0.08	0.11	0.10	0.21	0.61	0.43

Note: This is a hypothetical example involving a fictitious intervention.

We also examine the sensitivity of these findings to the shrunken impact estimate and to our choice of prior distribution (Exhibit 14). Specifically, we look at two alternative prior distributions for each finding. First, we look at the zero-centered prior based on all impact estimates in the WWC database that met standards ('Model' = 'Mirrored'; 'School level' = 'All; 'Outcome Domain' = 'All'). Second, we look at our original prior distributions but we set 'Model' to 'No small-study effects adjustment.'

Exhibit 14. Results of sensitivity analyses

Prior Distribution	Shrunken Impact Estimate (Effect Size)	Probability that the true effect is: Less than -0.05	Less than 0	Greater than 0.05	Greater than 0.10
Math test scores					
Middle school math, adjusted for small-study effects	0.16	0.02	0.05	0.87	0.73
Middle school math, unadjusted for small-study effects	0.17	0.01	0.03	0.89	0.78
Full WWC, zero-centered and adjusted for small-study effects	0.15	0.02	0.06	0.87	0.73
Reading test scores					
Middle school ELA, adjusted for small-study effects	0.09	0.10	0.21	0.61	0.43
Middle school ELA, unadjusted for small-study effects	0.10	0.06	0.16	0.66	0.48
Full WWC, zero-centered and adjusted for small-study effects	0.07	0.13	0.25	0.56	0.36

Note: This is a hypothetical example involving a fictitious intervention.

Incorporating sensitivity analysis when designing a study

Sensitivity to the choice of prior is lower (that is, findings are more robust) when the standard error is lower, as shown in Exhibit 6. At the design stage of a study, researchers can affect the standard errors of their ultimate impact estimates by altering their data collection plans based on the results of a sensitivity analysis. This type of sensitivity analysis can be conducted using the BASIE probability tool.

For example, we used the BASIE probability tool to assess sensitivity of posterior probabilities to choice of prior distribution for a study design that yields a small (standard error 0.20) and a large (standard error 0.10) sample size, given a hypothetical traditional impact estimate of 0.15 standard deviations. We see in Exhibit 15 that posterior probabilities are much more sensitive to the choice of prior distribution if the study has a small sample. For example, the posterior probability that the

effect is at least 0.05 standard deviation varies from 0.60 to 0.76 (a range of 16 points) in the small study and from 0.79 to 0.86 (a range of 7 points) in the large study.

Exhibit 15. Using sensitivity analysis at the design stage

Prior Distribution	Shrunken Impact Estimate (Effect Size)	Probability that the true effect is: Less than -0.05	Less than 0	Greater than 0.05	Greater than 0.10
Small study – anticipated standard error 0.20					
Middle school math, adjusted for small-study effects	0.13	0.10	0.19	0.69	0.57
Middle school math, unadjusted for small-study effects	0.16	0.07	0.13	0.76	0.64
Full WWC, zero-centered and adjusted for small-study effects	0.09	0.18	0.28	0.60	0.48
Large study – anticipated standard error 0.10					
Middle school math, adjusted for small-study effects	0.14	0.02	0.07	0.82	0.65
Middle school math, unadjusted for small-study effects	0.15	0.01	0.05	0.86	0.70
Full WWC, zero-centered and adjusted for small-study effects	0.13	0.03	0.09	0.79	0.61

Note: This is a hypothetical example involving a fictitious intervention. All calculations assume a traditional impact estimate of 0.15 standard deviations.

Future directions

With BASIE, we can answer the basic question—what is the probability that the intervention worked, given the findings from our study? This represents a substantial improvement over the NHST framework, which could not provide an answer to this basic question.

The answer to this basic question depends on a combination of prior evidence and Bayesian modeling, both of which have the potential to change and improve. As more findings are added to the WWC, the prior distributions provided in this guide can be updated. If more meta-data are recorded by the WWC (for example, more refined definitions of outcome domains or descriptions of intervention core components), more refined prior distributions can be synthesized. We have provided the methodological details in the Local Stops and Appendix to empower other researchers to develop these more refined prior distributions.

Local Stop: From prior to posterior probabilities: a closer look

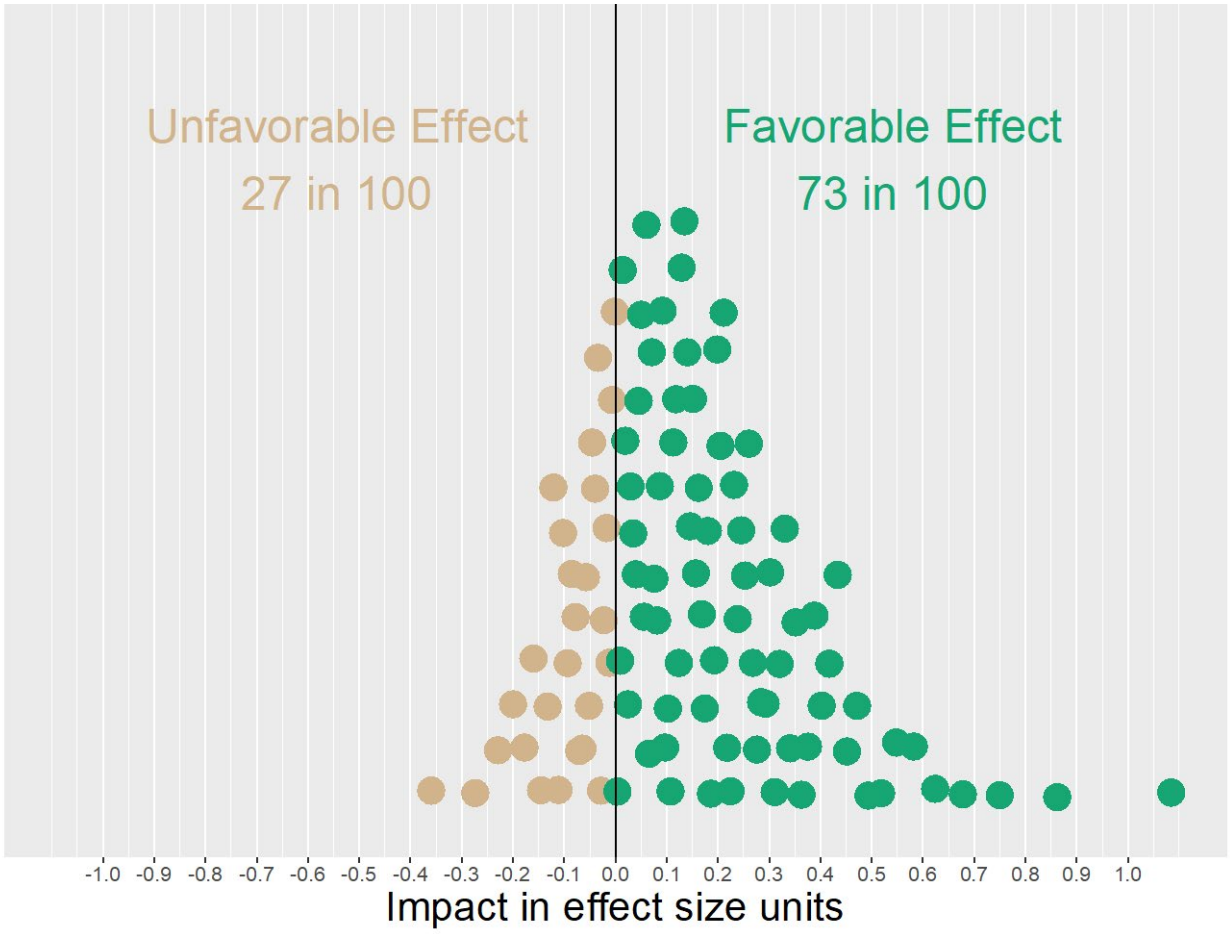
In this Local Stop, we describe how we conceptualize the three probability distributions (prior, impact estimate, posterior). We also use a numerical example, including a computer program written in R, to demonstrate the relationship between the prior distribution, the distribution of the impact estimate, and the posterior distribution.

Prior distribution

We use the superpopulation model (Deming & Stephan, 1941) to conceptualize the probability distribution of prior evidence. For example, we think of the intervention effects included in the WWC database as a random sample from a larger superpopulation of intervention effects. This superpopulation consists of the universe of all studies, and potential studies, that meet WWC criteria for review. We also think of the intervention effect of interest (the one for which we have an impact estimate that we would like to interpret) as being drawn from this same superpopulation of intervention effects.

The prior distribution based on our analysis of all estimates in the WWC database that meet standards (visit Local Stop: Bayesian meta-regression of prior evidence) is illustrated in Exhibit L1. This exhibit is a discretized version of a traditional density plot. We use a discretized representation of the density because Gigerenzer and Hoffrage (1995) argue that Bayesian algorithms are easier to understand when expressed in a frequency format.

Exhibit L1. The overall distribution of WWC intervention effects



Mean	Standard Deviation	Percentiles				
		10th	25th	50th	75th	90th
0.16	0.29	-0.11	-0.01	0.12	0.28	0.47

Note: This figure, which is an intentional duplication of Exhibit 3, is a discretized representation of the distribution of intervention effects. The distribution is represented by 100 dots corresponding to 100 quantiles (specifically, the 0.5th through 99.5th quantiles) of the prior distribution based on a meta-regression of all estimates in the WWC database that meet WWC standards. The dots are randomly scattered vertically so that they are visible. The density of the distribution is represented by the number of dots within a range. For example, there are 27 dots to the left of zero on the x-axis, meaning that 27 percent of the distribution is negative.

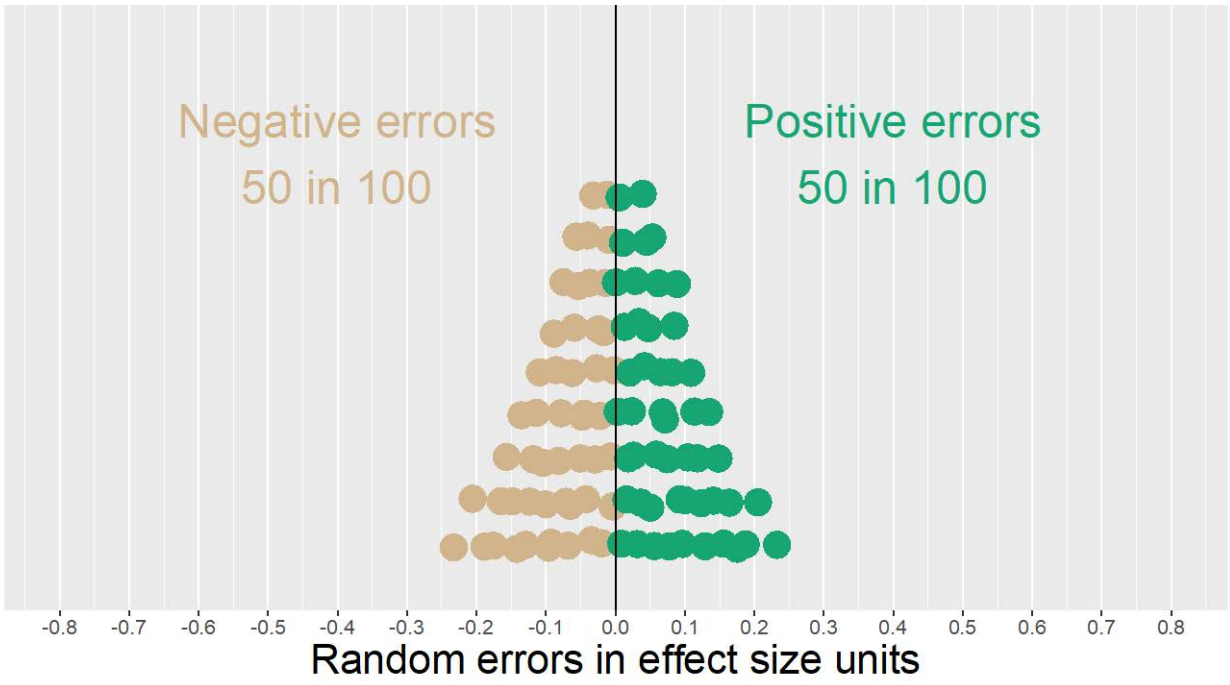
Probability distribution of the impact estimate

The probability distribution of an impact estimate $\hat{\theta}$ given a fixed true effect θ can be expressed as $p\left(\hat{\theta}\,|\,\theta\right)$. If we think of $p\left(\hat{\theta}\,|\,\theta\right)$ as a function of θ (that is, if we think of θ as a variable rather than a fixed value), then we have what is known as the likelihood function (which is the basis for maximum likelihood estimation).

We can use either the superpopulation model or the finite-sample design-based model (Schochet, 2016) to conceptualize the probability distribution of the impact estimate. Under the superpopulation model, we think of variability in the impact estimate as the result of randomly

selecting the treatment and control samples from larger treatment and control populations. Under the finite-sample design-based model, we think of variability in the impact estimate as the result of repeated re-randomizations of the study sample to either the treatment or control group. Under both models, the impact estimate is asymptotically normally distributed and centered at the true intervention effect with standard deviation equal to the standard error of the impact estimate.[6] The distribution of the impact estimate when the true effect is zero is illustrated in Exhibit L2 (deviations from zero are random errors).

Exhibit L2. Distribution of the impact estimate assuming the true effect is zero



Mean	Standard Deviation	Percentiles				
		10th	25th	50th	75th	90th
0	0.10	-0.13	-0.07	0	0.07	0.13

Note: This figure is a discretized representation of the distribution of the impact estimate when the true effect of the intervention is zero. Any deviation in the impact estimate from zero is due to random error arising from chance imbalance between the treatment and control groups. The distribution is represented by 100 dots corresponding to 100 distribution quantiles (specifically, the 0.5th through 99.5th quantiles). The dots are randomly scattered vertically so that they are all visible. The density of the distribution is represented by the number of dots within a range. For example, there are 50 dots to the left of zero on the x-axis, meaning that 50 percent of the distribution is negative. To create this figure, we simulated 100 experimental evaluations of an intervention where the true effect is zero and the standard error of the impact estimate is 0.10.

[6] A standard error calculated using the superpopulation framework reflects uncertainty due to both random sampling and random assignment, while a standard error calculated using the finite-sample framework reflects uncertainty due to random assignment alone. This means that the standard error is theoretically smaller under the finite-sample framework. However, the difference between *feasible estimates* of the finite-sample and super-population standard errors are smaller than in theory (Schochet, 2016). The theoretical implication is that the posterior distribution calculated using the feasible estimate of the finite-sample standard error will give relatively more weight to the prior distribution than if the posterior could be calculated using the true finite-sample standard error. Unless there is considerable impact heterogeneity, this issue is likely small in practice and has no actionable implications for using BASIE.

Posterior distribution

The posterior distribution of an intervention's true effect is a conditional probability distribution that tells us the likely range of true intervention effects that could have led to the specific impact estimate calculated in the evaluation of interest. We need both the prior distribution and the distribution of the impact estimate to calculate this conditional probability.

This probability distribution is calculated using Bayes' Rule. The equation for Bayes' Rule is $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$, where A and B are events, P(A) means "the probability event A occurs," and P(A|B) means "the probability event A occurs given that event B occurred." For example, event A could be "the intervention had an effect greater than zero" and event B could be "we observed an impact estimate of 0.25 standard deviations." We use Bayes Rule to calculate P("the intervention had an effect greater than zero" | "we observed an impact estimate of 0.25 standard deviations").

To illustrate this conditional probability distribution, we simulated 100 impact estimates (each with a standard error of 0.10, which is based on the distribution of the impact estimate) for each of 100 true intervention effects, which is a representative sample of effects drawn from the prior distribution. These 10,000 estimates are plotted in Exhibit L3, and the code that generated the figure is provided in Exhibit L4. For each of 100 true intervention effects (vertical axis), we plot 100 impact estimates (horizontal axis). For each true effect, the 100 impact estimates vary due to random estimation error.

The posterior distribution of true effects conditional on an estimated effect of 0.15 is represented by the true effects that fall within the bandwidth represented by the two vertical red lines. Holding the estimated effect constant at 0.15 (that is, 0.15 on the horizontal axis, which is indicated by a vertical dashed line), all of the true effects associated with an estimated effect of 0.15 lie between superimposed red lines. The probability distribution of the true effects lying between those red lines is the posterior distribution. That posterior distribution is also illustrated in Exhibit L5 (it includes the same points that are between the red lines in Exhibit L3, just zoomed in).

Exhibit L3. Illustration of the relationship between the prior distribution of true effects and the distribution of impact estimates

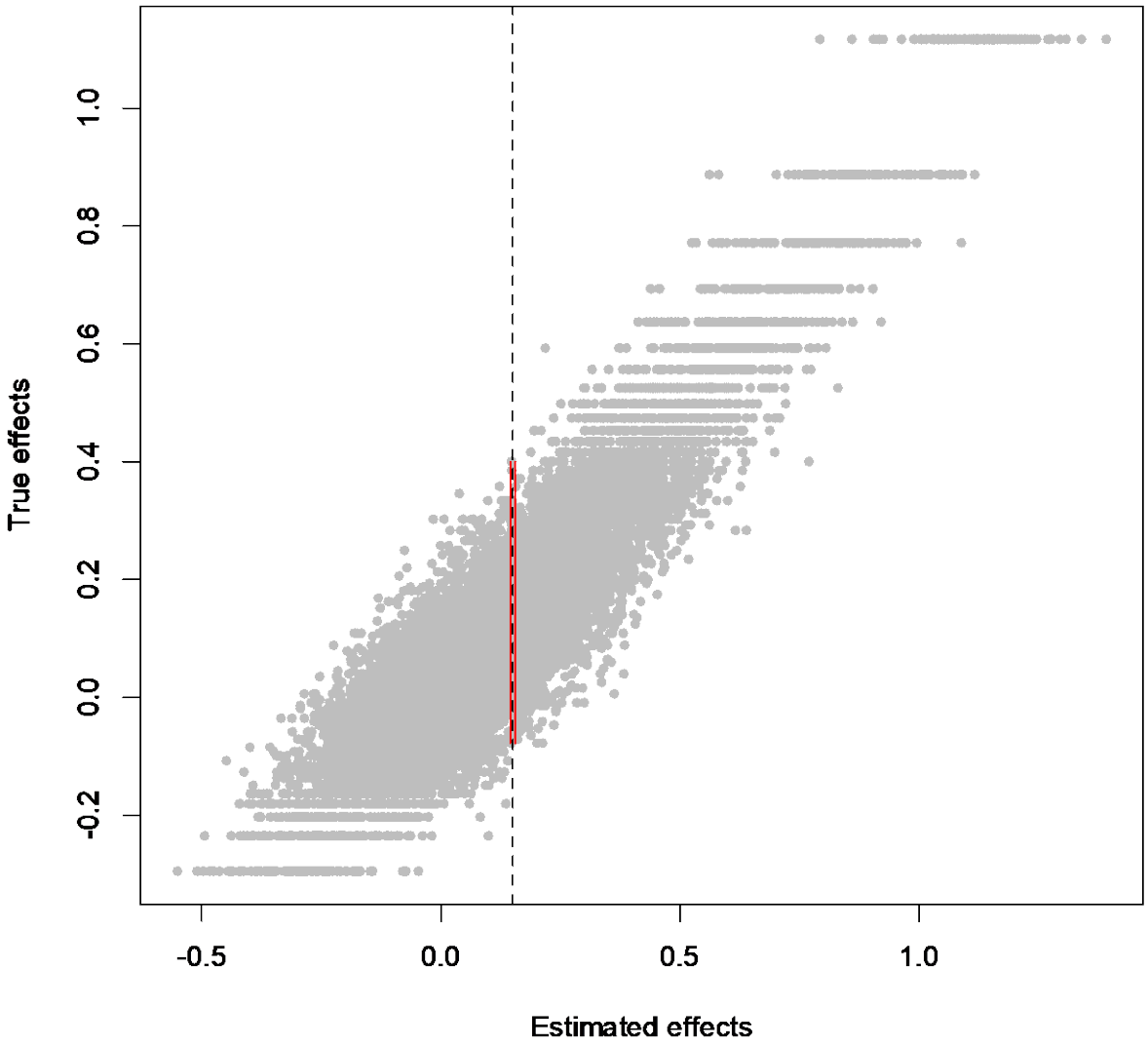
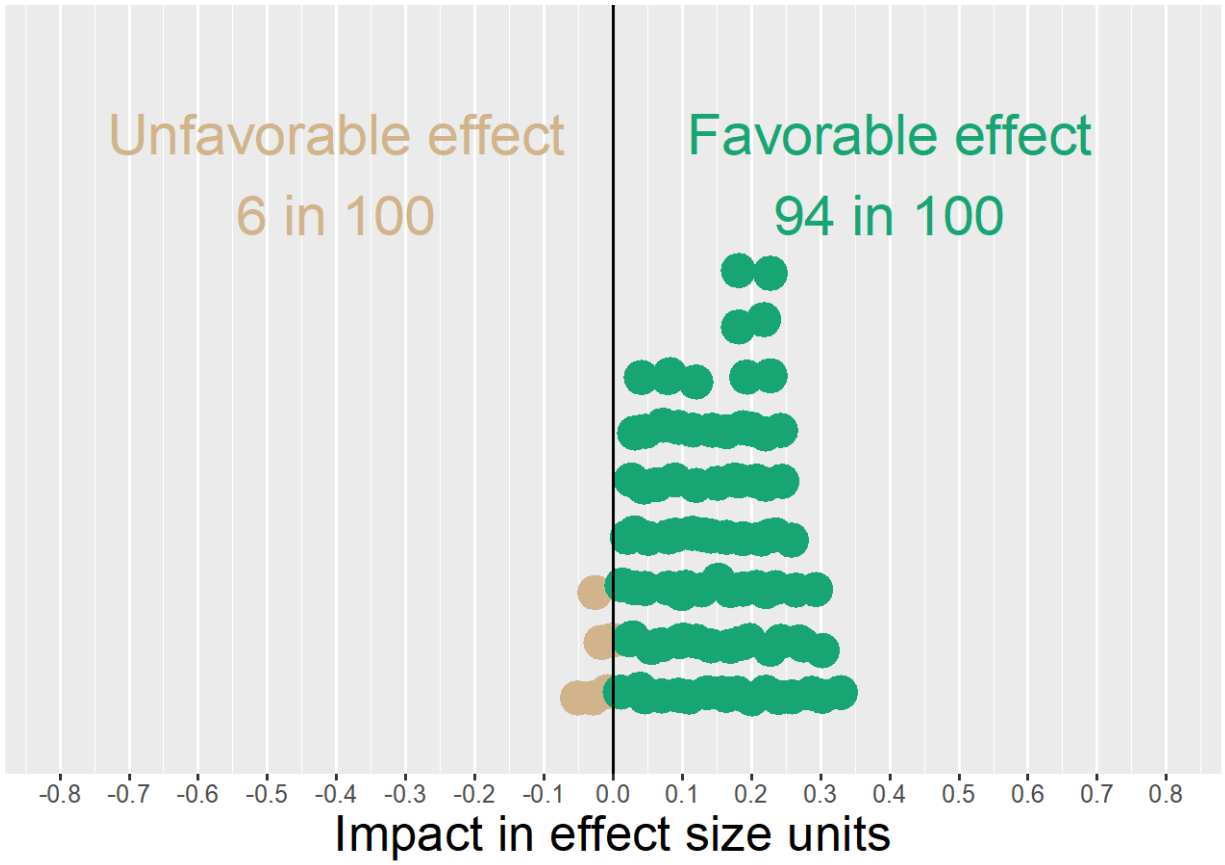


Exhibit L4. The R program used to calculate the posterior distribution illustrated in Exhibit L3

```
#first, create the distribution of prior evidence that matches
#the results of the meta-regression we conducted using data from the WWC
require(sgt)
prior <- 0.16 +
rsgt(n=1000000, mu=0, sigma=0.19, lambda=0.57,q=3.03/2,p=2) +
rsgt(n=1000000, mu=0, sigma=0.22, lambda=0.42,q=3.13/2,p=2)

#second, grab a representative sample of 100 true intervention effects
# from that prior (similar to what is presented in Exhibit L1)
N <- 100
true.effects <- quantile(prior, probs = seq(1/(N+1),N/(N+1),length.out=N))
true.effects <- array(true.effects,N*100)

#third, simulate 100 impact estimates for each of the 100 true intervention effects
# given a standard error of 0.10, thereby yielding 10,000 impact estimates
estimated.effects <- rnorm(n=N*100, mean=true.effects, sd=0.10)

#fourth, identify the subset of estimated effects that are
#approximately equal to 0.15
subset <- round(estimated.effects,2)==0.15

#finally, the posterior consists of the true effects corresponding to the subset of
#estimated effects
posterior <- true.effects[subset]

#Create Exhibit L3
plot(estimated.effects,true.effects,pch=20,xlab="Estimated effects",ylab="True
 effects",col="grey")
segments(x0=0.145,y0=min(posterior),x1=0.145,y1=max(posterior),col="red")
segments(x0=0.155,y0=min(posterior),x1=0.155,y1=max(posterior),col="red")
abline(v=0.15,lty='dashed')
```

Note: This program represents a brute force approach to calculating an approximate Bayesian posterior distribution. It is not a practical approach for Bayesian analysis in general. Rather, it is intended to illustrate the concept of the posterior distribution using a simple example. A more sophisticated version of this algorithm is used for the BASIE probability spreadsheet tool.

Exhibit L5. Posterior distribution of true intervention effects given an impact estimate of 0.15, a standard error of 0.10, and the distribution of prior evidence represented in Exhibit L1



Mean	Standard Deviation	Percentiles				
		10th	25th	50th	75th	90th
0.13	0.09	0.02	0.07	0.12	0.19	0.26

Note: This figure is based on the subset of true effects created by the R program presented in Exhibit L4. These are the same points that lie underneath the red line in Exhibit L3.

Return to BASIE Step 1

Local Stop: Why we do not recommend the flat prior

A prior that used to be very popular in Bayesian analysis is called the flat prior (also known as the improper uniform distribution). The flat prior has infinite variance (instead of a bell curve, a flat line). It was seen as objective because it assigns equal prior probability to all possible values of the impact; for example, impacts on test scores of 0, 0.1, 1, 10, and 100 percentile points are all treated as equally plausible.

When probability is defined in terms of belief rather than evidence, the flat prior might seem reasonable—one might imagine that the flat prior reflects the most impartial belief possible (Gelman et al., 2013, Section 2.8). As such, this prior was de rigueur for decades.

But when probability is based on evidence, the implausibility of the flat prior becomes apparent. For example, what evidence exists to support the notion that impacts on test scores of 0, 0.1, 1, 10, and 100 percentile points are all equally probable? No such evidence exists; in fact, quite a bit of evidence is completely inconsistent with this prior (for example, the distribution of impact estimates in the WWC). The practical implication is that the flat prior overestimates the probability of large effects. Following Gelman and Weakliem (2009), we reject the flat prior because it has no basis in evidence.

The implausibility of the flat prior also has an interesting connection to the misinterpretation of p-values. It turns out that the Bayesian posterior probability derived under a flat prior is identical (for simple models, at least) to a one-sided p-value. Therefore, if researchers switch to Bayesian methods but use a flat prior, they will likely continue to exaggerate the probability of large program effects (which is a common result when misinterpreting p-values).

Finally, we acknowledge that there exist other options for priors that are also not evidence based, including a bounded uniform distribution (for example, uniform(-10,10) on the effect size scale) and a weakly informative prior (for example, standard normal on the effect size scale). Those options are reasonable for some use-cases (for example, regularizing parameters that are not of direct substantive interest in more complicated Bayesian hierarchical models, including parameters in our meta-regression analysis of the WWC database). But we do not recommend these priors for the purpose of interpreting impact estimates since they are not evidence based. Though they will not be numerically equivalent to a one-sided *p*-value, they can result in inferential mistakes of similar magnitudes. For example, if we select the "mirrored; Main; Main" prior in the spreadsheet tool, the prior probability that an intervention effect is above 0.40 standard deviations is 10 percent. But if we use the improper uniform distribution the prior probability would be five times greater—50 percent. If we use uniform(-10,10), the prior probability of an effect greater than 0.40 is 48 percent. If we use the standard normal distribution, the prior probability of an intervention effect greater than 0.40 standard deviations is more than 3 times greater—about 34 percent.

Return to BASIE Step 1

Local Stop: Bayesian meta-regression of prior evidence

We recommend that researchers who wish to construct their own custom prior distribution follow the same conceptual approach we used to construct the 108 prior distributions provided in the BASIE probability tool. For example, researchers could analyze the same estimates from the WWC that we did but define school-levels and outcome domains differently. Researchers could also add their own estimates from another source to the estimates from the WWC. In this local stop, we describe our approach, including the specifications of our Bayesian meta-regression model.

Conceptual approach

To construct the prior distributions provided in the BASIE probability tool, we started by casting a wide net that includes the full set of estimates in the WWC database that meet WWC standards. From there, we narrowed the net to specific school levels and outcome domains. We did not narrow the net by discarding findings from the broader evidence base and estimating a separate meta-regression for the smaller subset of findings. Instead, we narrowed the net by fitting a Bayesian meta-regression to the full set of estimates from all school levels and outcome domains, with terms in the meta-regression that allowed us to estimate a separate prior for each combination of school level and outcome domain.

We recommend that researchers constructing custom prior distributions follow the same approach. This approach of narrowing the net using Bayesian meta-regression avoids the pitfalls of cherry picking and statistical noise from small samples because it uses the overall population of WWC intervention effects as an anchor. The more evidence that exists in a subpopulation, the more the Bayesian meta-regression can safely pull away from the anchor.

For example, we may hypothesize that the distribution of impacts on math achievement for middle school students is different from the distribution of all impacts in the WWC. However, a *hypothesis* that they are different is not sufficient justification for discarding all evidence involving other outcomes or school levels. Instead, we need evidence that impacts on math achievement for middle school students really are likely to be different. Thus, we would not simply estimate a prior based solely on WWC studies of math achievement for middle school students, but rather would use a Bayesian meta-regression to calculate the most likely distribution of impacts on math achievement for middle school students, given the totality of the WWC evidence.

An issue that affects all statistical modeling efforts is model uncertainty—that is, uncertainty about what predictors should be included in a model, the functional form of the relationships between outcomes and predictors, and (in the case of Bayesian modeling) the assumptions made regarding prior distributions for model parameters. Two important sources of uncertainty in our model are whether the model includes an adjustment for small-study effects and whether the model is constrained (via mirroring) to generate a zero-centered distribution of intervention effects.

Technical description of the Bayesian meta-regression model

The meta-regression model has a multilevel structure with individual impact estimates nested within the WWC-reviewed documents that reports the finding (we call these documents 'publications'). Because previous literature has found variation in performance gains by outcome domain and school level (Bloom et al., 2008), the model allows both the mean and standard deviation of intervention effects to vary across evidence subgroups defined by outcome domain and school-level. The model includes an adjustment for small-study effects (described in Appendix A). We fit the model using Markov chain Monte Carlo methods (MCMC) as implemented in the software Stan. The R and Stan programs that implement these analyses are provided in Appendix A.

Using all evidence that met standards as of July 2020, we model the findings from the WWC database as follows:

$$y_i \sim N\left(\theta_i + \beta_i s_i,\ h\left(s_i, d\right)\right)$$

In this equation, y_i is the reported impact estimate and s_i is the reported standard error, in effect size units, of finding i. We model y_i as coming from a normal sampling distribution. The mean of this distribution has two components: θ_i is the true impact that y_i seeks to estimate, and $\beta_i s_i$ is an adjustment for small-study effects. The variance of the distribution, $h\left(s_i, d\right)$, is the variance of the maximum order statistic obtained by taking d draws from the distribution of the impact estimate (see Local Stop: Adjustments for small-study effects).

We use regression to model the true impacts (θ) and small-study effects (β):

$$\theta_i = \theta^{Int} + \theta^{Mult} m_i + \theta^{Ach} a_i + \theta^{Gr}_{g[i]} + \theta^{Outc}_{o[i]} + \theta^{GrOutc}_{x[i]} + \theta^{Pub}_i + \theta^{Find}_i$$

$$\beta_i = \beta^{Int} + \beta^{ERIC}_{e[i]} + \beta^{Res}_{r[i]} + \beta^{Ach} a_i$$

In the following sub-sections, we detail each component of this model. The first sub-section explains the notation used in this model. The second sub-section describes the regression covariates in the WWC data and the indexing structure we use to represent the covariates in the model. The third sub-section describes the regression parameters to be estimated.

Notation

The notation used to describe this model conforms to standard practice in the field (Gelman & Hill, 2007). However, the notation may be confusing for readers unfamiliar with Bayesian hierarchical modeling. In non-Bayesian models, it is common to represent subgroup variables using so-called 'indicator variables' (also known as 'dummy variables'). For example, one might relate an outcome to indicators of grade level using the equation:

$$y_i = \alpha + \beta_1 P_i + \beta_2 E_i + \beta_3 M_i + \varepsilon_i$$

In this equation, the variables P, E, and M would be binary indicators of whether observation i belongs to the subgroup pre-K, elementary school, or middle school, with 'high school' as the omitted category.

With Bayesian models it is common to use a more compact notation. In this example, we would use the equation:

$$y_i = \alpha + \beta^{Gr}_{g[i]} + \varepsilon_i$$

In this equation, β^{Gr} is a zero-centered vector of parameters, and the subscript $g[i]$ denotes the grade level g that observation i belongs to, replacing the indicator variables for coding subgroup membership. For example, $g[i]$ could equal 1 for pre-K, 2 for elementary, 3 for middle, and 4 for high school. Note that an 'omitted category' is not needed in a hierarchical Bayesian model in which the vector of parameters is shrunken towards zero (Gelman & Hill, 2007).

Regression covariates and indexing structure

Every impact estimate recorded in the WWC database can belong to a variety of different subgroups. Subgroup variables included in our analysis include the grades of students included in an impact analysis, the outcome on which the impact of the intervention was estimated, ERIC publication type, and whether the finding met WWC standards with or without reservations.

In the list that follows we describe the indexing structure we use to represent subgroups in the WWC database in our model.

- $g[i]$ indexes the grade span studied in finding i; grade spans are
  - Pre-kindergarten
  - Elementary (kindergarten to grade 5)
  - Middle school (grades 6-8)
  - High school (grades 9-12)
  - Postsecondary
- m_i is an indicator for whether the finding pertains to multiple grade spans (see below for more on how this is implemented).
- $o[i]$ indexes the outcome domain of finding i; outcome domains are
  - Math achievement
  - English language arts (ELA) achievement
  - Science achievement
  - Other achievement (such as cumulative GPA, social studies)
  - Behavior (such as expulsion, arrests, social skills)
  - Attainment (such as attendance, high school graduation)
  - Miscellaneous (such as classroom observations, physical fitness)
- a_i is an indicator for whether the outcome domain of the finding is one of the four achievement-related domains.
- $x[i]$ indexes the interaction between the grade span and outcome domain of finding i; for example,
  - Pre-kindergarten math achievement
  - Elementary ELA achievement
- $e[i]$ indexes the ERIC publication type associated with finding i. The three publication types are
  - Appears in ERIC as a journal article
  - Appears in ERIC as a non-journal article
  - Does not appear in ERIC
- $r[i]$ indexes whether finding i met WWC standard without reservations (typically a high quality randomized controlled trial (RCT)) or met WWC standards with reservations (typically a randomized trial with high attrition or a quasi-experimental design).

Parameters

As described above, θ_i is the true impact estimated by finding i. The full list of the components of the true impact follows, with all prior distributions reported in Exhibit L6:

- θ^{Int} is the overall effect ("intercept") common to all findings in the WWC
- θ^{Mult} captures the effect of the finding pertaining to multiple grades
- θ^{Ach} captures the effect of the finding pertaining to one of the four achievement domains
- θ^{Gr} is a set of random effects of grade span studied in finding i
  - If a finding is reported as pertaining to multiple grade spans, the average of all applicable random effects is used (this is achieved using 'dummy' variables in the Stan code to handle this contingency)
  - They are constrained to sum to zero: $\theta^{Gr}_{PreK} + \theta^{Gr}_{Elem} + \cdots + \theta^{Gr}_{PostSec} = 0$
- θ^{Outc} is a set of random effects of the outcome domain of the finding
  - Because we also include θ^{Ach}, we require the outcome effects to be exchangeable only after taking into account the potential overall difference for achievement vs. non-achievement outcomes
  - They are constrained so that the achievement effects and the non-achievement effects each sum to zero, so that θ^{Ach} is determinable: $\theta^{Outc}_{Math} + \theta^{Outc}_{ELA} + \theta^{Outc}_{Sci} + \theta^{Outc}_{OthAch} = 0$ and $\theta^{Outc}_{Beh} + \theta^{Outc}_{Attain} + \theta^{Outc}_{Misc} = 0$
- θ^{GrOutc} is a set of random effects for the interaction of grade span and outcome domain
  - If a finding pertains to multiple grades, the average of all of the relevant random effects is used
  - They are not constrained to sum to zero
- θ^{Find} is a set of random effects capturing the idiosyncratic effect specific to each finding
  - These are distributed according to a skewed t-distribution, with degrees of freedom and skew parameter estimated by the model. The prior distribution for the degrees of freedom parameter is gamma with shape 2 and rate 0.1 (this implies a mean of 20 with a standard deviation of about 14). The prior distribution for the skew parameter is a $Beta(1.5, 1.5)$ distribution, rescaled (by multiplying all values by 2 and then subtracting 1) to have support between -1 and 1; the rescaled $Beta$ distribution has a mean of 0 and a standard deviation of 0.5. These weakly informative priors pull the distribution towards symmetry and thin tails.
  - These terms are heteroskedastic, with standard deviation of σ_i modeled as follows:
    $$\ln(\sigma_i) = \sigma^{Int} + \sigma^{Mult} m_i + \sigma^{Ach} a_i + \sigma^{Gr}_{g[i]} + \sigma^{Outc}_{o[i]}$$
    - σ^{Int} reflects the overall variation
    - σ^{Mult} allows for different variation if a finding pertains to multiple grades
    - σ^{Ach} allows for different variation if a finding pertains to one of the four achievement domains σ^{Gr} is a set of random effects for the grade span
      - As with θ^{Gr}, if a finding pertains to multiple grades the average of all relevant random effects is used
      - They are constrained to sum to zero: $\sigma^{Gr}_{PreK} + \sigma^{Gr}_{Elem} + \cdots + \sigma^{Gr}_{PostSec} = 0$
    - σ^{Outc} is a set of random effects for the outcome domain
      - They are constrained so that the achievement effects and the non-achievement effects each sum to zero (so that σ^{Ach} is determinable): $\sigma^{Outc}_{Math} + \sigma^{Outc}_{ELA} + \sigma^{Outc}_{Sci} + \sigma^{Outc}_{OthAch} = 0$ and $\sigma^{Outc}_{Beh} + \sigma^{Outc}_{Attain} + \sigma^{Outc}_{Misc} = 0$
- θ^{Pub} is a set of random effects capturing the overall effect associated with each publication

  - As with θ^{Find}, these are distributed according to a skewed t-distribution, with modeled degrees of freedom and skew parameter
  - Also as with θ^{Find}, these are heteroskedastic[7], with an analogous decomposition of their standard deviation, τ_i: $\ln(\tau_i) = \tau^{Int} + \tau^{Mult} m_i + \tau^{Ach} a_i + \tau^{Gr}_{g[i]} + \tau^{Outc}_{o[i]}$

- β_i is a multiplier on the standard error of finding i (s_i), to account for small-study effects. It is composed of the following terms:
  - β^{Int} is the overall small-study coefficient, i.e. the "intercept" of its regression
  - $\beta^{ERIC}_{e[i]}$ is a set of additive effects to allow for different publication types having smaller or larger small-study effects. These are constrained to sum to zero: $\beta^{ERIC}_{Journ} + \beta^{ERIC}_{NonJourn} + \beta^{ERIC}_{NonERIC} = 0$
  - $\beta^{Res}_{r[i]}$ is a set of additive effects to allow findings which meet WWC standards with or without reservations to have different small-study effects. These are also constrained to sum to zero: $\beta^{Res}_{Without} + \beta^{Res}_{With} = 0$
  - β^{Ach} is the impact on the small-study coefficient of the finding pertaining to one of the four achievement domains

To create a zero-centered prior, we merge the prior estimated using our Bayesian meta-regression with its mirror image. By mirror image, we mean that we flip all positive signs to negative and negative signs to positive. To create a prior without an adjustment for small-study effects, we remove the $\beta_i s_i$ term from the meta-regression model.

The output from fitting this model to the WWC database is a matrix with 4,000 rows (corresponding to 4,000 draws from the posterior distribution) and 21,433 columns (corresponding to model parameters). Most of those parameters are for the individual impacts in the WWC database and are not of direct interest to us. The parameters of primary interest to us are the parameters of the skewed generalized t-distribution, at both the finding and study level, that characterize the distribution of intervention effects for specific combinations of school-level and outcome domain. Those key parameters for the prior distributions we prepared using this model are presented in Local Stop: Prior distributions ready to use (Exhibits L8 and L9).

[7] To ensure θ^{Pub} is a publication-level effect, despite having a standard deviation that varies at the finding-level, we first estimate publication random effects on the unit scale, such that $\theta^{PubUnit}_j \sim N(0,1)$ is the unit-scale random effect of the publication j in which finding i was published. We then convert the unit-scale publication random effects to finding-level effects using the individual findings' values of τ_i. For example, a publication that appears particularly effective may have a $\theta^{PubUnit}_j$ of 0.8; if two findings from the publication have values of τ_i of 0.25 and 0.5, then their final values of $\theta^{Pub}_i = \tau_i \theta^{PubUnit}_{j[i]}$ are 0.2 and 0.4, respectively

Exhibit L6. Prior distributions for all model parameters

Parameter	Prior distribution	Parameter	Prior distribution
β^{Int}	$N(0,1)$	σ^{Gr}	$N(0,\sigma_{\sigma^{Gr}})$ with $\sum\sigma^{Gr} = 0$
β^{Eric}	$N(0,1)$ with $\beta^{ERIC}_{NonJourn} + \beta^{ERIC}_{NonERIC} = 0$	$\sigma_{\sigma^{Gr}}$	$N^{+}(0,0.5)$
β^{Res}	$N(0,1)$ with $\beta^{Res}_{Without} + \beta^{Res}_{With} = 0$	σ^{Outc}	$N(0,\sigma_{\sigma^{Outc}})$ with $\sum\sigma^{Outc} = 0$
β^{Ach}	$N(0,1)$	$\sigma_{\sigma^{Outc}}$	$N^{+}(0,0.5)$
θ^{Int}	$N(0,1)$	λ^{Find}	$B^{*}(1.5,1.5)$
θ^{Mult}	$N(0,1)$	ν^{Find}	$\Gamma(2,0.1)$
θ^{Ach}	$N(0,1)$	θ^{Pub}	$SGT(0,\sigma^{Pub},\lambda^{Pub},\nu^{Pub},2)$
θ^{Gr}	$N(0,\sigma^{Gr})$ with $\sum\theta^{Gr} = 0$	τ^{Int}	$N(-2,2)$
σ^{Gr}	$N^{+}(0,1)$	τ^{Mult}	$N(0,0.5)$
θ^{Outc}	$N(0,\sigma^{Outc})$ with $\sum\theta^{Outc} = 0$	τ^{Ach}	$N(0,0.5)$
σ^{Outc}	$N^{+}(0,1)$	τ^{Gr}	$N(0,\sigma_{\tau^{Gr}})$ with $\sum\tau^{Gr} = 0$
θ^{GrOutc}	$N(0,\sigma^{GrOutc})$ with $\sum\theta^{GrOutc} = 0$	$\sigma_{\tau^{Gr}}$	$N^{+}(0,0.5)$
σ^{GrOutc}	$N^{+}(0,1)$	τ^{Outc}	$N(0,\sigma_{\tau^{Outc}})$ with $\sum\tau^{Outc} = 0$
θ^{Find}	$SGT(0,\sigma^{Find},\lambda^{Find},\nu^{Find},2)$	$\sigma_{\tau^{Outc}}$	$N^{+}(0,0.5)$
σ^{Int}	$N(-2,2)$	λ^{Pub}	$B^{*}(1.5,1.5)$
σ^{Mult}	$N(0,0.5)$	ν^{Pub}	$\Gamma(2,0.1)$
σ^{Ach}	$N(0,0.5)$		

$N(\mu,\sigma)$ indicates a normal distribution with mean μ and standard deviation σ.
$N^{+}(\mu,\sigma)$ indicates a normal distribution truncated below at zero (i.e., positive-only) with location μ and scale σ.
$SGT(\mu,\sigma,\lambda,\nu,p)$ indicates a skewed generalized t distribution with location μ, scale σ, skewness λ, degrees of freedom ν, and kurtosis parameter p.
$B^{*}(\alpha,\beta)$ indicates a beta distribution with shape parameters α and β, rescaled (by multiplying all values by 2 and then subtracting 1) to have support between -1 and 1. The $B^{*}(1.5,1.5)$ distribution has a mean of 0 and a standard deviation of 0.5.
$\Gamma(k,\beta)$ indicates a gamma distribution with shape parameter k and rate parameter β.

Return to BASIE Step 1

Local Stop: Adjustments for small-study effects

In this section, we provide an overview of the issue of bias due to selection in what findings authors choose to report, and how 'small-study effects' can be indicative of this problem. We also describe the method we use to adjust for small-study effects.

The idea of bias due to selection in what findings authors choose to report is that the most favorable research findings are published and less favorable results are left unreported. Marks-Anglin and Chen (2020) provide a historical overview of this topic. Researchers posit multiple mechanisms by which findings that are unfavorable to an intervention are excluded from the literature:

1. *Publication bias*, in which editors of journals refuse to publish findings that are not statistically significant.
2. *p-hacking*, in which researchers, possibly responding to the negative incentives of publication bias, intentionally conduct multiple analyses until a p-value is low enough to be deemed statistically significant.
3. *The garden of forking paths*, in which researchers' unconscious biases lead to research decisions that suppress unfavorable estimates (Gelman & Loken, 2014). For example, if the first analysis a researcher conducts shows a favorable effect, they may not conduct as many sensitivity analyses as they would have if the first analysis showed an unfavorable effect. If it shows an unfavorable effect, they may conduct more analyses than initially expected.

Our adjustment method is motivated by the assumption that researchers report the most favorable impact estimate (regardless of the p-value or statistical significance). We focus on this motivation because even if researchers abandon the null hypothesis significance testing (NHST) framework, there may still be incentives and unconscious biases that lead researchers to promote more favorable impact estimates and leave less favorable estimates either unreported or never calculated. In other words, our adjustment method is focused on a future in which BASIE replaces NHST. However, our method also works well when researchers are hacking p-values because increasing the impact estimate is obviously a good way to get a smaller p-value.[8] In addition, our adjustment procedure is not focused on addressing publication bias because the WWC includes unpublished findings.

The statistical implication of this motivation is that the expected value of reported impact estimates is not the true effect of the intervention. Instead, the expected value of the reported estimates is actually that of a maximum order statistic. A maximum order statistic is the largest of multiple draws from a distribution. For example, consider rolling a six-sided die three times. The largest of the three rolls is a maximum order statistic. The expected value of a single die roll is 3.5. But the expected value of the largest of three die rolls (that is, the expected value of the maximum order statistic) is about 4.96.

Another implication is that there is likely a misalignment between the impact estimates and standard errors reported in the literature. We assume that the reported standard errors in the literature are not standard errors for the maximum order statistic but rather are the standard errors of the impact estimate. We think this makes sense because researchers are not typically aware that they are reporting maximum order statistics and so are not attempting to calculate and

[8] Another way to lower the p-value is to lower the standard error. However, the standard error is a less variable statistic than the impact estimate, so it is generally easier to hack a p-value by focusing on moving the impact estimate rather than the standard error. For example, consider a multi-site randomized controlled trial with equally sized sites. The variability in the t-statistic across sites is driven primarily by variability in the impact estimate, not variability in the standard error estimate.

report the standard error of a maximum order statistic. Our adjustment procedure addresses this misalignment.

The consequences of our adjustment procedure (relative to conducting a meta-regression without it) are to (1) reduce our estimate of the average impact in the WWC database and (2) increase our estimate of the standard deviation of impacts in the WWC database. Our adjustment procedure and evidence of its efficacy are described in Appendix A.

Description of our adjustment method

It turns out that the expected value of a random variable's maximum order statistic can be well approximated by a linear function of the standard deviation of that random variable (Royston, 1982). In our case, the random variable of interest is the impact estimate and the standard deviation of that random variable is the standard error. Thus, in the presence of bias due to reporting only the most favorable of multiple impact estimates, we would expect impact estimates in the WWC database with larger standard errors to be systematically more favorable to the intervention than impact estimates with smaller standard errors. That is, we would expect a correlation between favorable impact estimates and their standard errors. The more findings left unreported, the stronger this correlation. The presence of this correlation between favorable impact estimates and standard errors is sometimes called 'small-study effects,' because smaller studies have larger standard errors.

To correct for this bias, we fit a simple linear regression as part of our Bayesian meta-regression model: the dependent variable is the impact estimate in the WWC database, and the independent variable is its standard error. The constant term (intercept) in the regression is our bias–corrected estimate of the intervention effect. The coefficient on the standard error (which is not of direct interest to us but is needed for the standard error adjustment described in the next paragraph) is related to the number of less favorable estimates that were not reported.

Based on our estimate of the number of less favorable estimates that were not reported (which comes from the coefficient on the standard error in the regression described in the previous paragraph), and the assumption that the reported standard error is for the impact estimate not the maximum order statistic of the impact estimate, we calculate the standard error for the maximum order statistic.

To adjust for bias associated with small-study effects, we model the reported estimates from the WWC as

$$y_i \sim N\left(\theta_i + \beta s_i,\ h\left(s_i, d\right)\right)$$

$$d = \frac{\Phi\left(\beta\right)\left(1 - \frac{\pi}{4}\right) + \frac{\pi}{8}}{1 - \Phi\left(\beta\right)}$$

where i indexes reported estimates; y_i is the maximum order statistic resulting from d draws from the distribution of the i-th impact estimate (note, d is not directly estimable, but β is); N is the normal probability distribution; θ_i is the true impact (in effect size units); s_i is the reported standard error (which we assume is not the standard error of the maximum order statistic but instead the standard deviation of the distribution of the impact estimate); β represents the correlation between the reported impact estimate and s_i that is related to d as described in the

equation above (Royston, 1982); Φ is the standard normal cumulative distribution function; and $h(s_i; d)$ is the variance of the maximum order statistic (h is approximated using a twelfth-order polynomial, estimated via Monte Carlo, described in Appendix A).

In this model, the key parameters to be estimated are θ_i and β. We estimate these parameters in the context of the overall Bayesian meta-regression (see Local Stop: Bayesian meta-regression of prior evidence) using MCMC as implemented in the software Stan. The R and Stan programs that implement these models are provided in Appendix A. For the full population of effects in the WWC, we estimate $\beta = 0.47$. This estimate implies that for every impact estimate reported in the WWC, there are (on average) about 1.7 less favorable impact estimates left unreported.

Return to BASIE Step 1

Local Stop: Prior distributions ready to use

In this section we present all prior distributions that we estimated using data from the WWC (downloaded July 2020). We also present the results of analyses to understand whether our choice of the skewed generalized t-distribution in our Bayesian meta-regression yields a different understanding of the distribution of intervention effects than if we had made the more restrictive assumption that the prior is a normal distribution.

All prior distributions

The WWC database includes 98 outcome sub-domains which we consolidated into 7 domains for more parsimonious analysis and presentation. In Exhibit L7 we show which of the 98 WWC outcome domains are included in each of our 7 domains.

Exhibit L7. WWC outcome domains included in each of our outcome domains

Our outcome domains	WWC outcome sub-domains
Math achievement	Algebra; Data analysis, statistics, and probability; General Mathematics Achievement; Geometry; Geometry and Measurement; Number and Operations
ELA achievement	Alphabetics; Audience; Communication/ Language; Comprehension; Early reading/writing; English language arts achievement; English language development; English language proficiency; Genre elements; Language arts; Language development; Letter identification; Literacy achievement; Oral language; Organization; Overall writing quality; Phonological processing; Print knowledge; Reading achievement; Reading and listening comprehension; Reading comprehension; Reading fluency; Vocabulary development; Word reading; Writing achievement; Writing output; Writing processes; Writing quality
Science achievement	Science achievement
Other achievement	Academic achievement; Cognition; College academic achievement; College readiness; Conceptual knowledge; Functional abilities; General academic achievement (college); General academic achievement (high school); General academic achievement (middle school); Other academic performance; Primary school academic achievement; Procedural knowledge; Secondary school academic achievement; Social studies achievement; Technical skill proficiency
Behavioral	Behavior; Emotional/internal behavior; Executive functioning; External behavior; Knowledge, attitudes, & values; Problem behavior; Procedural flexibility; School engagement; Self-care/daily living; Self-concept; Self-determination; Social outcomes; Social-emotional competence; Social-emotional development; Student behavior; Student emotional status; Student engagement in school; Student social interaction
Attainment	Access and enrollment; Attainment; Attendance (high school); College and career preparation; College enrollment; Completing school; Credential attainment; Credit accumulation; Credit accumulation and persistence; Graduating school; Industry-recognized credential, certificate, or license completion; Postsecondary degree attainment; Progress in developmental education; Progressing in college; Progressing in developmental education; Progressing in school; School attendance; Secondary school attendance; Staying in school; Student progression
Misc.	Labor market outcomes; Medium-Term Earnings; Physical well-being; School leader retention at the school; Short-Term Earnings; Teacher attendance; Teacher instruction; Teacher retention; Teacher retention at the school; Teacher retention in the school district

Source: Files with outcome domains defined by the WWC are available at https://ies.ed.gov/ncee/wwc/reviewresources5.

In Exhibits L8 and L9, we report the prior distributions estimated with an adjustment for small-study effects (Exhibit L8) and without that adjustment (Exhibit L9). We also plot the densities of all

the prior distributions using a 'spaghetti plot' (Exhibit L10). For each combination of school level and outcome domain, the prior distribution is defined by a mean and two (zero-centered) skewed t-distributions (one for the study level, one for the finding level). Each prior has seven parameters, including the mean as well as the standard deviation, skew, and degrees of freedom parameters for the study level and finding level random effects. The first row in the table is based on a simplified model that does not include outcome domain or grade level effects; the remaining rows come from a common model that allows means and variances (but not skew or degrees of freedom) to vary across outcome domains and grade levels. In Exhibit L11, we show an example of R code that uses these parameter values to generate a large number of (synthetic) true effects from the prior distribution specified in the first row of Exhibit L8. From this distribution we report the mean, standard deviation, and key quantiles. The example R code also shows how we create a zero-centered version of the distribution by mirroring the distribution.

Exhibit L8. Prior distributions with adjustment for small-study effects

			Study level random effect			Finding level random effect		
School level	Outcome domain	Mean	Standard deviation	Skew	Degrees of freedom	Standard deviation	Skew	Degrees of freedom
All	All	0.16	0.19	0.57	3.03	0.22	0.42	3.13
PK	Math achievement	0.19	0.26	0.43	3.03	0.21	0.52	3.13
PK	ELA achievement	0.16	0.17	0.43	3.03	0.24	0.52	3.13
PK	Science achievement	0.17	0.22	0.43	3.03	0.26	0.52	3.13
PK	Other achievement	0.14	0.14	0.43	3.03	0.28	0.52	3.13
PK	Behavioral	0.17	0.21	0.43	3.03	0.30	0.52	3.13
PK	Attainment	0.17	0.29	0.43	3.03	0.16	0.52	3.13
PK	Misc.	0.15	0.16	0.43	3.03	0.25	0.52	3.13
Elementary	Math achievement	0.19	0.32	0.43	3.03	0.21	0.52	3.13
Elementary	ELA achievement	0.18	0.22	0.43	3.03	0.24	0.52	3.13
Elementary	Science achievement	0.22	0.27	0.43	3.03	0.27	0.52	3.13
Elementary	Other achievement	0.15	0.17	0.43	3.03	0.29	0.52	3.13
Elementary	Behavioral	0.20	0.25	0.43	3.03	0.31	0.52	3.13
Elementary	Attainment	0.21	0.36	0.43	3.03	0.16	0.52	3.13
Elementary	Misc.	0.19	0.20	0.43	3.03	0.26	0.52	3.13
Middle	Math achievement	0.16	0.22	0.43	3.03	0.20	0.52	3.13
Middle	ELA achievement	0.16	0.15	0.43	3.03	0.22	0.52	3.13
Middle	Science achievement	0.16	0.18	0.43	3.03	0.25	0.52	3.13
Middle	Other achievement	0.13	0.12	0.43	3.03	0.26	0.52	3.13
Middle	Behavioral	0.18	0.18	0.43	3.03	0.28	0.52	3.13
Middle	Attainment	0.18	0.24	0.43	3.03	0.15	0.52	3.13
Middle	Misc.	0.16	0.14	0.43	3.03	0.24	0.52	3.13
High	Math achievement	0.18	0.25	0.43	3.03	0.21	0.52	3.13
High	ELA achievement	0.13	0.17	0.43	3.03	0.24	0.52	3.13
High	Science achievement	0.18	0.21	0.43	3.03	0.27	0.52	3.13
High	Other achievement	0.17	0.14	0.43	3.03	0.29	0.52	3.13
High	Behavioral	0.18	0.20	0.43	3.03	0.31	0.52	3.13
High	Attainment	0.18	0.28	0.43	3.03	0.16	0.52	3.13
High	Misc.	0.17	0.16	0.43	3.03	0.26	0.52	3.13
Postsecondary	Math achievement	0.18	0.20	0.43	3.03	0.20	0.52	3.13
Postsecondary	ELA achievement	0.17	0.14	0.43	3.03	0.23	0.52	3.13
Postsecondary	Science achievement	0.19	0.17	0.43	3.03	0.25	0.52	3.13
Postsecondary	Other achievement	0.16	0.11	0.43	3.03	0.27	0.52	3.13
Postsecondary	Behavioral	0.19	0.16	0.43	3.03	0.29	0.52	3.13
Postsecondary	Attainment	0.19	0.22	0.43	3.03	0.16	0.52	3.13
Postsecondary	Misc.	0.18	0.13	0.43	3.03	0.25	0.52	3.13

Note: The prior distributions are composed of a mean and two skewed t-distributions. These skewed t-distributions are a subset of the skewed generalized t-distribution (Hansen et al., 2010), with the parameter p fixed at 2 and q = degrees of freedom/2.

Exhibit L9. Prior distributions without adjustment for small-study effects

			Study level random effect			Finding level random effect		
School level	Outcome domain	Mean	Standard deviation	Skew	Degrees of freedom	Standard deviation	Skew	Degrees of freedom
All	All	0.23	0.18	0.64	3.03	0.24	0.54	3.12
PK	Math achievement	0.28	0.22	0.49	3.03	0.24	0.62	3.14
PK	ELA achievement	0.26	0.14	0.49	3.03	0.24	0.62	3.14
PK	Science achievement	0.26	0.19	0.49	3.03	0.27	0.62	3.14
PK	Other achievement	0.22	0.12	0.49	3.03	0.28	0.62	3.14
PK	Behavioral	0.26	0.18	0.49	3.03	0.31	0.62	3.14
PK	Attainment	0.25	0.25	0.49	3.03	0.17	0.62	3.14
PK	Misc.	0.25	0.14	0.49	3.03	0.25	0.62	3.14
Elementary	Math achievement	0.27	0.30	0.49	3.03	0.25	0.62	3.14
Elementary	ELA achievement	0.27	0.19	0.49	3.03	0.26	0.62	3.14
Elementary	Science achievement	0.31	0.26	0.49	3.03	0.29	0.62	3.14
Elementary	Other achievement	0.23	0.16	0.49	3.03	0.30	0.62	3.14
Elementary	Behavioral	0.28	0.24	0.49	3.03	0.33	0.62	3.14
Elementary	Attainment	0.28	0.35	0.49	3.03	0.18	0.62	3.14
Elementary	Misc.	0.27	0.19	0.49	3.03	0.27	0.62	3.14
Middle	Math achievement	0.23	0.21	0.49	3.03	0.24	0.62	3.14
Middle	ELA achievement	0.23	0.14	0.49	3.03	0.25	0.62	3.14
Middle	Science achievement	0.22	0.18	0.49	3.03	0.28	0.62	3.14
Middle	Other achievement	0.20	0.11	0.49	3.03	0.29	0.62	3.14
Middle	Behavioral	0.26	0.17	0.49	3.03	0.31	0.62	3.14
Middle	Attainment	0.24	0.24	0.49	3.03	0.18	0.62	3.14
Middle	Misc.	0.24	0.14	0.49	3.03	0.26	0.62	3.14
High	Math achievement	0.25	0.25	0.49	3.03	0.24	0.62	3.14
High	ELA achievement	0.19	0.16	0.49	3.03	0.25	0.62	3.14
High	Science achievement	0.25	0.22	0.49	3.03	0.28	0.62	3.14
High	Other achievement	0.23	0.13	0.49	3.03	0.29	0.62	3.14
High	Behavioral	0.25	0.20	0.49	3.03	0.31	0.62	3.14
High	Attainment	0.23	0.29	0.49	3.03	0.18	0.62	3.14
High	Misc.	0.23	0.16	0.49	3.03	0.26	0.62	3.14
Postsecondary	Math achievement	0.24	0.20	0.49	3.03	0.24	0.62	3.14
Postsecondary	ELA achievement	0.24	0.13	0.49	3.03	0.24	0.62	3.14
Postsecondary	Science achievement	0.25	0.17	0.49	3.03	0.28	0.62	3.14
Postsecondary	Other achievement	0.22	0.10	0.49	3.03	0.28	0.62	3.14
Postsecondary	Behavioral	0.26	0.16	0.49	3.03	0.31	0.62	3.14
Postsecondary	Attainment	0.23	0.23	0.49	3.03	0.17	0.62	3.14
Postsecondary	Misc.	0.25	0.13	0.49	3.03	0.26	0.62	3.14

Note: The prior distributions are composed of a mean and two skewed t-distributions. These skewed t-distributions are a subset of the skewed generalized t-distribution (Hansen et al., 2010), with the parameter p fixed at 2 and q = degrees of freedom/2.

Exhibits L8 and L9 both have 36 prior distributions. We can also create 36 zero-centered priors by mirroring (Exhibit L11) the prior distributions in Exhibit L8. Combined, this is a total of 108 distributions.

Exhibit L10. Density plots of all prior distributions

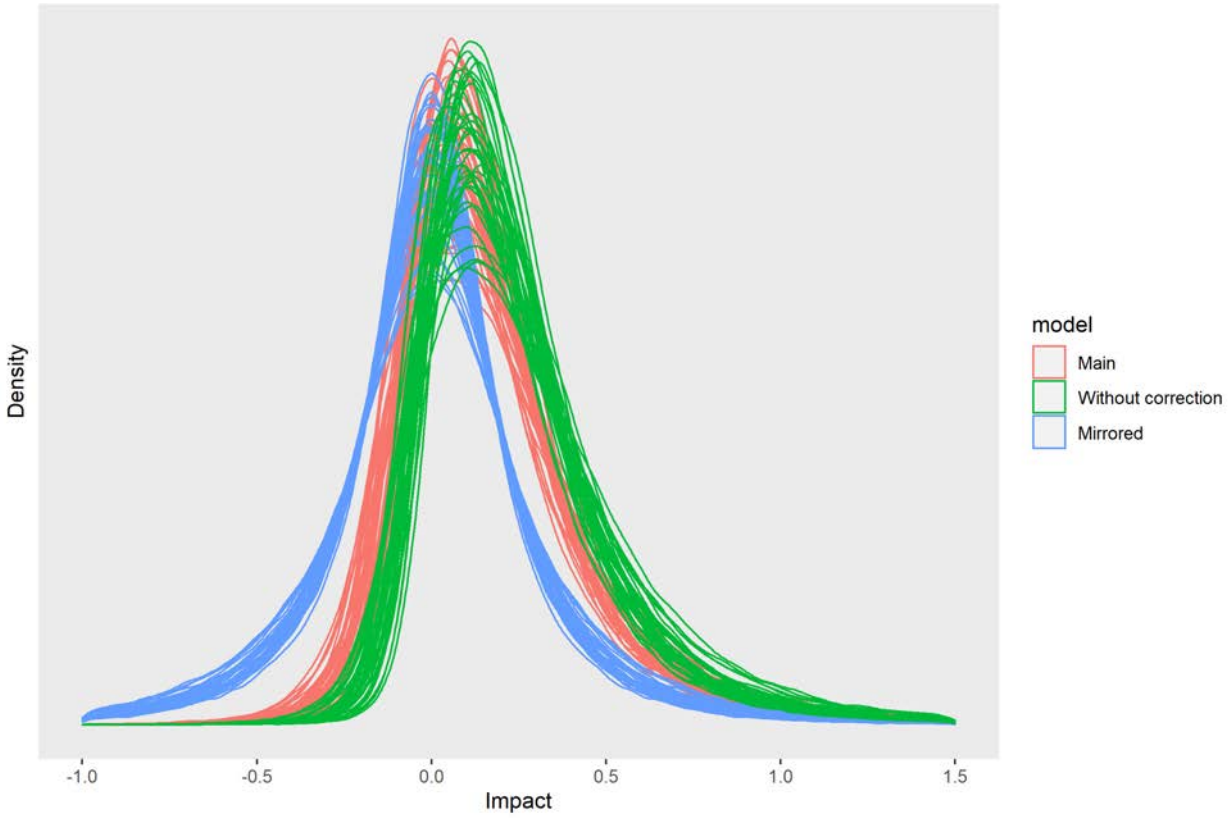


Exhibit L11. R code to create zero-centered prior distribution

```
#load the R package "sgt" (skewed generalized t distribution)
library("sgt")

#draw a large number of true effects from the distribution of all school levels
#and all outcome domains, with the file drawer adjustment (Exhibit A1, first row)
N <- 100000
y <- 0.16 + rsgt(N,mu=0,sigma=0.19,lambda=0.57,q=3.03/2) +
 rsgt(N,mu=0,sigma=0.22,lambda=0.42,q=3.13/2)

#examine characteristics of this distribution
round(mean(y),2)
[1] 0.16
round(sd(y),2)
[1] 0.29
round(quantile(y,c(0.1,0.25,0.5,0.75,0.9)),2)
  10%   25%   50%   75%   90%
-0.11 -0.01  0.11  0.27  0.47

#create the zero-centered, mirrored version of this distribution
y.mirrored <- c(-y,y)
round(mean(y.mirrored),2)
[1] 0
round(sd(y.mirrored),2)
[1] 0.33
round(quantile(y.mirrored,c(0.1,0.25,0.5,0.75,0.9)),2)
  10%   25%   50%   75%   90%
-0.33 -0.15  0.00  0.15  0.33
```

Comparing the normal and skewed generalized t-distributions

The benefit of using the skewed generalized t-distribution to model the distribution of intervention effects is that it allows for the possibility that large positive effects may be more common than large negative effects. We emphasize, however, that while using this distribution allows for that possibility, that possibility is not required or imposed. Using this distribution also allows for the possibility that the distribution of effects is symmetric and that large outliers are unlikely (that is, it allows for the possibility that the distribution is actually normal). In fact, we use a prior for the skewness and degrees of freedom parameters in our meta-regression that imposes a mild preference for a normal distribution—our model will only estimate a skewed, fat-tailed distribution if the data really support it.

Although using the skewed generalized t-distribution has the benefit of being able to capture a skewed distribution of effects (if such skewness exists), it comes with the cost of greater complexity. If the result of our meta-regression were that the distribution of effects is essentially normal, then it would be more convenient to just use the normal distribution. Specifically, it would make it easier to calculate both posterior probabilities and minimum detectable effects (MDEs) because the calculations could be done using formulas rather than simulations.

In this section we compare the distribution of intervention effects estimated using the normal distribution to the skewed generalized t-distribution to assess whether we really need to use the skewed generalized t-distribution. We make this comparison for the overall WWC population of intervention effects (all school levels and all outcome domains).

In Exhibit L12, we overlay the distribution we estimate if we assume a normal distribution (red) versus the distribution we estimate if we allow additional flexibility and use the skewed generalized t-distribution (blue). This exhibit shows that there does seem to be a genuinely positive skew in intervention effects.

Thus, we see that the added flexibility of the skewed generalized t-distribution is indeed leveraged—we do not estimate a symmetric, thin-tailed distribution. But is the distribution we estimate skewed enough and/or fat-tailed enough to actually affect the posterior probabilities that we care about at the end of the day?

In Exhibit L13, we examine the difference in posterior probabilities calculated using the two prior distributions shown in Exhibit L12. Each panel corresponds to a different posterior probability. Each point in the figures represents the posterior probability associated with an impact estimate and standard error (impact estimates are represented by color, standard error by symbol). The x-axis is posterior probabilities using the skewed generalized t-distribution; the y-axis is for using the normal distribution. If it made no difference which prior was used, all points would fall on the 45-degree line. Points above (below) the 45-degree line are cases when assuming a normal prior distribution would cause us to overestimate (underestimate) posterior probabilities compared to the posterior probabilities we would estimate assuming the more flexible skewed generalized t prior.

For example, in the top-left panel of Exhibit L13, we are looking at the probability that the true effect is less than -0.10 standard deviations. Looking at the red circle dot where the x-axis takes a value of 0.50, we see that the y-axis value is about 0.62. That means that the posterior probability estimated using the normal prior rather than the skewed prior is, in this case, about 12 percentage points too big. That is close to the worst case scenario—there are other examples where the probabilities are similar.

Our takeaway is that using the skewed generalized t-distribution makes a noticeable difference in our understanding of the distribution of intervention effects. We conclude that it is worth the added complexity.

Exhibit L12. Distribution of all intervention effects in the WWC, estimated using either the normal or skewed generalized t-distribution

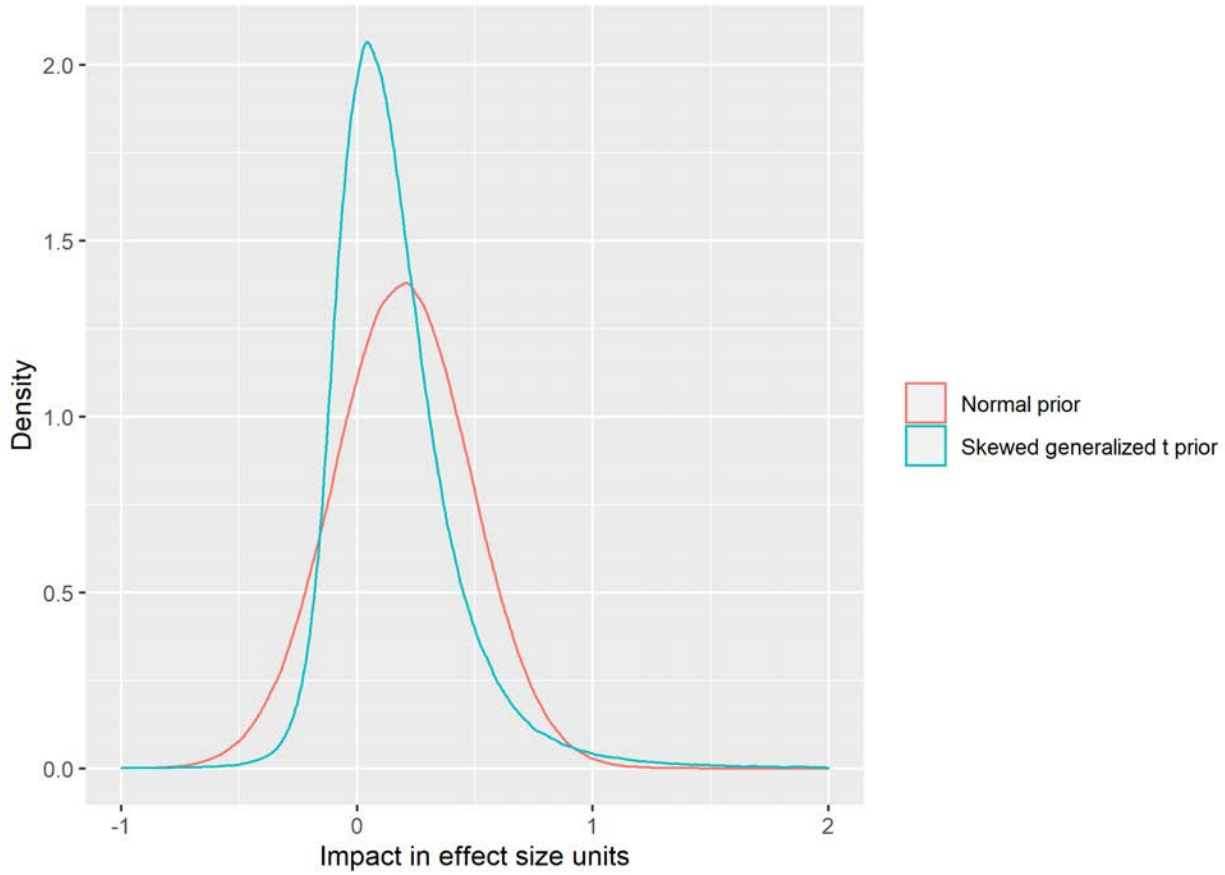
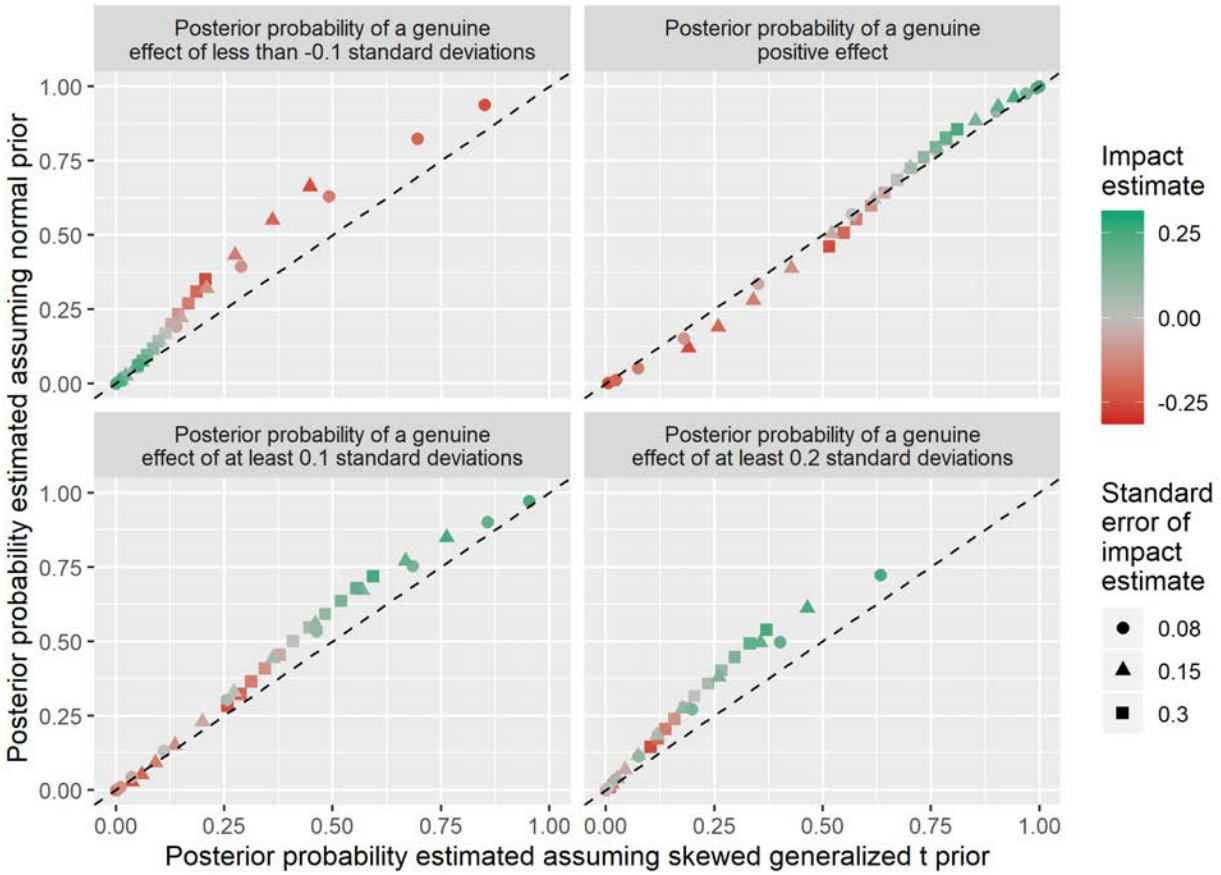


Exhibit L13. Sensitivity of select posterior probabilities to using either the normal or skewed generalized t-distribution



Return to BASIE Step 1

Local Stop: Misinterpretations to avoid

Though Bayesian posterior probabilities provide more useful information than p-values do, we do not mean to suggest that posterior probabilities are immune to misinterpretation. These probabilities have a specific meaning and can be misinterpreted. Misinterpretation may be especially likely when the reader has in mind a specific probability of interest to them that is different from the one reported. In the absence of a clear explanation of what the reported probability means, readers may assume that it means what they would most like it to mean (that often happens with p-values). We provide three examples of possible misinterpretations.

First, readers of an evaluation report might want to know the probability that an intervention is among the most effective of a specific population of interventions (for example, among the top 25 percent of interventions). If the population they are interested in differs from what is included in the prior, then they cannot calculate their probability of interest. For example, they might want to know the probability an intervention is among the most effective interventions designed to improve geometry test scores for 10th graders, but if the prior includes all interventions designed to improve math achievement outcomes for high school students, then their posterior probabilities will only be relevant to this broader population of evidence.

Second, readers might want to interpret findings from a study that is not drawn from the same population of findings represented by the WWC. For example, perhaps they would like to interpret findings from an education study in France. Readers can still use the prior distributions described in this guide and the spreadsheet tool to interpret such findings, but they should be clear that doing so requires the strong assumption that the findings they are interpreting are from a population like that represented by the WWC. In many cases, that assumption might lead to inferential mistakes of a smaller magnitude than using a prior distribution that is not evidence-based, such as the flat prior (visit Local Stop: Why we do not recommend the flat prior).

Third, readers might want a predictive probability of what an intervention's impact will be in the future. However, the posterior probabilities described in this report are retrospective statements regarding the impact of the evaluated intervention in the context it was evaluated. For example, the findings from a study conducted in the context of Chicago in 2010 might not apply in the context of Austin in 2020. Using Bayesian methods to make predictive probability statements is possible but requires more modeling and assumptions—it does not happen automatically.

Return to BASIE Step 3

Local Stop: Power analysis

In this section, we describe a simulation-based approach to conducting power analysis when impact estimates will be interpreted using posterior probabilities rather than p-values. We also provide example calculations, illustrating how the choice of prior affects study power.

The approach we describe is similar to traditional power analyses in that we can (1) calculate the probability of detecting an effect size of interest and (2) calculate the smallest effect that can be detected with high probability (the minimum detectable effect, or MDE). The difference is that we define "detecting an effect" in terms of a cutoff on a posterior probability rather than a cutoff on a p-value. For example, we could define "detecting an effect" to mean that there is a 95 percent probability that the true effect is greater than zero, given the impact estimate.

Simulation framework for calculating power and the MDE

We use a simulation-based approach to allow for flexibility in both the prior distribution and the distribution of the impact estimate. An R program implementing this approach, along with examples illustrating how to use the program, is available at this location.

By way of comparison, we could use a formula-based approach, but it would require that both the prior and the likelihood follow the same probability distribution family (such as the normal distribution).[9] Thanks to the central limit theorem, impact estimates are approximately normal with large sample sizes. But with small sample sizes, the t-distribution provides a more accurate representation of the distribution of an impact estimate. Regarding the prior, our meta-regression shows that intervention effects are most likely not normally distributed. The central limit theorem does not apply to the distribution of the prior, so we cannot expect that a larger number of findings in the meta-regression will lead to a normal distribution of true intervention effects.

The simulation algorithm for calculating the probability of detecting an effect size of θ_A, given a standard error based on assumptions about key design parameters such as sample size, intraclass correlation, and regression R^2 (all of which should be justified the same as with traditional power analysis), is the following:

1. For simulation replication i, randomly draw an impact estimate, $\hat{\theta}_i$, from the distribution of $\hat{\theta}$ (assuming that the true effect is θ_A). In the R program at this link, the probability distribution for $\hat{\theta}$ is the generalized t-distribution with mean θ_A, scale parameter $\hat{\sigma}$ (which is just the standard error of the impact estimate), and degrees of freedom ν based on the number of randomized units.
2. Calculate the posterior probability distribution for the impact, θ, based on the randomly generated impact estimate $\hat{\theta}_i$, the standard error $\hat{\sigma}$, and the prior distribution for θ. In the R program at this link, the prior distribution for θ is based on skewed generalized t-distributions (the user specifies the location, scale, degrees of freedom, and skew parameters).
3. Using the posterior probability distribution, assess whether an effect has been detected, which in this case means, for example, the posterior probability of a positive effect is at least 95 percent. Record whether an effect has been detected. (Under the NHST, an effect is detected if a p-value is smaller than a cutoff, for example 0.05.)

[9] By same family we mean that the distributions follow the same parametric form, making it possible to derive a formula for the posterior. Distributions that have this property are called conjugate.

4. Repeat steps 1–3 10,000 times and calculate the proportion of times that an effect is detected. That proportion is power.
5. To calculate the MDE, repeat steps 1-4 for different values of θ_A. The value of θ_A that yields the desired power is the MDE.

The most technically complex part of this algorithm is the calculation of the posterior distribution (Step 2). We provide two different options for calculating the posterior distribution. The first option uses Stan. The second option uses a more sophisticated (and faster) version of the simple algorithm shown in Exhibit L4. The second option does not require installation of Stan (or a C compiler, which Stan requires).

Examples

In Exhibit L14, we show some examples of MDEs for different priors, different criteria for detecting an effect, and two different study sample sizes. We examine the prior for the overall population of intervention effects in the WWC with our adjustment for small-study effects, centered at zero, and without the bias adjustment for small-study effects. We focus on varying these parameters because they are the ones that make the greatest difference in MDEs—varying outcome domain and school-level has negligible effects.

Some key takeaways:

- Sensitivity of the MDE to the prior is much smaller in larger studies. In a study that randomizes 40 schools,[10] the MDEs are virtually identical across the three priors examined. In a study that randomizes just 10 schools, the MDEs are more noticeably different. This is consistent with what we saw in Exhibit 4—posterior probabilities are more sensitive to the prior when the standard error is larger.
- In smaller studies, the zero-centered prior has the largest MDE and the prior without an adjustment for small-study effects has the smallest MDE. This makes sense: among the priors examined here, the zero-centered prior is most pessimistic regarding the potential for large effects while the prior without an adjustment for small-study effects is most optimistic.
- Holding study size constant, the most important factor affecting the MDE is the criterion used to detect an effect. Recall that the criterion used to detect an effect is a cutoff on a posterior probability. For example, we could define "detecting an effect" to mean "the probability of a positive effect is at least 95 percent."

Exhibit L14. Examples of how MDEs vary by prior distribution

	MDE for four effect detection criteria:			
Prior distribution	**80% chance impact is > 0**	**90% chance impact is > 0**	**95% chance impact is > 0**	**97.5% chance impact is > 0**
Large study—40 schools (20 treatment, 20 control)				
Overall WWC prior adjusted for small-study effects	0.10	0.13	0.15	0.17
Overall WWC prior, centered at zero	**0.10**	**0.13**	**0.15**	**0.17**
Overall WWC prior, unadjusted for small-study effects	0.09	0.12	0.14	0.16

[10] In this example, we imagine a study where schools were randomized because that is a common design in education evaluation research. As with MDE calculations conducted under the NHST, random assignment of students would yield a smaller MDE.

Prior distribution	MDE for four effect detection criteria: 80% chance impact is > 0	90% chance impact is > 0	95% chance impact is > 0	97.5% chance impact is > 0
Small study—10 schools (5 treatment, 5 control)				
Overall WWC prior adjusted for small-study effects	0.19	0.26	0.32	0.39
Overall WWC prior, centered at zero	0.22	0.29	0.36	0.44
Overall WWC prior, unadjusted for file small-study effects	0.16	0.23	0.28	0.35

Note: For all MDE calculations, we assume 100 students per school, an intraclass correlation of 0.15, school-level R^2 of 0.8, and student-level R^2 of 0.4.

Return to BASIE Step 4

Local Stop: Monte Carlo simulation approach used by the BASIE probability tool

The prior distributions included in the spreadsheet tool are not normal distributions, meaning that you cannot calculate posterior probabilities using formulas. For these prior distributions, we use a Monte Carlo simulation in which we draw a large number of values from the prior distribution, randomly draw a large number of impact estimates (each of which is centered at one of the previously drawn values from the prior), and then use local regression to calculate the posterior distribution conditional on an impact estimate. Unlike Markov Chain Monte Carlo (MCMC), which is typically used for Bayesian analysis, this simpler algorithm can be implemented in a standard Excel spreadsheet without any extra add-on software or Visual Basic programs. The spreadsheet will thus work on a wide range of devices, from smartphones to desktop computers. A simplified version of this algorithm is illustrated in Exhibit L3 and detailed in Exhibit L4. The full version of the algorithm is provided in this R program. In the full version, the bandwidth around the impact estimate (illustrated by the vertical red lines in Exhibit L3) has a half-width of 0.05 standard deviations.

We have validated this algorithm for all 108 prior distributions included in the tool. For each prior distribution, we calculated 15 posterior probabilities (the same 15 reported in the spreadsheet tool) using both the algorithm described above and the software Stan (Gelman et al., 2015) for each of five different impact estimates and for each of three different standard errors, for a total of 108*15*5*3 = 24,300 calculated posterior probabilities. The five impact estimates correspond to the 1st, 25th, 50th, 75th, and 99th percentiles of the selected prior distribution. The three standard errors are 0.01, 0.10, and 0.25. For each of the 24,300 probabilities, we recalculated 1,000 times using both Stan and our algorithm and compared the means of the recalculations (this provides an estimate of the expected value of the posterior probability from each method). Across the 24,300 comparisons, the largest difference in posterior probabilities between our algorithm and Stan was 0.009 (that is, nine tenths of a percentage point). The average absolute difference was 0.0003 and the 90th percentile was 0.001. From this validation exercise we conclude that education researchers can trust the accuracy of the spreadsheet tool. The R program used to conduct this validation exercise is available here.

Return to BASIE Probability Tool

References

Bloom, H. S., Hill, C. J., Black, A. B., and Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, 1(4), 289-328.

Carleton R. N. (2016). Into the unknown: A review and synthesis of contemporary models involving uncertainty. *Journal of Anxiety Disorders*, 39, 30–43.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences (1st ed.)*. New York, NY: Academic Press.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management,* 27(4), 724–750.

De Berker, A. O., Rutledge, R. B., Mathys, C., Marshall, L., Cross, G. F., Dolan, R. J., & Bestmann, S. (2016). Computations of uncertainty mediate acute stress responses in humans. *Nature Communications,* 7, 10996.

Deke, J., & Finucane, M. (2019). *Moving beyond statistical significance: the BASIE (BAyeSian Interpretation of Estimates) framework for interpreting findings from impact evaluations* (OPRE Report 2019-35). U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation.

Deming, W. E. & Stephan, F. F. (1941). On the interpretation of censuses as samples. *Journal of the American Statistical Association,* 36(213), 45-49.

Duval, S. J., & Tweedie, R. L. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-regression. *Journal of the American Statistical Association,* 95, 89–98.

Freedman, B. (1987). Equipoise and the Ethics of Clinical Research. *New England Journal of Medicine*, https://www.nejm.org/doi/full/10.1056/NEJM198707163170304.

Garvey, P. R. (2001). Implementing a risk management process for a large scale information system upgrade - A case study." *Insight*, 4(1), 1-12.

Gelman, A., & Weakliem, D. (2009). Of beauty, sex and power: Too little attention has been paid to the statistical challenges in estimating small effects. *American Scientist,* 97(4), 310-316.

Gelman, A. (2005, June 23). A Bayesian wants everyone else to be non-Bayesian. https://statmodeling.stat.columbia.edu/2005/06/23/a_bayesian_want/

Gelman, A. (2011). Induction and deduction in Bayesian data analysis. Special topic issue, Statistical science and philosophy of science: Where do (should) they meet in 2011 and beyond? *Rationality, Markets and Morals*, 2, 67–78.

Gelman, A. (2015, July 15). Prior information, not prior belief. http://andrewgelman.com/2015/07/15/prior-information-not-prior-belief/

Gelman, A. (2016, April 23). What is the "true prior distribution"? A hard-nosed answer. http://andrewgelman.com/2016/04/23/what-is-the-true-prior-distribution-a-hard-nosed-answer/

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6), 641–651.

Gelman, A., & Hennig, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 180(4), 967–1033.

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A., Lee, D. & Guo, J. (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Education and Behavioral Statistics,* 40(5), 530-543.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist,* 102(6), 460-465.

Gelman, A., & Shalizi, C. (2013). Philosophy and the practice of Bayesian statistics (with discussion). *British Journal of Mathematical and Statistical Psychology,* 66, 8–80.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis (3rd ed.)*. CRC Press.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review,* 102, 684–704.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, P., Goodman, S. N. &. Altman, D.G. (2016). Statistical Tests, p-Values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology,* 31(4), 337–350.

Hansen, C., McDonald, J. B., & Newey, W. K. (2010). Instrumental variables estimation with flexible distributions. *Journal of Business and Economics,* 28(1), 13-25.

Hedges, L. V. (1992). Modeling publication selection effects in meta-regression. *Statistical Science,* 7(2), 246–255.

Herrmann, M., Clark, M., James-Burdumy, S., Tuttle, C., Kautz, T., Knechtel, V., Dotter, D., Wulsin, C. S., & Deke, J. (2019). *The effects of a principal professional development program focused on instructional leadership* (NCEE 2020- 0002). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives,* 2(3), 172–177.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLOS Medicine,* 2(8), e124.

Kaplan, D. (2021). On the quantification of model uncertainty: a Bayesian perspective. *Psychometrika*, 86, 215-238.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4), 604–620.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. U.S. Department of Education, Institute of Education Sciences, National Center for Special Education Research.Marks-Anglin A, & Chen Y. (2020). A historical review of publication bias. *Res Synth Methods*, 11(6), 725–742.

Moss, J. (2020). publipha: Bayesian meta-regression with publications bias and p-hacking [R package version 0.1.1]. https://CRAN.R-project.org/package=publipha

Moss, J., & De Bin, R. (2019). Modelling publication bias and p-hacking. arXiv. https://arxiv.org/abs/1911.12445

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. Psychological Bulletin, 86(3), 638-641.

Royston, J.P. (1982). Expected normal order statistics (exact and approximate). *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(2), 161–165.

Rücker G., Schwarzer G., Carpenter J. R., Binder H., & Schumacher M. (2011): Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. *Biostatistics,* 12, 122–42.

Schochet, P. Z. (2016). *Statistical theory for the RCT-YES software: Design-based causal inference for RCTs (2nd ed.; NCEE 2015–4011)*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development.

Schwarzer, G., Carpenter, J. R., & Rücker, G. (2020). metasens: Advanced statistical methods to model and adjust for bias in meta-regression [R package version 0.4-1]. https://CRAN.R-project.org/package=metasens

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software,* 36(3), 1–48. https://www.jstatsoft.org/v36/i03/

Wasserstein, R.L., & Lazar, N.A (2016). The ASA's statement on p-Values: Context, process, and purpose. *The American Statistician,* 70(2), 129–133.

Weidmann, B. & Miratrix, L. (2021). Lurking inferential monsters?: Quantifying selection bias in evaluations of school programs. *Journal of Policy Analysis and Management.*

What Works Clearinghouse. (2014). *Assessing attrition bias—Addendum (version 3.0)*. U.S. Department of Education, Institute of Education Sciences.

Zurovac, J., Cook, T. D., Deke, J., Finucane, M. M., Chaplin, D., Coopersmith, J. S., Barna, M., & Forrow, L. V. (2021). Absolute and relative bias in eight common observational study designs: Evidence from a meta-analysis. arXiv:2111.06941v2

Appendix A. Additional details regarding our method for adjusting prior evidence for small-study effects

Our method to adjust for bias due to small-study effects was described in Local Stop: Adjustments for small-study effects. In this appendix we cover two topics First, we explain how we estimate the variance of the maximum order statistic that is needed for the Bayesian meta-regression. Second, we describe the simulations we used to assess the efficacy of our adjustment method compared to other available methods in the literature.

Estimating the variance of the maximum order statistic

As described in Local Stop: Adjustments for small study effects, we represent the variance of the maximum order statistic as a function $h(s_i, d)$, where s_i is the reported standard error of the impact estimate and d is the number of draws from the distribution of the impact estimate (if d equals one, then only one draw was taken from the distribution of the impact estimate, that one draw was reported, and so there are no findings left unreported). We need to calculate this variance as part of our Bayesian meta-regression analysis of the WWC database. Because this variance must be calculated a large number of times, we need a method for calculating it with low computational cost. In this section, we describe that method.

Our approach to estimating $h(s_i, d)$ involves three steps. First, we used a Monte Carlo simulation to calculate the variance of the maximum order statistic for each of 1 through 100 draws from the standard normal distribution. We used 10 million replications in this simulation to ensure precise estimates. Second, we regressed the precision (that is, the inverse of the variance) of the maximum order statistic on a high-order polynomial of d and saved the coefficients.[11] Third, we used the coefficients from the polynomial regression to calculate the variance of a maximum order statistic for a specified number of draws. The advantage of this approach is that the first two steps only needed to be run once. When we have the coefficients, we can calculate $h(s_i, d)$ for any value of s and d (the key variable is d; s is just a scaling factor). The first two steps are implemented in the R code in Exhibit A1. The regression coefficients estimated by the code below are then included in our Bayesian meta-regression.

[11] If d exceeds 100, then predictions made using coefficients from the high-order polynomial regression would be unreliable. With the WWC data, the implied estimate of d is about 1.7, so this is not an issue for our analysis. But if this approach is applied in a context with a much stronger correlation between impact estimates and their standard errors, then the range of support for the polynomial regression might need to be expanded beyond 100.

Exhibit A1. R code to estimate $h(s_i, d)$

```
library(parallel)
if(F){
#these commands are wrapped in if(F) to prevent accidental execution
#first, use Monte Carlo to calculate the variance of the maximum order statistic
#for 1-100 draws from the normal distribution with mean mu and standard deviation
 sigma
d <- 1:100
mc.data <- array(NA,c(length(d),2))
colnames(mc.data) <- c("d","v")
mc.data[,"d"] <- d
mc.data[,"v"] <- mcmapply(max.stats.mc,reps=10000000,d=d,mu=0,sigma=1,mc.cores=8)
mc.data <- as.data.frame(mc.data)

#second, we estimate a polynomial regression of the relationship between the
# precision of the maximum order statistic (1/v) and the number of draws (nd)
fit <- lm(1/v ~
 d+I(d^2)+I(d^3)+I(d^4)+I(d^5)+I(d^6)+I(d^7)+I(d^8)+I(d^9)+I(d^10)+I(d^11)+I(d^12)
 ,data=mc.data)

#using that estimate, we can quickly calculate the variance of the maximum order
 statistic
# for a specified number of draws, for example:
h(s=1,d=5,b=fit$coef)

}

h <- function(s,d,b){
 x <- d^seq(0,length(b)-1,1)
 Vinv <- b%*%x
 V_max <- (s^2)/Vinv
 return(V_max)
}

max.stats.mc <- function(reps=100000,d,mu,sigma){
 max.stats <- array(NA,reps)
 for(i in 1:reps){
    max.stats[i] <- max(rnorm(d,mean=mu,sd=sigma))
 }
 return(var(max.stats))
}
```

Simulation study

Several adjustment procedures address bias due to selective reporting of findings, but we prefer our method for three reasons. First, it is easily integrated into our Bayesian meta-regression model. Second, because it is focused on the impact estimate rather than p-values, we believe it will continue to be relevant as researchers move away from NHST. Third, our method appears to perform well compared to alternative methods based on a simulation study. Here we describe the alternative methods we examined, the setup of the simulation study, and report findings.

Alternative methods

We compare our adjustment method to six adjustment procedures from the literature:

1. **Duval and Tweedie (2000): Trim-and-fill.** Implemented in the R package metafor (Viechtbauer, 2010).

2. **Moss and De Bin (2019): A model for p-hacking.** Implemented in the R package publipha (Moss, 2020).
3. **Hedges (1992): A model for publication bias.** Implemented in the R package publipha (Moss, 2020).
4. **Rücker et al. (2011): Limit meta-regression [method = beta0].** Implemented in the R package metasens (Schwarzer et al., 2020).
5. **Rücker et al. (2011): Limit meta-regression [method = betalim].** Implemented in the R package metasens (Schwarzer et al., 2020).
6. **Rücker et al. (2011): Limit meta-regression [method = mulim].** Implemented in the R package metasens (Schwarzer et al., 2020).

Simulation setup

The simulation algorithm consists of three steps, which are repeated 1,000 times:

1. **Randomly generate genuine effects for a sample of 200 interventions.** These effects are drawn from a prior distribution. For this simulation, we draw the effects from the normal prior distribution with mean 0.05 and standard deviation 0.25.
2. **For each intervention, generate multiple impact estimates and select which one to report.** The number generated varies depending on the selection mechanism. The selection mechanisms are described in Exhibit A2. The standard errors of the impact estimates range from about 0.02 to 0.22 (in effect size units).
3. **Estimate the prior distribution of genuine effects by conducting a meta-regression of the simulated literature, applying separately each of the six adjustment methods described above.** Record the mean and standard deviation of intervention effects estimated by the meta-regression and adjustment method. A successful adjustment method will correctly estimate the mean and standard deviation of genuine effects to be 0.05 and 0.25, respectively.

Exhibit A2. Selection mechanisms

Selection mechanism	Description
Select the most favorable impact estimate among multiple independent samples	
Two unreported findings	Three independent impact estimates are calculated and the most favorable is reported.
Five unreported findings	Six independent impact estimates are calculated and the most favorable is reported.
Random (1-10) unreported findings	A random number (between 1 and 10) of independent impact estimates are calculated and the most favorable is reported.
p-hacking (up to 10)	The researcher keeps calculating independent impact estimates (up to 10) until one is found with $p < 0.05$. The impact estimate with the lowest p-value is reported.
Impact estimate hacking	The researcher keeps calculating independent impact estimates (up to 10) until an effect size of at least 0.20 is found. The most favorable impact estimate is reported.
Same sample, adjust for up to four different covariates[a]	
Select maximum impact estimate, covariates correlated with outcome	The correlation between each covariate and the outcome is 0.25; the researcher reports the most favorable impact estimate.
Select maximum impact estimate, not correlated with outcome	The covariates are uncorrelated with the outcome; the researcher reports the most favorable impact estimate.
Select maximum t-statistic, covariates correlated with outcome	The correlation between each covariate and the outcome is 0.25; the researcher reports the impact with the most favorable t-statistic.
Select maximum t-statistic, not correlated with outcome	The covariates are uncorrelated with each other and the outcome; the researcher reports the impact with the most favorable t-statistic.

[a] Using the same sample, the researcher calculates multiple impact estimates by varying the baseline covariates included in a linear regression analysis. The covariates are uncorrelated with each other. Every possible combination of four covariates is examined, excluding interactions or higher order terms.

Simulation findings

Simulation findings are reported in Exhibits A3 (tabular) and A4 (graphical). In these exhibits, we report the estimated mean and standard deviation of intervention effects estimated via meta-regression of simulated literatures subject to the selection mechanisms described in Exhibit A2. The estimates vary by the method used to control for selection bias. In all cases, the true mean and standard deviation of effects from these simulated literatures are 0.05 and 0.25.

Some key takeaways:

- **For all selection mechanisms considered here, the unadjusted means are too high.** The true mean is 0.05, but the unadjusted means range from 0.06 to 0.21, depending on the selection mechanism.
- **Unadjusted standard deviations tend to be too small.** The true standard deviation of intervention effects is 0.25, but the unadjusted standard deviation ranges from 0.19 to 0.25. The selection mechanisms that lead to the largest underestimate of the standard deviation are p-hacking and impact estimate hacking.
- **Two adjustment methods are more accurate than the unadjusted standard deviation.** For all the selection mechanisms considered here, Moss and De Bin (2019) and our method yield a more accurate standard deviation than the unadjusted standard deviation.
- **Our method and the Rücker et al. (2011) "betalim" method appear to perform very well.** These two methods are very similar. The biggest difference we see is that our method is more

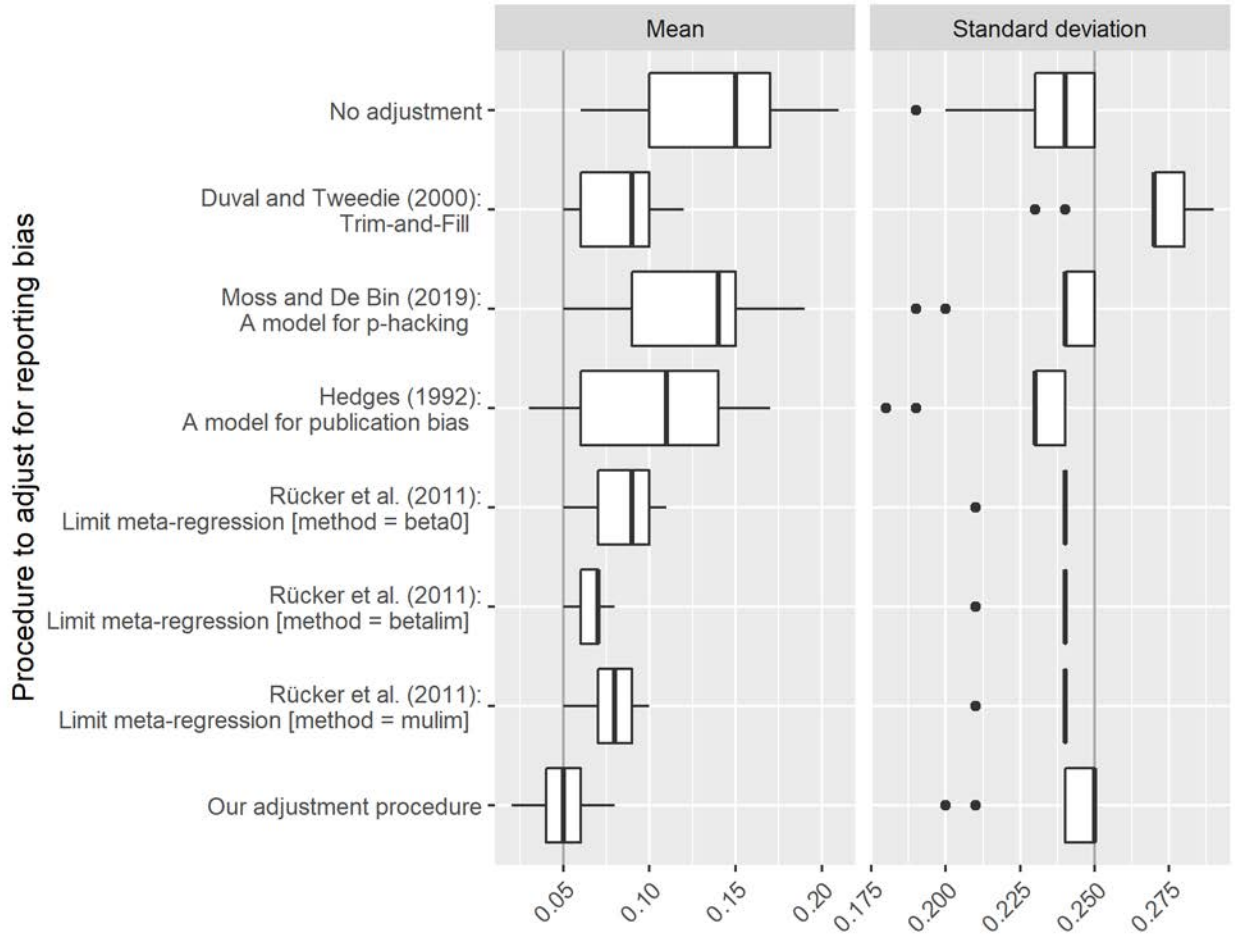
likely to understate the mean whereas Rücker's is more likely to overstate the mean. This is especially evident under the p-hacking and impact estimate hacking scenarios. Under these scenarios, we estimate means of 0.02 and 0.04, whereas Rücker estimates means of 0.06 and 0.07 (recall, the true mean is 0.05). With respect to the standard deviation, our method appears to be more accurate, but the difference is small.

Exhibit A3. Tabular comparison of procedures to adjust for reporting bias

True mean is 0.05; true standard deviation is 0.25	Selection mechanism								
	Select the most favorable impact estimate among multiple independent samples					Same sample, adjust for up to four different covariates			
						Select maximum impact estimate		Select maximum t-statistic	
	Two unreported findings	Five unreported findings	Random (1-10) unreported findings	p-hacking (up to 10)	Impact estimate hacking	Covariates correlated with outcome	Not correlated with outcome	Covariates correlated with outcome	Not correlated with outcome
No adjustment									
Mean	0.15	0.21	0.17	0.18	0.17	0.10	0.06	0.10	0.06
Standard deviation	0.23	0.23	0.24	0.20	0.19	0.25	0.25	0.24	0.25
Duval & Tweedie (2000): Trim-and-Fill									
Mean	0.09	0.12	0.10	0.11	0.10	0.06	0.05	0.08	0.05
Standard deviation	0.28	0.29	0.28	0.24	0.23	0.28	0.27	0.27	0.27
Moss & De Bin (2019): A model for p-hacking									
Mean	0.14	0.19	0.16	0.14	0.15	0.09	0.05	0.09	0.05
Standard deviation	0.24	0.24	0.24	0.20	0.19	0.25	0.25	0.24	0.25
Hedges (1992): A model for publication bias									
Mean	0.12	0.17	0.14	0.11	0.14	0.07	0.03	0.06	0.03
Standard deviation	0.23	0.23	0.23	0.19	0.18	0.24	0.24	0.24	0.24
Rücker et al. (2011): Limit meta-regression [method = beta0]									
Mean	0.09	0.11	0.10	0.08	0.10	0.07	0.05	0.09	0.05
Standard deviation	0.24	0.24	0.24	0.21	0.21	0.24	0.24	0.24	0.24
Rücker et al. (2011): Limit meta-regression [method = betalim]									
Mean	0.07	0.08	0.07	0.06	0.07	0.06	0.05	0.08	0.05
Standard deviation	0.24	0.24	0.24	0.21	0.21	0.24	0.24	0.24	0.24
Rücker et al. (2011): Limit meta-regression [method = mulim]									
Mean	0.09	0.10	0.09	0.08	0.09	0.07	0.05	0.08	0.05
Standard deviation	0.24	0.24	0.24	0.21	0.21	0.24	0.24	0.24	0.24
Our adjustment procedure: Reported impact estimates are maximum order statistic									
Mean	0.06	0.06	0.05	0.02	0.04	0.05	0.04	0.08	0.05
Standard deviation	0.24	0.24	0.25	0.21	0.20	0.25	0.25	0.25	0.25

Note: Findings based on a simulation with 1,000 replications.

Exhibit A4. Graphical comparison of procedures to adjust for reporting bias



Note: This plot summarizes the information in Exhibit A3. For each procedure (the sub-panes in Exhibit A3), we use a box and whisker plot to show the distribution of estimated mean and standard deviations across the selection mechanisms (columns in Exhibit A3). The boxes indicate the middle 50 percent of the distribution, and the whiskers extend from the boxes up to 1.5 times the interquartile range. Any points outside the range of the whiskers are plotted individually. The true values of the mean and standard deviation are indicated by a solid vertical line at 0.05 for the mean and 0.25 for the standard deviation.

Appendix B. Uncertainty arising from less rigorous designs

Findings from high quality RCTs are subject to random error—differences between the treatment and control groups that arise from the random assignment process itself. Because these chance differences arise from the randomized design of the study, statistical theory allows us to control the probability that we make mistakes based on these random errors. The errors in experimental findings are known unknowns.

The problem with most nonexperimental methods, and RCTs with sample attrition, is that they are subject to systematic (not random) errors that we do not typically fully understand. The technical term for this type of error is bias. These errors are (or at least seem to be) unknown unknowns—we typically cannot fully control the probability of making mistakes based on these errors.

For example, in a matched comparison group design (MCGD), we typically do not know why some people participate in a program while others do not. We can statistically adjust for pre-intervention differences between program participants and nonparticipants that we can see in data, but we cannot control for what we cannot see. Not only might there be errors in the findings from these studies, but we also do not know how large those errors might be.

Typically, the quantitative measures we use to assess uncertainty in research findings (standard errors, p-values, confidence intervals, and posterior probabilities) are based only on uncertainty due to the known unknowns of random errors. These measures typically ignore the unknown unknowns of bias. This means that a p-value or posterior probability from an RCT is not really comparable to a p-value or posterior probability from a MCGD unless we make the huge assumption that estimates from the MCGD are unbiased. But how useful is it to decision makers to report uncertainty measures that are accurate only if we make assumptions that we know are very likely wrong and wrong by an unknown degree?

If we truly know nothing about the potential for bias to affect estimates from nonexperimental studies, then there would be no value in nonexperimental studies. For example, if it is just as likely that bias in a nonexperimental effect size estimate could be -100 standard deviations, +100 standard deviations, or anything else, then the noise of bias in nonexperimental studies would completely drown out the signal of genuine program effects (almost all of which are likely smaller than one standard deviation, according to our meta-regression of the WWC database).

However, if we actually do know something about the likely magnitude of bias in nonexperimental studies, then it could be valuable to incorporate that information into quantitative measures of uncertainty. In this section, we describe sources of information regarding the distribution of bias in MCGDs and in RCTs with high attrition. We then show how that information can be incorporated into posterior probabilities.

The bottom line is—we have information about the distribution of bias and the methodological tools to make good use of that information. However, there is much more work to be done in this area. The evaluation field needs more and larger meta-analyses of bias estimates.

Sources of information regarding the potential magnitude of bias

Though we have much more to learn, it turns out that we actually do know something about the distribution of bias in MCGDs and RCTs with high attrition. In the case of MCGDs, we have the literature on within study comparisons, which very recently have become the target of important meta-analyses. In the case of RCTs with high attrition, we have the model of attrition used by the WWC and a ready source of data that could be meta-analyzed. We discuss each source of information in turn.

Within study comparisons

Converting the unknown unknowns in findings from MCGDs into something more akin to the known unknowns in experimental findings has been the focus of a growing body of methodological investigations known as within study comparisons (LaLonde, 1986; Cook et al., 2008). In a within study comparison, evaluators contrast the impact estimate from an RCT to that from an MCGD that shares the same treatment group. A single within study comparison tells us relatively little but, taken collectively, a meta-regression of bias estimates from within study comparisons can turn what was an unknown unknown into a known unknown. That is, we can estimate the distribution of bias. Weidmann and Miratrix (2021) provide an example of a meta-regression of within study comparisons in the context of English primary schools. They found bias centered at about zero (in effect size units) with a standard deviation of 0.04. Using a Bayesian meta-regression of 39 within study comparisons spanning multiple fields, Zurovac et al. (2021) calculated the distribution of bias in studies that include up to three design elements: (1) adjustment for a pre-test, (2) adjustment for a rich set of covariates (in addition to the pre-test), and (3) use a 'local' comparison group. For studies with all three design elements the distribution of bias was centered at 0.01 and the standard deviation was 0.07. As more meta-analyses of this type are conducted across a diversity of contexts, we will have an increasingly accurate understanding of the distribution of bias in MCGDs and how that distribution varies across contexts.

Distribution of differences on observable variables

One approach to estimating the distribution of bias in RCTs with high attrition could be to assume that differences on unobserved variables between those who remain in the treatment and control groups at the end of a study are exchangeable with differences on observable variables. This is the assumption underpinning the statistical model behind the WWC attrition standard (WWC, 2014). Specifically, the WWC attrition model assumes that the observed differences in baseline test scores between those remaining in the treatment and control groups at the end of a study are representative of the unobserved differences between those groups.

We are unaware of any meta-analyses of baseline differences between treatment and control group members who remain in the study samples of RCTs at follow-up. Were such meta-analyses to be conducted, they could inform an estimate of the distribution of attrition bias in RCTs analogous to the estimates of the distribution of selection bias in MCGDs enabled by the within study comparison meta-analyses described above.

Incorporating uncertainty due to bias into posterior probabilities

Once we have an estimate of the distribution of bias, we can treat bias as another error term, b, contributing uncertainty to our estimate of the treatment effect. Using the superpopulation framework, we imagine that we have randomly drawn our study from a large population of studies and that the bias affecting our study is also drawn from a large population of study biases. With this extra error term, the standard error of the impact estimate from the simple difference in means between a treatment and comparison group becomes the following:

$$se = \sqrt{\frac{\sigma_T^2}{N_T} + \frac{\sigma_C^2}{N_C} + \sigma_b^2}$$

where σ_T^2 and σ_C^2 are the variances of the outcome in the treatment and comparison groups, NT and NC are the sample sizes in the two groups, and σ_b^2 is the variance of bias. Note that unlike

uncertainty due to random error, the uncertainty introduced by bias is not diminished by sample size. This means that the uncertainty due to bias looms relatively larger in the largest studies.

As an example, consider a study with 800 individuals in the treatment group, 800 in the comparison group, and $\sigma_T = \sigma_C = 1$. In the absence of uncertainty due to bias, the standard error would be 0.05. Using the estimate of $\sigma_b = 0.04$ from Weidmann and Miratrix (2021), the standard error becomes $\sqrt{0.05^2 + 0.04^2} = 0.064$. If an evaluator sought to increase the sample size of the study enough to achieve a standard error of 0.05 despite the uncertainty from bias, the size of the study would need to be increased from 1,600 to approximately 4,400 (because $\sqrt{\frac{1}{2200} + \frac{1}{2200} + 0.04^2} = 0.05009$).

Incorporating the uncertainty from bias into a posterior probability is a simple matter of using this bias-corrected standard error in place of the usual standard error (assuming the mean of bias is zero; if the mean of bias is different from zero, then we can adjust the impact estimate by subtracting from it the mean of bias). Continuing the example, instead of entering a standard error of 0.05 into the spreadsheet tool described previously, we can enter 0.064.