INFORMING SPECIFIC DECISIONS WITH RIGOROUS EVIDENCE

DESIGNING AND ANALYZING DECISION FOCUSED EVALUATIONS

January 2021

IDinsight

DATA. DECISIONS. DEVELOPMENT.

Authors

Torben Fischer: Torben.Fischer@IDinsight.org
Doug Johnson: dougj892@gmail.com
Daniel Stein: Daniel.Stein@IDinsight.org

Acknowledgements

We thank Eva Vivalt, Richard Traunmüller, Sindy Li for helpful comments. We appreciate comments from our colleagues at IDinsight including Ruth Levine, Paul Wang, Jeffrey McManus, Heather Lanthorn, Valentina Brailovskaya, and Emily Coppel. We welcome further comments and thoughts to torben.fischer@idinsight.org. All errors remain our own.

About IDinsight

IDinsight uses data and evidence to help leaders combat poverty worldwide. Our collaborations deploy a large analytical toolkit to help clients design better policies, rigorously test what works, and use evidence to implement effectively at scale. We place special emphasis on using the right tool for the right question, and tailor our rigorous methods to the real-world constraints of decision-makers.

IDinsight works with governments, foundations, NGOs, multilaterals and businesses across Africa and Asia. We work in all major sectors including health, education, agriculture, governance, digital ID, financial access, and sanitation.

We have offices in Bengaluru, Dakar, Johannesburg, Lusaka, Manila, Nairobi, New Delhi, San Francisco, and Washington, DC. Visit www.IDinsight.org and follow on Twitter @IDinsight to learn more.

CONTENTS

EXECUTIVE SUMMARY

Impact evaluations of development interventions have increased dramatically over the past 20 years,[1] expanding from a research tool of academics that has recently been awarded with a Nobel Prize[2] to a decision-making tool by policy-makers. Despite this expansion in use cases, the methodological approach to design and analyze impact evaluations has remained mostly constant. This standard approach tends to test whether a program works, i.e. whether its effect is different than zero. Conclusions from this test implicitly assume that consumers of the research are an academic audience that is interested in generalizable knowledge and skeptical of any evaluation results. Therefore, the standard approach requires a relatively high level of certainty to convince the reader that results are "true."

We argue that in cases where the purpose of the evaluation is to inform a specific decision, researchers should consider alternative approaches to design and analyze impact evaluations. The unifying feature of the alternative approaches to design and analyze impact evaluations we discuss is that they explicitly consider the specific decision-makers' circumstances and decision framework. While this approach isn't necessarily new, we hope to provide practitioners with an accessible and practical treatment of the subject.

Specifically, we outline two approaches. In the first, we retain the standard frequentist statistical approach to impact evaluations, but outline how certain "default" parameters can be modified to take specific decision frameworks into account. For instance, in certain cases, decision-makers may be Ok implementing a policy even with relatively high uncertainty as to its effectiveness. Second, we show how Bayesian analysis may more directly account for a decision-maker's beliefs and preferences. We give an overview of a Bayesian approach to evaluation and illustrate how to implement it in practice, including hands-on guidance regarding sample size calculations, analysis, and interpretation of results. Finally, we discuss how an evaluator can choose between frequentist or Bayesian approaches.

The high-level takeaways from this paper are as follows:

- There are a number of common circumstances where the standard frequentist approach to impact evaluation is not ideal for decision-making. These scenarios include testing for "non-inferiority" of a new policy vs the status quo, testing two policies head to head, and making a decision based on a certain effect size threshold.
- When designing an evaluation for decision-making in the frequentist framework, the researcher should consider carefully the decision-relevant level of certainty (size) of the test, which may be different from the standard .05. They also may want to consider whether a one-sided test or multiple hypothesis tests are appropriate.
- Bayesian analysis has distinct advantages in a decision framework, as it can quantitatively incorporate the decision-makers' prior beliefs into the estimation. It also allows evaluators to make easily understood probabilistic statements such as "the probability of program A being better than program B is 60%."

[1] https://www.3ieimpact.org/blogs/impact-evaluation-still-rise
[2] https://www.nobelprize.org/prizes/economic-sciences/2019/press-release/

- Bayesian analysis tends to give the same results as frequentist analysis in conditions where priors are uninformative and models are fairly simple or sample sizes are sufficiently large. Gains to Bayesian analysis arise primarily in circumstances where interpretation of the conventional approach seems less intuitive, when priors elicited from decision-makers or existing studies from similar contexts contain relevant information, and where sample sizes are restricted.
- In both frequentist and Bayesian analysis, taking the needs of the decision-maker into account can, in many circumstances, result in studies with smaller sample size, while maintaining desired levels of Type 1 error and statistical power. In the frequentist approach, this can result from moving to one-sided tests or accepting lower levels of certainty to drive decisions. In the Bayesian approach, this can result from using tailored decision-rules that better reflect the decision maker's utility function in designing the study.
- Following good research practices, such as the pre-registration of the analytical approach, allows evaluators to maintain high levels of rigor for these adapted approaches to decision-focused evaluations.

The intention of this document is to serve as a guide to those designing evaluations for decision-makers with the intent of allowing for more directed evaluations to ensure maximum policy impact.

1. INTRODUCTION

As impact evaluations gain prominence, evaluators use them more and more not just to find out what policies "work" in a generalizable manner (Cohen & Easterly, 2010), but to help policy-makers make specific decisions.[3] Shah et al (2015) define a Decision-Focused Evaluation (DFE) as an impact evaluation that seeks to inform a "*specific* policy/programmatic decision of a *specific* implementer in a *specific* geography for a *specific* target population over a *specific* time horizon."

The limitations of the "standard" analytical approach to impact evaluations become apparent when the focus of an evaluation shifts from demonstrating "what works" to informing a more specific decision – that is, when doing Decision-Focused Evaluations.[4] For example, suppose a decision-maker commissions an impact evaluation to determine whether a promising, inexpensive new intervention should be scaled to cover a larger geography. The standard evaluation approach would assign individuals (or clusters) from within that geography to participate in the intervention in a random fashion. The decision to scale the program is based on an assessment of whether the program has a positive effect. In the standard approach, this assessment defaults to a two-sided null-hypothesis test with a "null" of no effect and a 5% level of significance.[5] Yet, evaluators might wonder whether these norms are applicable for this specific decision-maker. The following questions may be relevant:

- Would scaling up the intervention be justified with a different standard of evidence if it is low-cost and has a low risk of adverse effects?
- What if the decision-makers *needs* to choose one among multiple policy options (Kasy and Sautmann, 2020)?
- How should evaluators adapt their approach when collaborating with decision-makers' who have well-articulated priors about the effect of the intervention based on their understanding of the local context, previous results or expert opinion (Vivalt, 2017)?[6]
- How should such decision-makers act based on their priors and potentially conflicting evaluation results?

Given these types of questions, we argue that the design and inference of decision-focused evaluations need to be tuned to the needs of decision-makers. This paper illustrates how such adaptions to decision-makers' needs can lead to meaningful deviations from the standard approach to impact evaluation.

We describe three scenarios in which the standard approach may fall short in informing decisions and offer practical guidance on how to employ alternative approaches that produce more actionable evidence. For each of these three scenarios, we elaborate simple but widely applicable models for evaluators to customize and use. The three scenarios are:

[3] We will use policy-maker and decision-maker interchangeably.
[4] We use "standard" and "conventional" approach interchangeably. We recap the standard approach further below.
[5] Brodeur, et al. (2016) report that 85% of the 50,000 tests included in their study use regression coefficients and their associated standard errors.
[6] See Vivalt (Forthcoming.) for a discussion of generalizing results from impact evaluations using a Bayesian approach to meta-analysis.

1. **Testing for equivalence or non-inferiority**: In this scenario the decision-relevant finding is that a tested intervention is *not* different than the status quo.
2. **Head to Head Comparisons:** In this scenario the decision-maker is testing multiple options out of which she has to choose, but there is no "status quo."
3. **Comparing to a Benchmark:** In this scenario the decision-maker wants to take an action only if the effect size of a tested intervention is above/below a certain threshold.

We first consider smaller deviations from evaluation norms, by maintaining the standard frequentist statistical framework, and showing how it can be modified to address each of the three scenarios above. For instance, we consider conducting multiple hypothesis test, changing the certainty threshold for which evidence is deemed actionable (as in Abadie [2020]), or shifting to one-sided hypothesis test.

Next, we introduce Bayesian analysis, and argue that in many circumstances a Bayesian approach will better address a decision-maker's evidence need. We provide a practical guide to using Bayesian analysis to conduct impact evaluations, and also compare it to frequentist analysis. Bayesian analysis gives similar results to frequentist analysis when priors are weak or the sample size is large, but can result in large deviations of the estimate of treatment effect with strong priors and/or a small sample. As a result, evaluators should pre-specify and conduct sensitivity analyses to assess the effect of different priors on the results, akin to robustness checks usually conducted in quasi-experimental research. We also explain how to think about sample size in the Bayesian framework, and how this differs from the frequentist approach. Differences in sample size between Bayesian and frequentist approaches are driven by the distribution of priors, as well as the level of certainty ("sufficiency") used in the Bayesian hypothesis tests.[7]

We argue that Bayesian analysis can provide more actionable evidence because it allows the evaluator to make more easily interpreted probabilistic statements. For example, evaluators using a Bayesian approach would be able to say that "there is a 75% chance of the program reducing stunting by more than five percentage points." Such probabilistic statements can be directly linked to pre-specified decision rules. For example, a decision-maker may want to "scale a program if the probability of it reducing stunting by more than five percentage points is larger than 70%." This practicality of Bayesian approaches is oftentimes perceived to be limited in knowledge-focused evaluations due to the need to specify prior beliefs.[8] We argue that – in a decision-focused evaluation – Bayesian approaches are more feasible because the analysis can directly reflect the decision-maker's priors.[9] However, implementing a Bayesian approach requires gathering detailed information on priors as well as additional modeling assumptions, making it impractical in some circumstances.

This work builds on a long literature on how to design and analyze evaluations in development economics (Duflo et al 2007, Glennerster and Takavarasha 2013, Gerber and Green 2013). Additionally, it adds to the growing literature on using Bayesian analysis for causal inference. Much of the current literature comes from fields where Bayesian statistics are more frequently used, such as political science or clinical studies in medicine and psychology. For instance, Gelman et al (2013)

[7] Code for this example, and all other examples in this paper, can be found at https://github.com/IDinsight/dfe_methods
[8] Priors can be difficult to justify for a general audience, as readers may have a wide array of priors.
[9] We provide high level considerations about how to specify reasonable priors in the appendix.

provides an introduction to Bayesian analysis; Jackman (2004) provide an introduction to Bayesian analysis in political science; Berry et al (2010) discuss applications to clinical trials, and the US Federal Drug Administration (FDA) publishes guidance on how to use Bayesian analysis in medical device trials.[10] Further, Imbens and Rubin (2015) provide a brief introduction to the use of Bayesian models for impact evaluations using the potential outcomes framework.

Despite this ample guidance and numerous applications in other fields, the use of Bayesian analysis in economics and global development research is still in its infancy (BaŞtürk, et al., 2014). Examples of Bayesian analysis in global development include Meager (2019) who estimates the overall impact and heterogeneity of micro-credit interventions conducted in different contexts as well as McKenzie et. al (2018) who plan to compare results from a randomized control trial using frequentist and Bayesian methods.

This work differs from these previous texts in that we seek to offer practical guidance on when evaluators should consider using Bayesian methods to tailor evaluation designs to better inform specific decisions.

This paper will proceed as follows. We first outline the standard approach to impact evaluations, and then discuss the three example scenarios where the standard approach is not ideal. We then discuss how the frequentist approach can be modified to adapt to each of the three example scenarios. We then introduce Bayesian analysis, and provide a practical guide on how to implement Bayesian analysis in impact evaluations, including how to apply them to our example scenarios.

2. THE STANDARD ANALYTICAL APPROACH TO IMPACT EVALUATION

In a canonical example of an impact evaluation, a policy-maker or researcher wants to design a study to understand if a new program "works." For instance, they might want to know if a program providing access to credit increases farmer incomes. They carry out a randomized controlled trial (RCT), in which a randomly selected subset of farmers receives the intervention, to estimate the program's effect on income. If the evaluator observes a statistically significant increase in income among farmers receiving the intervention compared to those who do not, the RCT design would allow her to conclude that the program "worked."

In the standard analytical approach to impact evaluation, the evaluator is interested in the average treatment effect (ATE). The ATE is generally estimated using a simple model:

$$y_i = \alpha + \theta t_i + \varepsilon_i, \quad (1)$$

[10] See https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-use-bayesian-statistics-medical-device-clinical-trials

Where y_i is the outcome variable, t_i is the treatment dummy, and θ is the ATE. With data from a well-executed randomized controlled trial, the OLS estimate, $\hat{\theta}$, is an unbiased estimate of the ATE.[11]

In what we term the standard (or conventional) approach to inference, evaluators would use a two-sided null hypothesis test. The null of this test would be that the program has no effect, and the alternative hypothesis that the program has *some* effect. Formally, the null is that the treatment effect of the intervention is 0 and the alternative hypothesis that the treatment effect is not equal to 0. The level of significance, α, is set to its conventional level of 5% and represents the decision-maker's maximum willingness to tolerate false positives (Fisher, 1992).[12] In formal terms:

$$H_0 : \theta = 0;\ H_1 : \theta \neq 0;\ \alpha = 0.05 \qquad (2)$$

The evaluator analyzes data from the impact evaluation and conducts a test of whether the estimate $\hat{\theta} = 0$. The evaluator rejects the null hypothesis - i.e. concludes that the intervention "works" - if the p-value of this test is less than the specified level of significance, $\alpha = 0.05$. The p-value of the test captures the probability that the observed outcomes could have occurred in the case where $\hat{\theta} = 0$. Therefore, a small p-value means that the data would be unlikely if the treatment effect were actually zero. This approach is ingrained in researchers from years of practice (Brodeur, et al., 2016). We'll refer to this example as the "standard analytical approach to impact evaluation."

It is worth noting the following details about the standard analytical approach:

1. **The approach assumes that there is a status quo of "no program".** In other words, the comparison group is assumed to experience business as usual and the only difference to the intervention group is that – in addition to status quo – the latter participates in the program.[13]
2. **The approach implicitly assumes a high cost of adopting a program when it is no better than "business as usual".** This assumption is captured by the low tolerance of false positives, i.e. the low level of statistical significance conventionally required.
3. **The approach does not explicitly consider the size of the estimated treatment effect**, just whether it is statistically different than zero. Evaluators are encouraged to differentiate between statistically and economically significant effect sizes.[14] Yet, the economic significance of a given effect is generally difficult to quantify.[15] While Cohen (1988) provides broad

[11] The researcher could also calculate the maximum likelihood estimate (MLE) of $\hat{\theta}$ which would give the same coefficient in expectation under certain contitions namely normality of the error term. We will discuss the MLE estimator in further detail below.
[12] A false positive test result declares a program to "work" when in fact it does not. Formally, a false positive test rejects the null-hypothesis when it is in fact true. A false positive is also referred to as a Type 1 error.
[13] Business as usual may well mean that other programs may operate in the same geography. Random assignment to participate in the intervention would still allow to identify the average treatment effect.
[14] McCloskey and Ziliak (1996) provide early evidence on the lack of a differentiation between substantive and statistical significance. Miller and Rodgers (2008) provide guidance on how to describe the differences in research writing.
[15] Refer to Sterck (2018) for an innovative methodology applied to the growth literature.

generalizations, the substantive importance of an effect is often determined relative to empirical benchmarks (Hill, et al., 2008).[16]

The standard analytical approach works well when the goal of impact evaluations is to assess individual interventions and the intended audience is a general observer reading many such impact evaluations. Provided the impact evaluations adhere to good practices[17], this general observer can reasonably conclude that most of the interventions with statistically significant results do in fact "work." This can be especially important as a body of literature builds up on a certain intervention, and readers are trying to gauge whether it will work in a new context. Organizations such as the "What Works Clearinghouse", the Campbell Collaboration or the International Initiative for Impact Evaluations (3ie),"[18] which filter and aggregate impact evaluations, help readers sort through many such impact evaluations so that the general public can know which interventions "work."

We now turn to assess how closely this paradigm maps to the needs of decision-makers who are using evidence generated by an impact evaluation of "their program" to take specific decisions in their context.

3. DECISION-FOCUSED SCENARIOS

In this section, we outline three common decision-focused scenarios for which the standard analytical approach described above may not result in actionable evidence. We will return to these examples throughout the text.

1. TESTING FOR EQUIVALENCE OR NON-INFERIORITY

Imagine a Ministry of Agriculture runs a widespread farmer extension scheme, in which farmers receive a direct provision of inputs (such as fertilizer). The ministry believes this program improves farmer productivity and welfare – but the program is costly and cumbersome to manage. The ministry is considering switching from "business as usual" to a different, nimbler system, in which farmers are given vouchers for purchasing inputs at local shops. Given its lower cost, this new, nimbler system would be considered a success – and therefore worth switching to – if it is "just as good" as the current system. Thus, the decision relevant question is whether – or more precisely how likely - the new nimbler system is at least as effective in increasing farmer productivity and welfare.

In the medical-literature, assessing whether a new system is "just as good" as the current system is known as a "non-inferiority" test (Walker & Nowacki, 2011). As we will discuss below, questions about non-inferiority – or equivalence[19] – of programs compared to "business as usual" cannot be easily

[16] The authors differentiate between three types of benchmarks: normative expectations, policy-relevant changes, and findings from similar studies.
[17] Good research practices make decisions taken as part of the research transparent and result in reproducible results. While the evidence suggests these practices are not always followed, in line with Benjamin et al (2018), we believe that this is not a weakness of the traditional approach to evaluation but its application in practice. This may be improved by adhering to good practices such as pre-registering studies. For an in-depth discussion refer to Christensen, Freese and Miguel (2019).
[18] https://ies.ed.gov/ncee/wwc/; https://campbellcollaboration.org/better-evidence.html; https://developmentevidence.3ieimpact.org/
[19] As we illustrate further below, an equivalence test would – in addition to non-inferiority – also assess whether the new system not better than the existing system.

answered by the standard analytical approach. That is, the hypothesis test of the standard approach has a much more difficult time determining if outcomes are the same versus being different.[20] If the evaluator sets up the standard hypothesis test and fails to reject the null, this is not reliable evidence that the new program is "just as good" as the previous one. How should evaluators design a "non-inferiority" evaluation? Which hypotheses should they be testing? How would they be able to make the desired probabilistic statement about non-inferiority of the new system?

2. HEAD-TO-HEAD COMPARISONS

Suppose the Ministry of Agriculture has been allocated a fixed budget for the following year to implement a program to promote vegetable production. The ministry's officers have narrowed down their policy choice to two possible ideas: distribute free seeds or hire special vegetable-focused extension agents to provide training and advice. Both programs cost roughly the same per beneficiary. The officials need to choose between the two policy options but are unsure ex-ante which will work better. The decision relevant questions therefore are whether – or how likely – one intervention is better than the other – and how effective the preferred intervention is.

Other scenarios that commonly exhibit such "head-to-head" comparisons are A/B testing different implementation or marketing strategies (e.g. Channa et al 2019), benchmarking interventions to cost-equivalent cash transfers (e.g. McIntosh and Zeitlin 2018), or ranking different interventions according to their cost-effectiveness more generally (e.g. Horton, et al. 2017).

Most impact evaluations are designed to compare some new idea or intervention against the null of "business as usual" – and therefore implicitly assume that not changing the status quo is a viable policy option. But what if it is not and the decision-maker has to take some action? Setting up the corresponding hypothesis in the standard approach is challenging because there might be no clear null hypothesis to test against. What approach to inference should the evaluator take when making "head-to-head" comparisons? How can the results of such comparisons be easily interpreted by decision-makers?

3. COMPARING TO A BENCHMARK

Consider a donor who is interested in funding a school feeding program with the goal to increase young childrens' height-for-age z-score (HAZ). The donor is keen to provide funding if the implementer can demonstrate rigorous evidence that the program meets a pre-defined performance metric: an increase in HAZ of at least 0.2 standard deviations. Based on the effects of similar programs, the donor is somewhat skeptical that the intervention will be successful and plans to commission an impact evaluation to test the effect of the new intervention. The decision relevant question therefore is whether – or how likely – the intervention increases HAZ scores by at least 0.2 standard deviations.[21]

[20] Evaluators might consider addressing this problem by increasing the study's sample size to detect smaller effect sizes. For instance, suppose the decision-maker only cared that the new program effect was within 0.1 SD of the old program effect and the study was powered at 90% to detect effect sizes of 0.1SD or larger. If the evaluator failed to reject the null, she could then say that – for many studies replicating this study – at most 10% of studies (1 – Power) would not be "good enough". However, budget or logistical constraints frequently rule out such high-powered tests.

[21] Given existing evidence of the impact of cash transfers on child nutrition, this prior is fairly optimistic but could be justified if, for example, the donor has reason to believe that cash transfers would be particularly effective in the area in which they work. See Manley, Gitter, and Slavchevska (2013) for a review of the impact of cash transfers on child nutrition.

Alternatively, donors may only wish to fund programs which achieve a certain level of cost effectiveness." Similarly, pre-defined performance measures are at the heart of many performance incentive and results-based financing programs in global health and education, including social and development impact bonds (Eichler und Levine (2009), Nazari, Jenkins and Hashemi (2019), Drew and Clist (2015)).

While benchmarking program performance is critical to inform decisions around (not) funding, modifying or scaling programs, the standard analytical approach to evaluation first and foremost assesses whether an effect is statistically significant (different than zero), regardless of the size of the effect. Again, the question therefore is how evaluators should design decision focused evaluations that hinge on a comparison to a benchmark? Which hypotheses should evaluators be testing? For example, should the null hypothesis remain "no effect" or be changed to benchmark? How can the desired probabilistic statement be made in a rigorous way?

4. FREQUENTIST METHODS IN DFES

In this section we study frequentist adaptations to the standard analytical approach for each of the three decision scenarios described above. Across these scenario, we re-consider three critical decisions in setting frequentist hypothesis tests for decision-focused evaluations. These decisions are:

1. **Choice of the Null Hypothesis:** In the standard setup, the null hypothesis is that the program has no effect. We show that this hypothesis is not always appropriate in a decision-focused framework.
2. **Choice of the Alternative Hypothesis:** The standard analytical approach involves a two-sided hypothesis test. In decision-focused evaluations, we show that one-sided hypothesis tests are sometimes more appropriate, simultaneously reducing the resources required for the study.
3. **Level of Significance**: The standard analytical approach relies on a level of significance of 0.05. This is often referred to as the "size" of a test and prompts a particular rejection threshold or critical value. We show that higher (or lower) thresholds may be appropriate in decision-focused scenarios.

We illustrate each of these points drawing on the decision scenarios outlined above.

One thing to note is that by encouraging researchers to have more flexibility in their analytical approach, this allows more areas for researcher bias to affect the analysis (commonly known as p-hacking). Therefore, it's important that researchers outline their planned analysis in a pre-analysis plan. This is not only helpful for convincing external audiences that the analysis is sound, but also helpful for codifying an agreement between the decision-maker and the researcher on what findings will lead to what policy recommendations.

4.1. TESTING FOR NON-INFERIORITY

In the "Testing for Non-inferiority" example outlined earlier, a farm voucher policy will get scaled up as long as it is "just as good" as the status quo.

Although this seems like a relatively standard scenario, it is rather tricky to specify an appropriate hypothesis test. This is because standard statistical tests are much better at showing that things are different rather than similar.

We first consider the null hypothesis. It may be tempting to set the null hypothesis to be that the treatments have equal effect, and accept the other treatment arm as being equivalent as long as there is no significant difference. Formally, the null- and alternative hypotheses of the standard analytical approach are defined as

$$H_0 : \theta = 0; H_1 : \theta \neq 0, \quad (3)$$

where θ represents the treatment effect of the farm voucher compared to the status quo. Note that the above setup is known as a test of "equivalence" (as opposed to "non-inferiority, as described below), as we conduct a two-sided test, rejecting the null as long as $\theta > 0$ or $\theta < 0$. While the above hypothesis test maps well to our decision scenario, it can lead to misleading conclusions. This is because the absence of significant differences does not mean that treatment arms have the same effect, and instead might just be an indication of a poorly-powered test. Not rejecting the null just means that we cannot *rule out* that the effects are equivalent.

Instead evaluators need to manually set up a decision-relevant threshold based on a conversation with the decision-maker. For instance, assume that the decision-maker states she would be willing to move to the new voucher system long as it was not more than 10 percentage points worse than the status quo (assuming a binary outcome).[22] In this case, the null hypothesis can be formulated such that the alternative system is more than 10 p.p. worse than the status quo, while the alternative hypothesis is that the alternative system is not more than 10 p.p. worse than the status quo. Formally, we can formulate this as a one-sided hypothesis test:

$$H_0 : \theta < -0.1; H_1 : \theta \geq -0.1 \quad (4)$$

In the medical literature, this formulation is generally referred to as a "non-inferiority" test (Walker & Nowacki, 2011). For a more general test for equivalence of the two systems, the evaluator could specify a "two one-sided test" (TOST) procedure. The TOST approach rejects the null hypothesis based on the confidence interval of the estimate of θ. For instance, we may accept equivalence of the two programs if the 90% confidence interval of our estimate of θ does not include 0.1 or -0.1. Formally, the null and alternative hypotheses of the TOST approach are:

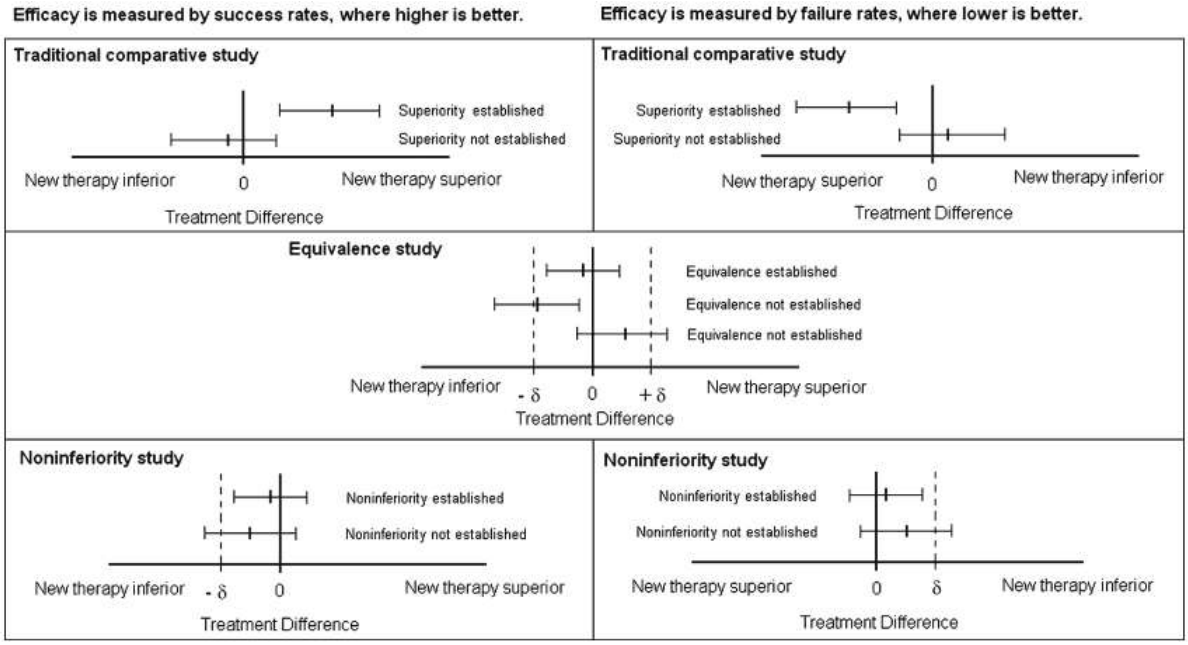
$$H_0 : \theta \in [-0.1, 0.1]; \; H_1 : |\theta| > 0.1 \quad (5)$$

Evaluators can tailor non-inferiority and equivalence tests to a particular decision scenario through selecting the parameters of the rejection threshold or adjusting the level of significance that determines the confidence interval. Figure 1 illustrates these trade-offs in terms of specifying the

[22] Among other things, the relevant decision threshold may be informed by the difference in program costs.

equivalence margin δ. As these tests introduce more choice parameters (rejection threshold, level of significance), it is critical that evaluators pre-specify their choices to maintain high credibility and rigor.

Figure 1 Two one-sided test procedure (TOST) and the equivalence margin in equivalence/non-inferiority testing



Source: Figure 1 in Walker & Nowacki (2011)

AN ASIDE ON ONE-SIDED TESTS

In the section above, we illustrated that one-sided hypothesis tests allow us to assess a program's non-inferiority or equivalence with the status quo. One-sided tests can be decision-relevant more broadly though (Cho & Abe, 2013).[23] The standard analytical approach involves a two-sided test, which allows us to detect "significant" findings for either positive or negative treatment effects. While this makes sense when evaluators are uncertain of the direction of the effect, in a decision-focused framework, in contrast, only one direction of the effect may be decisions-relevant. In such cases, evaluators should consider using one-sided test as they significantly increase statistical power or allow to reduce the resources required to conduct the study.

To illustrate these points, we go back to the "standard evaluation approach" scenario, in which a policy-makers is trying to decide whether to scale up a new project that provides farmers access to credit. The standard approach involves a two-sided hypothesis test:

$H_0 : \theta = 0$ ("The program has no effect on farmer incomes.")

$H_1 : \theta \neq 0$ ("The program has some (positive or negative) effect on farmer incomes.")

[23] Also refer to http://daniellakens.blogspot.com/2016/03/one-sided-tests-efficient-and-underused.html

Given the "standard" level of significance, α = 0.05, evaluators reject the null if $|\frac{\bar{\theta}}{s/\sqrt{N}}| > 1.96$.[24] While we reject the null if there is a significant negative *or* positive effect on outcomes, a negative effect and not rejecting the null would both lead to the same decision for the minister to not scale the program. That is, only a positive treatment effect is decision-relevant. For this reasons, evaluators should consider the following one-sided hypothesis test:

$H_0 : \theta \leq 0$ ("The program has zero or a negative effect on farmer incomes.")

$H_1 : \theta > 0$ ("The program has a positive effect on farmer incomes.")

Since the rejection threshold is now only positive, the cutoff for rejection at the same level of significance (α= 0.05) is lower. Now, the evaluator rejects if $\frac{\bar{\theta}}{s/\sqrt{N}} > 1.65$.

Choosing a one-sided over a two-sided test can make a moderate difference in the required sample size and thus for the resources required to conduct the study. For instance, consider the evaluator designs a study to detect an increase in the outcome variable by 0.2 SDs, with size of 0.05 and 80% power. Conducting a two-sided t-test, the evaluator would require a sample of 788 units.[25]. In contrast, for a one-sided t-test, the required sample size drops by around 20% to 620. (Note that we will speak further about sample size calculations in the Bayesian section of this paper.)

The main takeaway for evaluators of decision-focused scenarios is that one-sided hypothesis tests *may* be decision-relevant and that in such cases one-sided tests result in lower required sample size and thus lower required resources for an evaluation.

4.2. HEAD-TO-HEAD COMPARISONS

We now consider the "head-to-head comparison" scenario in which a policy-maker must choose between two new programs (seed distribution vs extension agents), and does not have a strong ex-ante belief of which program should be preferred.

As before, we first consider the null hypothesis. Usually, evaluators would set the null to be that the two programs have equal effects, testing against the programs having different effects. This null-hypothesis would not be particularly helpful in this case, since - if the evaluator fails to reject the null - the decision-maker is left right where she started. Yet, she needs to make a decision. The evaluator could possibly solve this problem by setting up a study with a very large sample size, allowing to detect even small differences in effectiveness. Using such high levels of resources is however seldom feasible – nor required as we will demonstrate below.[26]

[24] We assume a large enough sample that the distribution of our estimate of the treatment effect is approximately normal. $\bar{\theta}$ is the estimate of the average treatment effect. The figure on the left side of the equation is the measure of the treatment effect, normalized by the standard error of the estimate. 1.96 is the rejection threshold of a normal distribution for a test with a size of 0.05.

[25] Calculate using the 'power' command in STATA 14.2: power twomeans 1 1.2

[26] Another approach would be to pick one intervention as the null, and only choose the other program if the evidence suggests the one is better than the other. This approach is not satisfactory, though, if the decision-maker does not have a strong prior about which program is better to start with.

If this decision framework were taken literally, evaluators may recommend the decision-maker to choose the policy with a higher mean, without conducting a formal test. While this approach may make sense in some cases, in situations with low sample size there is a good chance that the decision will be driven by noise as opposed to any kind of meaningful signal.

The standard analytical approach is not able to provide satisfying evidence in this head-to-head comparison scenario, as it can only discern whether the difference between the two interventions is statistically significant. In the next section, we illustrate how a Bayesian approach to this scenario allows evaluators to estimate the probability that one intervention is better than the other, providing exactly what the decision-maker really wants to know.

One way to revive the frequentist approach for head-to-head comparisons is to tweak the scenario to assume the policy-maker has a very weak preference for one of the options (say, seed distribution), but will easily change her mind if there is any evidence that extension is more effective. In this case, the evaluator can set the null hypothesis to be that seed distribution is better than extension, and conduct a one-sided test. Seed distribution becomes the policy if the evaluator cannot reject the null.

But given this "weak" null hypothesis, does it still make sense to have a significant level of 5%? The Minister does not have strong priors that seed distribution is the best policy, and switching costs are relatively low. For these reasons, a lower standard of certainty may justify driving the decision. Of course, the evaluator will have to choose another arbitrary cut-off (say a level of significance of 20%) which is likely context-specific. Again, the evaluator's choice for this design parameters calls for a conversation with the decision-maker as well as pre-specification to transparently document the reasoning.

Adjusting level of significance – i.e. the rejection threshold – can have major implications for sample size, increasing or lowering the resources required to conduct the evaluation. Consider a study designed to detect a minimum detectable effect size[27] in the outcome variable of 0.2 SDs in a two-sided t-test to with 80% power. For a rejection threshold of 0.05, the required sample is 788 units. For a rejection threshold of 0.2, the required sample size decreases by more than 40% to 452 units.

Given the major implications on sample size, it is critical to clarify the interpretation of different levels of significance and under what circumstances evaluators may want to consider raising the cutoff above the conventional 0.05 (or perhaps lowering it). Using the standard 5% level of significance, evaluators reject the null-hypothesis when there is less than a 5% chance to observe the given test statistic if the null was actually true. In other words, when the null is true we would make the mistake of rejecting that null less than 5% of the time. For this reason, the level of significance is also referred to as the Type I error rate. In the seeds vs extension example above where seeds being superior is the null hypothesis, this means that there is less than a 5% chance of declaring extension to be superior when it was equivalent or worse than seeds. By specifying a relatively low level of significance (such as 0.05), evaluators restrict the statistical power of the test, which indicates the test's ability to reject the null when it is false. Thus, study designs that can be justified to have a higher rejection threshold (see below), allow evaluators to increase statistical power for the same sample, or lower the required sample and therefore resources for the evaluation while maintaining the same statistical power.

[27] The minimum detectable effect size is the lowest effect that would be deemed statistically significant given a sample size and thresholds for Type I and II errors.

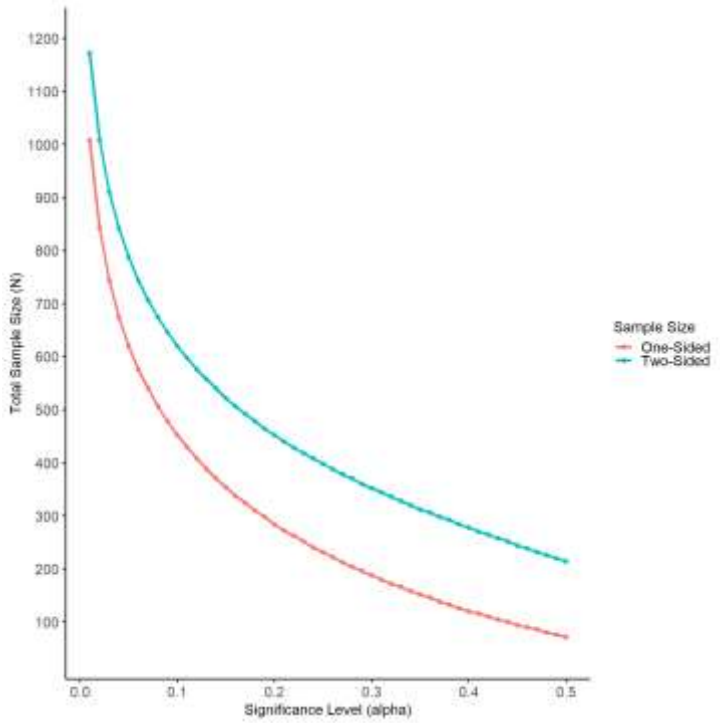
Figure 1 illustrates this trade-off by plotting how the required sample size decreases as the p-value cutoff increases while power is maintained at the conventional 0.8.

There are several circumstances where it may be justified to increase the statistical test's level of significance to a value larger than the conventional level of 5%:

- There is a specific audience for the evaluation results, who may be less skeptical/more risk tolerant than the general public.
- There are not large consequences if the null is incorrectly rejected. Most importantly, this requires sufficient certainty of the intervention doing no harm. In addition, adverse consequences are lower the lower the financial and societal costs of scaling a program based on a false positive test result.
- The decision-maker has relatively low confidence in the null hypothesis (similar to relatively uninformative priors in a Bayesian framework.)
- A decision must be made even though sample size (and therefore precision) is limited.

The main takeaway for evaluators of decision-focused scenarios is to not simply default to using a level of significance of 5%, but instead to carefully consider the audience and evidence needs when designing the rejection thresholds for decision-focused studies.

Figure 2: Tradeoff of Sample Size and p-value cutoff



Note: This graph illustrates the sample size required for evaluation designs that use two- and one-sided hypothesis tests respectively. Both types of design assume a level of significance of 5% and 80% power, same variance in treatment and control (standardized to 1), minimum detectable effect size (MDES) of 0.2 SD and control mean of 1.

4.3. COMPARING TO A BENCHMARK

In many impact evaluations, a program is defined as "successful" as long as the average treatment effect is positive, with discussion of the effect size as an afterthought (Miller and Rodgers 2008). While evaluators can set rules for statistical significance, how should they define a "policy-significant" effect size in a DFE?

The policy-significance or "substantive" or "economic significance" of an effect oftentimes depends on the policy-context (Hill, et al., 2008). In a decision-focused evaluation, a first step towards defining substantive effect sizes would be to ask decision-makers about their decision-making threshold. Returning to the "comparing to a benchmark" scenario we study a donor who only wants to fund a school feeding program if it increases HAZ (height-for-age z-score) by at least 0.2 standard deviations. The question therefore is how evaluators can conduct causal inference on this benchmark requirement.

Leveraging the guidance from above, the most straightforward way to inference would be to specify a one-sided hypothesis test with null hypothesis $\theta <= 0.2$ and alternative hypothesis $\theta > 0.2$, and the standard level of significance $\alpha = 0.05$. While there is nothing statistically wrong with this setup, this test requires a very high standard of evidence. Assuming sample sizes (and levels of precision) common in the literature, this approach requires a much higher point estimate than 0.2 to reject a null effect of 0.2, likely resulting in levels of statistical power well below what the evaluator may consider acceptable or sample sizes that would consume an enormous amount of resources.

As indicated above, an alternative evaluation design might change the rejection threshold and accept a higher Type 1 error rate. This would be justified if the decision-maker was ok with a higher amount of uncertainty as to whether the effect size was above the threshold. For instance, the evaluator could specify a level of significance of 20% instead of 5%. This specification opens a new avenue of potential criticism, though. This is because the statistical test might reject the null, $\theta <= 0.2$, while the treatment effect estimate might not be statistically different from zero at $\alpha = 0.05$. This approach is suboptimal because the evaluator likely wants to have a higher level of confidence that the effect is larger than zero.

Instead, evaluators could further adapt their evaluation design and conduct two hypotheses tests. In this case, the evaluator would only recommend the decision-maker to take the decision if both null-hypotheses are rejected. The first test requires a high threshold for rejection to ensures that the program has a positive effect on HAZ scores:

$$H_0: \theta <= 0, \qquad H_1: \theta > 0\,, \qquad \alpha = 0.05 \tag{6}$$

The second test assesses whether the program's effect is sufficiently large to receive funding, but requires a lower threshold for rejection:[28] [29]

[28] Another approach would be to have an extremely low standard of evidence on the effect size, and give their recommendation as long as the point estimate was above the threshold of 0.2 (without taking the standard error into account at all). This is similar to the solution in the "Head-to-Head" comparison above, but the decision-maker may desire more certainty.

[29] In a results-based financing scenario, such as a Development Impact Bond, outcome payers might be very sensitive to risk resulting from a "too large" probability of a false positive result. Lowering the threshold of rejection – or the level of significance – would mitigate this risk while requiring a larger sample size.

$$H_0: \theta <= 0.2, \qquad H_1: \theta > 0.2, \qquad \alpha = 0.2 \tag{7}$$

In Table 1, we summarize the analytical approach, decision rule, sample size, cost and other considerations regarding the different approaches to test the program intervention against a benchmark. The proposed "double-barreled" approach combines the results of two tests. This combination of tests allows us to reduce the resources required for the study, while providing the evaluator and funder with sufficient confidence that the intervention actually "works", i.e. that the average treatment effect is positive. At the same time, this approach is likely difficult to explain to a decision-maker who is not well-versed in statistics. For this reason, we will demonstrate below that a Bayesian approach to decision-focused evaluations that compare interventions against a benchmark may be a more natural approach.

Table 1 – Implications of Different Approaches to Compare a Program to a Benchmark

	Standard Two-sided Approach	Conservative One-sided Approach	Alternative One-sided Approach	Proposed Approach
Statistical Test and level of significance, α	• H_0: θ = 0.2 vs. H_1: θ ≠ 0.2 • **α=0.05**	• H_0: θ ≤ 0.2 vs. H_1: θ > 0.2 • **α=0.05**	• H_0: θ ≤ 0.2 vs. H_1: θ > 0.2 • **α=0.2**	• 1: H_0: θ ≤ 0 vs. H_1: θ > 0; α=0.05 • 2: H_0: θ ≤ 0.2 vs. H_1: θ > 0.2; α*=0.2
Decision rule	• fund the program if reject H_0	• fund the program if reject H_0	• fund the program if reject H_0	• fund the program if reject H_0 in both tests
Sample Size[+]	• 12,562	• 9,894	• 4,534	• 4,534
Cost in USD[x]	• 628,100	• 494,700	• 226,700	• 226,700
Considerations	• Very high bar to meet • requires $\hat{\theta} \gg 0.2$	• High bar to meet • requires $\hat{\theta} \gg 0.2$	• Higher α may leave evaluator unsure of $\hat{\theta} > 0$	• Additional test for $\hat{\theta} > 0$ • Evaluators should define α* with the decision-maker.

Note: [+] The sample size calculations assume a Minimum Detectable Effect Size (MDES) of 0.25. Given the 0.2 threshold, this results in an effective MDES of 0.05. We assume an average cost of USD 50 per unit. In the Proposed Approach, only one of the tests will bind when conducting power calculations. In our example it is Hypothesis Test 2, which is why the sample size is the same as in the Alternative One-sided Approach.

5. A BAYESIAN APPROACH TO DFES

In this section we introduce Bayesian Analysis in the context of decision-focused evaluations and illustrate how its application may lead to more actionable evidence in the scenarios above. Bayesian analysis is an approach to statistical analysis that uses Bayes' rule to update *prior* beliefs about unknown parameters and hypotheses in light of new *evidence* or data (Jackman, 2004). The resulting updated beliefs, the *posterior*, allow the researcher to make probabilistic statements about parameters or hypotheses of interest. Intuitively, Bayesian approaches investigate how likely a particular hypothesis is given our ex-ante knowledge about the matter and the additional evidence

generated by the research. Bayesian approaches are frequently used in a variety of fields, ranging from astrophysics, genomics, and computer science (Malakoff, 1999) to business marketing (Rossi and Allenby 2003) and clinical trials in health care (Spiegelhalter, Abrams and Myles 2004).

The Bayesian approach is also useful for analyzing data from decision-focused evaluations. First and foremost, Bayesian analysis allows evaluators to make easily understandable probabilistic statements about a hypothesis given the observed data. This means an evaluator could provide an intuitive description of the uncertainty involved in a particular action by saying "there is a 90% probability that program A has a larger effect than program B." In contrast, the frequentist approaches discussed in the previous section assess the plausibility of observed data conditional on a particular hypothesis. In fact, the p-value is defined as the probability of observing the given (or more extreme) data given a hypothesis. In other words, a p-value of 0.03 means that there is a 3% chance to observe results that are as or more extreme than the ones observed *if* the null hypothesis is true and the study were repeated many more times.[30] Yet, p-values are often misinterpreted to provide the probability that the null hypothesis is true (Amrhein, Greenland and McShane [2019], Wasserstein and Lazar [2016], Nickerson [2000]). In subsequent sections, we illustrate how probability statements following from a Bayesian approach provide more intuitive answers in the decision scenarios described above.

In addition, Bayesian analysis allows decision-makers to incorporate prior beliefs about the program to be evaluated.[31] If used appropriately, this feature allows evaluators to better tailor their analysis to the specific decision context. Beliefs can come from different sources: the decision-makers themselves, external subject matter experts, or the results from similar studies in the literature. Irrespective of their source, the evaluator captures these beliefs through a quantifiable distribution and integrates them into the analysis.[32] In the context of knowledge-focused evaluations (intended for a general audience), this attribute of Bayesian analysis is often framed as a drawback since it may introduce subjectivity into the results. We discuss below how to specify uninformative priors so that the priors have little impact on the findings. In the case of DFEs, though, evaluators aim to inform a particular decision in a particular context – and this decision-maker may have *some* measurable prior that can be incorporated in the analysis. At the same time, not all priors may be justifiable and approaches to assess how reasonable a given prior is exist (Schad, Betancourt and Vasishth 2019). For example, evaluators may fear stereotypical implementers to be overly optimistic about their programs. But what about funders who are skeptical about the (cost-) effectiveness of an intervention? In subsequent sections, we illustrate how incorporating existing beliefs into the analysis can make the final results more relevant to the decision context, while maintaining a level of rigor necessary for the results to be taken seriously by outside observers.[33]

Despite these merits, there are some disadvantages to Bayesian analysis. First, despite their increasing use by academics over the last years (BaŞtürk, et al., 2014), results from a Bayesian analysis are still perceived by many as less objective than those from a frequentist analysis (Sprenger, 2015). We hope

[30] In this case, given that the p-value is smaller than the standard rejection threshold of 5%, the evaluator would reject the null-hypothesis. Despite the statistical test rejecting the null, the evaluator will never know whether the null is in fact true.
[31] In fact, we discuss below that Bayesian analysis requires *some* belief about every unknown parameter in the model.
[32] We provide more detail on how beliefs can be elicited and quantified in the Appendix.
[33] Other merits of the Bayesian approach include for example the possibility to explicitly model decision-makers' utility or loss function that would allow to model more complicated decisions (Smith, 2010). Further, Bayesian Hierarchical Models oftentimes exhibit more power in sub-group analyses, especially if sub-group sizes are small (Chen and Lee 2020).

the discussion and guidance provided in this paper will contribute to broadening this acceptance further. Second, specifying a full probability model with priors can take significantly more time and cognitive effort than using the frequentist methods discussed above. This difference is mostly explained by the fact that writing code to fit Bayesian models, while much easier now than even ten years ago, still requires more effort than writing code to fit frequentist models.[34] For these reasons, the standard frequentist approach and its adaptations to decision-focused evaluations discussed above may remain optimal in some, but not all circumstances. We discuss how to choose between a frequentist and a Bayesian approach in a later section.

Even more than in frequentist analysis, there are many parameters in Bayesian analysis that must be chosen by the researcher. Therefore, we highly suggest that researchers adopting a Bayesian approach also pre-register their analytical framework, which ensures that readers of the research will not see parameter choice as an opportunity for researchers to choose specifications that correspond to their desired result.

In the following section, we provide a brief overview of the Bayesian approach. We then illustrate how it can be applied to the decision-scenarios introduced above and why it leads to more tailored evidence and recommendations in these scenarios.

5.1. OVERVIEW OF THE BAYESIAN APPROACH

Bayesian analysis uses Bayes' rule to update *prior* beliefs about unknown parameters and hypotheses in light of new *evidence* (Jackman, 2004). To illustrate this approach, we consider the simple model to evaluate a program introduced in the beginning of this this paper, reproduced below:

$$y_i = \alpha + \theta t_i + \varepsilon_i, \tag{8}$$

where y_i and t_i are individuals' (i=1,…n) outcome and treatment status respectively. Earlier, we discussed estimating the unknown average treatment effect of this model, θ, using ordinary least squares (OLS).[35] Turning to Bayesian analysis, evaluators still assume θ, to be unknown. Instead of θ, being a constant value, though, evaluators assume it – and all other parameters of the model – to follow some unknown distribution. The goal then becomes to estimate this probability distribution given the observed data. For this reason, evaluators express the model through probability distributions.

To illustrate these steps, we first recast the model as a maximum likelihood estimation (MLE) problem. Assuming the error terms are independent and identically distributed (iid) and follow a normal distribution, we formulate the distribution of outcome y_i given the model parameters and the characteristics, $p(y_i|\alpha, \theta, \sigma^2, t_i)$, as

$$y_i \sim N(\alpha + \theta t_i, \sigma^2); \;\; \sigma^2 = Var(\varepsilon_i), \tag{9}$$

Maximum likelihood estimation of this model provides the point estimates (and confidence intervals)

[34] More and more tools that make Bayesian analysis accessible to a larger range of analysts, such as the R package "rstanarm" (Goodrich, et al., 2020), have recently been developed.

[35] While the model assumes a constant treatment effect θ, estimation by OLS gives the average treatment effect even if the treatment effect differs by unit.

of the model parameters $(\hat{\theta}, \hat{\alpha}, \hat{\sigma}^2)$ that make the observed data y "most likely" to have occurred.[36] Formally, we would obtain these estimates by maximizing the **likelihood** function[37]

$$L(\theta, \alpha, \sigma^2|y) = \prod_{i=1}^{n} p(y_i|\alpha, \theta, \sigma^2, t_i). \quad (10)$$

Bayesian analysis is similar to maximum likelihood estimation in that it also requires a full probability model for the data, i.e. to specify a likelihood function that describes the outcome as a function of the model parameters. Given that we will not maximize this function for the model parameters, we will refer to the likelihood as $p(y|\alpha, \theta, \sigma^2, t)$ instead of $L(\theta, \alpha, \sigma^2|y)$.

Bayesian analysis differs from maximum likelihood estimation in that it relies on **Bayes rule** to combine the likelihood and a **prior** to generate a **posterior** for all parameters of the model. In other words, rather than calculating point estimates and associated confidence intervals, Bayesian analysis results in a joint probability distribution for all unknown parameters in the model. We define each of the bolded terms in turn.

First, **Bayes' rule** expresses the conditional probability of one variable (or set of variables) a given another variable (or set of variables) b in terms of the conditional probability of the latter given the former and the marginal probabilities of both (sets of) variables. In the case of two variables, that is:

$$p(a|b) = \frac{p(b|a)p(a)}{p(b)} \quad (11)$$

In Bayesian analysis, evaluators strive to estimate the probability distribution of the unknown parameters given the observed data. Applying the above equation to our model, a is the set of parameters we aim to estimate, $(\alpha, \theta, \sigma^2)$, b is the observed vector of outcomes y, and we assume the treatment status t is given.[38] Bayes' rule then becomes:

$$p(y, t|\alpha, \theta, \sigma^2)\, p(\alpha, \theta, \sigma^2) \quad (12)$$

[36] Evoking the same assumptions and fitting the model above using OLS leads to equivalent point estimates and confidence intervals as for estimating the model using MLE. Refer to the following source for a proof: https://bookdown.org/egarpor/PM-UC3M/app-ext-mle.html (shareable via creative commons license CC BY-NC-ND 4.0)

[37] Usually, we maximize the logarithmic transformation of the likelihood function to achieve computational efficiency. Given that the logarithm is a positive monotone transformation, it does not change the function's extrema.

[38] Note that we model the distribution of the outcome *y* (the likelihood), while we do not model the distribution of the treatment status *t*. This simplification allows to focus on the parameters of interest. In the application of Bayes rule, this implies that the conditional distribution of the treatment status equals the unconditional distribution, $p(t|\alpha, \theta, \sigma^2) = p(t)$. In theory, we could also model the treatment status *t*. That is, we could specify the joint likelihood $p(y, t|\alpha, \theta, \sigma^2, \psi)$ where ψ is some parameter(s) affecting the distribution of treatment status *t* and seek to calculate $p(\alpha, \theta, \sigma^2, \psi|y, t)$. The parameter ψ is not of interest on its own though and is unlikely to be related to the other parameters. Thus, this additional modelling is unlikely to shed light on the parameters of interest. The independence of ψ and the other parameters is trivially true in the case of an RCT and invoked under specific conditions in many quasi-experimental applications as well. Refer to Jackman (2004) for a more detailed discussion.

$$= \frac{p(y|\alpha, \theta, \sigma^2, t)\, p(t|\alpha, \theta, \sigma^2)\, p(\alpha, \theta, \sigma^2)}{p(y|t)\, p(t)}$$
$$= \frac{p(y|\alpha, \theta, \sigma^2, t)\, p(\alpha, \theta, \sigma^2)}{p(y|t)}$$

The term on the left-hand side of the equation is the joint probability distribution of the parameters given the data, commonly referred to as the **posterior**. The posterior allows evaluators to make specific statements about the probability with which a parameter - or any function of the parameters and data - takes on a range of values. For example, evaluators might say that the probability that θ is greater than 0.1 is equal to 0.05, $P(\theta > 0.1) = 0.05$.[39]

Moving to the right-hand side of the equation, the first term in the numerator is the **likelihood** as defined above, i.e. the full probability model of the outcome given the parameters and the treatment status. The second term in the numerator contains the joint distribution of the model parameters, $p(\alpha, \theta, \sigma^2)$.[40] This term is called the **prior** because it contains the evaluator's ex-ante beliefs about *all* unknown parameters of the model, including but not limited to the treatment effect θ.[41] As indicated above, prior distributions can be derived from many different sources, such as the researchers themselves, subject matter experts, the decision makers, or existing evidence from similar interventions found in the literature. We discuss several examples of how to specify and interrogate the effect of priors on the analysis below. Regardless of their source, priors contain quantified best guesses about all parameters of the model and how confident decision-makers are in them *before* seeing additional data from the study.[42]

Finally, the denominator contains the distribution of outcomes conditional on the treatment status, $p(y|t)$. This distribution of (observed) outcomes does not change the relative frequencies described in the numerator.[43] Intuitively, the purpose of this term is to normalize the posterior distribution to a proper probability function that integrates to one. Since calculating this denominator is often not possible because the conditional density function of the outcome is unknown (and might have a very

[39] In comparison, with frequentist analyses we can only make more complex statements such as "if the true value of θ is 0 and we conducted this experiment many, many times, we would observe estimates of θ greater than 0.1 roughly 5% of the time."
[40] In this example, the prior is a three-dimensional density function. To better characterize this density, we could factor the joint density into marginal densities, $p(\alpha, \theta, \sigma^2) = p(\alpha, \theta|\sigma^2)\, p(\sigma^2)$, and then specify prior distributions for these marginal densities. Oftentimes, we assume independence of parameter priors, see e.g. https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations. For more detail on specifying prior distributions refer to the appendix.
[41] Parameters such as α and σ^2, which we don't care about for their own sake are often called "nuisance parameters". Further, note that the prior distribution may nor may not be a *probability* distribution. For example, it would be possible to specify a "flat", uniform prior ranging from (-∞, ∞) that integrates to a constant larger than one, thus violating the axioms of probability. We will introduce more concepts and lingo about priors in the Appendix.
[42] We use the term confidence to vaguely describe the (relative) dispersion of the prior distribution. The more dispersed (or flat) a given marginal prior distribution, the less confident the researcher is about the plausible values for that parameter. The dispersion of priors is only meaningful in relation to the dispersion in the observed data (Gelman, Simpson and Betancourt, 2017).
[43] To calculate this term, we may use the property that the probability of parameters must sum to 1 over all of the possible values of the parameters: $p(y|t) = \int p(y|\alpha, \theta, \sigma^2, t) p(\alpha, \theta, \sigma^2) d\alpha d\theta d\sigma^2 = c$.

complex shape), researchers instead rely on advanced software such as JAGS or Stan to approximate this density.[44]

Abstracting from the normalizing term in the denominator, equation (12) states that posterior is proportional to the product of the likelihood and the prior:

$$p(\alpha, \theta, \sigma^2 | y, t) \propto p(y | \alpha, \theta, \sigma^2, t)\, p(\alpha, \theta, \sigma^2). \quad (13)$$

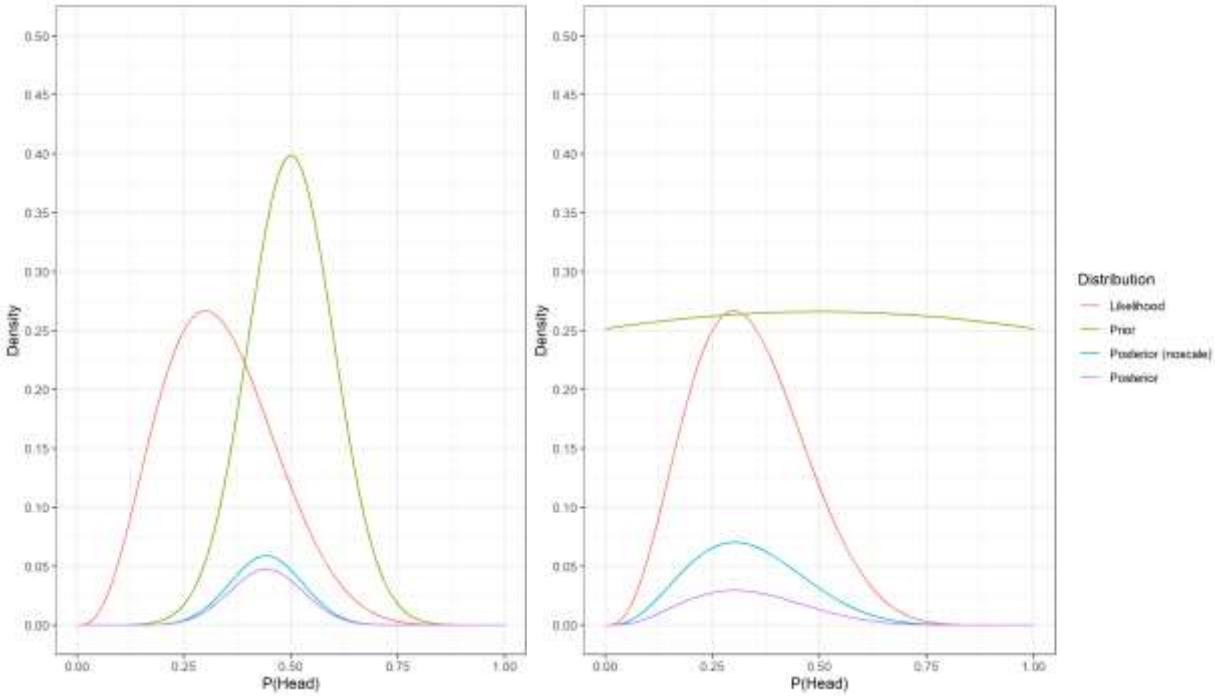
As prior information about the parameter(s) of interest becomes less precise, i.e. as the prior dispersion increases, the posterior density is increasingly determined by the information in the likelihood (Jackman, 2004).[45]

Figure 3 illustrates this relationship between the prior, the likelihood and the posterior using the example of a coin toss. In this example we have a coin but are not sure if it is fair (meaning 50% probability of heads, and 50% probability of tails.) We flip the coin 10 times, and find that it lands on heads thirty percent of time, generating the likelihood function displayed in Figure 3, with the maximum of the likelihood at 0.3. In the left panel we show an example where we had a strong prior that the coin is fair. In this case, our posterior distribution moves slightly towards favoring a biased coin, but with the maximum mean of our posterior suggesting a that P(Head) = 0.44. In contrast the right panel demonstrates a relatively flat (or uninformative) prior. In this case, the posterior mirrors the likelihood function, with P(Head) = 0.3. The figure also illustrates that the scaling of the posterior through the denominator in equation (12) leaves the characteristics of the posterior unaffected.

[44] Density functions can be approximated - or 'simulated' - through a computationally intense, iterative process called Markov Chain Monte Carlo (MCMC). Both JAGS (http://mcmc-jags.sourceforge.net/) and Stan (https://mc-stan.org/) use such MCMC methods to simulate posterior distributions. Stan uses computationally more efficient Hamiltonian Monte Carlo methods that often result in lower run time.

[45] We illustrate the numerical equivalence of the point estimates from maximum likelihood estimation and the mean of the Bayesian posterior in case of uninformative for one of the scenarios below.

Figure 3 – Illustration of the Bayesian Approach



Note: This figure illustrates the Bayesian posterior estimation process for a coin toss for an informative prior (left) and uninformative prior (right). The likelihood is identical in both cases and postulates that ~30% of coin tosses result in heads. The informative prior centers on 0.5 and is relatively precise. The unscaled posterior, i.e. the numerator of equation (12), is obtained by multiplying the likelihood and the prior. The posterior is then obtained by scaling this distribution to integrate to one. The informative prior dominates the likelihood function resulting in a posterior distribution that centers closer to 0.5. In contrast, specifying a relatively flat distribution, i.e. an uninformative prior, leads to the likelihood driving the posterior estimation.

In the following sections, we illustrate how the Bayesian approach can be applied to the three scenarios introduced above and how the resulting evidence can be more actionable than the in the conventional frequentist approach.

5.2. HEAD-TO-HEAD COMPARISONS AND TESTING FOR NON-INFERIORITY

In the head-to-head comparison scenario, the decision-makers needs to decide between two interventions: distributing free seeds and hiring special vegetable-focused extension agents. Above, we illustrated that informing this decision with evidence from a purely frequentist approach is challenging because results from a conventional hypothesis test might not provide sufficient evidence to support any recommendation. In this section, we demonstrate how evaluators could use a Bayesian approach to directly compare the effect of two interventions. We also show how the same analysis can be easily retooled to answer the question of whether one intervention is at least as good as another, i.e. the testing for non-inferiority scenario.

EXPERIMENTAL SETUP AND LIKELIHOOD

We assume that that the decision-maker seeks to learn about the effect of the two interventions through a randomized controlled trial. Let us suppose that the RCT includes two arms, one for each of the two new programs to be tested.[46]

To specify the probability model for the outcome, we start with the equation the evaluator would use to estimate impact in a frequentist analysis. Let y_i be (log) income for farmer i measured in USD and t_{2i} an indicator for whether the farmer receives vegetable focused extension.[47] We treat free seed distribution as the implicit control group and omit the treatment indicator t_{1i}. If the intervention was randomized at the farmer-level[48], we might estimate impact using the following equation:

$$y_i = \alpha + \theta t_{2i} + \varepsilon_i \tag{14}$$

where we assume that ε_i is iid normal. Mean farmer income under free seed distribution is captured in the constant α. The differential effect of vegetable-based extension (compared to free seed distribution) on log-income is captured by θ.[49] The goal of the analysis is to estimate the probability that one of the programs has a larger treatment effect. This objective can be achieved by estimating $P(\theta > 0)$. Because of the normality assumption, this model states that (log) income, y_i, is distributed normally with mean $\alpha + \theta t_{2i}$ or, using notation:

$$y_i \sim N(\alpha + \theta t_{2i}, \sigma_y^2) \tag{15}$$

Equation (15) specifies our data model for the outcome of interest, i.e. our likelihood.

PRIORS

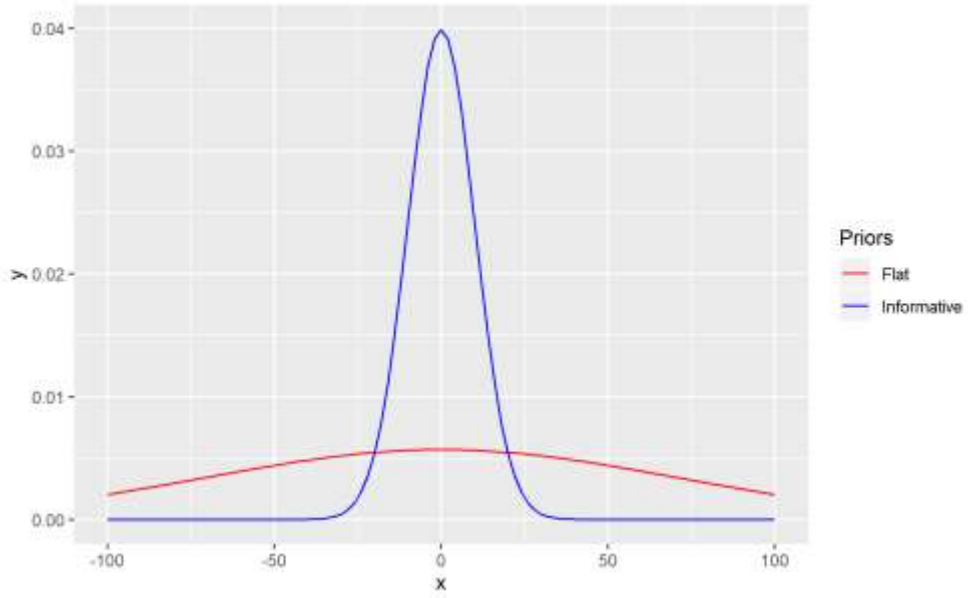
To complete the model, we need to specify priors for *all* parameters, i.e. α, θ and σ_y^2. We start first with the priors for α, θ. To recap, α corresponds to the average log income among target farmers receiving free seed distribution, and θ captures the difference in log income for farmers receiving extension services. In practice, we would likely have quite a bit of information to help us specify these priors. For example, we might have results from a recent survey of farmers in the area with which we could estimate average log income and results from other studies that estimate effects similar to α and θ.

At the same time, the decision-maker does not have a strong belief about which program is more effective ex-ante. For this reason, the evaluator may specify a relatively "**flat"** prior centered around 0 for both parameters. Specifying such flat priors is equivalent to saying that the decision-maker expects farmers' average (log) income after receiving free seeds to be close to 0 and expects no difference between the two programs. Further, flat priors exhibit relatively high variances and are so

[46] Alternatively, we could assume existence of a control group, which would simplify notation below. We do not choose this approach because we would not actually recommend including a control group in the study to reduce the required resources.
[47] We opt to use log-income rather than nominal income because we assume the outcome to follow a normal distribution further below. Generally, the *assumed* likelihood distribution should represent the actual distribution of the data fairly well.
[48] Randomization at the individual farmer-level would likely not be feasible. We adopt this simplification to make the presentation more straightforward. Spiegelhalter (2001) and Moerbeek (2019), for example, discuss Bayesian analysis for clustered randomized trials.
[49] This notation becomes more complex if the objective was to compare three or more programs.

named because their densities, when graphed appear flat. For example, in the graph below, the red probability density would be considered flat relative to the blue probability density.

Figure 4 Prior distributions - flat vs informative



Flat priors are also called **uninformative** priors because they encode little information about the parameter of interest (Gelman, Simpson and Betancourt, 2017).[50] In the chart above, the blue, informative prior assigns the parameter *x* a much higher probability of values between -30 and 30 whereas the red, uninformative prior assigns only a slightly higher probability to values in this region.

In the analysis, we assign the priors $p(\alpha) \sim N(0, 10^2), p(\theta) \sim N(0, 10^2)$.[51]

Specifying a prior for the variance term, σ_y^2, is more difficult. While evaluators and decision-makers may be comfortable with the idea of variance, specifying the variance of a variance is complicated. To complicate things further, by definition, σ_y^2 cannot be negative and specifying a prior for the variance, σ_y^2, is not the same as specifying a prior for the standard deviation, σ_y (Gelman, Simpson and Betancourt, 2017, Gabry et al., 2019). For higher-order terms like variances, we therefore recommend to use priors that other researchers have found to be reasonable. For example, to model variance parameter researchers usually specify an inverse-gamma, half-normal, half-Cauchy, or uniform distribution for the standard deviation (Gelman et. al., 2006, Klein et. al., 2016, Hodges 2013).

[50] Flat priors can be misleading, in particular when they represent information far from a known truth and effects of interest are expected to be small (Gelman, Simpson and Betancourt, 2017). In our example, a more reasonable prior could center closer to the average farmer (log) income in the context of the study – and restrict the dispersion of both prior distributions to the maximum increase and decreases in income observed in other studies.

[51] Note that there is nothing special about the normal distribution for the purpose of specifying a prior here. We use it only because of its familiarity. Proper modeling should employ prior predictive checks to assess the plausibility of simulated data based on these priors (Schad, Betancourt and Vasishth 2019).

Following, this convention, we assume that $p(\sigma_y) \sim Uniform(0,100)$. This means that any value for the standard deviation in farmers' log-income in the interval [0,100] is considered equally likely. Again, we could likely specify a more informative prior by obtaining information on the variation in farmer incomes from existing data.[52]

Summing up, our full model with both likelihood and priors is:

$$y_i \sim N\left(\alpha + \theta t_{2i}, \sigma_y^2\right)$$
$$p(\alpha) \sim N(0, 10^2), p(\theta) \sim N(0, 10^2)\ ;\ p(\sigma) \sim Unif(0,100) \quad (16)$$

SOLVING THE MODEL THROUGH SIMULATION

To provide an intuition for how to solve this model, let us look back at equation (12), where we noted that the posterior is proportional to the product of the likelihood and the prior and a normalizing term. In many applications, this product will not be known and - except for very simple models - closed form solutions to the posterior distribution do not exist.[53] The reason is that the posterior probability hinges on being able to integrate over an oftentimes unknown function in the denominator. Instead, we rely on statistical software, such as the Stan or JAGS probabilistic programming languages, that use Markov chain Monte-Carlo (MCMC) algorithms to describe the unknown posterior density function. MCMC algorithms have lately facilitated research advances because – under certain conditions – they allow researchers to *learn* from the posterior distribution of an unknown parameter through repeated 'sampling' from its density, even though the functional form of that density may not be known to the analyst explicitly (Guimerà, et al., 2020).[54] In other words, we can use MCMC algorithms to solve for any characteristics of the joint and marginal distributions of the model parameters in equation (16), including the mean and standard deviation of the average treatment effect θ.

To estimate the model in the example, we define a data generating process and all corresponding inputs in the statistical computing software R and then call Stan from within R (R Core Team, 2020).[55] Code for this example, and all other examples in this paper, can be found at

[52] For example, evaluators could consider informing their priors about contextual parameters, such as average income in a particular geographic area, with baseline information. At the same time, evaluators should avoid using the data they intend to use for inference to inform their choice of prior since this might artificially increase the precision of the information in the prior.

[53] In this example, a closed form solution to the model would exist. This reason is that a normal likelihood function along with a normal prior result in a normal posterior. This feature of normal priors is referred to as conjugacy. See Jackman (2004) for a formal discussion.

[54] MCMC algorithms combine two ideas. First, under specific conditions, it is possible to describe an algebraically unknown posterior density through a Markov chain. 'Sampling' from the posterior means simulating or drawing random numbers from the resulting (stationary) Markov Chains. Increasing the number of simulated draws allows the researcher to describe the distribution as precisely as needed. Second, the Monte Carlo principle states that researchers can learn *anything* about a random variable by repeatedly drawing from its density (Jackman, 2004). In combination, MCMC algorithms therefore allow to describe any characteristic the posterior distribution without providing a closed form solution for the density function. There are different MCMC algorithms. For example, Stan uses a specific form of a Markov-Chain Monte Carlo (MCMC) algorithm called NUTS. For more a more formal introduction to MCMC algorithms refer to Jackman (2004) or to http://elevanth.org/blog/2017/11/28/build-a-better-markov-chain/

[55] With the *rstan* package, R provides an interface to use Stan through R. For more information, refer to: https://github.com/stan-dev/rstan/wiki.

https://github.com/IDinsight/dfe_methods.[56] As indicated above, readers will notice that writing code to fit Bayesian models involves more and more complicated steps than writing code to fit frequentist models. For this reason, we encourage readers to use and modify the example code provided with this paper.

To define a simple data generating process, we set the parameters in the likelihood function of equation (16) as $\alpha = 2.3$ and $\theta = 0.05$. We generate outcome data for 200 observations (100 per treatment arm) by adding random noise from a standard normal distribution to the respective individual means.

RESULTS

In Figure 5, we display an excerpt of the default summary of the posterior estimate provided by R when fitting the model in equation (16) to the simulated dataset generated above. The columns describe various characteristics of the marginal posterior estimates for all parameters of the model. For example, the summary suggests that the posterior mean log-income for farmers receiving free seeds, α, is 2.26 and that 25% of farmers' log-incomes are lower than 2.20. Before interpreting results further, evaluators should assess whether the resulting posterior estimate has "converged."[57] We provide example assessments in the accompanying code. Moreover, evaluators should assess model fit to the observed data. The general idea of so-called "posterior predictive checks" is to use the resulting posterior distributions to assess whether they match characteristics of the observed data that has not been modelled explicitly.[58] Bandiera et al. (2016), for example, compare the similarity of several order statistics, such as the mean, minimum, and maximum, between data simulated from the resulting posterior distributions and the observed data. In Appendix B, we briefly describe an analogous approach.

[56] We recommend reading the README and walking through the short R notebook "Introduction Stan and RStan" before walking through the specific examples.

[57] Formally, evaluators should to assess whether the model has converged towards a *stationary* distribution and whether the samples that form the posterior estimate are drawn from the entire posterior support. The default summary includes one such convergence diagnostic, Rhat. It assesses whether the observed variance of the (marginal) posterior has converged to its true variance. As a rule of thumb, Rhat values are supposed to be smaller than 1.1. Further visual and numerical checks for convergence exist. For example, the Gewecke convergence diagnostic relies on a test for equality of the means of the first and the last part of a Markov chain. Finding differences in means, poses evidence against convergence. To assess whether samples "mix well", i.e. are drawn from the entire support of the posterior, evaluators should assess the autocorrelation plots of the Markov Chain. High levels of autocorrelation might indicate that the model is stuck in a local extremum. Gabry et al. (2019) provide a deeper discussion of using visual tools for posterior predictive checks.

[58] In addition, evaluators should assess how these posterior predictive checks are affected by particular modelling assumptions, such as the prior distributions.

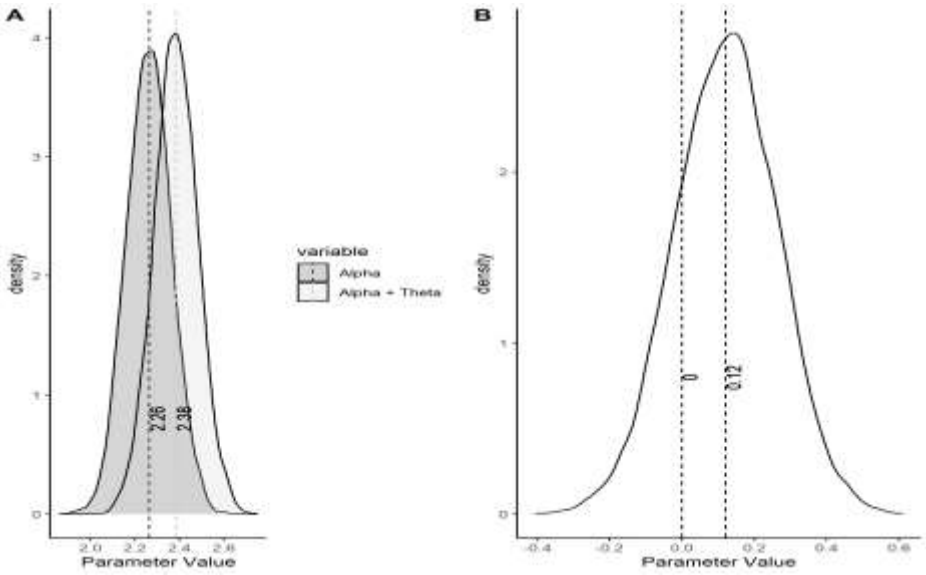
Figure 5 – Rstan output of Bayesian model estimation

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%
alpha	2.26	0.00	0.10	2.06	2.20	2.26	2.33	2.45
beta	0.12	0.00	0.14	-0.15	0.03	0.12	0.21	0.39
sigma	0.98	0.00	0.05	0.89	0.94	0.98	1.01	1.08
lp__	-95.20	0.02	1.22	-98.38	-95.76	-94.90	-94.30	-93.80

Note: This figure provides an excerpt from rstan's default summary of the posterior distribution when fitting the model in equation (16) to simulated data. The columns provide information about the respective parameter's posterior mean, Monte Carlo standard error (*se_mean*), standard deviation (*sd*), and quantiles (2.5%, …, 97.5%). The last row provides values for the log-posterior (*lp_*), which approximates the unknown density of interest.

In addition to this default summary of results, we can simulate draws from the multivariate posterior distribution. Figure 6 plots the marginal posterior distributions and means of farmers' log-income under free seed distribution, α, and vegetable-based extension, $\alpha + \theta$, as well as the difference in log-incomes, θ. We observe that the posterior mean log-income under vegetable-based extension is larger than the corresponding mean of income under free seed distribution. Further, the draws from the marginal posterior distribution for θ allow evaluators to estimate the probability that vegetable-based extension service leads to larger increases in farmers'(log) income than distributing free seeds, $\Pr(\theta > 0)$. This is simply given by the relative frequency of draws from the posterior for which this condition holds. In this example, we estimate $\Pr(\theta > 0) = 0.80$. In other words, according to this model, there is an 80% probability that vegetable-based extension services have a larger effect on farmers' income than distributing free seeds.

Figure 6 - Marginal Posterior Distributions for α, $\alpha + \theta$ and θ



Note: This figure illustrates the marginal posterior distributions for α, α+θ and θ, where α is log-income of farmers receiving free seeds, α+θ the log-income of farmers receiving vegetable-based extension and θ the difference in log-incomes.

TESTING FOR NON-INFERIORITY

We may also use the output from Stan to calculate other decision-relevant probabilities. Formally, this is equivalent to changing the decision-relevant question specified in the example above. For instance, let us assume that the first intervention (seed distribution) is much less expensive than the second (vegetable extension). In that case, the decision-maker may prefer implementing the seed program as long as the probability of it not being much inferior to the extension is large enough. We formalize this decision rule as: *"Implement free seed distribution if* $Pr(\theta < 0.1) > 0.7$*."* In other words, as long as we think there is more than a 70% chance that seed distribution is not 0.1 (in log-income terms) worse than extension services, the decision-maker will implement the seed distribution. Using the output from our estimation above, we calculate this probability as 44%. In other words, there is a 56% chance that vegetable extension is much more effective (by 0.1) than seed distribution. Based on this finding, the evaluator should recommend the decision-maker to implement vegetable-based extension services over free seed distribution.

The main takeaway for evaluators of decision-focused scenarios is that estimating the marginal posterior distribution of the average treatment effect allows to answer decision relevant questions about head-to-head comparisons and tests for non-inferiority with easy-to-interpret probabilistic statements.

COMPARISON TO MAXIMUM LIKELIHOOD ESTIMATION

We motivated Bayesian analysis by comparing it to maximum likelihood estimation (MLE). In this section, we therefore compare the results from estimating the Bayesian model with those from MLE.

Table 2 provides this comparison through two sets of columns. The first set of columns provides coefficient estimates and corresponding standard errors and 95% confidence intervals from MLE. The second set of columns provides the mean, standard deviation and 95% credible interval of the posterior estimates from Bayesian model estimation. The Bayesian 95% credible interval contains values that provide 95% of the probability mass of the posterior for the given parameter. Intuitively, the parameter is in this interval with 95% probability.[59] In comparison, the 95% confidence interval from MLE is constructed such that out of every 100 replication studies at least 95 of intervals would include the *true* value of the parameter. This means, for any given interval, it is impossible to say whether the true parameter is included in it.

In our example, we find the point estimates, standard errors and confidence intervals for α, θ, σ from MLE to be nearly identical to the means, standard deviations and credible intervals of the posterior from Bayesian estimation.

[59] The interval containing the parameter values resulting in the highest posterior density is also called the highest (posterior) density interval (HDPI). This interval will not be equal-tailed whenever the posterior distribution is not symmetric. In such cases, evaluators could instead report equal-tailed intervals (ETI). For more information refer to (Makowski, et al., 2019)

Table 2 Comparison of Maximum Likelihood and Bayesian Posterior Mean Estimates

	MLE			Bayesian		
	Coefficient Estimate	*Standard Error*	*95% Confidence Interval*	*Posterior Mean*	*Posterior SD*	*95% Credible Interval*
$\boldsymbol{\alpha}$	2.263	0.098	[2.070, 2.457]	2.263	0.099	[2.069, 2.458]
$\boldsymbol{\theta}$	0.119	0.137	[-0.150, 0.389]	0.118	0.140	[-0.160, 0.387]
$\boldsymbol{\sigma}$	0.969	0.048	[0.881, 1.072]	0.980	0.049	[0.888, 1.082]

Note: The first set of columns provides coefficient estimates, standard errors and 95% confidence intervals from maximum likelihood estimation. The second set of columns provides means, standard deviations, and 95% credible intervals from the Bayesian model estimation.

We can explain this observed similarity of results by looking again the posterior specification provided in equation (13) above. From this equation, we note that the posterior is a precision weighted average of the likelihood, $p(y|\alpha,\theta,\sigma,t)$, and the prior, $p(\alpha,\theta,\sigma)$. In this example, we used flat priors for all parameters, effectively resulting in multiplying the joint likelihood distribution with a (nearly) constant value (see Figure 4). For this reason, as sample size increases, the posterior from Bayesian analysis with a flat prior is proportional to the likelihood function.[60] That is, we have:

$$p(\alpha,\theta,\sigma|y,t) \propto p(y|\alpha,\theta,\sigma,t)\, p(\alpha,\theta,\sigma) \propto p(y|\alpha,\theta,\sigma,t) \quad (17)$$

As a consequence, maximum likelihood estimation, i.e. finding the parameters that result in the highest value of the likelihood, will result in the same values as finding the maximum of the posterior, as the latter approaches the likelihood function when using flat priors.[61]

There are still advantages of using Bayesian analysis with a flat prior compared to MLE, because the former provides the joint distribution of all parameters in the model. This feature allows evaluators to make easily interpretable probabilistic statements (like $\Pr(\theta > 0) = 0.8$). Such statements cannot be made in frequentist analysis since it never generates probability distributions of parameters.[62] That being said, in cases with a large sample size or uninformative priors, researchers may be tempted to conduct a maximum likelihood estimation (or OLS for that matter), and simply interpret its output as representing a probability distribution. This feels reasonable since one would obtain a similar posterior from conducting Bayesian analysis. We remain agnostic on whether this is acceptable, and leave this matter in the hands of statistical philosophers.

BAYESIAN ESTIMATION WITH INFORMATIVE PRIORS

We now illustrate further differences between Bayesian analysis and MLE by incorporating substantive ex-ante information about the intervention into Bayesian analysis. MLE is unaffected by

[60] The results are only nearly equivalent in our approach because of the relatively small sample size of 200. As sample size increases, the already negligible weight of the prior is expected to decrease further, making the results asymptotically equivalent. Further differences in results may occur because of differences in the optimization algorithms used by the two approaches.

[61] Generally, the posterior density asymptotically approaches the likelihood function, for any prior, as the sample size increases towards infinity (Jackman, 2004).

[62] If we were interested in making probabilistic statements about the relationship of model parameters, the comparison above suggests that maximum likelihood estimation could lead to nearly identical results as fully Bayesian analysis with uninformative priors.

such information as estimates are derived from maximizing the likelihood function. In contrast, the Bayesian posterior is akin to a precision weighted average of the likelihood and the prior.

In the example above, a policy-maker intends to decide between two programs with the goal to increase farmers' (log) income: distributing free seeds and hiring extension agents. The model in equation (15) required us to specify priors for the treatment effect of free seed distribution, α, and the difference in treatment effects, θ. The base model with uninformative priors is given by equation (16).[63]

To illustrate the effect of informative priors, we derive different scenarios that are based on different assumptions about available ex-ante information. First, we assume existing survey data indicates farmers' average log income in $'00 in the study area as 2.1 with standard deviation 0.5. In Scenario 1, we use this information without any additional information about the program effects and specify the priors

$$\text{Scenario 1:} \quad p(\alpha) \sim N(2.1, 0.5^2), p(\theta) \sim N(0, 10^2), p(\sigma) \sim fold.N(0.5, 2^2), \tag{18}$$

where $fold.N$ indicates the folded normal distribution.[64] Intuitively, the prior for α mirrors the observed income distribution in the region and suggests that under free seed distribution 90% of farmers earn more than $592 and 90% earn less than $1,125. The prior for θ is fairly vague (high variance) and states that we do not have a preference for either program (mean zero).

In the second scenario, we assume in addition that the literature suggests both interventions increase farmer incomes. Suppose, meta-analyses estimated the effect of free seed distribution programs on log-income at 0.3 with standard error of 0.1 and the effect of specialized extension programs at 0.1 with standard error of 0.3. In other words, the literature finds free seed distribution programs, on average, three times as effective as specialized extension. At the same time, the former program's treatment effects also vary more widely. We specify the prior in Scenario 2 as

$$\text{Scenario 2:} \quad p(\alpha) \sim N(2.4, 0.1^2), \; p(\theta) \sim N(-0.2, 0.3^2), \; p(\sigma) \sim fold.N(0.5, 2^2). \tag{19}$$

The prior mean for α now incorporates the treatment effect found in the literature. The prior for θ is also informative and assigns a 75% probability that the treatment effect of free seed distribution is larger than that of specialized extension.

We use the same simulated dataset as above, where we set $\alpha = 2.3$ and $\theta = 0.05$ and generate 200 observations of the outcome (100 per treatment arm) by adding random noise from a standard normal distribution to the respective individual means.

[63] The premise of the head-to-head comparison example was that the decision-maker did not have any ex-ante information about one program being better than the other. If such a preference existed, and we specified an informative prior for the difference in program effects, θ, we could also more easily specify a one-sided null-hypothesis test in a frequentist approach.

[64] If x is distributed normally, $y = |x|$ follows a folded normal distribution. The folded normal distribution accounts for the fact that the standard deviation parameter cannot take negative values. In addition, it does not impose a hard constraint on the upper limit for the standard deviation. The folded normal is also called half-normal if the mean is zero and the standard deviation is one.

Table 3 compares the estimation results from these scenarios with the Base Scenario in equation (16). As before, we illustrate the posterior mean, standard deviation and 95% credible interval estimates for all three model parameters. We observe that the differences in results between the Base Scenario and Scenario 1 are negligible. This is not surprising because the prior does not contain any information regarding the effectiveness of the two programs. In both cases, the results suggest specialized extension increases (log) farmer incomes more than free seed distribution. Put differently, including contextual information such as average income and standard deviations in the study area do not influence the results. In contrast, Scenario 2 takes into account evidence from the literature, which finds free distribution more effective than specialized extension. While our data provides evidence of the opposite (the posterior mean for θ is larger -0.2), that evidence is not strong enough to result in a positive posterior mean estimate for θ. This finding is also reflected in the lower probability with which specialized extension is more effective than free seed distribution (13% compared to 80%). These findings suggest that informative priors might have a strong influence on the analysis and that cautious decision makers keen to take into account existing information may need to seek stronger evidence.

Table 3 also provides results for a Scenario 3, which uses the same priors as Scenario 2 but assumes a significantly larger study (2000 observations instead 200). Comparing the results of this scenario to those of Scenario 2, we observe that the posterior mean for θ turns positive and that $P(\theta > 0)$ increases to 55%. The larger sample size evidently changes the analysis by putting relatively more weight on the likelihood over the prior (see equation (13)).[65] Being aware of the effects that priors may have on the results, evaluators should pre-specify and conduct a sensitivity analysis to assess how alternative sets of priors would change their findings.

Table 3 - Posterior Mean Estimates for Models with Informative and Uniformative priors

	Base Scenario	**Scenario 1**	**Scenario 2**	**Scenario 3**
$\boldsymbol{\alpha}$	2.26 (0.1) [2.07, 2.46]	2.26 (0.1) [2.06, 2.45]	2.37 (0.08) [2.22, 2.53]	2.32 (0.03) [2.26, 2.37]
$\boldsymbol{\theta}$	0.12 (0.14) [-0.16, 0.39]	0.12 (0.14) [-0.16, 0.39]	-0.09 (0.08) [-0.25, 0.07]	0.01 (0.04) [-0-07, 0.09]
$\boldsymbol{\sigma}$	0.98 (0.05) [0.89, 1.08]	0.98 (0.05) [0.89, 1.08]	0.98 (0.05) [0.89, 1.09]	0.97 (0.02) [0.95, 1.01]
$\boldsymbol{P(\theta > 0)}$	80%	82%	13%	55%
N	200	200	200	2000

Note: The table provides means, (standard deviations), and [95% credible intervals] of the Bayesian posterior estimates from the scenarios specified in equations (16), (18) and (19), respectively. N indicates the studies sample size.

[65] In practice, we would want to supplement the above steps of estimating the specified Bayesian model with various additional steps to ensure the assumptions underlying our model are reasonable and that the output from the software can be trusted. First, as discussed in a comment above, prior to using output from Stan, we should always check that the software "converged" through posterior predictive checks (Jackman, 2004). Second, we would want to perform additional sensitivity analysis on our priors to ensure that the final results are not sensitive to seemingly small changes in the prior specification. Third, we would want to perform various checks on the overall model to ensure that the model is a reasonable approximation of the data. In the online code repository accompanying this example, we provide some simple advice on how to check that a model has converged and demonstrate one approach to checking the model. Also refer to (Schad, et al., 2019) for more information on prior predictive checks and setting up a Bayesian workflow.

Finally, at first glance, it seems natural that the decision-maker would adopt one intervention if there is more than a 50% chance that it is better than the other intervention. Yet, if there is a 51% probability that one is just a tiny bit better than the other, but a 49% probability that the former is much worse, she would likely not adopt it. For this reason, we had introduced the concept of decision rules, such as "Implement free seed distribution if $\Pr(\theta < 0.1) > 0.7$." Yet, evaluators may want to incorporate more complex utility (or loss) functions that better captures such trade-offs into the analysis. Intuitively, utility/loss functions quantify the decision-maker's utility/loss when opting for the an intervention for various values of the parameter of interest, θ. Bayesian analysis would then allow evaluators to maximizes decision-makers' expected utility. While exploring the depths of Bayesian Decision Analysis is out of scope of this paper, being able to estimate decision-maker's expected utility for any given choice is key additional benefit of Bayesian analysis (Smith, 2010).

As in the standard approach to impact evaluation, the findings above suggest that it is important for evaluators to be able to judge whether Bayesian DFE's are well-designed. We will provide further insights on criteria that indicate how well Bayesian DFE's are designed further below.

The main takeaway for evaluators of decision-focused scenarios is that incorporating informative priors may influence the analysis and that evaluators need to be aware of the direction of such influence. Evaluators should pre-specify and conduct a sensitivity analysis to assess how different priors affect their results.

5.3. COMPARING TO A BENCHMARK

The "comparison to a benchmark effect size" scenario from the introduction described a common but tricky situation for evaluators. To recap the example, a donor wants to know whether a school-feeding program increases height-for-age scores by at least 0.2 standard deviations. She is somewhat skeptical that the new intervention has an effect this large and commissions an impact evaluation to find out.

Above, we illustrated how a frequentist analysis does not allow us to directly estimate the probability that the effect of the new intervention is larger than a pre-determined benchmark. Instead, we proposed a set of two hypothesis tests: a first hypothesis test at a high level of significance to confirm that the effect is positive and a second hypothesis test with a lower level of significance to confirm that the effect is indeed greater than the benchmark. If we reject both of these hypotheses, we may be somewhat confident that the intervention meets the threshold. But the frequentist approach does not allow us to state the probability that the intervention meets this threshold.

Bayesian analysis allows us to directly estimate the probability that the effect of an intervention is larger than a desired threshold while taking into account the donor's skepticism. Analogously to the previous section, we first build a simple model for the outcome. We then assign priors to all of the parameters in the model. Lastly, we fit the model and perform inference using R and Stan.[66]

EXPERIMENTAL SET-UP AND LIKELIHOOD

We start with the basic equation that we would use to estimate the average treatment effect in the standard approach to impact evaluation. Let y_i be the height-for-age z score (HAZ) for child *i* and t_i an

[66] Code for this example, and all other examples in this paper, can be found at https://github.com/IDinsight/dfe_methods

indicator for whether the child received the intervention. Assuming the intervention is randomized at the child-level[67], we would estimate the impact using the following regression equation:

$$y_i = \alpha + \theta t_i + \varepsilon_i \quad (20)$$

We use this model to specify our data more, i.e. our likelihood as follows:

$$y_i \sim N(\alpha + \theta t_i, \sigma^2) \quad (21)$$

To simplify the calculations, we assume $\sigma^2 = 1$. While we would rarely want to assume fixed variance in an actual model, the standard deviation of height-for-age is generally around 1 for most populations and thus this is a fairly benign assumption.[68]

PRIORS

To complete the model, we need to specify priors for the remaining parameters α and θ. Suppose that we elicited the donors' prior and she believes there is a ~16% chance that the intervention has an effect as large or larger than 0.2 *and* that there is a 16% chance that the effect is less than 0.[69] This corresponds roughly to the prior:

$$\theta \sim N(0.1, 0.1^2) \quad (22)$$

We also assign a normal prior for α:

$$\alpha \sim N(-1.7, 2^2) \quad (23)$$

The mean of this prior is based on the prevalence of stunting among school-aged children in the context.[70] Our prior for the variance of α is relatively high which means that the data will drive our estimates for this parameter. The full model is summarized as follows:

$$y_i \sim N(\alpha + \theta t_i, 1^2);\ \theta \sim N(0.1, 0.1^2);\ \alpha \sim N(-1.7, 2^2) \quad (24)$$

RESULTS

As in the previous section, we solve for the parameters using the model in equation (24) and using simulated data from 150 observations.[71] The estimation result is an R object that contains simulated draws from the posterior distribution of the model parameters, which we can use to make *any* probabilistic statement about them. In this example, the evaluator is interested in the probability that the average treatment effect is larger than the threshold, $P(\theta > 0.2)$. Therefore, we calculate

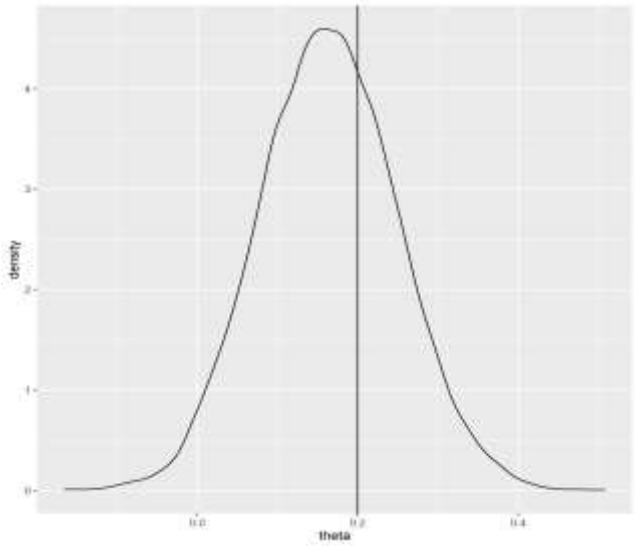
$$\mathrm{P}(\theta > .2) = \sum_{i=1}^{iter} 1(\theta_i > .2), \quad (25)$$

[67] Randomization at the individual child-level would likely not feasible in this example of a school feeding program. We adopt this simplification to make the presentation more straightforward.
[68] HAZ scores are standardized with reference to WHO's global database of child growth and malnutrition data (https://www.who.int/nutgrowthdb/en/). Thus, unless the variance of height-for-age differs significantly from this reference population the standard deviation should be approximately 1.
[69] Prior elicitation is an art. There is more and more interest in forecasting and prediction methods in social science. See for example https://socialsciencepredict ion.org/, http://www.tonyohagan.co.uk/shelf/index.html, DellaVigna and Pope 2018, Haaland, Roth and Wohlfart 2020, O'Hagan, et al. 2006, O'Hagan 2019.
[70] If $\alpha = -1.7$ and $Var(y_i) = 1$ (see above) the prevalence of stunting, defined as $HAZ < -2$, would be roughly 38%, which is approximately the level of stunting among children aged below five in Nigeria at the country level.
[71] The data generating process is found in the accompanying code for this example.

where 1(*condition*) is the indicator function, *iter* is the total number of draws, and θ_i is the *i*th draw from the posterior for θ. The resulting probability is 33%.

Alternatively, we can look at the entire distribution of θ by graphing the posterior.

Figure 7 – Posterior Distribution of θ



The resulting posterior and probability estimate directly informs the donor's pre-defined decision rule. For instance, the donor could have agreed to scale up the program if the posterior showed a larger than 70% probability of the treatment effect being greater than the desired threshold of 0.2.

The main takeaway for evaluators of decision-focused scenarios is that a Bayesian approach allows to directly estimate the probability of an intervention surpassing a decision-relevant benchmark. This approach provides a better tailored answer to the decision-relevant question than the corresponding frequentist adaption of the standard approach to impact evaluation.

5.4. SAMPLE SIZE CALCULATIONS FOR BAYESIAN APPROACHES

In the section above, we have illustrated three common scenarios in which Bayesian approaches to decision-focused evaluations can provide better-tailored insights to decision-makers compared to the standard approach to impact evaluation or frequentist adaptions. In this section, we discuss desirable characteristics of Bayesian evaluation designs. We illustrate how evaluators interested in applying Bayesian approaches to decision-focused evaluations could think about the sample sizes required to result in reliable and trustworthy evidence. We also compare how these considerations and sample sizes compare to the standard approach and the frequentist adaptions discussed above. First, we recap the standard approach to calculating sample size for impact evaluations. We then illustrate one approach to calculate sample size for the Bayesian approach, compare the results and provide general insights. To illustrate these approaches, we consider the "Comparing to a benchmark effect size" scenario. We hope this guidance will be useful for evaluators who consider implementing Bayesian approaches to DFEs.

Regardless of the approach to inference, evaluators should propose a sample size that makes both implementer and funder sufficiently confident in the findings of the study. Sample sizes which are too large waste valuable resources, while sample sizes which are too small may result in inaccurate or non-informative study findings.

RECAP OF THE CONVENTIONAL APPROACH TO CALCULATE SAMPLE SIZE

In frequentist approaches, evaluators consider the evidence provided by a particular study design credible if the following two conditions hold:[72]

1. A sufficiently large proportion of (hypothetical) replication studies reject the null hypothesis if the null is indeed false *(sufficiently high statistical power or sufficiently low Type 2 error).*
2. A sufficiently low proportion of (hypothetical) replication studies reject the null if the null is indeed correct *(sufficiently low Type 1 error; level of significance).*

A "successful" study is one that rejects the null if the null is false or one that does not reject the null if the null is true. Since evaluators never know whether the null is actually true, they generally consider a *program* "successful" when they reject the null-hypothesis (of no effect) in the *study*. The above conditions therefore provide probability thresholds for what generally constitutes a credible study design. Inherently, this approach is based on the assumption that decision-makers are only interested in funding/implementing programs that are successful in the long-run. For this reason, decision-makers desire studies with sufficiently high statistical power and sufficiently low Type 1 error.

To calculate a study's required sample size, the evaluator has to specify thresholds or error rates for both of the above conditions. As stated above, in the standard approach to impact evaluations, the corresponding levels of significance and statistical power are 0.05 and 0.8, respectively. However, we have made the point that a different level of significance might be reasonable for a DFE. In contrast, we recommend evaluators to abstain from conducting studies with drastically lower than 80% statistical power.[73]

In addition, the evaluator will have to specify the minimum (standardized) effect size relevant to inform a decision.[74] Evaluators should align on this effect size with both the implementer and the funder. For example, suppose a funder is only interested in supporting programs that increase height-for-age scores by 0.2 standard deviations. This benchmark effectively defines the decision relevant effect size because programs with smaller effects will not be considered eligible to receive funding.[75]

[72] Note that these conditions refer to the study design. There are other characteristics of a study that determine whether the ultimately produced evidence should be considered credible. For more information on best practices in research credibility and transparency refer to the BITSS library, https://www.bitss.org/resource-library/.
[73] Higher statistical power provides us more confidence to detect the hypothesized (or larger) effect sizes and results in a larger probability to conclude that the program is "successful" when it indeed is. A study with low statistical power has a high chance of generating inconclusive results and is therefore typically not worth its cost. In cases where there are high costs of having an inconclusive study, higher than conventional power may be warranted.
[74] Standardized effect sizes are applicable to all types of outcome variables (binary, continuous, count). Usually, specifying standardized effect sizes requires to also understand initial distribution of the outcomes of interest. This understanding also allows to translate standardized effect sizes into level effects.
[75] Effect sizes of similar programs in comparable contexts might be found in the literature and could inform the decision relevant effect size.

The three ingredients – level of significance, statistical power, and decision relevant effect size – form the basis for sample size calculations.[76]

We now consider the "Comparing to a benchmark effect size" scenario. Let us assume that we agreed with the funder and implementer on a decision relevant effect size of 0.2 standard deviations and that we design a study with 80% power. To assess the intervention's effectiveness, we propose a "double barreled hypothesis test". The hypothesis tests are:

$$\begin{aligned} &\text{Test 1.} \quad H_0{:}\,\theta <= 0,\ H_1{:}\,\theta > 0\,,\ \alpha = 0.05 \\ &\text{Test 2.} \quad H_0{:}\,\theta <= .2,\ H_1{:}\,\theta > 0.2,\ \alpha = 0.2 \end{aligned} \tag{26}$$

To obtain the appropriate sample size for this study, we would calculate the sample size required for each of these tests and go with the larger one. To do this, we have to assume an actual effect size we want to power for, known as the minimum *detectable* effect size (MDES). Even though our decision-relevant threshold is 0.2, we can't use that as our MDES, because our second hypothesis test requires that we find an effect larger than 0.2.[77] For this reason, we use a MDES of 0.25 for these tests. This MDES implies that the effective MDES of Test 1 is 0.25 and that of Test 2 is 0.05. Due to this difference in effective MDES's, the first hypothesis test requires about 400 students (200 per group), whereas the second test requires about 4,600 students (2,300 per group). As a consequence, to complete both tests with sufficient power we require a sample size of 4,600 students.

A BAYESIAN APPROACH TO SAMPLE SIZE

How should evaluator think about sample size when employing a Bayesian approach to evaluation design? In a Bayesian design, evaluators' recommendations will be based on (some characteristic of) the posterior distribution of the parameter(s) of interest. Evaluators can therefore formalize the quality of these recommendations by collaborating with the decision-maker to pre-specify a decision rule that translates (the characteristic of) the posterior into specific actions the decision-maker should take and investigate how "reliable" it would be to follow this decision rule for a particular sample size.

The goal of Bayesian sample size calculations is to find the lowest sample size for which stakeholders feel confident that following the decision rule when presented evidence from *one* particular study (with that sample size) is *generally* the "right" thing to do. In general, decision-makers will want to avoid supporting a program when the program is not as effective as desired (Type I error). At the same time, decision-makers will require study designs to exhibit a sufficiently large probability of indicating support for a program if the program is indeed as effective as desired (Statistical Power). Evaluators will recognize the familiarity of these conditions with the concepts of frequentist Type 1 error and

[76] There are additional considerations to be taken into account. For example, sample size calculations for clustered designs will have to account the expected intra-cluster correlation (ICC).
[77] If the MDES was 0.2, we would need an infinite sample to reject Test 2.

statistical power.[78] These two conditions therefore imply that the analog of Type 1 error and statistical power are also key study design characteristics for Bayesian evaluation designs (Chen, et al., 2011).[79]

We illustrate a Bayesian approach to calculate sample for the "comparing to a benchmark effect size scenario". In this scenario, assume the funder is willing to support the school feeding program if the (estimated) probability of its effect on height-for-age scores being larger than the desired threshold is greater than 80%, or $\hat{P}(\theta > 0.2) \geq 80\%$. This is just another way of saying that the decision maker tolerates a 20% risk that the program is less effective than the desired benchmark.[80]

How should evaluators think about these study characteristics in a Bayesian evaluation design? We define a Bayesian evaluation design to consist of a likelihood function, a prior, a decision rule, and a given sample size. We assume that the evaluator designs the comparing to benchmark effect size study above. Given the Bayesian evaluation design, the evaluator can generate a large number of simulated study results from this design by sampling repeatedly from the corresponding data generating process. We will walk through this process with an example below. To conceptualize the Type 1 error for a Bayesian design, let us assume that the studied intervention has no effect on the outcome of interest, i.e. that the decision rule should not be followed.[81] The frequentist Type I error of the Bayesian study design captures the proportion of simulated studies for which the results suggest the decision-maker should follow the decision-rule - despite the program being hypothesized to have no effect.[82] To conceptualize statistical power of the Bayesian design, let us instead assume that the intervention was *somewhat* more effective than the minimum desired threshold. Statistical power of the Bayesian design is then given by the proportion of simulated studies with that *hypothesized* effect size for which the results actually support the decision. To gauge the minimum required sample size that meets the desired thresholds for both of these criteria, evaluators can simulate a sufficiently large number of studies from a given Bayesian design.[83]

[78] Moreover, decision-makers are likely engaging in multiple research studies and are therefore expected to also care about the long-term quality of these studies. This notion of ensuring long-term quality through avoiding false positives and false negatives is formalized in the frequentist concepts of *Type 1 Error and statistical power.*

[79] Gubbiotti and De Santis (2008) compare Bayesian and Frequentist approaches to determine a study's statistical power. Also, if the researcher is willing to specify the decision-maker's utility function, they can optimize sample size the maximize utility.

[80] Evaluators utility/loss function may require evaluators to formalize different decision rules for other scenarios. For example, evaluators could require the highest posterior density interval (HPDI), which usually captures 95% of the posterior density, to exclude the threshold value. In the testing for non-inferiority scenario, evaluators could also specify an interval in which the decision-maker considers the intervention effectively equivalent to status quo. For more details, refer to Chapter 2 of Berry et. al (2010). Alternatively, decision rules could capture the width of the highest posterior density interval to not exceed a certain threshold. Such criteria are useful in applications in which a particular margin of error should be not exceeded (e.g. a sample survey). Lastly, evaluators could model a decision rule by formalizing and maximizing the decision maker's utility/loss function (Smith, 2010).

[81] This assumption is only valid for the "comparing to a benchmark" scenario. The "head-to-head" comparison and "Testing for non-inferiority" scenarios may require a different "null" effect size.

[82] The decision-maker would be "wrong" because the action implied by the decision rule opposes the right decision under the hypothesized *zero* effect size. Note that assuming a fixed effect sizes seems at odds with a Bayesian approach as such approaches generally treat parameters as stochastic, i.e. to follow a probability distribution. See the following footnote for more detail.

[83] To simulate the data, evaluators should draw from the "design" prior for all parameters that are not hypothesized as "fixed". Evaluators could also obtain the Bayesian equivalent of frequentist Type 1 Error and statistical power by instead drawing from the prior distributions for *all* parameters, including those hypothesized as fixed in the main text. The Bayesian equivalent of statistical power is called *assurance* and provides the proportion of simulation studies in which the decision should be taken

We illustrate the process to assess these two characteristics of a Bayesian design step-by-step for the "comparing to a benchmark" scenario. To recap, in this scenario a funder will support a program if its effect is greater than 0.2 standard deviations. First, we assess frequentist Type I error rates for Bayesian evaluation designs that only differ in sample size. We start by selecting a set of possible sample sizes for the study. For illustrative purposes, we choose a wide range of sample sizes between 50 and 5000. For each of these sample sizes, we simulate 1000 datasets from the likelihood, i.e. the data generating process of the design, which assumes that there is no effect of the intervention. We estimating the results for each of these studies and derive the corresponding decision according to the pre-defined decision rule of the design.[84] Here, the decision rule is to support the program if the probability of its effect being larger the benchmark is large enough. Formally, we assess whether $\hat{P}(\theta > 0.2) \geq X$, where X is the "sufficiency threshold" – the lowest probability that needs to be reached to still support the program. We then obtain the frequentist Type 1 error for a given sample size and sufficiency threshold as the proportion of simulated studies for which the results suggest to support the intervention despite the imposed program effect to be zero.

The left panel of Figure 8 illustrates Type 1 error rates for the designs of varying sample sizes and decision rules, i.e. sufficiency thresholds.[85] For example, the results of approximately 275 of the 1000 simulated studies (27.5%) with sample size 50 and 50% sufficiency threshold, i.e. requiring 50% of posterior probability above the 0.2 threshold, suggest to fund the program despite it having no effect. For a given sample size, we observe a reduction in the type-1 error rate as the sufficiency threshold increases from 50% to 80%. The increased sufficiency threshold is a stricter decision rule in which the decision-maker requires more "certainty" of a study design. Stricter decision rules are less likely to lead to a "wrong" decision. Analogously, for a given sufficiency threshold or decision rule, we observe a reduction in the Type 1 error rate as the sample size increases. Increasing the sample size leads to more "precise" posterior distributions around the true null effect.[86] We indicate the designs that meet the conventional criterion of 5% or lower Type 1 error rate through the horizontal line. At the same time, we do re-iterate our message from above (in the section on frequentist analysis) that evaluators should critically investigate their Type 1 error rate threshold.

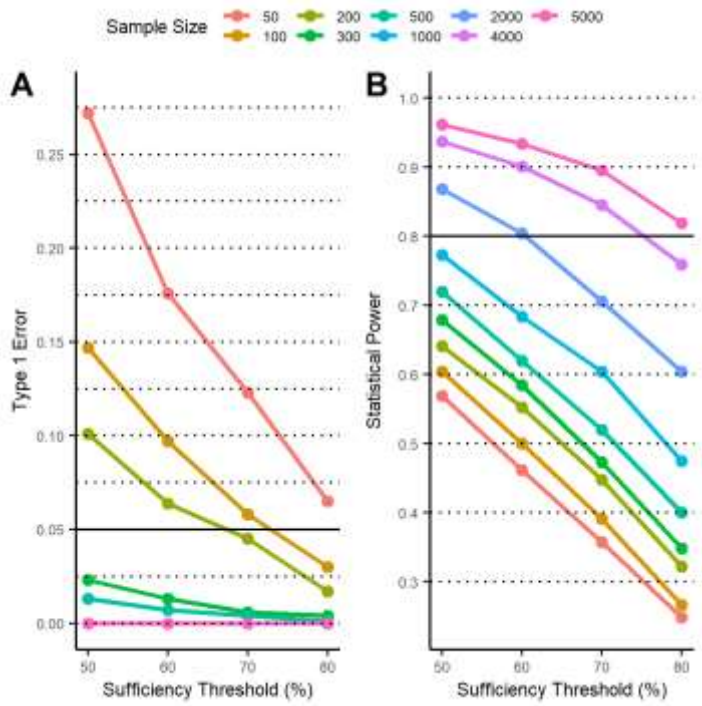
given the current prior. Since the prior is incorporated in full, it is more difficult to delineate between an objective right or wrong decision in this case. There may also be situations in which evaluators may want to use a different prior in study design as compared to the prior used in the analysis (Gelman, Simpson and Betancourt, 2017). For example, many evaluators will prefer the analysis being driven by the evidence (rather than the prior) and therefore use a weakly/uninformative prior in the analysis stage. At the same time, evaluators might want to integrate a funder's skeptical prior about the intervention's effectiveness when determining the ideal sample size for the study. These differing priors are usually referred to as "analysis" and "design" priors, respectively (Berry, et al., 2010).

[84] In this example, the estimation is based on uninformative analysis priors. We specify the prior for the constant as N(-1,10) and the prior for the average treatment effect as N(0.1,100).

[85] The results are based on the approach described in the preceding paragraph and based on uninformative analysis priors.

[86] This increase in precision is analogous to the increase in precision observed in maximum likelihood estimation as sample size increases.

Figure 8- Frequentist Type-1 error rates and Statistical Power of Bayesian study designs



Note: This graph provides Type 1 error rates (left) and statistical power (right) of Bayesian impact evaluation designs. Each dot this graph provides the Frequentist Type-1 error rate/statistical power of a Bayesian design with the respective sample size and sufficiency threshold. The horizontal line indicates designs with lower than 5% Type 1 error rate and larger than 80% statistical power respectively.

In addition to the Type 1 error rate, evaluators should investigate the frequentist statistical power of these Bayesian designs. To this end, evaluators will generally assume a *fixed* effect size that (just) supports the decision rule, and then conduct simulations analogous to those above to obtain the proportion of simulated studies for a particular design that supports the specified decision rule.[87] In the running example, let us assume an effect size of 0.25. SDs, which is larger than the benchmark of 0.2 SDs. In the accompanying code, we simulate 1000 datasets from the likelihood of the design (which assumes this treatment effect) for each of the sample sizes. Next, we estimate the posterior for each of the simulated datasets.[88] Finally, we obtain frequentist statistical power of the Bayesian designs as the proportion of studies for a given sample size and sufficiency threshold for which the decision rule is met (as it should be given the assumption).

The right panel of Figure 8 illustrates the frequentist statistical power for the designs of varying sample sizes and decision rules, i.e. sufficiency thresholds. For example, the results of approximately 600 of the 1000 simulated studies with sample size 100 and sufficiency threshold 50% indicate to fund the

[87] The closer this chosen effect size is to the decision threshold (of 0.2 in our case), the lower – and the more accurate – evaluators should expect that proportion of studies to be. Therefore, clearly stating the effect size used in the simulations is a key requirement for evaluators to assess the credibility of the frequentist statistical power of Bayesian designs.
[88] The estimation is based on the same uninformative analysis priors as for the estimation of the Type 1 error rate.

program (despite the evaluator knowing that the treatment is larger than desired benchmark). For a given sample size, we observe that requiring a higher sufficiency threshold reduces a design's statistical power. Analogously, for a given sufficiency threshold, increasing the sample size of a design increases statistical power. Again, this patterns can be explained with the estimated posterior becoming more precise around the true effect for larger sample sizes.[89] We indicate designs that meet the conventional criterion of 80% or more statistical power through the horizontal line.

Given both the Type 1 error rates and the statistical power of these Bayesian designs, evaluators should choose the smallest sample size that meets *both* of the pre-specified criteria thresholds. For example, for Bayesian designs with sufficiency threshold 60% and conventional levels of 5% and 80% of Type 1 error and statistical power, respectively, the evaluator should choose a sample size of 2000.

This approach to Bayesian sample size calculations poses two questions. First, how would the results differ if evaluators used informative priors to design the study? Second, how do the results compare to the corresponding frequentist approach in the "comparing to a benchmark" scenario? We answer each of these questions in turn.

The main takeaway for evaluators of decision-focused scenarios is that the familiar concepts of "Type 1 Error" and "Statistical Power" also define the quality of Bayesian study design. Evaluators should simulate studies from a given design with varying sample sizes to identify the lowest sample size that still meets the required threshold for both of these characteristics.

HOW DOES THE CHOICE OF PRIOR AFFECT BAYESIAN SAMPLE SIZE CALCULATIONS?

The sample size calculations above are based on estimation results for a large number of simulated studies from a given Bayesian design. So far, we used uninformative priors in this estimation process. This means that the results were mostly driven by the (simulated) data. How does the required sample size change if evaluators plan to specify an informative prior for (some of) the parameters of interest?

Generally, informative priors are either more optimistic or more skeptical of the intervention.[90] Below, we illustrate how both types of priors affect the required sample size. In our running example, let us assume that we elicited priors from the implementer and several potential funders of the school feeding program. The implementer is very optimistic of the program's impact. We specific this very optimistic and certain prior for the ATE as *N(0.3,0.1)*. One funder is somewhat optimistic of the program but less certain of its belief. We specify this optimistic prior as *N(0.3,0.5).* In contrast, other funders are more pessimistic of the intervention. One funder is very pessimistic and thinks that the intervention will on average decrease HAZ scores. We specify the prior as *N(-0.1,0.3)*. Another funder is less pessimistic but relatively certain the intervention will not achieve the desired 0.2. benchmark. We specify this prior as *N(0.1,0.1)*.

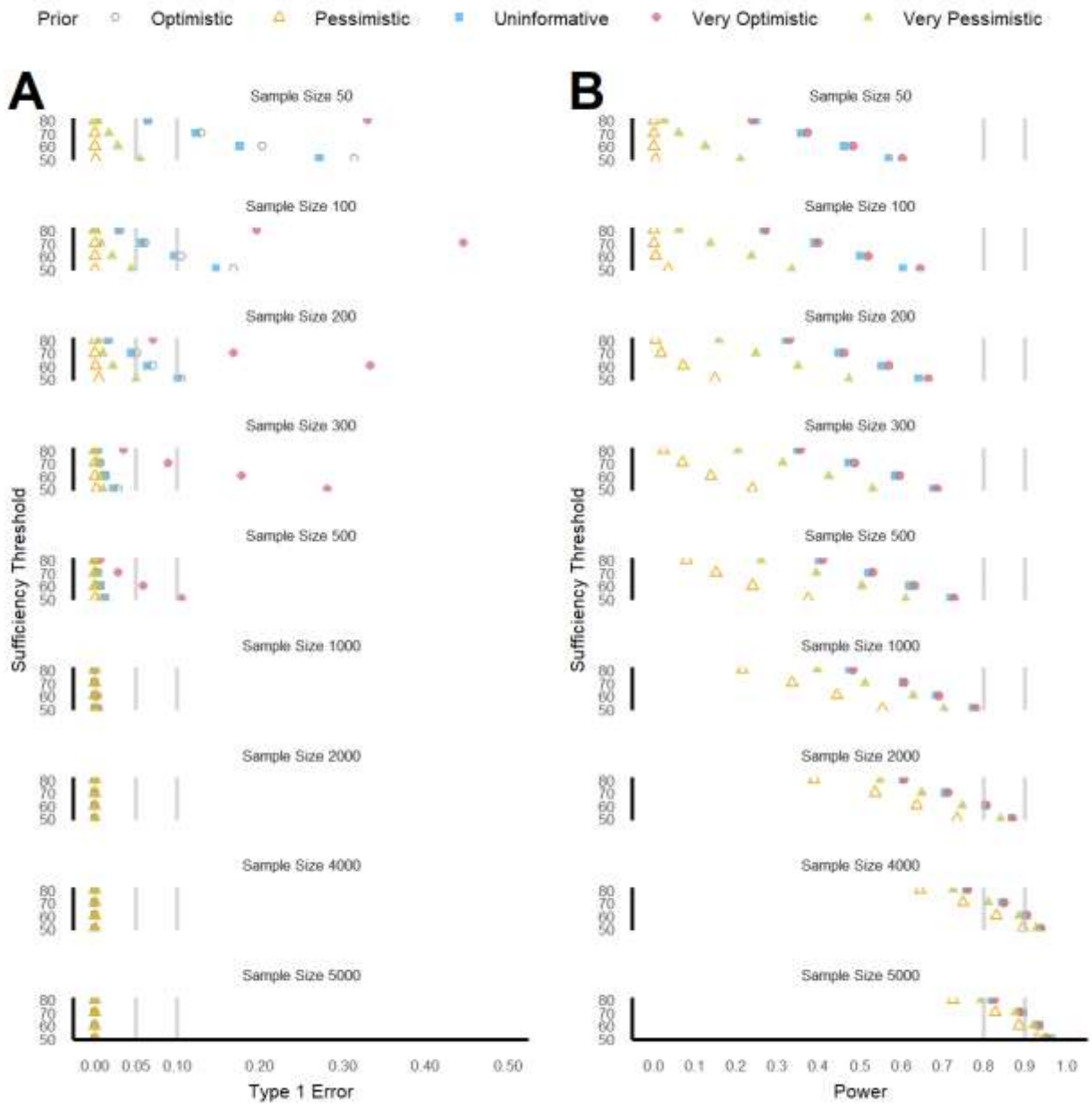
[89] Similar to frequentist designs therefore, we'd expect statistical power to converge to 1 as sample size increases further.
[90] We provide some general remarks on eliciting priors in the Appendix. Further, Carlin and Sargent (1996) and Greenhouse and Waserman (1995), for example, propose approach to assess the sensitivity of the posterior distribution with regards to different classes of prior distributions.

Figure 9 illustrates the results of simulations that use the respective priors specified above. There are several generalizable takeaways from these simulation results:

1. For any prior and sample size, the stricter the decision rule (here the higher the sufficiency threshold), the lower Type 1 error and the lower statistical power. This trade-off between Type 1 error and statistical power is similar to the tradeoff evaluators encounter in conventional evaluation designs.
2. The effect of the prior choice on both design characteristics is more pronounced, the smaller the sample size. Intuitively, this is explained by the increasing relative weight of the likelihood in equation (13) as the sample size increases. An interesting question is therefore what an appropriate sample size would have to be for evaluators not worry about formally incorporating strong priors into the design of a study. We illustrate that this question can be solved through simulation.
3. An optimistic (pessimistic) prior leads to larger (smaller) type 1 error rates than an uninformative prior. The effect is less pronounced as the sample size decreases for the same reason as above.
4. An optimistic (pessimistic) prior leads to larger (smaller) statistical power than an uninformative prior. The effect is less pronounced as the sample size decreases.
5. The stronger the prior, i.e. the more the mean deviates from the truth, and the more precise the prior, i.e. the lower its standard deviation, the stronger is the effect on the design characteristics.

The main takeaway for evaluators of DFE's is that incorporating decision-maker's ex-ante beliefs will affect the properties of a study. Generally, for a given sample size and the decision rule, optimistic priors tend to increase statistical power and Type-1 error. These effects of informative priors weaken as the sample sizes increase. Evaluators should therefore simulate the effects of the prior choice on the study characteristics during study design.

Figure 9 - Frequentist Type 1 error and Statistical Power of Bayesian Designs with different Analysis Priors



Note: This figure illustrates the Frequentist Type 1 error rates (left) and Statistical Power (right) for designs of the "Comparing to a Benchmark" scenario. For each sample size (denoted by the panels) and sufficiency threshold (denoted by the rows), the figure illustrates how the prior affects the design characteristics. We use five illustrative analysis priors: i) uninformative: *N(0.1,100)*, ii) very optimistic: *N(0.3,0.1)*, iii) optimistic: *N(0.3,0.5)*, iv) pessimistic: *N(0.1,0.1)*, v) very pessimistic: *N(-0.1,0.3)*.

HOW DOES THE REQUIRED SAMPLE SIZE FROM BAYESIAN DESIGNS COMPARE TO THAT OF THE ADAPTED FREQUENTIST APPROACH?

In the "Comparing to a benchmark" example in the section above, we proposed to tweak the standard approach to inference by conducting two one-sided hypothesis tests. The first test assesses whether the intervention's effect is statistically larger than zero, the second one whether the intervention's effect is larger than the desired benchmark, albeit with a lower level of significance. Evaluators might be interested in how the frequentist Type 1 error and statistical power of that adapted approach compare to that of the Bayesian approach proposed above.

Generally, such comparisons are complicated by two factors that may easily lead evaluators to compare apples to oranges. First, the two approaches are based on inherently different decision rules. In this example, the Bayesian approach is designed so that the decision to fund the program should be taken if the posterior probability above the benchmark value is large enough. In comparison, in the frequentist approach, the decision should be taken if the results of the study reject the null hypotheses of both tests. Both decision rules incorporate the notion of risk the decision-maker is willing to take through the sufficiency threshold and the level of significance, respectively. Below, we illustrate that the particular decision rules in this example in combination with similar levels of risk tolerance leads to comparable results design characteristics. Yet, we also demonstrate that these findings do not generalize by illustrating how different levels of risk tolerance lead to different design characteristics. Moreover, depending on the decision context, evaluators may choose a different decision rule in the Bayesian approach, leading to different results altogether.[91] Second, we illustrated in the previous section how the choice of prior can affect the characteristics of Bayesian designs. For this reason, the choice of prior may generally also influence any comparison with frequentist approaches.

We illustrate these points in Table 4. Columns (1) and (5) shows Type-1 error rates and statistical power for frequentist study designs that are based on two hypothesis tests for different sample sizes. For a sample size of 100, for example, 26 of the 1000 simulated studies (2.6%) would reject both hypothesis tests, even though the program had no effect. As expected, we observe that the Type 1 error rate decreases and that statistical power increases as sample size increases. Columns (2) and (6) illustrate analogous results for Bayesian designs that require a sufficiency threshold of 80%. Comparing these to the results of the frequentist approach, we observe that the Type 1 error rates and statistical power are nearly identical, and that similarity increases with larger sample sizes. As mentioned above, this finding follows from the coincidental choices regarding the decision rule and risk tolerance of the decision-maker in this example. Changing the decision rule, leads to incomparable results. This is illustrated by the differing results in columns (3) and (7) that show the characteristics of Bayesian designs that require a sufficiency threshold of 60%. Analogously, the choice of prior affects the comparison with the Frequentist approach. This is illustrated in columns (4) and (8) that show characteristics for Bayesian designs that require a sufficiency threshold of 80%, but instead of an uninformative prior use a very pessimistic prior in the analysis.[92]

[91] For example, evaluators could have required the 95% credible interval to exclude the desired benchmark of 0.2. Given this alternative decision rule, the evaluator would assess a different characteristic of the posterior distribution that may no longer be comparable to the asymptotic distribution(s) that determine(s) the decision in the Frequentist approach.

[92] The definition of priors is identical to the section before.

Table 4 - Comparison of Type 1 Error Rates of Double-barreled test and a Bayesian Design

	Type 1 Error Rates				Statistical Power			
Model No	1	2	3	4	1	2	3	4
Sample Size								
50	**0.06**	**0.065**	**0.123**	0.003	0.214	0.249	0.462	0.026
100	0.026	0.03	0.058	0.000	0.271	0.267	0.500	0.061
200	0.018	0.017	0.045	0.004	0.324	0.322	0.553	0.157
300	0.004	0.004	0.006	0.001	0.348	0.349	0.585	0.204
500	0.001	0.001	0.004	0.000	0.397	0.400	0.620	0.260
1000	0.000	0.000	0.000	0.000	0.478	0.475	0.684	0.395
2000	0.000	0.000	0.000	0.000	0.603	0.604	**0.804**	0.548
4000	0.000	0.000	0.000	0.000	0.756	0.759	**0.901**	0.724
5000	0.000	0.000	0.000	0.000	**0.823**	**0.819**	**0.934**	0.793

Note: This table provides Type 1 error rates and statistical power for different evaluation designs.
Model 1: Adapted frequentist approach based on two hypotheses tests.
Model 2: Bayesian model using uninformative prior and sufficiency threshold of 80%
Model 3: Bayesian model using uninformative prior and sufficiency threshold of 60%
Model 4: Bayesian model using pessimistic prior and sufficiency threshold of 60%

WHEN TO USE BAYESIAN VS FREQUENTIST ANALYSIS

In the previous sections, we have shown how a Bayesian approach can provide better tailored answers to some types of decision-focused impact evaluations. At the same time, frequentist approaches are (still) more likely to be accepted by evaluators and academic audiences. An important question is therefore how evaluators should decide between these two approaches when designing a DFE.

We recommend that evaluators consider a) the audience and b) the types of questions that the evaluation seeks to answer. If the primary audience of an evaluation is a single, internal decision-maker whose priors and decision rule can be clearly articulated then a Bayesian design is likely more appropriate. Following a Bayesian approach, evaluators will incorporate these priors into the analysis and use the resulting posterior distribution to directly inform the specified decision rule. Even if the primary objective of the evaluation is to inform the general knowledge of the public or academic audiences, a Bayesian approach may still be appropriate. The main reason is that results from Bayesian approaches may be easier to interpret for decision-makers and for large parts of the general public. To overcome concerns towards Bayesian approaches, in the previous section, we demonstrate how - regardless of the prior - they can be designed to meet the required thresholds of statistical power and Type 1 error. In the Appendix, we furthermore emphasize the fact that evaluators should conduct sensitivity analyses to assess the extent to which the prior affects the results, similar to conventional robustness checks.[93]

In contrast, for evaluations designed for a general audience (knowledge-focused evaluations), the audience may be sufficiently skeptical of explicitly incorporating any prior information into the

[93] In addition, evaluators should use prior predictive checks to assess whether their prior choice can be justified (Schad, Betancourt and Vasishth 2019).

analysis. In such cases, the standard approach to impact evaluation – that requires a high bar to consider evidence noteworthy – may be more suitable. Most evaluations do not fall into one of these two extremes. Although researchers may design evaluations specifically to inform specific decisions that implementing organizations have to make, they also may want to use the evaluation results to demonstrate the impact of their program to a wider audience. In some circumstances, researchers may therefore want to consider pre-specifying and conducting different analyses for each target audience.

Just as important as the audience are the types of questions that the evaluation seeks to answer. Frequentist null-hypothesis testing is most useful when evaluators intend to compare two options and are very skeptical of one option (i.e. the program "working"). We have demonstrated how evaluators can tailor the standard frequentist approach to better suit the decision scenario. For example, evaluators should consider conducting one-sided instead of two-sided hypothesis tests if only one direction of the effect is decision-relevant. Further, evaluators may consider adjusting the required level of statistical significance. Both approaches reduce the resources required to conduct an evaluation while maintaining high levels of technical rigor to inform the given decision. Yet, the examples discussed in this paper show there are several circumstances where the decision relevant question cannot easily be framed in terms of a null hypothesis versus an alternate hypothesis. In these situations, Bayesian approaches may be more practical and more feasible, leading to better tailored answers to the original decision to be informed.

Table 5: Choosing Between Bayesian and Frequentist Analysis

More suitable for Frequentist	More Suitable for Bayesian
Evaluation designed for a general audience	Evaluation designed for a specific decision-maker
Testing a new, unproven approach, compared to an acceptable status quo	No strong status quo option
Decision rests on a parameter being significantly different than zero	Decision rests on the distribution of one or more parameters
	Detailed Priors can be elicited and agreed upon by all parties

6. CONCLUSION

The starting point of this work is the observation that impact evaluations of development interventions have transitioned from being an academic method to study what programs work to being *a* tool that helps evaluators guide decision-makers' actions in a particular context with reliable evidence. In this paper, we investigate the implications of this change in target audience for the design and analysis of decision-focused evaluations.

We argue that alternative approaches to design and analyze impact evaluations may result in better tailored evidence to inform specific decisions, especially in certain scenarios where the standard approach tends to result in unclear recommendations. We advocate evaluators to broaden their

evaluation toolkit and attempt to provide practical guidance for evaluators seeking to better tailor evaluations to inform specific decisions, i.e. "decision-focused evaluations." In many cases, better-tailored designs may lower evaluation costs while maintain a high level of technical rigor. In other cases, though, better tailored designs may further increase costs to be more decision-relevant.

We illustrate scenarios in which following the standard evaluation approach may be sub-optimal to inform specific decisions. For each scenario, we show how an alternative design or inference may lead to a better tailored analytical approach. These alternative approaches include both minor tweaks of the standard approach, such as reconsidering the hypotheses to test and/or adjusting the level of significance, as well as larger deviations that rely on Bayesian analysis. Bayesian analysis allows to integrate decision-makers' ex-ante beliefs, known as "priors," about the state of the world into the analysis. These priors may or may not carry substantial information about the studied intervention and are combined with the evidence collected in the study to form a posterior, which captures an updated understanding of the world. Bayesian analysis is particularly well-suited for decision-focused evaluations because it allows evaluators to make easily understood probabilistic statements such as "the probability of program A being better than program B is 60%." We discuss when evaluators should consider taking a Bayesian approach to evaluation design and illustrate how to implement it in practice, including hands-on guidance regarding sample size calculations, analysis, and interpretation of results.

The unifying feature of the alternative approaches to design and analyze impact evaluations we discuss is that they explicitly consider the specific decision-makers' circumstances of the decision framework. For instance, in certain cases, decision-makers may be fine to implement a policy even with relatively high uncertainty as to its effectiveness. Other decision makers may only be interested in one direction of the effect. For instance, a decision maker may only be interested in whether an intervention has a positive effect, while finding a negative effect would lead to the same decision as not finding any effect. Yet, other decision-makers may have considerable priors on the effectiveness of programs under consideration. Funders may for example be skeptical of a given intervention given its track record in the literature and would want to incorporate this skepticism in their funding decision.

One worry about moving analysis away from research norms is that this would open up the door to spec searching and p-hacking, therefore decreasing the reliability of results using alternative methods. We argue that this worry can be mitigated through detailed and transparent pre-analysis plans.

By taking the decision framework of their study into account in their evaluations design, researchers can ensure that evaluation is as useful as possible to its audience of decision-makers.

REFERENCES

Abadie, A., 2020. Statistical Nonsignificance in Empirical Economics. *American Economic Review: Insights,* 6, Volume 2, pp. 193-208.

Amrhein, V., Greenland, S. & McShane, B., 2019. *Scientists rise up against statistical significance.* s.l.:Nature Publishing Group.

Bandiera, O., Fischer, G., Prat, A. & Ytsma, E., 2016. Do women respond less to performance pay? Building evidence from multiple experiments.

BaŞtürk, N., Çakmaklı, C., Ceyhan, S. P. & Dijk, H. K. v., 2014. On the rise of Bayesian econometrics after Cowles Foundation monographs 10, 14. *Œconomia. History, Methodology, Philosophy,* p. 381–447.

Benjamin, D. J. et al., 2018. Redefine statistical significance. *Nature Human Behaviour,* Volume 2, p. 6.

Berry, S. M., Carlin, B. P., Lee, J. J. & Muller, P., 2010. *Bayesian adaptive methods for clinical trials.* s.l.:CRC press.

Brodeur, A., Lé, M., Sangnier, M. & Zylberberg, Y., 2016. Star wars: The empirics strike back. *American Economic Journal: Applied Economics,* Volume 8, p. 1–32.

Carlin, B. P. & Sargent, D. J., 1996. Robust Bayesian approaches for clinical trial monitoring. *Statistics in Medicine,* Volume 15, p. 1093–1106.

Channa, H. et al., 2019. What drives smallholder farmers' willingness to pay for a new farm technology? Evidence from an experimental auction in Kenya. *Food policy,* Volume 85, p. 64–71.

Chen, M.-H.et al., 2011. Bayesian design of noninferiority trials for medical devices using historical data. *Biometrics,* Volume 67, p. 1163–1170.

Chen, N. & Lee, J. J., 2020. Bayesian cluster hierarchical model for subgroup borrowing in the design and analysis of basket trials with binary endpoints. *Statistical Methods in Medical Research,* p. 0962280220910186.

Cho, H.-C. & Abe, S., 2013. Is two-tailed testing for directional research hypotheses tests legitimate?. *Journal of Business Research,* Volume 66, p. 1261–1266.

Christensen, G., Freese, J. & Miguel, E., 2019. *Transparent and reproducible social science research: How to do open science.* s.l.:University of California Press.

Cohen, J., 1988. Statistical power analysis for the social sciences.

Cohen, J. & Easterly, W., 2010. *What works in development?: Thinking big and thinking small.* s.l.:Brookings Institution Press.

DellaVigna, S. & Pope, D., 2018. Predicting experimental results: who knows what?. *Journal of Political Economy,* Volume 126, p. 2410–2456.

Dobson, A. J. & Barnett, A. G., 2018. *An introduction to generalized linear models.* s.l.:CRC press.

Drew, R. & Clist, P., 2015. Evaluating development impact bonds. *A study for DFID, department for international development uk.*

Eichler, R. & Levine, R., 2009. *Performance incentives for global health: potential and pitfalls.* s.l.:CGD Books.

Fisher, R. A., 1992. Statistical methods for research workers. In: *Breakthroughs in statistics.* s.l.:Springer, p. 66–70.

Gabry, J. et al., 2019. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* Volume 182, p. 389–402.

Gelman, A. et al., 2013. *Bayesian Data Analysis, Third Edition.* s.l.:Taylor & Francis.

Gelman, A. & others, 2006. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis,* Volume 1, p. 515–534.

Gelman, A., Simpson, D. & Betancourt, M., 2017. The prior can often only be understood in the context of the likelihood. *Entropy,* Volume 19, p. 555.

GLENNERSTER, R. A. C. H. E. L. & TAKAVARASHA, K. U. D. Z. A. I., 2013. Statistical Power. In: *Running Randomized Evaluations: A Practical Guide.* s.l.:Princeton University Press, p. 241–297.

Goodrich, B., Gabry, J., Ali, I. & Brilleman, S., 2020. *rstanarm: Bayesian applied regression modeling via Stan..* s.l.:s.n.

Greenhouse, J. B. & Waserman, L., 1995. Robust Bayesian methods for monitoring clinical trials. *Statistics in medicine,* Volume 14, p. 1379–1391.

Gubbiotti, S., 2009. Bayesian methods for sample size determination and their use in clinical trials.

Gubbiotti, S. & De Santis, F., 2008. Classical and Bayesian power functions; their use in clinical trials. *Biomedical statistics and clinical epidemiology,* Volume 2, p. 201–211.

Guimerà, R. et al., 2020. A Bayesian machine scientist to aid in the solution of challenging scientific problems. *Science Advances,* Volume 6.

Haaland, I., Roth, C. & Wohlfart, J., 2020. Designing information provision experiments.

Hill, C. J., Bloom, H. S., Black, A. R. & Lipsey, M. W., 2008. Empirical benchmarks for interpreting effect sizes in research. *Child development perspectives,* Volume 2, p. 172–177.

Hodges, J. S., 2013. *Richly parameterized linear models: additive, time series, and spatial models using random effects.* s.l.:CRC Press.

Horton, S. et al., 2017. Ranking 93 health interventions for low-and middle-income countries by cost-effectiveness. *PLoS One,* Volume 12.

Imbens, G. W. & Rubin, D. B., 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* s.l.:Cambridge University Press.

Jackman, S., 2004. Bayesian analysis for political research. *Annu. Rev. Polit. Sci.,* Volume 7, p. 483–505.

Kasy, M. & Sautmann, A., 2019. Adaptive Treatment Assignment in Experiments for Policy Choice.

Klein, N., Kneib, T. & others, 2016. Scale-dependent priors for variance parameters in structured additive distributional regression. *Bayesian Analysis,* Volume 11, p. 1071–1106.

Lemoine, N. P., 2019. Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. *Oikos,* Volume 128, p. 912–928.

Makowski, D., Ben-Shachar, M. S. & Lüdecke, D., 2019. bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software,* Volume 4, p. 1541.

Malakoff, D., 1999. Bayes offers a'new'way to make sense of numbers. *Science,* Volume 286, p. 1460–1464.

Manley, J., Gitter, S. & Slavchevska, V., 2013. How effective are cash transfers at improving nutritional status?. *World development,* Volume 48, p. 133–155.

McCloskey, D. N. & Ziliak, S. T., 1996. The standard error of regressions. *Journal of economic literature,* Volume 34, p. 97–114.

McIntosh, C. & Zeitlin, A., 2018. Benchmarking a child nutrition program against cash: experimental evidence from Rwanda. *San Diego: University of California.*

McKenzie, D., Meager, R., Iacovone, L. & PÃ©rez, D. R., n.d. *What are the Effects of Improving Management Practices on Exporting among SMEs in Middle-Income Countries? A Comparison of Bayesian and Frequentist Impact Evaluation Approaches,* s.l.: s.n.

Meager, R., 2019. Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics,* Volume 11, p. 57–91.

Miller, J. E. & Rodgers, Y. v. d. M., 2008. Economic importance and statistical significance: Guidelines for communicating empirical research. *Feminist Economics,* Volume 14, p. 117–149.

Moerbeek, M., 2019. Bayesian evaluation of informative hypotheses in cluster-randomized trials. *Behavior research methods,* Volume 51, p. 126–137.

Nazari Chamaki, F., Jenkins, G. P. & Hashemi, M., 2019. Social impact bonds: Implementation, evaluation, and monitoring. *International Journal of Public Administration,* Volume 42, p. 289–297.

Nickerson, R. S., 2000. Null hypothesis significance testing: a review of an old and continuing controversy.. *Psychological methods,* Volume 5, p. 241.

O'Hagan, A., 2019. Expert knowledge elicitation: subjective but scientific. *The American Statistician,* Volume 73, p. 69–81.

O'Hagan, A. et al., 2006. *Uncertain judgements: eliciting experts' probabilities.* s.l.:John Wiley & Sons.

R Core Team, 2020. *R: A Language and Environment for Statistical Computing,* Vienna: s.n.

Rossi, P. E. & Allenby, G. M., 2003. Bayesian statistics and marketing. *Marketing Science,* Volume 22, p. 304–328.

Schad, D. J., Betancourt, M. & Vasishth, S., 2019. Toward a principled Bayesian workflow in cognitive science. *arXiv preprint arXiv:1904.12765.*

Shah, N. B., Wang, P., Fraker, A. & Gastfriend, D., 2015. Evaluations with impact: decision-focused impact evaluation as a practical policymaking tool. *International Initiative for Impact Evaluation (3ie),* p. 16.

Smith, J. Q., 2010. *Bayesian decision analysis: principles and practice.* s.l.:Cambridge University Press.

Spiegelhalter, D. J., 2001. Bayesian methods for cluster randomized trials with continuous responses. *Statistics in medicine,* Volume 20, p. 435–452.

Spiegelhalter, D. J., Abrams, K. R. & Myles, J. P., 2004. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* s.l.:Wiley.

Sprenger, J., 2015. The objectivity of subjective Bayesian inference.

Sterck, O. & others, 2018. *On the economic importance of the determinants of long-term growth,* s.l.: s.n.

Vivalt, E., 2017. *How Much Can Impact Evaluations Inform Policy Decisions?,* s.l.: s.n.

Vivalt, E., n.d. Forthcoming."How Much Can We Generalize from Impact Evaluations?". *Journal of the European Economics Association,* Volume 5.

Walker, E. & Nowacki, A. S., 2011. Understanding Equivalence and Noninferiority Testing. *Journal of General Internal Medicine,* 01 2, Volume 26, p. 192–196.

Wasserstein, R. L. & Lazar, N. A., 2016. *The ASA statement on p-values: context, process, and purpose.* s.l.:Taylor & Francis.

APPENDIX

APPENDIX A – SOME REMARKS ON PRIORS

One of the main differences between frequentist and Bayesian approaches to impact evaluation is the latter's requirement to specify prior distributions. In this section, we make a few general remarks about how to specify priors in a Bayesian model.

First and foremost, evaluators will have to specify priors for *all* parameters in their model. In the head-to-head comparison scenario, for example, we specified prior distributions for the respective treatment effects *and* the variance of the outcome. This is not just a coincidence, but in every Bayesian model the evaluator will have to specify a distribution for *every* parameter of the model.[94]

A prior can be classified based on the characteristics of its distribution.[95] In the head-to-head comparison scenario, we first assumed that the decision maker does not have any additional, ex-ante information or context about the parameters. As a consequence, we specified a flat distribution over all possible values for these parameters, assigning (near) equal probability to all values.[96] Such priors are also known as *uninformative priors*. As shown in the main text, results from models with uninformative priors often mirror the results from conventional maximum likelihood estimation (and simple linear models such as OLS). In other words, the data will drive the results of the Bayesian model.

Informative priors capture additional information in the ex-ante beliefs and therefore generally do affect the analysis.[97] The additional information could come from (expert) opinions or findings from previous studies. Suppose the decision-maker read a large amount of literature and is aware that all studies found a particular intervention to increase the binary outcome of interest between 5 and 25 percentage points. The evaluator could then include this information in the prior by assigning a larger mass of the distribution to values in that range and a lower mass to values outside that range. A different funder might be very skeptical of the proposed intervention. The researcher might therefore want to account for this explicitly by specifying an informative prior that puts a significant amount of mass around the null effect region. If the resulting posterior still suggests that the intervention has a sufficiently high probability to be effective, the decision-maker might be more inclined to provide additional funding.[98] Explicitly incorporating ex-ante beliefs into the design of a decision focused evaluation can therefore provide additional value to a decision-maker. While incorporating subjective

[94] In fact, the line between the prior and the likelihood can be fairly arbitrary since the choice of likelihood often encodes prior info as well. In the head-to-head comparison scenario, for example, we assumed that the outcome is distributed normally.

[95] There are three major dimensions along which priors are often classified: proper vs improper, conjugacy and degree of information. A proper prior represents an actual probability distribution. An improper prior usually integrates to infinity (or a fixed number larger than one) and does therefore not represent an actual probability distribution. Conjugacy, describes whether the functional form of the prior - along with the functional form of the data - implies the same functional form of the posterior. Conjugate priors allow the analyst to find nice and tractable solutions to a given problem without having to rely on sophisticated computer algorithms. Beta and normal priors along with normally distributed likelihoods are conjugate.

[96] The specified prior took into account that a variance term cannot be negative, thereby reducing the range of possible values. Accounting for uncontentious information (e.g. variance terms being non-negative or treatment effects for binary outcomes being in (-1,1)), sometimes leads analysts to refer to their priors being *weakly informative*.

[97] Generally, informative priors will affect the analysis less, the larger the sample size. This pattern can be observed in the sample size calculations with informative priors.

[98] In spirit, specifying a skeptic prior feels very much like lowering the level of significance applies in a conventional null-hypothesis test.

beliefs of decision-makers in the analysis is clearly a contentious topic among evaluators, scientific approaches to elicit and pressure test such elicitation approaches do exist (O'Hagan 2019; Schad, Betancourt and Vasishth 2019).[99]

TECHNICAL GUIDANCE ON HOW TO SPECIFY PRIORS

First and foremost, the evaluator should be transparent about what prior is chosen, why it is chosen, and how this prior affects the results.[100] The most contentious prior choice will most certainly with regards to the average treatment effect(s) of the intervention(s) studied. Evaluators seeking to specify informative priors for the average treatment effect should therefore follow a scientific elicitation process, such as a meta-analysis or guided elicitation approach. In addition, evaluators should ensure and report that their evaluation design with this informative prior exhibits good operating characteristics, i.e. sufficiently high statistical power and sufficiently low Type 1 error rates.

Second, if the evaluator knows that some parameters follow a specific distribution, they might want to specify a 'flat' version of that distribution as a prior. For a mean or an OLS estimate, for example, evaluators could specify a weakly informative prior as a normal or t-distribution with a rather large variance parameter.[101]

Third, specifying priors for variances is hard. The evaluator has oftentimes less intuition about these parameters and has therefore even less intuition about how to specify the variation of a variance term. In addition, uninformative priors are often improper. This means that they are not probability distributions in a narrow sense, but rather integrate to infinity or a constant larger than one. While it is fine to specify improper priors, such choices might lead to improper posteriors. If the posterior estimate is improper, this is oftentimes a flag for our model not having converged to a stable distribution. One way to overcome this problem would be to specify a slightly different, but similar prior. For variance terms, the evaluator can also specify proper priors over non-negative values e.g. through a half-Cauchy distribution. Alternatively, the analyst could specify an inverse-gamma or a constrained uniform distribution.

Fourth, if a parameter is known to be constrained to a particular interval, the prior should reflect this feature. For example, if the evaluator aims to estimate a proportion, the prior should only take values between zero and one. This could be achieved through a flat Beta(1,1) or a constrained uniform distribution. In case the parameter is unlikely to take values at the extremes of the distribution, a flat normal distribution might be fine too. Alternatively, when interested in the treatment effect on a binary variable, the evaluator knows that it can only take values in $[-1,1]$. This means the absolute, maximum change in a proportion can be 100 percentage points. Such priors are often referred to as weakly informative (Gelman, Simpson and Betancourt, 2017).

[99] Further, priors are rarely uninformative in a strict sense. First, changes in the unit of measurement of the parameter (a re-parameterization) will often require changes to uninformative priors. Take a flat prior from a uniform distribution with lower bound zero and upper bound 500, $\boldsymbol{U}[\mathbf{0},\mathbf{500}]$. This prior might be uninformative for hourly wages measured in US$, but very informative for wages measured in Zimbabwean dollars. Second, even in cases where little is known about the intervention, the evaluator can typically exclude or assign low probability to certain values from the parameter space.

[100] Similar to a conventional robustness check section, Bayesian impact evaluation reports could include a sensitivity check section that discusses the effects of different prior choices on the analysis results.

[101] Note that a "large" variance parameter is relative. For a given variance of the prior, the larger the variance in the data, the more informative the prior will be.

Finally, the evaluator may need to do some math to transform ex-ante beliefs into a reasonable prior about the model parameter. For example, the evaluator may be interested in a binomial parameter and have a prior for that parameter (say flat over the unit interval). For analysis, it is oftentimes computationally more efficient if the posterior can be approximated by a normal distribution. The evaluator may therefore want to transform the parameter to the logit scale and do inference on that scale. This transformation, in turn, would also require the evaluator to specify a prior on the logit scale.

More technical guidance on how to specify priors can for example be found on the Stan Dev Github page.[102]

[102] https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

APPENDIX B – MODEL CHECKING USING A TEST QUANTITY

In addition to testing the sensitivity of results to the choice of prior, it is useful to test that the overall model (including the prior) is a reasonably good fit for the data. One way of doing this is through the use of Bayesian test quantities (Gelman et al, chapter 6). Broadly, Bayesian test quantities are used to compare actual data to simulated data from the model. In contrast to frequentist test statistics, Bayesian test quantities may be functions of parameters themselves. In addition, when comparing test quantities from actual data and simulated data we average over values of our parameters rather than assume that the parameters are fixed (as we do when calculating a frequentist p-value).

This last aspect of Bayesian test quantities may appear confusing at first. To clarify the difference between frequentist test statistics and Bayesian test quantities, we provide an example based on the head to head comparison example. Suppose that we wished to test whether our model is a good fit for the data. In particular, we may be concerned that the normal distribution is not appropriate for use in the likelihood because the actual data may be distributed with very "fat tails." A simple way to test this would be look at the actual difference between the maximum and minimum for each of the treatment groups (and control) and compare this value with the values we obtain when we calculate this quantity for simulated data. (Note that there are many other possible test quantities that we could define and there is nothing special about the one defined here.) In other words, we first calculate the following test statistic for the actual data:

$$T(y) = \max(y_{i\epsilon c}) - \min(y_{i\epsilon c}) + \max(y_{i\epsilon t1}) - \min(y_{i\epsilon t1}) + \max(y_{i\epsilon t2}) - \min(y_{i\epsilon t2}) \quad (27)$$

In our case, the value for the observed data is 14.718. Next, we iterate through our draws from the posterior for all of the parameters. For each draw from the posterior, we generate simulated data using our likelihood and values for the parameters and calculate the test quantity for each replicated dataset. Finally, we calculate the proportion of times the test quantity from our replicated dataset is greater than the test statistic for the observed data.[103] In this case, the Bayesian p-value is 0.4 indicating that this test statistic has not provided evidence that the normal model is a poor fit.

[103] Code for this example, and all other examples in this paper, can be found at https://github.com/IDinsight/dfe_methods