Bayesian Interpretation

September 29, 2022

3:00–4:00p.m. ET

Webinar Transcript

John Deke: Today, we're going to talk about the Bayesian interpretation of estimates [BASIE] framework, which is used to provide a Bayesian interpretation of impact estimates in impact evaluations—specifically, [our subject today is] in the field of teen pregnancy prevention. And I am John Deke. I am a senior fellow at Mathematica. And I will be joined by my colleague, Mariel Finucane.

Mariel Finucane: And our plan for today is to first quickly remind you where we left off last time. A tiny bit of motivation about when we want to use BASIE and why we think it's helpful. And then we'll spend more time digging into these two topics. First, on the theory side of things, we want to walk you through the different components of a BASIE analysis. And second, we're going to get nice and practical with a spreadsheet tool demo and show you how this works.

John Deke: All righty. So, first let's just (inaudible) on that refresher on when we would want to use BASIE and why we would want to use BASIE. The use case scenario we have in mind here is: you're conducting an impact evaluation, you calculate the outcome for the treatment group, the average outcome for the control group, you've got an impact estimate. And you've got that in hand, and you're wondering what to make of it. Is this the result of a genuine effect of the program, or is this due to random chance? And then relatedly, maybe you're not doing the evaluation yourself, but rather you're reading findings from another evaluation, you're trying to make sense of it. So you've got an impact estimate in your standard area, and you're trying to figure out, what's the probability this thing worked? Those are the two use case scenarios that we have in mind for BASIE. (Inaudible.)

Mariel Finucane: And I think oftentimes in this moment, when we're wondering how to make sense of study findings, we would often turn to *p*-value (inaudible) statistical [significance testing]. But one of the primary motivations why we (inaudible) BASIE instead is because statistical significance has, itself, really been roundly rejected. In 2016, the American Statistical Association released the statement here about how widespread and pervasive the misinterpretation of *p*-values is. That was followed up a few years later with the special issue of *The American Statistician* on the same topic and really starting to look forward to this, well, if we're not going to use *p*-values, then what? What does the post-*p*-value world look like? And in that same year, the journal *Nature* had this paper with more than 800 signatories really urging for statistical significance to be retired altogether. So, it's a real groundswell of consensus against statistical significance in *p*-values, [which] was a big motivation for me and John developing BASIE.

John Deke: All righty. So, a big chunk of our talk today is going to be focused on how to actually do this, but we are going to spend a little bit of time talking about BASIE theory. [Let's] go to the next slide. We're not going to spend a ton of time on the theory or the underlying statistics, and so I just wanted to make sure that you guys had a variety of resources that you could turn to for additional details beyond what we can cover in our hour talk today. So, on the slide here we have several resources for you. The first resource is a brief that we wrote for OPRE [Office of Planning, Research, and Evaluation] introducing BASIE. The second is a more in-depth report which we wrote for the Institute of Education Sciences, the Department of Education. And that really provides a lot more methodological detail than either the brief or what we'll cover today. So, if anybody wants to get into more nitty-gritty, that guide would be a good bet. Even though it's in a somewhat different field where it's focused more on things like test scores as outcomes, the concepts are very applicable here as well. So, for a lot more detail, you can go there. We also have a webinar version, or a webinar that accompanies that guide, which is the third bullet here on our slide. We have an example of using BASIE in a teen pregnancy prevention context, which was the evaluation of

Making Proud Choices, which I believe Russ was involved with. And then we have another paper that goes into even more technical depth, and it is focusing on Bayesian interpretation of cluster of subgroup impact estimates where you're looking at a bunch of subgroup estimates from one study. And so for anybody who wants to dive more deeply, those are some resources available to you. But we won't be getting into that level of detail in this talk.

Mariel Finucane: [The] level of detail that we will be getting into is as follows. We're going to quickly walk through these four different steps. First, we're going to select prior evidence. The next thing we need to do is report (inaudible) impact of the intervention. And then most important is Step 3. This is really where we [think] BASIE shines bright. We need to interpret that impact estimate, decide what to make of it. As John was saying, it's (inaudible). And then lastly, we'll want to conduct sensitivity analyses for that prior evidence that we selected (inaudible).

John Deke: All righty. So, selecting the prior evidence. The first step … you guys won't actually need to do because it's already been done for you, but if you wanted to do a deeper dive and do it yourselves, you certainly could. But the first step is to use meta-regression, meta-analysis, to synthesize prior evidence from a literature to form prior distributions. And if you want to see technical details about how we would recommend doing that and how we have done that in the past, you can see that IES [Institute for Education Sciences] guide, which we described before. But for the rest of today, we're going to be kind of taking for granted that we already have some prior distributions that have been synthesized through this type of meta-analysis. So, [we'll talk about] two prior distributions used in the evaluation of Making Proud Choices, and which are also provided in the Excel tool that we will show later today. One is a domain that we can call the Sexual Behavior Domain. And this is a meta-analysis of evaluations of teen pregnancy prevention interventions which was conducted by Juras et al. (2019). And that prior distribution is assumed to be normal, and it has a mean of .03 and standard deviation of .15. That's all an effect-size unit, so in the context of meta-analysis, all the impacts are converted into effect-size units. So, that's the (inaudible) for that domain, which is more a domain that involves more distal outcomes, things that are harder to move the needle on. The other domain that we're looking at is an outcome domain for outcomes that are more proximal to an intervention. Outcomes where it is maybe a little easier to move the needle. And we're calling that the Risk and Protective Factor domain. Now, this is based on a meta-analysis of findings from the What Works Clearinghouse, which is not a teen pregnancy prevention context. But the reason that we were drawing on the Clearinghouse is because that is a source of information that includes impact[s] on a lot of outcomes that are more proximal to interventions. And so that was the motivation for picking that. So, the first step is to pick which of these prior distributions that you want to use to interpret your impact estimates. And we will give examples of this in the very near future.

Mariel Finucane: John, can I say one more thing about these priors before we go on? Is it okay?

John Deke: Absolutely.

Mariel Finucane: Cool. So, I want to give just a little bit of intuition behind what John was just saying about these distal outcomes in the first bullet that are a bit harder to move and then these more proximal outcomes in the bottom bullet that are a little easier to move. And I just want to draw you guys' attention to the values, the means, and the standard deviations and to show you that that's showing up here in these priors. So, what you see in the top bullet is that the mean, that's the kind of average impact across all of the interventions in that first meta-analysis. It's .03. It's very close to zero. Whereas in the second bullet, you see it's .16. That's further from zero. That's showing on average, in the second meta-analysis, things were having larger effect sizes. But then you also see that the standard deviation is larger in the second bullet. Right? That [it's] .29 instead of the .15. And what that's saying is that not only on average were the interventions in the second meta-analysis more—not only did they have larger effects on average, but also there was more spread across them in terms of how big their effects were. So, some interventions in that second analysis presumably had really whopping impacts. Like, if we take that mean of .16, and then we add the .29, you can see that just one (inaudible) you're going to see very large impacts, you know, north of .4 standard deviation. So, it can happen for those more proximal outcomes that we see very big impacts. You see in the first bullet that that's quite unusual for those (inaudible) outcomes.

John Deke: Yeah. No, that was really helpful. That was a great point. And maybe it's worth just also clarifying that when we say proximal outcomes, what we have in mind are outcomes that are really tightly aligned with the logic model of an intervention. So, if you think about like a sequence of dominoes that you expect to fall after pulling the lever implementing an intervention, those are the big dominoes that are closest to the lever. They're the things that you would think, well, duh, of course the intervention is going to affect that. It would be like you get to the end of a class and you ask kids, what did I just tell you? Whereas the things that are further out, which we're calling distal, those are like longer-term behaviors that are much harder to move. And so it makes sense if we would use a prior distribution that has a higher meaning and fatter tails is—well, not fatter tails, but a bigger—more dispersion, as Mariel was pointing out, for those proximal things. So that's an important choice as to which of those types of prior distributions you use. And we'll provide some more guidance on that as we go along.

Mariel Finucane: Super. Thanks, John, that's great. Okay, so we've completed Step 1. We've chosen which of these two priors we want to use. In Step 2, we now need to decide which impact estimate are we going to report. And we have two options here. I'm about to describe them for you guys. And we're going to choose which one we want to give more emphasis to. Maybe which one are we going to present in our executive summary or our abstract, for example. And we want to be nice and rigorous and pre-specify which one of those it's going to be. We don't want to check both of them to see which one is bigger and then [specify] that one. So, we're going to pre-specify which of the following two things we want to focus on. The first option is the traditional estimate that everybody knows and loves based only on the data from your study that you're running right now. And a reason that John and I think it's very important to support this one is that it's nice and transparent. It shows your readers what was found in your study data. Another kind of down-the-road benefit is that it's helpful for future meta-analyses. John and I believe strongly in meta-analysis because, as we were just saying, it's useful for developing (inaudible). The other option, which might be a bit more novel for some folks, is this Bayesian estimate of impact. And that's going to take the traditional estimate from the previous bullet and combine that with the prior evidence. So, if your traditional estimate says, whoa, this thing had a huge whopping impact, but you're using that very, kind of conservative, (inaudible) prior that's helpful for distal outcomes, centered close to zero, then the Bayesian estimate is going to be pulled, you know, it's going to be a weighted average of those two things. Here we saw really exciting things in our data, but our prior is pretty skeptical. So, (inaudible), the Bayesian estimate is going to … kind of split the difference depending on how big your study is. And the reason that it's very important, we think, to report that estimate as well is that it's less susceptible to statistical noise, which is a particularly important problem if you're conducting a very small study where you can be in that scenario where you get a whopping estimate in your study data just by chance alone. So, we think these are both important. We think you should report them both. And we think you should decide … which one you're going to focus on.

John Deke: That's great, Mariel. Thanks. So, once we've selected our prior distribution, and we've recorded an impact estimate, we now want to interpret that impact estimate, and we want to make probability statements about how to interpret it. An example of the type of probability statement that we can use is here, color-coded on the slide. We say we estimated 75 percent probability that our intervention increased knowledge of contraception methods. Now we're going to change colors given our estimate. And now we'll change colors again [given] prior evidence on the impacts of educational intervention. So, to diagram this sentence, the first component, in blue, that is talking about what Bayesians call the posterior distribution for the impact. The second little phrase there in red is what Bayesians call the likelihood, that is the new information that we collected in the study. And then the final phrase, in green, that was the prior information, the prior distribution. Those three components come together in this little sentence here. And there are a few things that are worth pointing out about statements like this because they really do have a pretty precise meaning. Typically, these probability-type statements have three characteristics. One is they apply to the effect of the evaluated intervention on the sample included in your study. So, this is really about your study. It's not about a different context. They are not statements about the chances of the intervention having an effect in the future. [It] isn't saying what the effect would be for your sample if you repeated it again in the future. It's not a probability statement about what would happen if you tried to implement this intervention [for] other group of kids in the future. This is the probability that the thing worked. Past tense. Not the probability that it will work. And that's not a difference from statistical significance or *p*-values. *P*-values and statistical significance are also

retrospective, trying to figure out what happened in your study. It's just that they don't actually tell us what we wanted them to tell us. This is actually telling us the probability the thing worked. Whereas *p*-values and statistical significance couldn't even do that. So, I'm pointing out the caveat, but I want to make sure you understand that it does still have value. And then also these can only be interpreted in the context of the prior. If you are using the prior distribution that applies to more distal outcomes, then that's going to affect your interpretation as opposed to if you use that prior distribution that includes more proximal outcomes. So, that is Step 3, interpreting the impact estimates. Mariel, do you have any additional thoughts on that?

Mariel Finucane: No. that was great. Thoughts on Step 4, though; all-important sensitivity analysis. As John was just saying, our posterior does depend on our priors. We recommend calculating posterior probabilities again for at least one pre-specified sensitivity analysis prior distribution. And one that we recommend you consider is this: a zero-centered prior. And what I mean by zero centered is that we've literally taken that distribution, remember in both of the examples John gave of priors, they were centered north of zero. They had a mean greater than zero, so [we're] guessing that on average, these interventions have benefits. And just take that whole distribution and shove it over so that it's centered now right on top of the zero line. So, that's a nice prior. It kind of reflects this whole point of going in about whether the thing is going to work or not. And then I additionally think it's especially relevant because potentially it can be elevated to the role of your kind of primary analysis prior, not just sensitivity analysis, for studies where we're comparing two active treatments to each other. So, instead of comparing some new treatment to some business-as-usual control arm, if instead we're saying, how about Curriculum A versus Curriculum B. then a zero-centered prior could make sense. Because we don't know, obviously, which one of those is going to be more effective. One other thing I wanted to say about sensitivity analysis is that it can be useful, even at the design stage. So, whereas typically we run our analysis and then we analyze sensitivity, in the Bayesian world it can actually be helpful to flip that on its head and do the sensitivity analysis and then run your study. And the reason is that sensitivity to the prior distribution, so how much your posterior probabilities move around depending on what prior you used, that's going to decrease the bigger your study is. So, you'll have less sensitivity if you run a nice big study. And that's what I mean by taking sensitivity into account at the design stage. You can decide how much sensitivity you're willing to tolerate, and then design your study to be big enough that you won't have those issues.

John Deke: All righty. We have covered what BASIE is and a little bit about the theory of the thing. Now we want to go into examples of how to actually do this. What we're going to do is, we're going to set up an example. We're going to look at two hypothetical studies with pretend impact estimates. We're going to go through the exercise of choosing the prior distribution. We're going to hop over to a spreadsheet if I can figure out how to share my screen, and we'll show you how to calculate these probabilities using the spreadsheet tool. And then we'll run back to the safety of PowerPoint, and we'll continue the discussion with sensitivity analysis. And then we can discuss what we've learned today and do Q&A. That's where we're headed now. In this example, we are imagining that we have two different studies. So, the big difference between these two studies is just the sample size. Study A, we have 400 randomized students …. 200 in the treatment group, and 200 in the control group. And we're looking at impact on two different outcomes. One is a knowledge outcome, so that's more proximal to the intervention. And the other one is a behavioral outcome, which is a more distal outcome. And we see impact estimates on these two outcomes. And these are being reported in effect-size units. And the reason we do everything in effect-size units is so that we can make the impact estimates compatible with the prior distribution. But I realize that in many cases it's more friendly to users, to readers of reports, if you report impact in sort of the natural units. And so you can move back and forth. You can convert to the effect sizes, calculate the probabilities in interest, and then when you present them, convert back to the natural unit. So, that's just a disclaimer that we're going to be focused on effect sizes for convenience, but you can convert back to natural units to make things more accessible. Anyway, that aside. Here we have impact estimates. We've increased knowledge by .25 standard deviation, so that was a favorable impact estimate. And we reduced actual initiation by .10 standard deviation, so that's also a favorable impact of the intervention. And because this first study has 400 students, it's got a standard error of .10. And here we've reported bad old *p*-value just as a frame of reference. Now, in the second study, amazingly, the second study has exactly the same impact estimates as the first study, but, because it's a smaller study, the standard errors are quite a bit larger. And you see the *p*-values are much larger. And so, of the four impact estimates

reported across these two tables, only one of them is statistically significant, and we would have to ignore the other three. So, those are the two hypothetical studies that we are looking at. And I'll turn it over to Mariel to help us choose prior distributions.

Mariel Finucane: Great. If we want to use BASIE here, we've got to choose a prior. As we've already discussed, we think this can be a good prior to use for the more proximal outcomes. This one can be more useful for the distal outcomes. Just to review. What's new in this slide is which priors we want to use for sensitivity analyses for these hypothetical studies. As I already mentioned, we like the idea of trying zero-centric versions of each prior. And then also for the more proximal outcomes where in general the prior tends to be a little more lenient, we want to do a sensitivity analysis where we use the more conservative (inaudible) prior (inaudible).

John Deke: Okay. Now I'm going to wrest control of the presentation from Mariel, and I'm going to try to share Microsoft Excel. Let's see. Did that work? Do you see that, Mariel?

Mariel Finucane: It looks great.

John Deke: This is a spreadsheet tool that we will be sharing with you guys. And it includes several different tabs. Right now I am on the Prior Evidence tab, and we have some prior distributions that are already filled into the spreadsheet. You don't have to fill them in, but what you can do here is you can look, and you can see what they are. We have four different sources of evidence, or four different prior distributions. The one is the Juras et al. (2019). That's a meta-analysis of more distal outcomes, behavioral outcomes in the TPP literature. And we can see what the characteristics of this prior distribution are. The mean of the distribution is .03 standard deviations, which is pretty small. And then there is a standard deviation of effect of .15. So, what does that mean in terms of actual probability? We have some pre-calculated probabilities over here. The prior estimate of the proportion of intervention effects that are less than zero is 42 percent. So, we're saying 42 percent chance of unfavorable effects. I should note that in these meta-analyses and everywhere in the spreadsheet tool, we need to flip signs so that a positive sign always means a favorable effect … we need to do that so that all the numbers are comparable and hang together. So, if it's a good thing to not do something, then we need to switch it so that we're doing the opposite of that so that we're not [using] initiating sex [as a positive], for example. But anyway, so what this is showing is that about 58 percent of intervention effects are favorable. About 32 percent of those effects are greater than .10 standard deviations. And just 13 percent of effects are greater than .2 standard deviations. So, that's the prior distribution that we would use for more distal outcomes. For more proximal outcomes, this is where we have to go outside of the TPP literature because we don't have current meta-analyses with a lot of proximal outcomes. We're going to go borrow information from our friends in the education literature, where they do have more proximal outcomes. And we see that there is a much higher prior probability of big positive effects. Instead of a 40—I'm sorry, instead of a 13 percent chance that an effect is bigger than .2, our prior probability here is 45 percent. So, that's a very different prior probability. We also have versions of these two distributions that are centered at zero. And as Mariel indicated earlier, a motivation for doing that would be if instead of comparing a treatment to a business-as-usual control condition, if you are comparing Treatment A to Treatment B, then we don't have a prior hypothesis about which condition is better than the other. Most of the studies in the meta-analysis, they have a belief, you know, they have a reason, they have a hypothesis to believe that the treatment group is better than the control group because it's comparing the business-as-usual control group. But if that's not your situation, if it's really no difference—known difference between the two, then you might want a zero-centered prior distribution. So, we have zero-centered versions of both of these. We just peg the center at zero and keep the standard deviations. And so you could see, for example, with that distribution focused on proximal outcomes, the prior probability that the effect is greater than .2 goes from 45 percent down to 25 percent. It's a more skeptical version of that prior. So these are the prior distributions available in this spreadsheet. Now we're going to go to a different tab, the Study Findings tab, and this is where we will enter the impact estimates and standard errors from our study. These are the regular traditional, non-Bayesian impact estimate inputs you get just from using your study data. And you can see that these are the studies that I described in the PowerPoint slide. It's Study A and Study B. But you can see that I flipped that minus .10 to a positive .10 so that a positive sign always means favorable, as we just mentioned. So, here, on this tab, you would enter your impact estimates and your standard errors from

your study on this tab …. that's pretty straightforward. And then we go to the Interpretation tab. And here on the Interpretation tab we have these little pop-up menus in the cells, and those pop-up menus key back to the earlier tabs. So, you can pick which prior distribution you want to use. And you can choose—for example, say I want to use that one. You can pick your prior distributions. And it gives you a little reminder of what the mean and standard deviation of those prior distributions are. And then you can go over here, and you can pick which of the impact estimates you had entered on the other tab—what do you want to interpret? And then those will automatically populate here in the Impact Estimate and Standard Error fields. The purple was the prior, the green is your data, and the blue over here is the posterior distribution. This is the information about what's the probability that this thing worked given the impact estimate that I entered and the prior distribution that I selected. And so you can also [use] this Bayesian impact estimate, that's the shrunken impact estimate that Mariel referenced when she was saying there were two different impact estimates to report, that's the Bayesian impact estimate that's created by this spreadsheet. And then you can specify what probabilities are you interested in. I'm interested in the probability that the true effect is greater than some value. Here, I've got it set at zero, but if I change this in my live demo, then this probability over here changes. If I'm interested in a probability that's greater than .10, then that's a 93 percent chance. If I'm interested in [whether] the probability is greater than .25, that falls down to a 46 percent chance. So, you can basically calculate just about any probability you want right here in the spreadsheet tool. I'm going to pause for a moment and just see if Mariel has any thoughts on anything I've described here.

Mariel Finucane: That actually sounded good to me in all regards. I was wondering, should we pause and see if anybody else has questions while you still have this screen share up?

John Deke: Sure. Sounds good. Any questions on the line?

Russell Cole: So, before you even get questions, I think it would be super helpful, John, Mariel, would you mind just spending one or two seconds just talking through the interpretation, the probability statement in particular, on that far right column for a handful of estimates? I think it would be really helpful to kind of ground this and help people understand especially that idea of, it's greater than zero as it's kind of representing this is the probability that it's a favorable effect. I think it would be really helpful if you could just give a couple of quick talks about that. Thanks.

John Deke: Sure, Russ. In this first line, let's look again at what we're interpreting here. This is an impact estimate from Study A. It's impact on STI knowledge. And we're thinking of that as a more proximal outcome. And the impact estimate was .25 with a standard error of .1. So, when we look at an impact estimate, we know that that impact estimate is influenced by two things. If we're in an experiment anyway, it's influenced by two things. One is random chance. Just by random chance your treatment group might do better or might do worse than the control group, so that's one thing that influences that impact estimate. The other thing that influences that impact estimate is a genuine effect of the program. What we want to calculate is given this impact estimate and this standard error, what is the probability that the true effect is greater than zero in this case? And what the spreadsheet is telling us is, you've got a pretty big impact estimate there. And it's reasonably precisely estimated. And given what we see in the prior distribution, it's actually pretty common for interventions to have big effects on proximal outcomes. And given all of that information, your impact estimate, your standard error, and how common it is for us to see large effects in the literature, we estimate that the probability that there is genuinely a favorable effect, which is just, you know, crossing it's-better-than-nothing line, is 99 percent. So in this case, it's a very high percentage. Now say we decide, well, you know, I've got some other interventions over here in the closet, and they can do better than zero, too. Like it's not such a big deal to just have a favorable effect. But what I would really like to know is what is the probability that the true effect is at least bigger than .10 standard deviation. So, something a little more meaningful. And then that automatically calculates, and it's 93 percent. So that's saying, given our impact estimate of .25 standard deviations, our understanding that sometimes random quirky things happen but also our understanding that sometimes, you know, interventions are truly effective, what's the probability that the true effect is bigger than .10? Well, it's 93 percent. And then, as we'll discuss, to do the sensitivity analyses that Mariel was describing, you can go down the spreadsheet rows here, and you can choose different priors. So, you can choose the distal prior distribution. Or you can zero center it. And you can see how the probabilities change as you change the prior distribution.

And you'll notice that for Study A, where the standard error is always .10, that posterior probability doesn't change much as you switch priors. It just goes from 99 percent to 98 percent as we go across those different prior distributions. But as Mariel has mentioned, if you have a bigger standard error, which is the situation with Study B where the standard error is .2, then you start to notice more variation in these probabilities. Instead of going from 99 percent down to 98 percent, it goes from 91 percent down to 77 percent. So there the prior starts to matter a lot more because you have a less precise impact estimate; you're relying more heavily on the prior distribution. So, when Mariel was saying you can do sensitivity analysis at the design stage, she was saying—and remember these are not real studies, I just made these up. You can do the same thing. You can poke these numbers into the spreadsheet and see, well, if I had, you know, 400 students in my study as opposed to 100, how sensitive would these key probabilities be? And this can show you that ahead of time. So, that's what she was talking about. So … Russ, I don't know if I'm doing what you had in mind.

Russell Cole: No, I think it's super helpful.

John Deke: Okay.

Russell Cole: And if you wouldn't mind … there might be other stuff. Do the next one, the next row, Row 14, because I think that's a really critical one to hit home for everyone who is here.

John Deke: Sure.

Russell Cole: Behavioral impact that I think was non-significant in your example.

John Deke: All righty. So, Row 14. This is that behavioral impact. It's the impact on sexual initiation. And it's an effect size of .10. Standard error of .10. And what we see is that the probability that there is a favorable effect, given our impact estimate and the prior literature, is 83 percent. Which is a fairly high percentage, you know. If you were going to place a bet on a football game, and I said I'll give you an even-money bet and there's an 83 percent chance your team will win, would you take that bet? I would. I think that's some pretty good odds. It's much better than 50 percent. So, it is an example of how in a study that would traditionally be regarded as underpowered because you can't cross the *p*-less-than-.05 threshold, we can still say useful and informative things. We can still report the probability that this worked, and there can be some meaningful differences in those probabilities. So, maybe I'll see if there's any questions from folks on the line before I hop back over to the PowerPoint slides. All right. I've going to give up control back to you, Mariel.

Mariel Finucane: So, notice anything else about this slide or do you feel like we've covered it in the spreadsheet?

John Deke: Well, I'll just say some summary things here. On this slide we've pulled out some of those things that we calculated over in the spreadsheet. We put them in our, you know, in our PowerPoint report where we're going to our conference and reporting on our findings. And we can notice several things. So, one thing, which is what Russ was alluding to, in the second row for Study A, the *p*-value is .32. And with a *p*-value of .32, there is no way you can say that's statistically significant. It's not going to cross any statistical significance bar. And so, what people would say is, oh, the intervention had no effect. There was a null effect here is what they would conclude. But when we actually calculate the probability that the intervention had a favorable effect, which is not what a *p*-value is, we see that there is an 83 percent chance that the intervention had a favorable effect. And an 83 percent chance of a favorable effect is really pretty different from saying that it didn't work. Those are two very different statements and very different interpretations. And so, you know, it's not a guarantee that it worked. You know, we can't say it definitely worked. There's a 17 percent chance that it didn't. But still, saying that there is an 83 percent chance is really very, very different. And I think more correct than to say that it just didn't work. So I think that's one value of this approach. And then in the second panel here where we have the smaller study that had the same point estimates, the probability of a favorable effect on sexual initiation is a lot lower. Instead of 83 percent, it's 68 percent. But that's still something. It's better than a coin toss. And you can still see that there is a 91 percent chance that there was an impact on knowledge which, again, if we were relying on statistical significance, we would say something silly like, oh, there was no effect, it didn't work.

As if we were very confident that it didn't work, which is really not right. In fact, there is only a 9 percent chance that it didn't work, and we are saying it didn't work. So, I think that, at least in my mind, this is a nice example of how these probabilities—I can provide us a more accurate, nuanced understanding of the findings from the study.

Mariel Finucane: Last but not least, this is Step 4. Now we want to do sensitivity analysis. And much like the last slide, in this slide I'm just presenting numbers that John already showed you in the spreadsheet, but it's been arranged in a nice tabular PowerPoint form, but this is not new information. But just like he did, I'm going to do the same thing and kind of talk through them again just to drive home some of the main points. First, to orient you to the table, we've kind of flipped things around now, and in the first table we have impacts on STI Knowledge, so that's more proximal outcomes. And the bottom table, Sexual Initiation, that's the more distal, harder to move the needle. And now we have to study the columns. For the first one is Study A, the big guy, and the last column is Study B, the smaller guy. And as you go down the rows of each table, what you see first is that main impact estimate using our primary priors. That's the first row in each of the two tables. And then the subsequent rows in each table show the different sensitivity [analyses] we wanted to run. So, let's just quickly remind ourselves first what those sensitivity analyses were. So, looking at the first table you can see that, since this is the more proximal outcomes, in the first row our main prior distribution has that larger mean, that larger standard deviation, more permissive. The next row shows the idea that we would take that same prior standard deviation of .29 but shove the mean over to zero point zero. That's our zero-centricity analysis. Third row in the first table, you recall that we wanted to also check the sensitivity using this more conservative prior even though this is the more proximal outcome. So that's the mean of .03 with a standard deviation of .15. And then for good measure, we'll zero center that as well in the last row. So those are the four different priors we're considering. And then just to quickly repeat some things that John had already mentioned. The (inaudible) to really notice when you compare across the two columns in that first table is that for the larger study, with the n of 400, Study A, there's really not a lot of sensitivity to which prior you choose. You see those probabilities of 98 or 99 percent in all cases. And as John was saying, that's because we have more data that makes us trust our database estimates from this study more and makes us less reliant on the prior. So the prior moves around, but we don't pay much attention to it. By contrast, in the second column, in Study B, when we have just a quarter of the sample size, which doubles our standard error and makes us a lot more uncertain about the impact estimate from our particular small study, as John already showed, you're getting this much higher sensitivity, with probabilities all the way down to 77 percent. So, that's the first table. In the second table, we tell a very similar story, where for this more distal outcome, sensitivity is less when the sample size is more. That's the main take-home point here. Great. So, I'll turn it over to John now for some discussion.

John Deke: All righty. These are really just two topics to try and get some discussion going among the group. We could think about how our perception of study findings changes when we're using BASIE instead of statistical significance. And what do we make of the sensitivity analyses? But we can also use this time for questions and answers, talking about whatever you guys would like to talk about. Maybe I'll go to Q&A first to see if there's any questions that folks have on the line.

Katie Henley: John, this is Katie Henley. I have one interpretation question in terms of how you might see this being reported in the abstract of a manuscript. So, to convey a message to readers who are more familiar with frequent statistics and relying still on that *p*-value, the bad *p*-value, and combining with the Bayesian analysis that we've conducted. Just … a couple of sentences of what you would see would be included in the results section, or maybe in the discussion section of the conclusion section, implications, whatever you want to call it, of the abstract.

John Deke: Sure. Let me just acknowledge that there are a lot of different contexts that people may find themselves in when they are writing abstracts or conclusions to either journal articles or reports or whatever the vehicle might be. And so, depending on exactly what your context is, and what constraints you face, and what you can say, my answer might change. My ideal, my preferred, would be to say, just to go back to like the examples we had, we had an impact estimate of .10 standard deviations reducing sexual initiation, and we estimate there is an 80 percent chance that that is truly a favorable effect. You know, I would ideally couch it entirely in those posterior probability statements. I would talk about that impact

estimate that comes from the study data. That's the one I would typically choose to emphasize although I think, Mariel and I aren't exactly the same person, we have different tastes, and so I, think she might tend to emphasize the shrunken estimate. But … I think I would report that impact estimate. We reduced sexual initiation by .10 standard deviations, and there is, whatever it was, an 83 percent chance that that effect was due to the program. That it was a genuine effect. I would ideally want to say it something like that. And the reason I would prefer to just completely keep the *p*-values and statistical significance out of it is that those are completely different numbers. They're not really comparable. And it is only through the misinterpretation of *p*-values and statistical significance that people might conclude that there is some kind of an inconsistency between the two, but people do, indeed, misinterpret them, and so I think it could potentially confuse readers if they are combining misinterpretation of statistical significance with the thing, we actually want, which is the probability that the thing worked. Now, having said that, there are plenty of contexts still where people are required to report *p*-values and statistical significance. And so, if you are in that context, then you have to do it. And I guess I would keep several things in mind. One is keep in mind and really hold onto the actual definition of these terms. Like really try to avoid falling into the trap of misinterpreting *p*-values and statistical significance. And let me say just a little bit more about what I mean. *P*-values and statistical significance are characteristics of your impact estimate. They are not telling you anything about the true effect of the intervention. Very explicitly, very by-definition. You know, you can cite textbooks, they don't tell you anything about the true effect of the intervention. To say anything about the true effect of the intervention in a quantitative, probabilistic sense, you have to use the posterior probabilities. There's really no other choice. Really by definition. And so, keeping that in the back of your mind as to what these things really are can help, perhaps, avoid confusion in presentation. So, if you have to report *p*-values and statistical significance, I would put it in parentheses. I would downplay it. I would, try to make it clear that [what] you're talking about is the characteristic of the estimate. But then we're actually interpreting the thing we want to interpret, which is the probability that they worked and the probability that the effect was greater than .05 standard deviations or whatever is relevant to you. I would recommend focusing on that. Mariel, do you have any thoughts on that very tricky question?

Mariel Finucane: Yeah. I'll first acknowledge indeed that it is a tricky one. And then I'll add one point to John's great discussion of context mattering. He says, you know, in his ideal world, we would talk about the probabilities and eschew the *p*-values altogether, but in reality, sometimes you have to report those. I'll just say that we're not alone in thinking that that's ideal, that just focusing on the probabilities is ideal. But remember, we showed those very high-profile papers at the beginning really urging folks to get rid of the *p*-values altogether. So this is a movement, it's not just us. And in fact there are some journals, I think mostly in the field of psychology still, that do not let you report *p*-values anymore. They've really done away with them altogether. So, John and I long for and predict the time will come when that becomes more commonplace in our field as well.

John Deke: And I suspect, although if folks know of counter-examples, I'd love to hear it, but I suspect that when it comes to journals, I'd be surprised if there were journal editors who still require or push you to report *p*-values and statistical significance. And if they do, it seems like there is a ton of ammunition to push back on that. The main context in which I, myself, am still constrained and need to report *p*-values and statistical significance is in a journal context. But there are some cases where legislation has references to statistical significance written into the legislation. And that constrains what you can report in government reports. You have to reference statistical significance due to some, I think, misguided but well intended legislation that was frankly written before the ASA statement on *p*-value misinterpretation came out. And so, it was a misinterpretation. But it's there, and it's the law, and you have to deal with [it]. But that's really the main scenario where I have really run into a tough constraint on that. I think when it comes to journal editors, I'd be surprised if somebody really went to the mat in defense of *p*-values. But if somebody has had that experience, I'd love to hear it.

Russell Cole: I don't want to take this too far afield, but I do want to shove in really quick just for folks' expectations as they're preparing their impact analysis plans and ultimately their final reports for this grant. I do just want to quickly say that there is a need to report … your traditional impact estimate and standard error …like folks have said here, it's important for … reporting and for setting the stage for (inaudible) analyses. But I also want to just put out that reporting the statistical significance or a *p*-value, it's necessary for the

current team pregnancy prevention evidence review effort. Because the team pregnancy prevention evidence review effort currently focuses on statistical significance when determining evidence of effectiveness. But I do want to just say this. As Mariel and John have said, this is a moving target. This is like—this is a field that's moving. You know, the users of TPPR have voiced a need for alternate ways of thinking about evidence, particularly for small, underserved populations where it's really hard to lift off well-powered evaluations that are going to get statistically significant results. I know that folks on the line here are doing studies with small, hard-to-serve populations. So there's definitely a call for an alternate approach, and I think that Bayesian interpretation is really a good way to attempt to address that request. So, I think I'm going to try to say, please do this. Please create that groundswell to help push the needle and create more of a need for the standards to kind of reflect where information should be headed. But for the sake of achieving where the teen pregnancy prevention evidence review is now, please do plan on reporting statistical significance or *p*-values somewhere. And you can certainly feel free to come up with different ways to interpret or focus your interpretation. I just want to quickly say one last thing … That Making Right Choices report that John referenced does have some ideas where we kind of highlight both the traditional inferential statistics as well as the Bayesian interpretation, so that might be a good resource for you all to look at to see how to present some of this information and how to kind of thread that needle of how to package the findings and interpret it correctly.

John Deke: That's really helpful, Russ. And, you know, I just want to—that gives [me] a nice opportunity to offer a little clarification. So, I will always say, like from now until the end of time which maybe is a little bit too confident, but I think I will always say that you should report your traditional impact estimate, and that's what we advise in BASIE. And you absolutely should always report your standard error. That will always be the case. I would never envision a time when you wouldn't want to report those things. Those are important because they are actually necessary components to using BASIE. We can't do it if you don't have those things. So I would never say don't do that. And I think it's totally fine—and if you do those two things, if you report the impact estimate and standard error, then anybody can calculate a *p*-value if they are interested in it because a key statistic is impact estimate divided by standard error. Then you are just one more hop away, and you've got a *p*-value. So, you should never not report those things. The question that was asked was, what do you say in your abstract? You know, what do you say is like your high-level takeaway? And that's where I would really focus, in my ideal world, on the probability that the thing worked, in that takeaway discussion, abstract kind of context. But it is absolutely always the case that in your paper, in your report, whatever it is, and not because you have to, because you should, you should report the impact estimate and the standard error. That's good stuff. Definitely agree with that. So, I guess I should say, does anyone else have any questions? Well, we're two minutes away from time, so you've got a short period of time if you are teetering on the brink of asking a question to just jump in and do it. All righty. Well, it was great talking to folks this afternoon. I hope it's helpful. Please feel free to email us if you have any further questions.

Mariel Finucane: Great. Thanks so much, everybody.