Bayesian Interpretation and Core Components Introduction

Thursday, November 18, 2021

12:00–1:00 p.m. ET

Webinar transcript

Russell Cole

This is an opportunity to introduce some ideas with participants from all four of the grantees in a group setting. We're going to do some more of these group TA meetings in the future. We'll have some additional webinars that help to bulk out some of the introductory ideas that we're going to present today.

I'm going to just spend a quick moment on housekeeping. Everyone was muted on entry. We are hoping to keep lines muted today to minimize any issues with feedback and echoes. If you do have a question, please submit it into the chat at the bottom of the WebEx screen. We will have team members watching that, and we'll also have a Q&A session at the end. If you have any tech issues, we'll try to address those immediately via the chat. And we'll save the more substantive questions about content until the end. That's when we'll read aloud the submitted questions and answer them.

Again, please stay muted throughout the presentation to limit any tech interruptions. As you might have heard a moment ago, I'm recording today's meeting so that we can share it with any of your team members who couldn't attend or for folks who wanted a recording to turn back to. With that, let's begin. Emily, could you go to the next slide.

Here's a high-level presentation of what we're going to try to do today. In a moment, I'll spend a couple of minutes doing introductions, then we'll get to the main content. About half of the time, we'll introduce and motivate Bayesian interpretation of impact estimates, and then the latter half of the time we'll focus on components of TPP programs. At the close of the hour, we'll have a Q&A block, and we'll follow up with links to the slides and a recording of the presentation in the future. Next slide, Emily.

All right. Here's who's presenting today. You can see our faces here and in the video. John Deke is a senior fellow at Mathematica, with over 20 years of experience designing impact evaluations, developing evidence standards, and providing technical support to other evaluators. Recently, he's focused on using Bayesian methods in the context of impact evaluations.

Mariel Finucane is a principal statistician at Mathematica. She has seven years of experience leading quantitative policy analyses, with a particular focus on Bayesian methods, impact evaluation, and primary care delivery.

Emily LoBraico is a new researcher at Mathematica and a member of the TPP Evaluation Technical Assistance team. She has experience using component analyses to identify the drivers of substance use prevention programs and is helping to develop a series of resources for describing components of TPP programs.

I'm Russ Cole. I'm a PI on this project. I've been doing evaluation technical assistance for over a decade with OPA TPP grantees. I've been working with Emily on the program components work for OPA, and have developed some of the earlier versions of work in TA briefs that we'll include links to at the end of the presentation. With that, let's begin.

Why are we here today? Our primary goal as TA providers is to help you conduct compelling, rigorous effectiveness evaluations. That's obviously your goal, too. We're also hoping that the impact evidence that you produce is statistically significant and favorable, assuming the program works, since many audiences focus on this, including the old Teen Pregnancy Prevention Evidence Review, which was recently revived.

There's more that you can do in your impact evaluation than estimate program impacts and do a traditional inferential test. We're hoping that, in today's webinar, you see some opportunities to go beyond what you might have done in the past. You can enhance your standard presentation on the magnitude of the observed impact and the *p*-value for the inferential test with a Bayesian posterior probability statistic to provide additional decision support around your impact estimate. And you can plan on describing the components or the ingredients of your intervention to better articulate what your program is, the effective contrast that you're testing, and potentially do some supplemental analyses to link individual components to outcomes. Both of these ideas are things that may play a role in the future of the evidence review. We want to get them on your radar now so that you can plan accordingly.

We will loop back and do additional webinars on these topics next year in advance of when your analysis plans are due so that you have everything you need to lay out how you might potentially go beyond the basis of doing a traditional inferential analysis. Again, you'll hear more about this in the remainder of today's introductory presentation. With that, I'm going to pass this to John and Mariel to begin their work with Bayesian interpretation.

Mariel Funicane

Thanks so much, Russ. Really appreciate it. Really glad to be here. As he said, this first half is going to be John and me presenting on this framework that we have developed. We call it BASIE, and BASIE stands for BAyeSian Interpretation of Estimates. It's really for interpreting impact estimates for rigorous policy evaluations. Slide.

Today, we will be talking about when and why we think this framework can be helpful. It's going to be a brief introduction. And then we'll come back with more details in the future, tell you more about the theory underlying the framework, and get very practical and present to you the spreadsheet tool that we've developed that will let you use this framework, if you're interested in your own research. Slide.

First, when is this a framework that might be helpful to you? We have two use case scenarios in mind. In the first one, you can imagine that you're the one conducting an impact evaluation and you've already done a lot of the legwork. You've carefully designed your study. You've collected the data. You've figured out exactly how you want to specify your regression, and you've run that regression on a computer. And now you're holding the output from that computer program in your hands. You have an impact estimate and a standard error.

This step is where BASIE comes in. Remember that the I in BASIE stands for interpretation. You have this impact estimate, and you need to interpret it. You're wondering what to make of this impact estimate and, in particular, the probability that you've found an intervention that really does meaningfully move the needle for the population that you're aiming to serve. That's the first use case.

The second use case is very similar except that, instead of conducting the impact evaluation, you're now a consumer of this research. Perhaps you're reading an evaluation report or a manuscript about an impact evaluation, and you find yourself in the same position of staring at an impact estimate and a standard error and wanting to interpret that evidence, wondering what to make of it. Next slide.

We think that, at this crucial juncture, you're going to want to use this BASIE framework. In the next part of the presentation, we're going to tell you why we think that is, in particular why we think statistical significance testing, which is a kind of alternative way to interpret impact evaluation findings, why we think that's perhaps not the best way to go. Then John will give an example of why that is not a good way to go, what kind of information can get left on the table when that's the approach you take. And then he'll give, on the other side of the coin, an example of what you can learn if you instead use BASIE. Next slide, please.

John Deke

Thank you, Mariel. The use case that Mariel described—the situation in which we would want to use BASIE—is the situation where previously we've been using statistical significance. Just to restate: We've got this impact estimate, and we're trying to figure out what's the likelihood that this represents a genuine effect of the intervention, as opposed to just statistical noise that arises from random assignment or random sampling.

We used to use statistical significance for that purpose, but—you can go to the next slide—what we were taught—and hit it again—by the American Statistical Association in 2016 is that we were misinterpreting statistical significance. If we want to know that there is a high probability that an intervention worked, statistical significance isn't the thing that's going to tell us that.

Now, a lot of organizations put out a lot of statements all the time, and it's usually not a big deal and goes into the ether; nobody pays attention. But this statement from the American Statistical Association was pretty unusual in that the association never makes statements. They're a very circumspect, introverted kind of group, and they don't talk much. It took them a long time to gather up the courage to make this big, bold statement, and it was based on a lot of work in the journals over many years, with many researchers; they finally got together, reached consensus, and put out this statement. So, it's not a small deal. And the purpose of the statement was to say you've been doing this wrong, it's not what you think it means. You can just go through the next two animations.

This was followed up in 2019 with a special issue of the *American Statistician*, talking about statistical inference in the 21st century, like how are we going to move beyond this world of statistical significance, of *p* less than .05? And in 2019, there was a big commentary in *Nature* in which 800 researchers signed a statement saying we need to stop using statistical significance.

Now, the challenge, of course, is that you can't turn the entire ship of research on a dime. It's going to take time. And one of the problems with statistical significance actually makes it hard to replace; it was this universally applied standard thing, used in all types of different situations. So, now we have to figure out what to do in our specific context. That's what BASIE is about, trying to figure out what to do in our context. It will take a little bit of time to move from where we were to where we want to go, but we're setting the groundwork to do that. I will hand it back to Mariel.

Mariel Funicane

Great. Next slide, please. John described for you this kind of groundswell of—really nearly a scientific consensus that statistical significance testing was not telling us what we thought it could and really can't offer the

appropriate tool for interpreting evidence. I want to spend one slide telling you about some of the major concerns that supported that groundswell of publications.

Why do we and why do they reject statistical significance testing? The first and perhaps the main problem is that statistical significance testing leads to overconfidence in our conclusions, where we really divide the world into black and white, into thumbs up and thumbs down. If we see a *p*-value less than this totally arbitrary threshold of .05, we say Eureka! And if we see a *p*-value greater than that threshold, we despair and throw our research in the trash.

And the problem with doing that, with declaring Eureka for *p*-values less than .05 and despairing over *p*-values greater than .05, is that a small *p*-value does not actually imply a high probability that the intervention you're evaluating works. I'm going to say that again because it's incredibly important and, at least for me, was very counterintuitive the first time I heard it. A small *p*-value does not actually imply a high probability that the thing you're evaluating works.

And it's really important to realize that that misinterpretation of *p*-values, where, when we see a small *p*-value, we think that it implies a high probability that the intervention works, that's not just a semantic mistake we're making. That can actually be a largemagnitude, meaningful mistake. Conversely, we shouldn't despair when we see *p*-values greater than .05, which is, of course, especially important in the context of a low-powered study where sample sizes tend to be smaller. John will give some examples in a moment where even when the *p*-value is bigger than .05, there's really still a lot that can be all right. That's our first reason for rejecting statistical significance.

A second one, which I'm sure you have all heard about and thought about, is that a *p*-value doesn't reflect the size of the impact, which, of course, is incredibly important for determining whether something matters or not. A third problem is that the pervasive use of *p*-values has created this incentive system in research that can have some bad consequences. For example, using *p*-values can lead to publication bias wherein only statistically significant findings show up in the literature. And then it can also lead to problems like key hacking and data mining. I have to admit that this is something that I have done. It's not – it's just kind of baked into the current system where you try different regression specifications and cross your fingers to attain significance with one of them.

These problems together—publication bias, key hacking, and data mining—lead to an evidence base where what we see reported in the literature is hard to make sense of and hard to trust because we're really

only seeing the things that made it through these various problems. Those are three big problems with statistical significance testing. And before I give it back to John, I want to mention quickly that all of these problems are made more problematic when we're in the context of small studies, when sample sizes are small. In our future chats with you, we'll go into that in more detail, but today, keep it in the back of your mind that all of these problems are particularly important in small studies.

John Deke I'm going to talk about an example that focuses on one specific aspect of the problems with statistical significance. I want to set aside some of the issues. I want to set aside the issue that the *p*-value doesn't exactly mean what we want it to. And let's just assume that we can calculate the probability that one group, like a treatment group, had better outcomes than the other group. Let's assume we can do that. And instead, let's focus on the issue of that arbitrary bright line of saying *p* has to be less than .05 or, conversely, that we want there to be a 95 percent chance that one group is better than the other. Let's focus on that issue, and let's move to the next slide.

From a completely different world than program evaluation, let's look at sports betting and the NFL. I grabbed this screenshot from fivethirtyeight.com a few weeks ago. It was Sunday, November 7th. And FiveThirtyEight helpfully puts up these probabilities that different teams are going to win a game. The thing I want to draw your attention to is that there is no probability reported on this screen that is bigger than 95 percent. There's no probability on this screen that is less than 5 percent, conversely.

What this means is that if you were to use statistical significance to inform your decisions about placing a bet on a football game, you wouldn't place any bets at all. You would say the difference between the expected scores of these teams, they're all statistically insignificant, and so I don't think any teams are winners. Everybody's a loser. I'm not going to bet on anything.

If you take that view, which I don't think anybody actually does in the real world, then you're going to leave money on the table. You're going to miss out on an opportunity because, if I were to offer you an even-money bet on these games, you'd, in expectation, be well advised to take that bet. If I said that you can bet on only one of these things, if you look across the numbers, there is important variation in these probabilities—74 percent is importantly different from 86 percent, which is importantly different from 59 percent.

If I said, pick one of these to bet on, there's useful information for you to make that decision, and it would be much less useful if, instead of

showing you these probabilities, all we said was that there's no statistically significant difference in our expectations for the scores among these teams. That would not be useful at all. This is just to ground you in the idea that there can be more useful information than the dichotomy of thumbs up or thumbs down, and you want to pay attention to your context. How is this information going to be used? In this context, it's for placing a bet on a football game, but, in a research context, of course, there are going to be some other use cases in mind. That's important to keep that in mind rather than that arbitrary cutoff.

Mariel Funicane Thanks, John. Just to orient you quickly, we told you when to use BASIE for interpreting impact estimates. We've given you a lot of reasons why we think statistical significance testing is not giving you what you need. Now, we want to tell you a little bit briefly about BASIE and why we think it's a good alternative.

BASIE is this framework to help you interpret impact estimates and, very importantly, it's going to answer that key question that *p*-values can't answer. It's going to tell us what is the probability that an intervention was actually effective. It's going to do that using three ingredients. The first is that impact estimate and standard error from your study. The next is Bayes rule, which is just an incontrovertible mathematical theorem. And then the third, this is very important, is prior evidence.

We're going to get these probability statements that we're really interested in. That's the big win. The big cost is that we need to ground our particular study in the context of a broader, prior evidence base. This might look like a set of studies from a similar field or from similar interventions or targeting similar outcomes in similar populations, some relevant set of prior evidence. And when you put those three things together—your impact estimate, Bayes rule, and this prior evidence from the literature—that's what's going to allow you to calculate this probability instead of using *p*-values.

One thing we wanted to quickly point out is that there's nothing methodologically new in the framework. We haven't proved any new theorems. We haven't derived any new estimators. However, we did give BASIE a name. We call it BASIE, and we did that to package it up and set it apart from this broader world of Bayesian statistics. The word Bayes gets used a lot in a lot of contexts and fields and in popular reporting, and we wanted to be clear that BASIE is specifically targeted and tailored for use in really rigorous, high-stakes policy evaluation.

We think of it as defined as much by what it is not as what it is. What we mean is that BASIE in particular is not a framework that has any place for squishy personal beliefs, and we just wanted to be very clear about that

because, for some of you, it could be the case that when you hear the word Bayes you think that there are these squishy priors baked in. We want to be extremely clear that in BASIE that is not the case. We do need prior evidence, but we use hard, rigorous evidence from the literature. Next slide, please.

John Deke

I'm going to go through a different example. I did football before, but now I'm going to zone in on something a little more near and dear to all our hearts, which is interpreting findings from a study. This is a completely hypothetical, made-up example, but it might look familiar. Here we have an intervention that I just entitled Play It Safe. We're looking at the impacts of this intervention on two different outcomes, and it's in effect size units. These are the typical impact estimates and standard errors that we calculate all the time by comparing the mean outcome of the treatment group to that of the control group.

When we're using *p* less than .05 as our benchmark for statistical significance, we're going to find ourselves in a situation where we have to say the evaluation finds no impact—no impact—of Play It Safe, thumbs down, despair from Mariel's earlier slide, total failure. Even though the intervention was well implemented, the estimated impacts were statistically insignificant. Let's go to the next slide and see what happens when we have what I think is a much more nuanced and much more informative and useful—potentially—interpretation.

Here, instead of focusing on that bright line cutoff of 95 percent, which you will remember from the football analogy made it impossible for us to understand which teams were more likely to win than others, and we lost out on the opportunity to make a good bet, we can now say the evaluation finds a 90percent chance that Play It Safe reduces intentions to have sex without condoms. And we can continue to talk about even more results that would be considered statistically insignificant, but now we can actually talk about them, understand them, and interpret them with meaningful probabilities.

We can say strong implementation may explain why there's also a 71 percent chance that the program reduced intentions to have sex without condoms by at least .05 standard deviations. And we can say that there is a 69 percent chance of lower rates of positive STI tests, though any effect was likely small. This is, I hope, an example of how BASIE can be useful in the context of actual evaluations, particularly smaller evaluations, where it's really hard to get statistically significant findings. That's it for us today. We'll have a lot more to say about this on a future call.

Russell Cole

Thanks, John and Mariel. There will be time at the end to ask questions about what they just covered. For now, we're going to shift gears and talk

about our second topic, program components and component analysis. Before we jump in, I want to be clear that we just want to illustrate some thinking about components. This isn't meant to be any kind of in-depth instruction about how to do this kind of work. We just want to lay some foundation and plant some seeds for people to start thinking about if this is something you want to pursue a bit more deeply.

Emily LoBraico

To start, we want to address what a component analysis is. A component analysis differs from traditional impact evaluations in a lot of ways. In a traditional impact evaluation, we assess the effects of the whole program. Usually, we want to see if, in its entirety, it's related to better outcomes for people who receive it versus some kind of comparison group.

For example, in this hypothetical logic model, an impact evaluation might test the impacts of this whole program in the gray box on the proximal and distal outcomes that we expect a program to impact based on some underlying theory or some other evidence. This is valuable. This is important. We definitely don't recommend that anyone stops doing these types of analyses, but we don't learn everything that there is to know about a program when we do traditional impact evaluations.

One thing we don't learn is exactly what about a program is related to the different outcomes and even what are the different pieces of a program. For example, in this program, there were three separate program components that make up the whole program. Instead of looking at the whole program, a component analysis is a way to learn about these smaller pieces of the program, which are called components. And we are able to look at how these different things might differ from the counterfactual and then eventually look at perhaps the relationships between these different components and specific impacts on outcomes that we're interested in. We're talking about doing this in an exploratory way, not a rigorous way right now. This is meant to be more descriptive and exploratory because it's definitely in its earlier stages.

In terms of the definition of program components, you've probably gathered from what we've discussed so far that these are the ingredients of a program. They're the different pieces of the program that together make up the larger whole. They're usually defined or described somewhere, like in a program manual or some other written documentation.

More specifically, there are different categories and types of program components. We're definitely not the first people to talk about program components. There was a recent National Academy of Sciences report which dove into this topic. We used that to expand on and articulate seven different types of program components. I'm going to walk through each of

the seven types and give an example of them to put them in context of TPP programs.

The first one is the content, and this is the subject matter that's provided in a program. In a TPP program, this might look like information about condom use. The delivery mechanism is the way that the content is provided. This could be something like a lecture. The format of the content being delivered could take on an in-person format, virtual format, or small- group format. The staffing component is the intended training and characteristic of the people who deliver that content. It could be something like a specific training requirement of a program developer, like a two-day facilitator training, but it could also involve experiences that qualify people to deliver programs. Being a teacher is a common staffing component of programs.

The dosage component covers all of the different types of durations and intensities of the program content. This could be something like six two-hour lessons, which is a full dosage of a program, or it could be more specific, like this activity should take 25 minutes. The environment is the intended setting or location where a program occurs; a health class is a common environment. The target population features are the intended features of the people who are meant to receive the program. For instance, high school students or LGBTQ youth are two examples of target population features. Together, the seven components produce the intended experience for youth in TPP programs.

Even though I just described these as seven things that are completely separate from each other, the truth is that it's the combinations of these program components that describe how a program is meant to be implemented. If you think about a single activity from a multisession TPP program, you might have something like a 20-minute small-group activity with high school students during health class featuring a discussion about communication and healthy relationships.

In this one 20-minute activity, we have different components. We have a dosage component. We have format. We have a target population, an environment, a delivery mechanism, and content. Even though this is one thing, there are actually six components going on at the same time. And even though we would want to implement the program that way, certain stakeholders, like program developers, might have thoughts about some of these six components being more essential than others to the program's impacts on program participants.

Out of the whole list of program components that might be in a program, there's a subset that we might think of as core or essential to program impacts, but we don't know for sure if this is true until we have evidence

for this. One way that we can produce evidence is through rigorous effectiveness evaluations, like some kind of randomized study where we have all the components, and we randomize people to different combinations of them. But let's say you're in the middle of implementing a program; you can still produce some preliminary evidence about promising components by disaggregating programs into their components, like I did in the example last slide. And this can help produce some preliminary evidence about these components that is useful because we're definitely in a new field, so any information is helpful.

If we do this, it will actually provide some preliminary evidence to perhaps do some of those more rigorous effectiveness evaluations. If we do this again and again across programs, eventually we'll have enough information to perhaps make changes to improve programs or even come up with new programs that are very effective in a new way of looking at program components. Russ, I think you're up.

Russell Cole

Thanks, Emily, for helping us lay out how to think about program components, the subsets that are more influential on participant outcomes, and then the subset of those that have been demonstrated to be effective and that we can label as evidence-based components. There's definitely a lot there, and we're excited to share more detail about that with everyone in the future. For now, in the spirit of just trying to get folks interested in this, we want to hint at ways that this information might be useful to you in terms of future communication.

First, I'm going to talk a little bit about how a description of components can help provide some additional information that can be useful in your study reporting. Presenting the components of your program, particularly the ones that you think to be core, helps your reader understand the conceptual drivers of your program's logic model or theory of change. Often, the descriptions of interventions in journal articles or reports are really thin, really uninformative. And being more thoughtful and laying out these components and the subsets that are core can help audiences better understand the nature of the programs that are being evaluated.

Second, in the context of an impact study like the ones that you're doing, laying out the components of both the treatment and control groups' programs is going to help you have a much better sense and articulation of the effective contrast being tested. That is exactly what the difference is in the services or the components that are being offered or received across conditions. Better understanding of this contrast may help you understand which outcomes are most or least likely to be affected or, equivalently, have the largest or smallest observed impact estimates.

As more and more programs start laying out their components, practitioners may be able to use presentations of components to help them select programs that fit their needs and fit within constraints of their intended implementation settings. We think that this is a real contribution and is potentially something that will be integrated into the next round of the Teen Pregnancy Prevention Evidence Review.

Beyond simply creating a more transparent presentation of your intervention and the contrast for your impact studies, thinking about components may help you answer additional research questions that go beyond whole-program effectiveness that Emily was talking about. This would be like an opportunity for you to answer research questions that can supplement your impact paper or serve as a standalone dissemination product. When Emily introduced this topic, she showed a logic model that outlined how a program could be conceptualized as having three individual components and demonstrated that those individual components were expected to influence different outcomes of interest. The idea is that if you have data on variation in youth experiences of components, you can conduct analyses to capture how variation in those component experiences is associated with variation in outcomes.

To try to make this a bit more concrete, you almost certainly are going to have attendance data, which will likely vary across youth, to some extent. For example, if you're in schools, delivering a 10-day program, hopefully you have attendance data for each of those 10 sessions. Those attendance data can be used to show how youth in your treatment group had different exposures to different components of interest. The idea is that you can empirically examine the extent to which those different component experiences—for example, each day of lessons is considered a separate component—are associated with different outcomes as hypothesized by your logic model. In this exploratory space, you can find out which components appear to be most influential in pushing particular outcomes, showcasing an interesting or innovative finding. Next slide, please.

Going back to what this gets you, let's start with the first thing, the basics. What might you be able to say about your program? This is like a course illustration. The idea is that you'll be able to enumerate a variety of types of content, for example, that are offered as part of your program. You can see here content about goal setting, content about consent or healthy relationships, condoms as a type of contraception, and so on. The idea here is that it may be possible to fully list the different types of content that make up your program.

I'm going to talk about a tool we're working on that enables users to disaggregate the content and other program component types into well-defined, individual items on a checklist. This checklist captures all of the

program component types that Emily introduced, like content, delivery mechanism, dosage, staffing, everything that's shown here. It's a way for you, as a researcher or a project director, to reliably report on everything that makes up your program as it's intended and, in doing so, be able to communicate all of these ingredients of your program. Next slide, Emily.

And then there's that second benefit of components, linking variation in component experiences to variation in outcomes. Again, essentially, it's taking your program – taking your project logic model, breaking it into components, and doing some exploratory analyses around it. The working example from before showed this logic model, where a short program was made up of three broad components: a discussion about personal safety and consent, condom demonstration, and a lecture about STIs. The expectation was that these three different components might be differently linked to outcomes.

The idea is that you can explore a naturally occurring variation in component experiences and try to do some analyses around them. Perhaps in this situation, in this study, there was variation in the attendance at that key condom demonstration, that second puzzle piece. Some youth in the program attended, and some didn't. If we can convince ourselves that those who attended aren't terribly different from those who didn't attend, for example, and we see that they look pretty similar to each other on survey data that we collect at baseline, there might be an opportunity to explore hypotheses around what that condom lesson does.

Our logic model assumed that the condom lesson was likely to improve knowledge around pregnancy and STIs. Let's explore and potentially uncover an exploratory finding like the one shown in this slide. We can say that individuals who attended the condom lesson had scores on the pregnancy and STI knowledge scale that were 11 percentage points higher than those who did not attend. I feel like I'm about to have my hand slapped because I reported a *p*-value. But you can actually incorporate some Bayesian interpretation, even around these components, but more for the future.

That's the idea. You can do some exploratory analyses to capitalize on variation in component experiences and see if your logic model is right that certain components do appear to play a role in particular outcomes. There's definitely more to it than what I presented, but the goal for today's presentation isn't to get into the weeds. It's more to show the promise and get folks excited about this moving forward.

You might be saying to yourself, all right, this sounds great, I'm interested, but you haven't presented enough information for us to do this

now, and that's okay. The first thing to do is to plan on attending the future webinar where we're going to get more into the weeds on this topic.

During that future webinar, we're going to introduce three documents that are currently in process and are being refined. One is a frequently asked questions document that lays out some of the framing that Emily shared in writing. Second is a component checklist. It's an Excel worksheet that enumerates a large number of potential components of teen pregnancy prevention programs and uses all the categories that Emily presented. It's got lots of different types of content, lots of different delivery mechanisms, format, staffing approaches. The idea is that you can go through this checklist and indicate all the components of your program, the subset that you believe to be core or related empirically or hypothetically to outcomes of interest. Finally, there are instructions to guide you on using the checklist to enumerate the components of your programs.

In this future webinar, we'll also talk about fully articulating the components of your program, including the subsets that are core in your dissemination around your program and your evaluation findings. We'll talk more about doing additional types of analyses that enable you to explore variation and component exposure and how that's associated with variation in youth outcomes to identify those components that appear to have some influence on participant outcomes.

What can you do now? I think the key thing is to start thinking about the components, particularly the core components of your program. What are the most important things that define your program? What are the components that you believe to be linked to outcomes? Those are the core components. And start to think about logic models that show how your program's individual components influence outcomes. You probably already have logic models about how your program works as a whole, but now we're proposing a thought exercise that goes a little bit finer in terms of level of detail. It's thinking about how individual components of your program are associated with outcomes, both proximal outcomes that are well aligned with the individual components as well as the more distal, potentially behavioral outcomes.

The other important thing to think about now is, to the extent possible, attendance data by, for example, collecting implementation data around components and planning on building linkages around those implementation data to outcome surveys. So, for example, in that classroom setting that we talked about before, hopefully you'd have attendance data for each student in the class and, in doing so, you might be able to link the profile of attendance data to youth outcome surveys to enable you to explore how variation in youth attendance is associated with

variation in youth outcomes. And we'll definitely attempt to connect those dots in a future webinar, but this is the kind of planning that you can do now.

That's it for us for today. We're going to stop and turn to your questions. If you have questions, we encourage you to submit them via the chat. Here is a question for Mariel and John. It's a broad question. "Can you talk more about how the Bayesian interpretation can guide decision support? Don't audiences or stakeholders want something simple, like a yes or no answer about whether a program works or whether a program should be funded in the future?"

John Deke

Sure, Russ, I can take that question. And it's a great question. It's a question we often get. And I think it's an important distinction that we should probably be more conscious in making when we're talking about BASIE that decisions are binary. Decisions are often yes or no, either I'm going to implement this intervention, or I'm not going to implement this intervention. Or, in my football analogy, I'm either going to place a bet, or I'm not going to place a bet. Decisions are often yes or no, a binary decision, but that doesn't necessarily mean it's statistical significance just because a thumbs up or thumbs down is the right way to answer that question.

I could ask a yes or no question, "Should I bet on the Chiefs this weekend?" If Russ says no, but he's answering an entirely different question such as "Do you like anchovies on your pizza?" then just the fact that he's giving me a yes or no answer doesn't mean that he's answered my question. So, if you're asking should I implement this intervention, and if you're asking someone to help you answer that question using research evidence, the way to answer it isn't to say, "Oh, is it statistically significant or not?" The way to answer is what to ask back. What are your criteria? What are you trying to achieve in making the decision to implement this? What are your benefits? What are your costs? What are your risks? What are your considerations?

And then you can go into something called decision analysis or Bayesian decision analysis, and you can develop an answer that is targeted and relevant to that question at hand. In the football example, I gave the example of suppose that you can only bet on one team, so what's the criterion there? Well, it's whichever team has the highest probability of winning. That would be the right criterion for that decision. But if you have another situation? It could be a completely different criterion.

I'll give an example from the field of education; maybe you've got a school that is picking between multiple math curricula. Well, not teaching math is not an option. They're going to teach math. And so then it's just a

question of which of the curricula is most likely to work. And so then, again, you're going to compare different probabilities of intervention effectiveness and pick the one that's most likely to work.

If it's 52 percent versus 48 percent, that's far away from the 95 percent cutoff, but that's more relevant to that decision. So, it's a great point. Decisions are quite often, almost always binary yes or no, but if you're going to make a binary yes-or-no decision, you need to have the statistics and the analysis lined up with the criterion used to make that decision, not some completely arbitrary criterion that somebody literally a hundred years ago came up with in the context of agriculture experiment. That doesn't make any sense. So, great question, Russ.

Russell Cole — We did have a couple more questions that came in. John, Mariel, these are both for you. We'll start with the first one that Jenn asked. "Will you have examples of using BASIE for studies that may be underpowered? We've had lower recruitment and attendance than planned during the pandemic and are exploring options."

John Deke — Yes, we will definitely have examples for studies that are smaller and that would traditionally be regarded as underpowered. I don't know if you have any thoughts, Mariel.

Mariel Funicane — Absolutely. I think that is one of the settings where this framework can provide the best value- added. And I think, although it was made up, John's last example in our deck was a good one because in his original table neither of those two *p*-values met statistical significance, which is often the big problem that we run into when we have smaller, underpowered studies where recruitment has been an issue. Hopefully, that example gives some flavor of how you can still make useful conclusions from studies that don't attain statistical significance, perhaps because of recruitment issues. But in terms of concrete, not made-up examples that we could give right now, John, I'm thinking back to your principal professional development work. Is that an example?

John Deke — Well, strictly speaking, no, because that was a pretty big study. But they did have findings that were not statistically significant, but they were substantively important, and it was important to understand that something might have been going on even though it wasn't statistically significant. I suppose that's an example of how underpowered is in the eye of the beholder. Most people would have considered that study to be pretty well powered, but that doesn't mean that there still aren't some findings that you're going to look at differently if you have a more nuanced interpretation.

Russell Cole — And another question that came in on the BASIE side was, "To what extent does BASIE align with WWC [What Works Clearinghouse]" —I'm going to fill in another gap—"and the Teen Pregnancy Prevention Evidence Review Standards?" John, Mariel, do you want to speak to either of those?

John Deke — I can definitely say something about that. I would say it's completely orthogonal in some sense to evidence review standards. I've worked on evidence review standards before, especially for the What Works Clearinghouse. I developed the attrition standard. I led the development of the regression discontinuity design standard. I know something about that process. And necessarily and inevitably, it is a slow process that is focused on reviewing the existing literature. There's kind of a chicken-and-an egg problem for evidence reviews in which they're going to focus their standards development in order to review the literature that exists so that they can put things into different buckets.

One kind of misinterpretation that people often have of evidence review standards is they treat them as if they are textbooks or guides for best practice in research, and they're not the opposite of that, but that's not what they're intended to do. They are intended for the evidence review to classify the existing literature into different buckets. That's job one for evidence review standards. The evidence reviews probably will take some time to evolve in response to both the ASA statement in 2016 and then in response to ideas that people have for dealing with that statement over time, like BASIE.

I think that one of the nice things about BASIE is we're not replacing your traditional impact estimate or standard error. You can still report *p*-values and statistical significance. It is an augmentation to that. It is complementary to that. And as folks add this to their studies over time, then the evidence reviews, which move kind of slowly, will respond to that and evolve. I think that that is just a reality of how these things work. I think that this is a good thing to add to your study. It's not going to hurt you with respect to the evidence review in any way, but it can make your findings hopefully more useful to broader audiences. There is a world outside of evidence reviews. There are other readers of studies beyond evidence reviews. And I think this can be useful for those broader audiences, and ultimately we can nudge the evidence reviews in a new direction, I think.

Mariel Funicane — One more quick point on that, Russ. Sorry. I almost never disagree with John, but I would say it's not completely orthogonal, BASIE and evidence reviews, because one really exciting thing about using BASIE in the TPP context is that there is an evidence review, which is going to be extremely useful for developing these evidence-based priors that we were talking

about earlier. I come from the field. I focus most of my work on primary care delivery, and there is no evidence review of previous interventions to improve primary care delivery. For us, developing a prior is way harder than it's going to be in this context.

John Deke That's a super great point, looking at the question from a different angle. That's a great point. This is why I like giving presentations with Mariel. She's balancing me out.

Russell Cole I wanted to loop back to John's point of pushing the field forward to make an argument for evidence reviews to update their standards to take this into account. There is an increasing body of research in the teen pregnancy prevention field that is starting to show these Bayesian posterior probabilities as complementing or supplementing the traditional inferential tests. And we're hoping that everyone on this call is going to take whichever pill it is, I don't know if it's the red pill or the blue pill, but also incorporate this into your impact studies to push the field and help push the evidence review to acknowledge that this is an important thing to consider, to go beyond statistical significance in the determination of whether a program has evidence of effectiveness or not.

And we're going to try to make it easy by giving it a nice little spreadsheet to do it so it won't hopefully be a whole lot of work.

And there's one final question that's come in, at least thus far. We've got five more minutes, so, folks, if you do have other questions, please add them to the chat. The question is, "Are BASIE studies publishable with *p*-values greater than .05?"

John Deke Well, so I cannot speak for all journal editors, clearly. And if there is a journal editor out there who continues to enforce statistical significance as a filter in what is published, I can't control it. But I am aware of many journals moving in a direction of completely ignoring statistical significance and definitely removing it as a filter. I think that if there is a journal—I mean, if somebody gets a referee comment or an editor comment, "Oh, we're not going to publish this because it's not statistically significant— there are very strong grounds for pushing back very hard against that type of feedback. And some of those references we gave in the slides from *Nature,* from the American Statistical Association, are really strong supports for pushing back on that kind of thing. I'm curious, Mariel, what your experience has been with journals.

Mariel Funicane Yeah, plus one.

Russell Cole Yep.

And just to say this, Emily, I don't want to put you on the spot, but I remember in the brief that you were recently working on, we pointed out a number of journals that specifically highlight that they publish findings that are not traditionally statistically significant. I think *Evaluation Review* was one. Do you, by any chance—

Emily LoBraico I don't remember the other one, no. But *Evaluation Review* is correct.

Russell Cole Yeah. There certainly are journals. There's also the *Journal of Articles in Support of the Null Hypothesis*. Again, I can certainly follow up if folks have any questions about that.

Russell Cole It's looking like the questions are drying up. Oh, excuse me. "Curious if other TPP Tier 2 Phase II grantees will compare virtual and in-person implementation. Seems like it fits under component analysis." Thumbs up. Totally agree. It seems like exactly the kind of research question that a component analysis would enable you to explore. And we strongly encourage folks to do exactly that type of exploratory analysis. Is the program more effective under a virtual implementation setting versus a standard in-person implementation setting? That seems like a pretty straightforward subgroup type of analysis that you can do, and I think that it absolutely provides a useful answer to a question that many of us are asking ourselves.

I encourage folks to unmute themselves to weigh in on that question. It looks like Heather is saying that, yes, if we're able to do in-person implementation, we will definitely look into this. Well, folks, this is great. Thanks, everyone, for making time today to listen to this. Again, this is kind of the appetizer. There will be a main course next year, and we'll get into the weeds for both of these topics as two separate presentations. If you have questions about any of this stuff, you can reach out to those of us shown here on our emails. Again, we will be posting a recording as well as the slides to Max in the future. Thanks, all, for making time for this. And happy Thanksgiving, everyone.