**An Overview of Meta-analysis and Effect Size**

*Presenters:*

*Ryan Williams, PhD & Joshua Polanin, PhD*

A webcast originally recorded on September 20, 2018 and re-aired on June 19, 2019

Sponsored by AIR’s Center on Knowledge Translation

Disability and Rehabilitation Research (KTDRR)

[https://ktdrr.org/training/webcasts/webcast69](http://ktdrr.org/training/webcasts/webcast69/index.html)

Transcript for audio/video file on YouTube:

<https://youtu.be/M8h05QPMJb8>

JOANN STARKS: Hello, and welcome to today's webcast, brought to you by the Center on Knowledge Translation for Disability and Rehabilitation Research, or KTDRR, at American Institutes for Research, in coordination with the Campbell Collaboration. The Center on KTDRR is funded by the National Institute on Disability, Independent Living, and Rehabilitation Research, known as an NIDILRR, in the US Department of Health and Human Services, Administration for Community Living.

The Campbell Collaboration is an international organization that promotes positive change through the production and use of systematic reviews and other evidence synthesis for evidence-based policy and practice. The Center on KTDRR partners with Campbell's Disability Coordinating Group to help increase the number of Campbell reviews in the disability field.

The KTDRR Center and the Campbell Collaboration have worked together to offer a five-part training course that focuses on high quality methods for synthesis of evidence, including the procedures and methods for conducting systematic reviews, as well as software, tools, and strategies for analyzing and reporting data. This final session was aired and recorded in September, 2018.

I'm Joann Starks with the Austin office of American Institutes for Research, or AIR. I also want to thank my colleagues, Shoshana Rabinovsky and Ann Outlaw, who are helping with the logistics today.

In today's webcast, An Overview of Effect Sizes and Meta-analysis, our presenters will discuss the concept of effect size, which is a quantitative indicator of a relation between two variables. They will also explain the basic processes of a meta-analysis and how the technique can be used to answer complex questions.

Now I'd like to introduce our speakers. Dr. Ryan Williams is a principal researcher at AIR, and leads large-scale evaluations and research syntheses. Dr. Williams' work focuses on improving generalizations in education research through research synthesis. He is currently the principal investigator of several IES projects that focus on meta-analysis. Dr. Williams is Associate Methods Editor of the Campbell Collaboration Education Coordinating Group.

Also presenting today is Dr. Joshua Polanin, a principal researcher at AIR, who has experience in the application and use of quantitative methodology in criminal justice, education, and behavioral health. He is the principal investigator of two National Institute of Justice-funded systematic reviews and meta-analyses, and is the co-PI of one IES-funded systematic review and meta-analysis. He has a leadership role with the What Works Clearinghouse project and is also co-PI of an IES methods training Institute for advanced meta-analysis. Dr. Polanin is also active in the Campbell Collaboration and served as the Associate Editor of the Methods Group.

Well now, let's get started. If you have any questions during the presentation, put them in the chat box and we'll hold them to answer at the end. I will now hand things over to Josh Polanin. Thanks, Josh.

JOSH POLANIN: Thanks, Joann. And thank you all for attending the webinar and listening to the live cast. Both Ryan and I are very excited to be a part of this. I'd like to start off the presentation today just with a brief overview of what we're going to talk about and discuss.

As Joann mentioned we're going to go over a little bit about an effect size. What it is and why it's important, especially in meta-analysis. Then we're going to dive into how one might calculate an effect size, especially depending on what the underlying data are that you're seeing in individual studies.

Then we're going to actually go through a high level overview of three different varieties of effect sizes, standardized mean-difference, the odds ratio, and the correlation. And after that we're going to get into a discussion on meta-analysis and why we synthesize effect sizes.

We're going to talk a little bit about the basics of a fixed versus a random-effects model, how to assess heterogeneity among those effect sizes, may explain that heterogeneity in various ways. And then at the end if there's time we'll take some questions. As Joann mentioned, if you have a specific question about a response please feel free to put in the chat box and we'll get to it at the end.

With that, I'll kick things off and start with what is an effect size and why it's important to us, especially in meta-analysis. And to set the context here I'd like to start off with a thought experiment. And I'd like you to imagine that you're somebody who is interested in understanding the effects of various programs that are all using about the same basic tenets, underlying intervention effects, and you're interested in what those basic programs, the effects that they have on a certain outcome domain.

And as a way to contextualize this thought experiment I pulled a recent systematic review and meta-analysis on therapeutic effects of horseback riding interventions. And this is an interesting citation there and you're welcome to go find it.

It's an interesting review for a number of different reasons. But for our context, the interest is both the interventions themselves. I find the idea of therapeutic riding one that's of great interest. And every study included in the review included individuals with neuromotor, developmental, or physical disabilities. So a population of particular interest.

In this particular review the purpose of it was to determine whether the whole scope of therapeutic riding therapies improved balance. Balance was one of their outcomes among many other outcomes. And we can imagine in this thought experiments that the therapeutic horseback riding interventions have about the same tenets. But balance across each of the included studies might be measured a little bit differently within each study.

And if you're interested in synthesizing all those programs to come up with one basic answer to, are the horseback riding interventions effective at improving balance, there needs to be a way to group all those effects into one synthesis. And that's where the effect size comes in and that's why it's important to us in meta-analysis.

If I go to the next slide here, an effect size is a representation of the effect or magnitude of the relationship of interest and an effect size is standardized so that it's comparable across the studies. So, you can imagine that in our thought experiments study A measures balance on, say a five point scale, but study B measures balance on a 10 point scale.

What an effect size does is it standardizes each one of those treatment effects and puts it on the same scale. And once we put those effects onto the same scale, then we can combine those effects into one model and come up with one single answer to the question of interest. And the meta-analysis, at the end of all this, after we synthesize all these effects, expresses the results of that synthesis in this quantitative index of the effect size. So that's why, when we are talking about meta-analysis, we really start by thinking through what is effect size of interest, how do we get those from each one of those studies, and then how do we go about synthesizing those effects using our meta-analytic model.

Now, calculating the effect size is something that's of particular interest and is always dependent on the underlying data. But with that, I'm going to turn it over to my colleague Ryan Williams, who's going to talk through various effect sizes.

RYAN WILLIAMS: Hi, everyone. Thanks, Josh, and thanks, Joann. Before I jump into more of the specifics around computing effect sizes, we're going to go through a few different families of effect sizes here that Josh alluded to. I want to kind of situate the larger scope of it into the systematic review, a kind of general enterprise, which many of you may be familiar with.

Conducting a systematic review of the literature may involve a meta-analysis at the end of quantitative synthesis at the end, or it may not. So, we're certainly talking about situations in which you're looking for specific types of evidence that have quantitative indicators that might lend themselves to quantitative synthesis or meta-analysis. But that shouldn't indicate that there aren't other ways of reviewing literature that are out there. But if you are in quantitative effects land, we're certainly a fan of the methods that we're talking about today.

Josh also mentioned there are several different families of effect sizes. The families kind of correspond to, or are defined, by the research design that they might be applied to as well as the characteristics or the nature of the outcome. So are we talking about continuous outcomes, are we talking about categorical outcomes, are we talking about simple bivariate relationships? Things like that where there's no, maybe obvious, outcome. And so we'll talk through at least three families today.

But another thing that Josh mentioned was kind of why we use effect sizes, and that's to put everything onto the same scale. And the two basic features of an effect size, which are an indexing of the direction of a relationship as well the magnitude of the relationship. And it's important to pause here and think about what those two things mean. And I think it's going to be almost always the case that you do have to do some standardizing as a meta-analyst, but there may be a situation in which the research that you are reviewing and synthesizing is represented in outcomes that are already on a common scale. This is, I think, probably more typical in medical settings where there are types of outcomes, maybe binary outcomes, of contraction of certain diseases, or illnesses, or not that are in kind of on an actual scale. They can be interpreted and also have those key characteristics of an effect size and in the direction and magnitude of a relationship.

So just as a side note where the research is represented on an interpretable scale across the studies that are there, we probably encourage you to use that. It's something that's really easy to automatically interpret as the researcher, and stakeholders can also more easily make sense of things that are on natural, pre-existing scales rather than on things like standard deviation. Which isn't to say you can't translate things back onto other scales later on, and we can talk a little bit more about that later if you'd like.

So, we have standardized and unstandardized effect sizes. We're going to talk through kind of a few basic ways to compute effect sizes and how to compute estimates of their precision, or their standard errors, which is going to be an important aspect of doing a meta-analysis of those effects at the end.

The first family of effect sizes that we're going to talk about is what's called the d family, or the standardized mean difference family. And this is going to be appropriate where you have outcomes that are continuous. So they're not yes, no, or not on some kind of categorical nature, but you also have multiple groups. And you're looking at differences in the means between these two groups, or it can be more than two groups.

And this is going to be pretty typical for a lot of experiments or quantified experiments that are looking at the effectiveness of certain intervention, but the same family of effect sizes can be applied to situations where you're looking at differences in continuous outcomes where you have boys and girls or urban schools versus suburban schools, things like that that aren't necessarily manipulable. But if you want to compare differences, you can use this family of effect sizes.

And a little bit on the notation. I'm going to go through a few formulas here just to get at the very basics of what's going into this type of effect size. So the things that we're going to hone in on as meta-analysts here are the means for, we'll say, the two groups that we're focusing on. Treatment and control is probably going to be the most typical application here.

We have our group means X sub G1 for group 1, X sub G2 for group 2. We also need those group sample sizes. This is really important when you're doing a meta-analysis. It's going to be really hard and probably impossible to move forward unless you know something about the sample sizes. The good thing is those are typically reported.

And we also are hoping to find something about the unconditional group standard deviation. What's the standard deviation for the outcomes in group 1 or treatment, and what's the standard deviation for the outcomes in group 2? This is going to help us kind of figure out what an overall standard deviation might look like and will be kind of the vehicle through which we standardized those differences in the X bars, or the means.

And so we can put these different pieces into our basic formula for standardized mean difference effect size, where we have the differences between X bar sub G1 and X bar sub G2 in our numerator, and then we have, again, some measure of kind of an average or a pool of standard deviation in our denominator. So, this allows us to express those two group differences on whatever scale that they happen to be in terms of standard deviation units. So do they differ by 0.2 standard deviation, 0.1 standard deviation, half a standard deviation? And the nice thing about that is it's going to be a scale that's interpretable. It takes, maybe, a little bit more effort, but it's interpretable, and it tells us something about the direction and the magnitude of the difference.

We can take that standardized mean difference, and we can apply-- in fact, we should. We should always apply what's called a Hedges' g correction. This is just a small sample bias correction. If we multiply our effect size times 1 minus 3 over 4 times the total group size, the total sample size, minus 9, that's going to adjust for any small sample bias that may creep in there.

Unless we're dealing with really, really small studies, this transformation or this adjustment isn't going to have a huge impact, so we won't see any differences between our unadjusted standardized mean difference and the adjusted standardized mean difference. But it's never going to hurt you, and it's always a good idea to do. And it's usually done kind of automatically with most of the contemporary software packages that are out there.

And as I mentioned earlier, as part of the enterprise of doing a meta-analysis, we need to capture kind of two basic pieces of information. One is the effect size, and another is the measure of the precision of that effect size. This is important, because when we get the meta-analysis, the actual synthesis and multiple effect sizes, we're going to give weight to studies that's proportional to the amount of information we think is in there, which is to say the amount of information that's in the effect sizes. Studies that are larger and have lower variance, we're going to give those more weight. Studies that are smaller, less precise, a little bit more error, we're going to give those less weight. And this helps to kind of maximize the overall efficiency and the power of meta-analysis in general.

Our standard error for the standardized mean difference is going to be really a function of the group sample sizes. Remember, I said earlier you're going to really need to capture those sample size to move on in a typical meta-analysis. So, we need our two group ends, and we need our effect size that we already computed, and we can get our standard error using this formula. So nothing too bad.

So here's a quick example of what-- it's almost kind of an ideal example of kind of extracting the information that we would need to compute our effect sizes. So this is from a paper by Henggeler looking at the effects of multisystemic therapy on various outcomes, arrests, incarcerations. I forget what SRD is. So with the randomized trial they supplied in the first table, table 1, the means and the standard deviations of their various outcomes for the treatment group, multisystemic therapy, and their control group, which was just usual services.

And I noticed here that I kind of outlined, I think, the section of the output that doesn't quite correspond to the output that I'm going to use in the following example. So I think the incarceration data here with the means of 5.8 and 6.2 point are going to be what we use. Let me see if I can actually try to highlight those here for us. Not too great, but I'm working on a track pad. So those are the two basic pieces of information that we're going to try to use.

So here's a basic example run through of some manual calculations that you would do using this information. Again, if you're doing this on a large scale, you can input a lot of that raw information about the means and the standard deviations into your computer program and get your effect size kind of automatically. But that is what's happening on the back end, which is always good to have a sense of.

So the first thing that we're going to do is get our pooled standard deviation. As I mentioned earlier, that's just kind of a weighted standard deviation between the treatment and the control group. When the two groups are roughly equal sized, then it's going to look something like a basic average between the two groups. But often times, the two groups are not exactly the same like here. We have 43 and 41, so we're going to give a little bit more weight to the standard deviation that's 13.9 than the 19.1 standard deviation.

In any case, we get a pool, or an average, standard deviation of about 16.6 here, and this is going to be our denominator. We have our two group means, which are 5.8 and 16.2, which produce an effect size here of negative 0.63. So we can say that these two groups were different on average by 0.63 standard deviations, which some might characterize as at least moderate or large even.

We can apply the Hedges by the corrections. And again, this is small sample correction, and it adjusts the effect size down only very slightly to negative 0.62. Again, this isn't going to have a big effect once you get beyond, say, 30 individuals in each group. And then we can get our standard error for our effect size. So we plug-in our sample sizes, we plug-in our effect size, and we get a standard error of about 0.22. And we can use that standard error to calculate confidence interval that we would in any other situation.

And now, I'm going to toss it back to Josh, and he's going to move on to the next family of effect sizes.

JOSH POLANIN: Thanks, Ryan. It was a good overview of the standardized mean difference.

We're going to shift slightly to what we call the r family, or the correlation coefficient. And the correlation coefficient-- for those of you who have conducted a correlation meta-analysis, you know that it's a little bit different when it comes to clinical calculating it from a primary study. And that's because, unlike the presentation that Ryan just presented, where you need to find various pieces of information in order to actually estimate an effect size, the correlation itself is an effect size, and therefore, a lot of the work has been done for us.

But we need a correlation appropriate, and when would you be interested in using it? As the sign says, we're mainly interested in the correlation coefficient if the study has a continuous outcome measure, if a study is interested in the relationship between a continuous predictor and a continuous outcome, or if a study uses some sort of regression model. Those are three sort of large indicators that the correlation would be needed. And if that is the case, if those studies are of interest to you, then you're in luck, and your meta-analysis is at least relative to standardized mean difference a little bit simpler, because when you go in to study, what you're looking for is the correlation reported by the authors themselves and some sort of sample size indicator. And that's the all the information you're looking for.

Now, what you do with that information is a little bit of the rub, and that's where the next slide comes in. And for that, we'll move here. And just like in the standardized mean different world, we are interested in calculating a standard error around that correlation. Ryan mentioned that we're going to use that standard error in various ways. And of course, you can calculate a confidence interval from it, and in a few minutes, we're going to talk about weighting effect sizes by a measure of that standard error.

So in any case, you need to calculate it. So what you're going to do, if you're in the correlation world-- and actually, what your computer is going to do-- is once you have that correlation, you're actually going to transform it or translate it into Fisher's z using this formula. And in practice, while the formula is important, what you'll often do is tell your computer program that you're interested in translating it from a correlation to a Fisher's z, and then it will apply this formula.

And once you transform or translate it from correlation to Fisher's z, then you're able to calculate an asymptotic standard error using the simple formula in step 2. And the formula is just 1 over the square root of the n minus 3. And then that will give you the appropriate information to go on and conduct your meta-analysis. So the r family really, in the scheme of effect sizes you can calculate, is really a pretty simple one.

That's the r family overview. And that's some more information that I just repeated there, so I'm going to move on to the odds ratio.

So we sort of started out with the premier effect size standardized mean difference, the one that gets used most often and is fairly complicated, moved on to one that's relatively simple, and now, we'll talk about the odds ratio, one that is used quite often in medicine and in other disciplines. And it's a little bit more complicated than the correlation coefficient mainly because you have to derive it yourself again.

But let's start off from the idea that you have two groups, and the outcome of those two groups is some binary outcome. Maybe it's something like passing a test, or graduating from some program, or perhaps being cured of some disease. And the odds ratio is one of a family of effect sizes that can actually be calculated when the data is represented in this way. That's why I'm calling this the odds ratio family-ish.

So I'm talking about calculating the odds ratio, because it's the effect size that we use most often in meta-analysis, and that's the one we're going to talk about in a second. But when you have this type of data, you can actually calculate a few different effect sizes. Those effect sizes we're not going to talk about today, but there are others that you can estimate.

So again, in our example, you've got two groups, and the results and the outcome of each of those two groups is some binary indicator. And what do you do with that information? Well, the simple answer is that you're going to calculate the odds of success in each group and then divide the odds of the success for the treatment group by the odds of success by the comparison group. It's exactly what it says in odds ratio.

So we try to make this as simple as possible here. And let's imagine that you have 37 total people in your trial, 19 of which were assigned to the treatment group, 18 of which were assigned to the comparison group. And five of those 19 were successful in the treatment group, and six of the 18 were successful in the comparison group. So if you work out the odds of success in each group, you get 0.36 of odds of success for the treatment group and 0.50 odds of success for the comparison group.

And to calculate the odds ratio, you're literally going to take the 0.36 and divide it by 0.50, and that gives you an effect size of 0.72. And so to interpret this, you're saying that the odds of success in the treatment group are 0.72 times the odds of success in the control group. So the control group is doing a little bit better in this scenario.

Now, just like the correlation, in practice, really you're going to input this information into a computer program and let it do the hard work for you. But on the back end, what it's trying to do is what you're going to need to do. What your computer program is going to need to do is translate or transform that effect size again, the odds ratio into the log odds ratio. And when you do that, that allows you to simply calculate the standard error of that log odds ratio.

And the calculation for the standard error of the log odds ratio is down here at 0.2. And you add up each cell. You take the inverse of each cell, and add up those inverses, and take the square root of that, and that's the standard error. So the cells here, as you'll typically see them represented, represent the cells A, B, C, and D over in that right-hand corner box.

But the point here is that, in practice, you'll probably be telling a computer program, here are the cell frequencies for A, B, C, and D. And behind the scenes, it will transform that odds ratio into the log odds ratio, and then take the standard error of that log odds ratio. And once you have done that, then you have enough information to go ahead and estimate a meta-analytic model.

Before we move on to meta-analysis and hand it back over to Ryan, I'll just mention that there are formulas available that allow you to transform effect sizes across metrics, meaning you can transform an odds ratio into a standardized mean difference, an odds ratio into a correlation, and so on, and so forth. That information is readily available, although it is available on the Campbell collaboration website, and I'll make sure that we post that information when we follow up with this webcast.

And with that, I'll turn it back over to Ryan. He's going to talk a little bit more about meta-analysis.

RYAN WILLIAMS: Sure. So now, we've got a very fundamental understanding of what kind of goes into these effect sizes. As Josh mentioned, there are ways that we can convert among the different effect sizes. And we also presented kind of the most ideal or the most straightforward ways of calculating those effect sizes. You can get at them in a bunch of different ways. So if you don't automatically see a table of means or a table of frequencies, don't be too discouraged, because chances are there are a number of other ways that you can extract an effect size and a variance from other reported pieces of information.

So then, once we have all of these pieces of information, our effect sizes, and our standard errors of our effect sizes, we can put them together in a meta-analysis. And there's two big, basic families of synthesis here. The first is what's called a fixed-effect model-- and I'll talk through what that is-- and the other is the random-effects model.

Both are common in that they both compute a weighted average effect. So we're going to combine all the means. Then we're going to use some measure-- either measure the precision and weights-- in the meta-analysis. How those weights are conceptualized and additional information that may or may not be relevant is what distinguishes the two approaches. And so I'll talk through those two strategies in the next few slides.

So the fixed-effect model considers one source of variation, and that's sampling variance. And we already computed this as part of our standard error. Remember, we threw those sample sizes into our standard error formula as well as the direct measure of our effect size, and we estimated what those standard errors were for each of the effect sizes in the synthesis. And if we think that the fixed-effect model is what's right here, we have to believe that the only thing that distinguishes them are that they have, different people, in our case probably, in the different studies. It's only sampling that variant that determines why we don't get exactly the same result in every single study. So that's the underlying premise here.

And if you were to kind of visualize what the fixed-effect model looks like in a synthesis, it might look something like this, where you have different rows and confidence intervals around different effect sizes here, and they're all hovering around the average. But you can draw a straight line through their confidence intervals and basically capture all of them. And that's kind of the underlying image of what a fixed-effect model looks like.

So combining everything together, it's really simple. It's a weighted average. So we're going to compute our weights, which are going to be 1 over the squared standard error, which is 1 over the variance of the effect size, and we're going to take sum of the product of the effect size and their weights over the sum of the weights, weighted average. And we can also get a standard error around that average, which is going to be the square root of 1 over sum of the weights. Again, the only thing going into our weights here is the standard error based on the sampling variance that we already computed.

So some notes on the fixed-effect model. It's helpful when the effect sizes are homogeneous, when they're coming from pretty tightly controlled situations, and we expect that there's really only one true population effect. The only reason we don't get the same effect, as I said earlier, is because they have different samples. These tend to be strong assumptions, and when we don't quite believe these assumption pools, we need to turn to something like a random-effects model.

And so where the fixed-effects model kind of used one source of variation to characterize the weights that go into the meta-analysis, the random-effects model is going to have two sources of variance. One is the sampling variance, so that's still important. We still need to get those sample sizes, and we still need those to compute the standard errors, of course, by each of our effect sizes.

So that's kind of our sampling variance there, but then we also have this other piece of information called tau squared. And this is our estimate of the between effect size of the between study variance, or the heterogeneity in the effects. So where as the fixed-effect model kind of assumed that there is one population effect and all of the sample effect sizes are sampled around that and vary around that one true effect, the random-effect model assumes that there are potentially multiple population effects. And you have variance in these sub-population effects that are sampled around a super population mean.

So then we have these two sources of variation, sampling variation and variance among the study group, variance among these kind of sub-population averages. And that's what is going to drive the difference.

This has some implications for our imprints. So under the fixed-effects model, where we're assuming a homogeneity of effect sizes, we only have one of variance. Our standard errors are going to tend to be smaller under a fixed-effect model than under a random-effects model when there is some heterogeneity in the effect sizes. So because we're incorporating the two sources of variance, the standard error around the average effect is going to tend to be a little bit larger.

And so going back to the assumptions for the fixed-effect model, unless we think that the conditions generating the effects are coming from tightly-controlled conditions-- maybe the same lab team, same context, same outcome, nearly same participants under the same experimental conditions-- unless we think that those are kind of the data generating mechanisms, it's going to be hard to rule out a random-effects model. And it's also, in our experience and in most of the meta-analysis that you're likely to encounter, there's not going to be a whole lot of statistical evidence in a given synthesis to support a homogeneity model, where all the effects are revolving around a single population effect.

Usually, there's some non-trivial heterogeneity in those effects. That may be due to some of several characteristics in the study, which Josh is going to talk about, but there may be some others sources of variance that are unexplained or potentially random sources of variance. And so for these reasons, we tend to recommend that people default to the random-effects model when doing a synthesis. If you think that the fixed-effect model is the right model, it's never going to hurt you to test out the fit of a random-effects model to see if that affects your inferences in any meaningful way. And if you're effect sizes are truly homogeneous, the random-effects model isn't really going to hurt you at all.

And so with that, I'm going to kick it back over to Josh to talk about heterogeneity.

JOSH POLANIN: Thanks, Ryan. Great overview of the meta-analysis process.

Now, let's move into what you might do once you've synthesized the effect sizes using those various models. And the first thing that you're probably going to be interested in, once you synthesize the effects, is the variation among those effects in your included review.

And what we're presenting right now is a graph of a review of effects of brief alcohol interventions on alcohol use among college-aged students. And each one of those rows represents a different study, and each one of those points on this graph represents a different effect size. And those bars around the effect sizes are each studies' 95% confidence interval.

And the difference, as you move down the graph, of where those effects land across the horizontal axis represent variation among the effect sizes. So we can see, at the top, the study with ID number 7 has a slightly negative effect size of negative 0.36, and the one all the way at the bottom, study ID number 16, has a really, really large effect size of 3.38. In the end, we have an overall effect size of 0.34, represented by that diamond row there at the bottom.

So this is a really good representation of what the results of a meta-analysis look like. And although, in this example, we find a fairly large effect size of 0.34, we can see, just visually looking at this graph, that there's a fair amount of variation among those effects. And so we're not only usually interested in what does the overall effect of an intervention say, but how much variation is around that effect and then what actually explains the differences in these effects across studies.

And that's what effect size heterogeneity is. And while this is a great visual representation of effect size heterogeneity, we're often interested in quantifying that information and understanding of what our eyes are telling us match what the statistics behind effect size heterogeneity . And for that, we need to actually quantify it.

And there's three primary statistics that you'll see in the literature being used to quantify heterogeneity. Now, because this is an overview of effect size and meta-analysis, I'm not going to go into how these are derived or calculated, but I will give just a brief overview and tell you that the Q statistic gives us some indication if that variation is just due to chance alone. There's a lot of studies here, and it could just be that, due to the population or something about the way the intervention was calculated, the way it was implemented, that there's variation among effects, and that's what the Q statistic primarily tells us.

The statistic that we're often interested in the most though, however, is the second one there, and that's the tau squared. That tells us the magnitude of the variation. It's really the effect size of that effect size heterogeneity, and that quantifies how much difference there is across all of the effects. We can actually turn that tau squared into another variable, and it will give us a 95% credible interval as well if we're really interested in the spread of the effects. But for now, I'll stop by saying that that provides a nice magnitude of the variation of the effects.

And then the I squared tells us the proportion of true variation among the effects. So it accounts for that random variation that the Q value does and tells us how much of that variation is actually true. So those are the three primary ways that we can quantify heterogeneity. And again, this turns that visual information that we saw in the previous slide into something that's quantitative. It allows us to give us some evidence to what we're seeing in visual capacity.

So after you calculate and estimate that meta-analytic model, the next thing you do is you quantify this heterogeneity. And assuming that you find a significant or a reasonable amount of heterogeneity, the next thing you're going to ask yourself and you're going to want to do is what actually explains that heterogeneity. Why do those effects differ across studies? And for that, we're going to use either a series of moderator analyses, or more likely, a meta-regression.

So there's a couple of approaches here that you'll find in the literature. And the two I've listed at the top are what I'm calling the old approaches, and then there's a relatively new one that we'll talk about in the second. But what do you see in the literature to begin with? The classic example of explaining heterogeneity and talking about differences among effect sizes is what we call a one-way moderator analysis.

And in this example, imagine that you have two groups of studies. Maybe you've got a group of studies that randomly assign participants to groups, and then you've got a group of studies that non-randomly assigned participants to groups. A one-way moderator analysis would put those studies into those two groups, recalculate the effect size within each of those groups-- so within the group that was at random assignment and an effect size that had non-random assignment-- and then compare those two groups of studies across each other and see if the difference between those two groups is potentially statistically significant. And that's one way to go about testing for effect size heterogeneity, and in an exploratory sense, that can be a useful tool for us.

The other thing that you see quite often, at least in the past, has been a one-variable meta-regression approach. Now, this is where you have a moderator that's of a continuous nature. Maybe you've collected the average age of the participants in each study, and you'd like to try to explain the variation among effects by regressing the age of the participants in your study on the effect sizes. So you're looking for, now, really a correlation between the age of the sample and the studies and the effect sizes themselves. The old approach would take into account just that one variable and see if there was a relationship between those two variables.

So those are both sort of the old, standby traditional approaches, and they had a lot of use, and they're certainly great for an exploratory analysis of the variation among effects. But what we think is a better approach is what we're calling the multiple variable meta-regression approach. And essentially, what this is doing is taking that one variable meta-regression approach and applying more co-variates to it.

So instead of just being interested in the relationship between age and effect size, now, perhaps you're interested in controlling for the assignment of participants to a group. So if we combine the two, we want a control for the assignment of participants to the group, but still look at the relationship between the age of the participants in the sample and an effect size. And in a multiple variable meta-regression model, we can do that.

This is analogous to what primary studies do all the time in multiple regression models. You are interested in the effects of multiple variables or the relationship between multiple variables on some outcome variable here. Same idea in this set up, just now, your outcome is the effect size in each study.

We advocate for this approach for a couple of different reasons. You're allowed to control for confounding factors. The one I mentioned, assignment of participants to groups, is a great one. You're able to reduce your type 1 errors. By simultaneously testing moderators, you remove the traditional approach of testing one moderator at a time, and that's going to reduce the number of type errors. And this is much easier to interpret.

In the past, meta-analyses have tended to pass one moderator at a time. And when statistically significant results were found for each one of those moderators, it was difficult to know which of the moderators actually mattered. Instead, if you use a multiple meta-regression analysis, you can simultaneously enter those variables into your model, and that model will give you an indication of which of the variables actually represent a relationship between the two variables of interest.

And the final note I'll say there is, in the past, multiple variable meta-regression wasn't used as often, because it was difficult to implement. Just in practice, the tools weren't as readily available to practitioners of meta-analysis. That has changed with the increased use of tools like R and Stata. There's now plenty of opportunity and really an ease of use of tools that make these approaches available to you.

So with that, I will turn it back over to Joann, who I think will maybe wrap things up for us and provide some questions if there are any.

JOANN STARKS: Well, thank you very much, Ryan and Josh. I hope the audience will have some questions for us. I think you did a great job. It was really very interesting. I'm not a statistician by any means, but I think you did a wonderful job of making it seem understandable.

And so a question that did come up for me was is it really important for anyone who wants to be part of a systematic review team or doing like an analysis of a synthesis of your evidence-- is it important to be able to do these things, that everyone on the team should be able to do it, or do we just have to have a few experts like you that can help us out to make these things happen?

RYAN WILLIAMS: So-- Josh, you can jump in here and add your perspective on that too. Both of us have kind of worked on larger teams of people doing both systematic reviews and meta-analysis.

I know I'm a firm believer in a well-rounded team, one that has a strong amount of content expertise, library science, information retrieval expertise, and statistical expertise for doing a meta-analysis. So I don't think it is necessary that everybody on the team is expert in everything and certainly not in meta-analysis. But if you are going to get involved with one of these, having somebody on the team who can credibly claim some expertise in statistics and meta-analysis specifically, it's certainly a big help.

And, Josh, if you have a different perspective.

JOSH POLANIN: I absolutely share your perspective. I think it's important to have lots of different voices represented just like Ryan Sheridan. Certainly, it's unusual for someone to have the content knowledge and possess the skills to conduct systematic reviews and meta-analysis. So it's probably good for all the members of the team to have an understanding of where the project is going and sort of a broad view of the idea of effect sizes and how they'll be combined. But a deep knowledge of the stats and the meta-analytic models is probably not something that everybody on the team needs. As long as somebody does, but maybe not everybody.

JOANN STARKS: Well, thanks. That's kind of reassuring. Like I said, you really did do a great job of making it understandable, I think, to someone who's not an expert in statistics, so that's really helpful. Sometimes you throw these terms around, and yet inside, you're going, I'm not really sure what that means. But I feel like I do understand a lot better after listening to you today, and I'm sure those people that are statisticians got a lot more out of it even than that.

I'll go ahead and let people know that we do have an evaluation survey, and we'd really appreciate if you could respond to that for us. Just take a couple of minutes to give us some feedback about the webcast. And so [the link](https://www.surveygizmo.com/s3/4552615/Webcast-Eval-Effect-Sizes-Meta-Analysis) is there on the slides. If you download the slides, it's also going to be available there. And we'll also be sending out an email to everyone who did register for this webcast.

We are about out of time now, and so I just want to thank Dr. Ryan Williams and Dr. Joshua Polanin for sharing this information today about some important concepts for systematic reviews and synthesizing evidence. I've already mentioned, please fill out the evaluation form. I want to thank everybody for coming today, and I want to thank all the staff from AIR that helped planning and logistics. And of course, we want to thank NIDILRR for their support to offer these webcasts and other events.