Good afternoon or good morning, depending on where you're joining us from, and welcome to the "Designing Strong Studies" webinar. We're really glad that you joined us today. For the past decade, the field of education research has been undergoing a rather remarkable change. That change has to do with what is considered adequate evidence for deciding whether a program or a policy or a practice has made a difference for an education outcome. In 2002, as we'll hear in more detail in a minute, the Department of Education's What Works Clearinghouse set a high bar for what it would consider to be strong evidence of effectiveness. It did so, even though at the time there were relatively few studies in education that met the criteria for the highest study rating of "Meets What Works Clearing House Standards Without Reservations." Since that time, the education research field has risen to the challenge of providing strong evidence for causal claims. More and more studies reviewed by the What Works Clearinghouse, or the WWC for short, have met WWC standards with or without reservations. In fact, there are almost 800 studies that now meets What Works Clearinghouse standards. The education research field is now seen as a leader in upholding high standards for studies of effectiveness. More researchers want to know how to conduct studies that can meet WWC standards. The Department of Education is asking larger numbers of its grantees to contribute to our knowledge of what works by conducting evaluations that will at least meet WWC standards with reservations. At the Department of Education we want to keep this good work going, and that's why we're offering this webinar, and I'm very glad that you have joined us today. There is still much more work to be done. In some areas of education research, there is very little strong evidence at all, and practitioners and policymakers are flying by the seat of their pants when making decisions. In these areas especially, researchers have an incredible opportunity to advance our understanding of the programs, policies, or practices that make a difference for children and for young people. We hope this will help you learn about designing strong studies and reporting your findings in such a way that others can easily assess your results. When the webinar is over, please take a moment to peruse the WWC website. I think you'll like what you see. You'll see the different ways that we have been making research evidence accessible and engaging to practitioners and policymakers, including new videos, new practice guides, and new topic area pages. Thank you very much. And now over to Jill Constantine.

Thank you, Ruth, and thanks so much to the audience for joining us today. I'm Jill Constantine. I'm a vice president at Mathematica Policy Research. I'm a principal investigator on the What Works Clearinghouse, and I have been involved in the What Works Clearinghouse, in some form or another, since 2005. First, I want to give you an overview of the formal presentation today. I'm going to pick up on the goals of the webinar that you heard from Ruth, and then I'm going to go into a brief overview of the mission and some of the processes of the WWC. I'll then turn it over to my colleague, Roberto Agodini, to provide an overview of WWC standards, as well as tips on designing and reporting on strong studies. Also on the webinar today is Neil Seftor, the project director of the What Works Clearinghouse, and he will be moderating the questions. The webinar is scheduled to run about an hour-and-a-half. The formal presentation will only last about 45 minutes, and we hope to leave about 45 minutes to take your questions. As our producer noted, you're all be in listen-only mode so we can better manage this very large audience that we're delighted to have. However, you can submit your question any time it occurs

to you through that Q&A box. You don't have to wait until the formal Q&A period, although we will begin answering questions in that formal period. And all questions will be read out loud so the whole audience can hear them, so you don't have to worry that you didn't submit a question, you can hear the questions and answers to the questions that others have submitted. And, of course, you'll continue to have an opportunity to submit questions during the Q&A period. So let me start. To reiterate a bit, the motivation for this webinar, we hope to help researchers design and execute studies that are more likely to meet WWC standards on the research, on the effectiveness of interventions. Let me describe -- Ruth mentioned just briefly, any time we say "interventions," that's our word for any program, curriculum, policy, or practice you might try in a classroom or school or school district, that intervention is shorthand for all of those things. But we also want to provide best practices in reporting on your studies. And we'll end with a reminder about the different resources available to you on the WWC website. We know that some people in the audience are very familiar with the What Works Clearinghouse, but those who aren't, I'm going to give a quick overview of its mission and some of the processes. As Ruth mentioned, the What Works Clearinghouse was established in 2002 as an initiative of the National Center for Education Evaluation within the Institute of Education Sciences, which is within the U.S. Department of Education. NCE itself is an independent research arm within IES, and it commissions projects and research that are both credible and sincere, and, since they're independent, free and clear of political influence. And within this sits the What Works Clearinghouse. This is our schematic that we use to show how we -- and every time I say "we" on the webinar, I'll mean the Department of Education, IES, Mathematica, and our literally hundreds of partners across the country who work on the WWC. This diagram shows how we intend for the WWC to influence research and practice, and it informs a lot of the decisions that have been made with respect to standards and processes and dissemination. In the middle is what the WWC does, which you'll hear a little bit more about in a minute. But our mission above really sums up what the WWC is. It's to be a central and trusted source of scientific evidence for what works in education. On the right-hand side of the diagram you can see that all of our processes and standards are designed to disseminate information on what works in education to education decision-makers, and that can be teachers, school and district leaders, and state and federal policymakers. The goal is for this to support decision-makers to implement more evidence-based interventions, which will improve student achievement. The ultimate goal of improving student outcomes is really the core of the WWC and it is why you in the audience work on developing conducting research on effective interventions, and you're trying to test those interventions in a way to see what really might work for children. And it's certainly why many of us are involved in the WWC. Improving student outcomes is really our highest priority. The stakes are high, and that's why the Department of Education has invested a lot in the independence and the rigor and the transparency of the WWC. The left-hand side of the diagram reflects the crucial roles of the producers of the information play, and researchers, and that's researchers from all types of organizations, academic researchers, developers, think tanks. All researchers are obviously crucial partners in generating high quality effectiveness research. The WWC has developed high standards, as Ruth mentioned, high standards for this research, and we make all that information available on our website, along with other methods of dissemination to support you to generate more of the type of research that meets standards, which means more and better information going out to educators. So the goal is to get a virtuous cycle, not a vicious cycle but a virtuous cycle going of research and dissemination and getting that all flowing to improve student outcomes. This

diagram also reflects the information that flows back to the WWC. Decision-makers give us feedback on what they need to know, and researchers are our partners in the development of standards. This webinar is largely focused on producers of information, since it's focusing on designing and carrying out strong studies. If you're from a school district or you represent the users of information, we hope you also find this webinar useful as you think about designing studies in your district or being a consumer of research. However, we will schedule another webinar in the fall that's targeted more directly at the users of information and how to become better consumers, and what the WWC offers to assist you in your decision making. Now I want to take a minute just to describe the WWC process, just sorting out what it is that the WWC does do and what it doesn't do; okay? So as part of its mission, the WWC reviews research that tests the effective interventions of student outcomes or asks the question "what works". But we don't just summarize the finding from all those studies, we review them against standards for high-quality effectiveness research and only report on findings from the studies that meet that high quality standards. As part of our reporting, we document study details so readers have the appropriate context for the findings and details such as was the intervention implemented as intended with fidelity, where did the studies take place, for what populations of students. All these contextual factors are important for helping decision-makers at all levels understand whether this intervention might work in their own context. It's important to clarify what the WWC does not do. We don't directly test or study interventions or commission new research on interventions. We summarize existing research. We don't recommend interventions. We let decision makers use information to decide for themselves. And we don't officially approve the use of an intervention on behalf of any U.S. Department of Education programs. Now this slide, kind of getting down to our lest set of definitions to make sure everybody understands what we mean by different terms, this slide depicts the subset of all research that is the focus of the WWC. Education research, right, we think of that as our larger green circle there, education research examines all types of issues from implementation, to measurement and assessment, to effectiveness of interventions. The WWC focuses on effectiveness research and reviews all that research against the standards for a well-designed study. A well-designed study is one where you can be confident that any improvement in student outcomes is due to that intervention being studied and not some other characteristics of the district or the schools or the teachers or the students themselves. Roberto is going to talk quite a bit more about this, but we sometimes call these well-designed studies "causally valid studies," meaning, you have identified the cause of any student improvement and it is the intervention you tested. Right now the WWC has reviewed more than 10, 300 studies, and counting, on effectiveness. We've sifted through literally hundreds of thousands of the universe of all research, that larger bubble, to identify those effectiveness studies. Other types of research and information are important; for example, information on how well an intervention was implemented. To the extent that authors report this information, we also report it. The WWC has developed a study author guide with the type of information that helps decision-makers understand the context for study findings. It's designed to help study authors with best practices in reporting, and Roberto is going to describe that in some more detail and show you some screen shots of it, but that's also going to -- a link to that is also going to be available on the resource page. Just a word about evidence standards that Roberto will describe in more detail. The WWC assesses the quality of the research designs in the studies we review. To make sure we do that consistently, we have developed standards for the types of designs that can meet our highest standards, as well as our standard with some reservation. Standards have been

developed by panels of national experts in evaluation design and methodology. Again, they focus on effectiveness studies and have been developed to ensure that studies have been designed and executed in a way that maintains the causal validity of the study. Each study is reviewed against those standards by reviewers that have been trained and certified to apply those standards consistently. We don't describe that training process in our prepared remarks here, but feel free to submit a question about it if you have one. Review of effectiveness studies results in one of three ratings for that study: It can meet WWC standards without reservations, meet WWC standards with reservations, and that suggests there was something either about the study design or execution that made it less causally valid than the high standardization; or a study may not meet WWC standards, and that means the study design or the execution of the study compromised causal validity too much to meet the WWC study design standards. So in my last prepared remarks here is a schematic of what we encounter sometime in the clearinghouse that we hope to improve by offering this webinar and working together with research partners. This is an example of the outcome of a review of an intervention target at adolescent readers. In this particular review we identified over 118 studies, identified 115 of them that made some sort of claim about the effectiveness of that particular program. After we reviewed the studies, only six met our standards with or without reservation, just about 5%. So that's a larger body of effectiveness studies than we find for many of our interventions, but, unfortunately, only 5% meeting standards is still not so unusual. As Ruth says, it's getting better over time, and sometimes we do much better in looking at the research on some intervention, but we still have an awful lot of this going on, more than we would like. So the hope in this webinar is to describe to researchers the type of study design that meets standards and the type of things that can happen as you're executing a study that can undermine standards, and some practices for trying to avoid those problems. We also describe the type of reporting authors should do to ensure that the WWC can conduct a thorough review. And so now I'm going to turn it over to Roberto Agodini, and he's going to carry on with talking more specifically about the standards and study designs.

Thanks, Jill, and hello to everybody out there in webinar land. Thanks for taking the time to join us today for this webinar. Today we'll talk about two types of study designs, randomized control trials, or RCT for short, and quasi experimental designs or QEDs for short. These designs are frequently used in the kinds of research that Jill was mentioning. This is research that examines the effectiveness of educational interventions. That's really the focus of the What Works Clearinghouse. And we see a lot of studies that rely on these type of two designs. And today we'll talk about the key features of these studies that are examined in What Works Clearinghouse reviews and how researchers can design those features to meet What Works Clearinghouse standards and, in effect, strengthen those features of the designs. But before we jump into those details, I'd like to briefly mention what we mean by an RCT and a QED to set the context for the presentation. By an RCT, what we're referring to here is a design that identifies a study sample at baseline before the intervention is implemented and randomly divides that study sample into groups. And in this presentation I'll refer to an RCT that divides the study sample into two groups randomly; one that I'll refer to as treatment and the other one as comparison. The different study groups are then offered access to an intervention or not. And in my example the treatment group

will receive access to the intervention, the comparison group does not. RCTs can also divide their study sample into more than two groups to test perhaps multiple interventions relative to nothing, or multiple interventions against each other. By a quasi experimental design or a QED, we're referring to a design that also has a treatment in comparison group, but the key difference between that design and an RCT is that the two study groups are not created by randomly dividing the study sample into the groups. Instead, the researcher constructs the groups so that they're similar at baseline where one of them, the treatment group is offered or already experienced the intervention, and the comparison group has not. For example, if this was a retrospective study and the researcher identified a group of individuals or sites that already experienced an intervention, then the researcher looks for other individuals or sites that seem to be comparable to the intervention sites, and then attracts the outcomes of the two groups. One of the key features that the What Works Clearinghouse examines when reviewing an RCT is the extent of attrition in the study sample, and, again, this was the sample that was defined at baseline. And the clearinghouse's focus is on two measures of attrition. One is the extent of overall attrition, that's in the pooled treatment and comparison group sample when they're combined together, and the other one is the difference in attrition between the two study groups, treatment and comparison. Now the individual topic areas that are reviewed by the What Works Clearinghouse select either a liberal or conservative attrition boundary. A liberal standard or a liberal attrition boundary is selected if it's reasonable in that topic area to assume that the studies on that particular topic area that attrition was not due to study group assignment. In that case the liberal boundary is used. So, for example, in the elementary math topic area we use the liberal boundary because we believe it's reasonable to assume that attrition in studies of interventions that are aimed at increasing, say, math achievement in grades kindergarten through eighth, that attrition would not be due to a study group assignment. However, some topic areas adopt the conservative boundary; for example, the dropout prevention topic area adopts the conservative boundary because there is more concern there that attrition could be due to which group individuals or sites were assigned to, treatment or comparison. So starting at the top of the flowchart on this slide, the Clearinghouse first asks itself the question, do the levels of overall and differential attrition meet the topic areas attrition boundary, and if they're below that boundary, then that study falls down the yes column, and the highest potential rating that study can receive is meets standards without reservations, which Ruth and Jill already mentioned, are the highest. It's the highest rating that's issued by the Clearinghouse. And if the attrition crosses the boundary, then the study is considered to be high attrition RCT, and then the study has to assess baseline equivalence. And the clearinghouse wants to see if the treatment and comparison groups are comparable, because now attrition is too high, particularly if there's a big differential between the treatment and comparison group, and there's concern that the results may not support causal statements about the intervention. So working down the right side of the flowchart, we then look at baseline equivalence of the treatment and comparison groups, and this is in the analytic sample. This is the sample that's used to calculate impacts, so these are individuals who have the outcome data that's needed to calculate impacts, and who also have the data that's needed to assess baseline equivalence. If the two groups are comparable, then the highest potential rating the study can receive is meets standards with reservation, because there's a lot of concern that their causal statements may not be a hundred percent supported, and if the two groups are not equivalent, then the study is rated does not meet standards. And I mentioned that the meets standards without reservations and meeting standards with reservations, when we look at attrition, those are the high potential ratings,

and say "potential" because the Clearinghouse looks at other features of RCTs that I'll talk about in a minute. With QEDs and high attrition RCTs, one of the key features that examined is baseline equivalence of the treatment and comparison groups. Again, here, just like it was on the previous slide, this is all about baseline equivalence in the analytic sample, and if the treatment and comparison groups are comparable at baseline then the highest rating that the study can receive is meets standards with reservations, and if they're not comparable it will be rated does not meet standards. So given the huge role that attrition plays in reviewing RCTs and the huge role that baseline equivalence plays in high attrition RCTs and QEDs, we'll talk today about attrition and how to strengthen or how to prevent studies from experiencing attrition, and also about baseline equivalence. But for RCTs, we'll also talk about the random assignment process and what researchers can do to strengthen that process to meet Clearinghouse standards. And, also, in cluster-level RCTs, those where schools, for example, are randomly assigned to treatment and comparison group, and then individuals within those schools, such as teachers or students, are analyzed so the unit of random assignment isn't a school but either teacher-level or student-level outcomes are examined. There we'll talk about the kinds of consistency between designs and analyses that the Clearinghouse looks for. And for both designs, the What Works Clearinghouse also examines whether there are any confounding factors, whether any confounding factors exist, whether appropriate outcomes were analyzed, and the way in which missing data were handled. So we'll talk about those key study features as well. So turning to RCTs, I'll talk a bit about the random assignment process. The What Works Clearinghouse accepts random assignment at any level and at multiple levels; for example, random assignment of schools, random assignment of teachers, students. When researchers are deciding the level at which the random assignment will be conducted in a study, an important consideration of course is the level at which the intervention is implemented. So if you're studying a program that is a whole-school program, ideally the randomized control trial would conduct school-level random assignment so that implementation in the study mimics the way the program is implemented outside of the study, you know, in reality. This way the evidence that's produced by the study is useful for understanding the effect of the program and the way it operates in reality. It's also really useful because it can help to prevent contamination of the comparison group. For example, again, if you are studying a school-level program, if you were to conduct random assignment of teachers within schools, such that you had treatment and comparison teachers in the same school, there's a possibility that the comparison teachers could end up accessing the intervention, and in effect, start being treated just like the treatment teachers are, and that will undermine the -- or that will distort the true effect of the program and will more than likely provide less useful evidence on the effectiveness of that program. When you're conducting the random assignment, make sure its units are assigned entirely by chance. This, on some levels, sound obvious, but, for example, if you are having the random assignment operationalized by, let's say, educators are district-level folks, if they're the ones who is need to operationalize it, make sure that it's being operationalized in a way that is entirely by chance. Make sure that each unit or each individual, depending on the level of random assignment, that each individual has a non-zero probability of being assigned to each study group. So, you know, the classic example is each individual has a 50% chance of being assigned to treatment, 50% assigned to control or comparison, you know, the proverbial coin flip. There can be different probabilities across conditions, so you can assign 70% to treatment, 30% to comparison, and either have consistent assignment probabilities within group or use an appropriately analytical approach when you're calculating impacts.

It can be very useful to conduct random assignment within strata, and let me explain this by example. If you were conducting a school-level random assignment design and the total study sample was going to be 20 schools, 10 of which are assigned to treatment, 10 of which are assigned to comparison, with those relatively small school sample sizes, the baseline characteristics of the treatment in comparison schools could differ because you don't have enough sample within each group. So if it was important, for example, to balance the treatment and comparison groups along school achievement, one thing you could do before you conduct a random assignment is divide your 20-school sample into, let's say, two groups, one that has a lower achieving schools and one that has higher achieving schools, then conduct the random assignment within each of those two strata, and that will help increase your chance that the distribution of school achievement at baseline ends up being comparable across the treatment and comparison group. Another important point to make here is that when you're calculating impacts, you must maintain the assignment status in those analyses, even if non-compliance occurs. So, for example, if some treatment group members end up behaving more like from what you'd expect from controls or comparisons, in that they do not participate in the intervention that was assigned to them or was offered to them, you still need to keep those individuals in the treatment group when you're calculating impacts, and, similarly, if some comparison group members end up accessing the intervention and, therefore, behave more like what we'd expect from treatments, we still need to keep them in the comparison group when calculating impacts. That's important for maintaining the internal validity of the results to support causal statements. To support a What Works Clearinghouse review, it's really useful if researchers report several pieces of info about the random assignment process to make our review smooth. One is simply to indicate that a random process was used. If that's not indicated, then we have no choice but to assume that the study is a QED. And also, consistent with the points I mentioned on the previous slide, describe the assignment process in detail to make clear the unit of random assignment. You know, was it school random assignment, teacher random assignment, et cetera? What was the probability of assignment to the treatment and comparison groups? Did you do any stratification when conducting the random assignment? And, if appropriate, how were the assignment probabilities accounted for in the impact calculations? Turning to attrition, this attrition occurs when some of the sample members that were initially assigned to the treatment or comparison groups are not included in the analysis because we don't have the key data that's used to calculate impacts. And here, for a low attrition RCT, the key data is outcomes. If it's a high attrition RCT, then we also need characteristics used to assess baseline equivalence. And when we think about attrition we often think about individuals who have dropped out of a study and that's why we can't get the outcome data for them. But, in fact, attrition also includes individuals or units who are still part of the study but, for whatever reason, we couldn't collect their outcome information. So in other words, attrition -- non-response to data collection for individuals who are still in the study still counts as attrition. And this sort of brings up an important point here; that when you're conducting an RCT it's useful to collect the data that are needed to assess baseline equivalence, because if the study ends up experiencing high attrition, that baseline data are needed to assess baseline equivalence so that the study could potentially be rated meets evidence standards with reservations. So, if possible, it's a good safeguard to make sure that you collect the baseline data in an RCT. Of course if you end up with low attrition, the baseline data isn't needed to assess equivalence. So some suggestions about ways researchers can minimize attrition in RCTs, one very useful activity is to make sure that everyone who is involved in the study understands what study

participation entails. If you are conducting a study of teacher professional development, obviously teachers have to be informed of the teacher professional development program if they're in the treatment group, and both treatment and comparisons need to know what kind of data collection, direct data collection might be conducted with teachers. But, for example, if the professional development program is in math, it can also be very important for the math coaches in the schools to be aware of the study and what participation activities entail, because the professional development that's provided, and even some of the data collection, you know, could affect the work that the coaches do. It could dovetail with the work they do. Or they, at a minimum, need to be aware of the activities. But it's also really useful for higher-level staff to also be aware of the study. The principles of the schools should be aware, and even all the way up the chain to the superintendent and assistance superintendent. You know, it's happened before where a study is put into place and halfway through the study, a superintendent who didn't know about it finds out about it and decides that the study, what it's doing is inconsistent with the district goals and pulls the plug on the study. So it's best to make sure that everybody who is a key decision-maker is aware of the study. Another useful way of minimizing attrition is conduct random assignment after participants consented to study participation. And this is, of course, the case for a study where consent is required. And the reason is that non-consent counts as attrition. So, for example, if you conducted the random assignment of, let's say, teachers, and then asked them about their willingness to participate in the study after the random assignment results, any non-consent counts as attrition, counts against the study. Another suggestion is to conduct the random assignment -- it can be useful to conduct the random assignment as close to the start of the implementation period as possible, because this could help to minimize attrition that's due to turnover. So, for example, suppose a school-level RCT of teacher professional development program is going to begin implementation this coming school year, this September, and suppose the study team recruited the schools to participate in the study during the previous school year, the one that just past, and suppose they conducted the random assignment after they recruited them and informed the schools of their study group status before the previous school year ended, so let's say by the end of May or June. Well with this kind of a study the study's baseline teacher sample would need to include all the teachers who were in the schools before the random assignment results were provided. So it would have to include the teachers that were in the schools around May or June. Well between May or June and when implementation begins this fall there could be some teacher turnover. That turnover will count as attrition. It will count against the study. And so thinking about that, there's one more suggestion that comes to mind right now as I'm talking, that's not on the slides, and that is, it could be useful to recruit sites that are likely to have low attrition, particularly if you need to inform the sites about their random assignment status well before implementation begins so that study activities can occur. For example, you know, with my example where the random assignment is conducted by May or June, and then the schools are informed of their status, it could be important to do that so the teachers in the treatment schools could participates in, let's, say a summer institute where they're being delivered professional development. Well, in that situation any teacher turnover that occurs during the summer would count against the study. But if the study were able to enroll sites that tend to have low turnover, that could help with the attrition problem. Of course the results of that study would be particular for that type of a sample, you know, a sample where there is low attrition -- sorry -- where there's low turnover, so you need to sort of balance the desirability of that sample with the desire to maintain low attrition in an RCT. In terms of

supporting a What Works Clearinghouse review, it's useful if studies report separately for the treatment and comparison group the number of units that were randomly assigned and the number of units in the analytic sample. That helps us assess the amount of attrition that occurred between the baseline sample and the one that's used -- the analytic one that's used to calculate impacts. For cluster-level RCTs, ones where, for example, schools are randomly assigned but student-level data are analyzed, for those, also report by study group the number of individuals at baseline but only in the clusters that did not end up dropping out, if, in fact, any clusters did drop out, and also report the number of individuals in the analytic sample. And this is important because with a cluster-level RCT the clearinghouse examines both cluster and sub-cluster attrition. So to give you an example, if a school-level RCT was calculating impacts by using teacher-level outcomes, it's useful if the study reports separately for the treatment and comparison group the number of schools that were randomly assigned, the number of schools in the analytic sample, that teacher cluster-level info, and for the sub-cluster info, also report the number of teachers at baseline in the schools that did not drop out, and the number of teacher in the analytic sample. So if you're conducting a cluster-level RCT, the What Works Clearinghouse accepts analyses of both cluster-level data and sub-cluster outcomes, but it's important to be aware of the way the Clearinghouse uses the two sets of results. When cluster-level outcomes are analyzed the What Works Clearinghouse considers that analysis as providing evidence about cluster-level effects. So, for example, in a school-level RCT, if impacts are based on, say, school-level average achievement at follow up, you know, after the intervention was implemented, the What Works Clearinghouse considers that analysis to provide evidence about school-level effects, and it does not consider that analysis as providing evidence of student-level effects because the students in the schools may have changed between baseline and follow up, and, therefore, the effects of the intervention are confounded with any student mobility effects, particularly if there are differences in mobility across the treatment and comparison group. It's also important to be aware that in some What Works Clearinghouse reporting cluster-level effects factor into the qualitative conclusion about an intervention's effect but not its magnitude. For example, if you look at an intervention report that's been issued by the What Works Clearinghouse about a particular intervention, it summarizes the evidence about that intervention, and at the bottom of the first page of the intervention report you will see a table that indicates how many studies met standards, either with or without reservations, it indicates the qualitative conclusion about the intervention's effect; that is, whether it's positive, negative, or no effect, and it also indicates what magnitude of the effect, and that magnitude, if there are two or many studies in that intervention report, it will be based on an averaging of the results across the studies. So a study that analyzes cluster-level effects or that analyzes cluster-level outcomes, it contributes to the qualitative conclusion about an intervention's effect; that is, whether it is positive, negative, or no effect. But if that study is being lumped in with other studies in that intervention report that analyzed a sub-cluster effects; for example, student-level effects, then it does not contribute to the magnitude of the effect. Only the studies that examine the sub-cluster level outcomes contribute to the magnitude. In order to provide evidence of the magnitude of an interventions effect, sub-cluster outcomes have to be analyzed, and to provide such evidence, the sample must include the sub-cluster units that were identified before the results of the random assignment were revealed to those who were randomly assigned. For example, if you're conducting a school-level RCT and you're examining teacher-level retention, the sample must include the teachers in the schools before the random assignment results were provided to the schools. It

cannot include any teachers who joined the schools after the random assignment results were provided because the effect of the intervention could be confounded with that type of teacher mobility. One last point about cluster-level RCTs, the statistical tests of differences in treatment and comparison group outcomes or statistical test of the impacts should be adjusted for the extent of clustering, and if they're not adjusted, the What Works Clearinghouse will make an adjustment, and it will assume interclass correlations of . 2 for academic outcomes and .1 for behavioral ones. So now let's turn over to QEDs, and I already mentioned that baseline equivalence must be demonstrated for QEDs and for high-attrition RCTs, and this has to be based on the units or individuals in the analytic sample, the ones who were going to be used for the impact calculation, and it also has to be based on baseline characteristics. The review protocols in the various topic areas that are reviewed by the What Works Clearinghouse lists the necessary characteristics to check. When to support a What Works Clearinghouse review, authors should calculate the treatment and comparison group difference at baseline and express that in standard deviation units. If they see that the difference between treatment and comparison group is less than and equal to .05, then they do not need to adjust their impacts for any of the baseline measures that they examined. But if the treatment and comparison group differences are greater than .05, or less than or equal to .25 standard deviations, then the impacts require statistical adjustment. And if there is a difference that's greater than .25 for any required characteristics, then no outcomes in the domain where that measure falls can meet standards. Let me elaborate on that a little bit. The beginning reading topic area examines outcomes in several domains. It includes alphabetics, reading fluency, comprehension, and general reading achievement. Now suppose a study examines three outcomes in the alphabetics domain. If the baseline measure of one of those outcomes differs by more than .25 standard deviations across the treatment and comparison groups, then none of the other outcomes that were measured in the alphabetics domain meets standards. Looking across the topic areas that are reviewed by the What Works Clearinghouse, the types of baseline characteristics that often are required include prior measure of the outcome; for example, in the math topic area, we require a pretest of math achievement. Sometimes other related measures are acceptable, for example, the science topic area math pretest is acceptable, given the very high correlation between science and math achievement. In some cases an aggregate measure is to be reported; for example, if you're conducting a study that's examining teacher retention or an intervention that's hoping to improve teacher retention, then to get a sense of whether the treatment in comparison groups are comparable along this domain, the topic area requires that school-level retention during the prior study year, during the pre-study year is reported. As Jill mentioned, the What Works Clearinghouse includes a reporting guide for study authors, and that guide includes these two tables, which summarize some of what I've already talked about. Here, the first table, Table One, is a table shelf for what's useful for authors to report about the baseline sample, and this is all separate. You can see separate in the intervention group and comparison group. Intervention group here is the treatment group that I've been referring to. And Table Two presents similar information but for the analytic sample, the ones who have the data that's needed to calculate impacts, again, separately by intervention and comparison group. As I mentioned earlier, the What Works Clearinghouse examines several other features of RCTs and QEDs that are actually common to both of the designs, and one of those features is whether there's an existence of confounds or confounding factors. This is an issue for both RCTs and QEDs. An example of a confound, if a study conducted school-level random assignment, if it was a school-level RCT, and only one school was

assigned to, let's say to each of the study groups, there is a confound there because the study can't isolate the effect of the intervention because the treatment and comparison schools, of which there's only one in each group, can differ in ways related to the outcomes. Another example is a confound within the intervention itself, and say you're conducting a QED focused on determining the effect of a professional development program, if the professional development program that was provided to the treatment group, if, in addition to that, a new curriculum was provided to the treatment group, one that was not provided to the comparison group, there is an existence of a confound here because this study, which was focused on the effects of the PD, can't isolate the effect of the professional development. So studies that have confounds end up being rated as does not meet standards. The Clearinghouse also examines the outcomes that were analyzed to make sure that they're acceptable and reviews the outcomes along a number of features. One is that the outcomes have to have face validity. And specifically the measure has to be clearly defined. It has to have a direct interpretation, and must measure the construct it was designed to measure. The Clearinghouse also requires at least one -- that the measure meets at least one of the Clearinghouse's requirements for reliability, and there are three reliability measures that are accepted; either internal consistency, temporal stability of test and retest reliability, and inter-rated reliability. Of course, for some outcomes some of these reliability measures apply and others don't. Another thing we look for is whether the outcome is potentially over-aligned with one of the study groups, and let me explain this with an example. Suppose a study is examining the effectiveness of a curriculum and suppose that the study uses an end-of-chapter test in the textbook that was provided to treatment students and use that test to assess achievement of both the treatment and comparison students. And suppose that in the course of the treatment group using the textbook, the treatment students already took the end-of-course test before it was administered by the study. This outcome would likely be considered over-aligned because the treatment group was already familiar with the test by the time it was administered for the studies purpose, whereas the comparison group was not familiar with the test. And the last feature we look at is how missing data were handled. And the What Works Clearinghouse accepts several methods for handling missing data. One is a very straightforward approach where authors analyze units or individuals that have complete data without adjusting for missing data. Another method that's accepted is, again, individuals that have complete data that's needed to calculate impacts but also have the data that's used to adjust those impacts for any covariates like, for example, student gender, race ethnicity, et cetera. Studies can also work with complete case data but adjust for non-response using weights, and then in the case of low-attrition RCTs, and only with low-attrition RCTs, multiple imputation and maximum likely techniques can be used to impute outcomes. If studies do use multiple imputation or maximum likelihood, we ask that they state the software package and procedure that was used or provide a citation for a peer-reviewed article or textbook that describes the technique that was used. But it's important to keep in mind here that while imputation techniques can be used in low attrition RCTs, it can also be used to impute data for the covariates that are used in impact calculation, but imputed data cannot be used to meet the attrition standard. So there are several resources that are available to folks that will elaborate on what I've covered here. The What Works Clearinghouse's Procedures and Standards Handbook covers what I talked about here, and also in greater detail, as I mentioned earlier, and as Jill mentioned, there is a reporting guide for study authors that indicates the information that make it easy for us to do a study review. There are also study review guides and instructions. If you have questions, please send them to

the What Works Clearinghouse help desk and there's many ways you can follow the What Works Clearinghouse, and we hope you do.

Thanks, Roberto. At this time, we're going to turn it over to one of our participants who has been behind the scenes at this point, Neil Seftor, the project director for the What Works Clearinghouse, is going to come from behind the scenes now and moderate the questions and answers. As I said, I'm pretty sure we've received many questions already, but you can continue to send them in, and Neil will read them out loud and direct them to Roberto or me.

Thank you, all. The first question is just a procedural one. And someone wants to know whether the slides will be provided after the talk, and, yes, we will provide the slides after the presentation. And also, as Roberto mentioned, all of the information that was presented here, and a lot more detail, can be found in the WWC Procedures and Standards Handbook. That's a wealth of information for background on WWC standards. Okay, the next question is for Jill, and that is, "Other than the resources presented on the last slide, are there other ways to learn more about the standards?"

Thanks. That's a very good question. So, first, your first resource and best resource is that WWC Procedures and Standards Handbook. It is quite a meaty document, so it's very detailed, there's lots more information and detail than presented here, including all the research and methodology behind the standards. So when you click on that link, you will see a very large document, I think a few hundred pages with all the appendices, so there will be a great deal of information there. However, if you're very interested in learning more about the standards and applying the standards, we frequently get questions about actually becoming a certified reviewer, and that's a great question. The WWC sponsors reviewer certification training for people to become certified reviewers, when an upcoming training is becoming available to the public it will be announced on our website, so you'll see it up there on the rotator in what's new. And it will be a link for you to actually apply. You do have to apply to get into a training. You provide information about your education and qualifications, and that helps IES select who might be the appropriate candidates for a training, because we are looking – IES and the WWC is looking towards who is most likely to complete the certification process successfully. We try to meet all the demands for people who would like to become certified reviewers through annual trainings. Anybody may apply. The WWC does not require, for example, an advanced degree in education, but you should have very strong training in research, design, and methods. And if you're accepted to a training, it isn't just about attending the training, you have to attend the training and pass a couple different sorts of tests, a multiple choice test and a practice study review to actually become certified. So if you want to actually conduct or use that are consistent with WWC standards, then you must take the training and be successfully certified.

Great. Thank you, Jill. Another item about certification and reviewers, on the WWC website we have a list of certified reviewers, and we particularly note those who are certified on the different versions of the standards. The most current versions are the versions who meet standards. Just a reminder, I was reminded by our producer that the slide deck and the webcast navigation slides are already available in the resource list widget, indicated by the green folder icon at the bottom of your screen. This next question is for Roberto. "How do WWC standards address school consolidation when considering attrition in cluster randomized control trials?"

So I can't recall ever coming across the situation, at least not in reviews, but having in my own work, and my sense here is that if two schools, for example, were consolidated when assessing attrition, we still count, in a sense, the two schools that were randomized at baseline, and then if they were consolidated into one and they're still in the sample, then they would be considered to not have attrited; whereas if the consolidate, the school dropped out, then the two schools that were used to create the consolidated schools, they would then be considered to have left the sample.

Okay, great. Thank you. For Jill, "Does the WWC have any reviews based on disabilities, and if so, does that include language-disabled children? Also, is the WWC focus only on K through 12?"

Oh, great questions, separate questions, but that's okay, you used your question space efficiently. So we do have several topic areas focusing on children with special needs, young children with special needs, children with learning disabilities, emotional, behavioral disabilities. So you can go to our website and search by -- you can look at our different topic areas or search by a particular disability. And if we you don't see your area of interest covered, you should certainly submit something to the help desk telling us what area you work in, and that you think it would be really useful to have an area looking at that issue. And the WWC does not only focus on K-through-12 education. There is a postsecondary topic area, and there have been reviews of individual studies, reviews of reports, and even practice guides, all of our products, that have looked at postsecondary programs. It's a little bit more -- it's more recent than our K-through-12 work, so there's not as much of it as our K-through-12 work, but it's growing rapidly, and you can now look under that topic area.

Thanks, Jill. The next one is for Roberto. "How are the attrition standards defined, and who is defining the liberal and conservative standards?"

So the details about the attrition standards are covered in the What Works Clearinghouse Procedures and Standards Handbook, which is the first link on the slide on your screen, and a lot of work was done

to identify what constitutes the liberal and conservative boundaries, and what that basically amounts to is the liberal boundary includes combinations of overall attrition and differential between treatment and comparison group that are acceptable if the study has attrition below those values, and also what is considered unacceptable, that goes beyond that boundary. So, for example, you know, a certain amount of overall attrition, and a certain amount of differential attrition at various values is acceptable, but once it goes beyond that, they're not. And like I said, the handbook describes the details as to how we came up with the liberal and conservative standards. In terms of how they're defined, in each of the topic areas, the principal investigator for the topic area works the content expert in the topic area to decide whether a liberal or conservative standard should be used. And as I mentioned earlier, it really has to do with what's a reasonable assumption in that area. So going back to my example in the elementary math topic area, we've adopted a liberal boundary because we assume that we can tolerate more attrition than, let's say, a conservative boundary would indicate, because we think it's reasonable to assume that attrition won't be due to the assignment of either units or individuals in a study.

Great. Thank you. For Jill, the next question is, "Is there an appeal process for WWC reviews?"

Oh, yes. Excellent question. So the short answer is, yes, there certainly is. And what happens, if you have a question, just even a question about a review of a specific study or the review of an entire intervention, or if you have a complaint, right, you think the WWC missed something or interpreted something incorrectly, you can submit the question or the complaint or the, you know, different information into our help desk. We take everything in through the help desk, and identify the study, and you write down what your question or your issue is, and be as specific as possible. That helps us figure out what needs to be done. And if it a question or a concern about a particular study, it will go through a quality review, and that means a review team outside of the original review team that reviewed the study, an independent one, will look at the study itself and relook at all of the information and all of the processes for reviews the study, and they will draft a report. And in the report, if the quality review team finds that all the proper procedures were followed, a report will be issued back to the person who submitted the question indicating that. If that independent quality review discovers there was a mistake made, we will still issue the report back to the person who asked the question, but we will actually change your report, and if there is an error in one that's out there, we'll take it down, we'll correct it, we'll repost it, and then, depending on the nature of the error, we'll actually have a footnote saying this was corrected to deal with, you know, whatever it was so that people know the earlier version had an error and it's been corrected. Neil, did you submit another question?

I'm sorry, I was muted. Okay. Yeah. The next question is for Roberto, and that is whether baseline equivalence has to be established before randomizations or can it be after a randomization, such as in the spring but before implementation in the fall?

And that's the measurement of baseline or the assessment of baseline equivalence?

The measurement of baseline, yeah.

The measure of baseline, yeah. It has to occur before -- well the randomization results may have been provided and the baseline -- and the study might be using baseline data from, let's say, the prior school year. For example, if a study was using administrative data it could use student achievement from the prior school year, so the baseline data would be coming from an existing data source, and the random assignment really couldn't affect those data in any way. Is that the nature of the question?

Well let me -- this is Jill. Let me offer one clarification. A random assignment does not actually require any measurement of baseline; right. So you don't have to show that groups were equivalent at baseline, so you can conduct your random assignment and there's no requirement ahead of time that you show that your groups are equivalent after the randomization. Sometimes it's nice to do that. Sometimes people do it anyway. Where you would need to show that your groups are equivalent is if now your random assignment study, if you've had too much attrition, either overall or differential or a combination, from your study. Okay? So then if you've had too much attrition from your study, then you will have to show that your remaining treatment and control students were equivalent at the point that your intervention began. So I think that might answer the question. I know it's a little bit hard because we only see what's written, so keep that in mind. It's only about for randomization. You don't have to establish or measure baseline at all. We advise you to because you never know what might happen in terms of attrition, at which point you would have to establish equivalence.

Right. Thank you. There's a lot of details in those random assignment studies. For Jill, I know that the WWC also has standards for the review of single-case designs and regression discontinuity designs. Are there similar resources, or will there be a webinar focused on these designs?

Oh, thank you, that's a great question. We probably should have said this at the beginning, so we apologize for that. So this webinar has been focused on randomized control trials and match comparison groups, or what we call quasi experimental designs. But two other types of study design can meet WWC standards; regression discontinuity designs and single-case designs. So there are resources for those standards also. They are also in the Procedures and Standards Handbook, so you can see how you have to design your study, how you have to execute, and what you have to report. And there's a lot about the execution and reporting for both of those types of study designs. So the resources are there. Regression discontinuity designs when the standards were developed were still relatively rare in education, but

they're becoming more common because they're so -- they really are facilitated by the spread of administrative data. And then single-case designs are very -- are common, sometimes very common in some of our special education areas. So right now we don't have a similar webinar. We have done webinars and presentations in the past focused on those samples. Right now, in our next webinar, we don't have a presentation on those types of study designs planned. But that's a great question for us, and if that would be a useful webinar for you, absolutely, send us a question into the help desk, and send a comment into the help desk saying, "I would really like to see a webinar on this." And, you know, sometimes a little context about, you know, who you are and how you use it and how you would use that kind of standard is very helpful too. So thank you, that's a very good question.

Okay. For Roberto, "If classrooms cannot be randomly assigned but the students are randomly assigned to the classrooms, is that considered an RCT or a QED?"

I mean here, you know, it will be important for us to look at the intervention that's being studied. But if the study can, in effect, document that students were randomly assigned to the classrooms, and this was a useful design for examining the intervention, that's the focus here, then it would be considered an RCT. And an analog here, although it's not a classroom-level assignment, might if you think about the way charter school studies are often conducted, where the study examines charter schools that have more kids who want to go to them than the number of slots that they can fill, and so they hold a lottery with all the children. And if someone studies the kids who were selected for charter school entrance and the kids who were not selected, and if that selection process was random, then the study did not create the groups and did not randomly assign who gets the charter school and who doesn't, in fact, that was done by the schools themselves, in that case it's effectively a randomized control trial. Again, provided that the researchers can document that the random assignment was conducted in the ways that I outlined in the talk, and sometimes on those studies they even attend the lottery process.

Okay. For Jill, "Can a study be submitted to the WWC for review before being submitted for publication to find out if it might meet standards or get recommendations so that it will?"

Okay, I'm going to answer that question in two ways. So a study can be submitted to WWC for review before it's been published or been submitted for publication, because being submitted for publication or being published is not a requirement for a WWC study in that -- this actually is an important point that comes up sometimes -- we search for all research that will be made publicly available. It does not necessarily have to be published in peer-reviewed journals. It can be made available through other ways. It can be reports. So we will certainly review things that aren't published or haven't been submitted for publication, because we want to get at all the literature, including the gray literature. But

it's not because we're going to tell you if it looks like it's on the right track or, you know, it might meet standards. We really only review studies to determine if they meet evidence standards, in other words, if you look at the findings and they meet the standards for causal validity. So you can submit something that's still interim findings for example. It can be a first-year follow up or something, and, you know, you may have further follow ups down the line, but right now we don't give the kind of, yeah, it looks like it's going okay or not review. You know, and that's something, if that's the kind of thing that would be useful to the public, you can certainly submit that as a suggestion, but it's not something we do right now.

Okay, thanks. For Roberto, "Teachers who are in both the treatment and comparison groups, if they have students in both groups, allows for the control of the teacher effects but might result in some contamination. Is this better than having teachers separate in the two groups?"

This is a tricky one. You know, I can appreciate the goal of controlling for the teacher effect. It's a much more powerful design in terms of statistical power if teachers can have both treatment and comparison students. For example, if a teacher is teaching at the middle school level, assign some of the teacher sections to treatment and some to comparison, but as the person who submitted the question noted, the big issue here could be contamination, where the teacher can't separate what they've learned in the intervention and only limit it to the treatment classrooms or the classrooms that were assigned to treatment and those that are assigned to comparison. I think our advice here would be that this could be a useful design if the intervention that's being delivered really does not involve the teacher very much. So, for example, if the teacher -- if technology was going to be delivered in the classroom and the technology amounted to students getting some time during class time to work with computers, let's say, that were this the back of the classroom and didn't require the teacher really to be involved in that intervention, then it seems reasonable that a teacher can have treatment and comparison sections, and the thread of contamination should be much smaller in that setting. But if the intervention is really striking at the core of what the teacher does, let's say the intervention is really influencing the teacher's practices and instructional approaches, it's hard to imagine that contamination won't be an issue, and in those cases, you know, it's likely that the study may not find an effect if, in fact, there was one, because of contamination.

So this is Jill. Let me just add onto that. So contamination doesn't necessarily mean it wouldn't meet standards.

Right.

It depends. This is a very contextual sort of question, so not the ones we can do a one-size-fits-all kind of answer, because it's so dramatically, as Roberto was suggesting, really depends on your context; right. If it's a supplement, it would be great just to have one teacher, and some kids are randomized out to go out of the classroom and get a supplement and other kids aren't. So it really depends on the nature of the intervention you're studying, so it's a little bit hard to give a general answer on that, as Roberto suggested. So he's listed all the things to consider when you're thinking about that, but I would say that one would depend on the intervention being studied.

Okay, thank you. For Jill, "How closely is the information provided in this webinar aligned with standards for IES proposals?"

So that's one of those good, it's hard to know exactly what the question is about. If it's about how WWC standards are aligned with, for example, some of the grant programs offered through NCER, I believe that answer is it depends on what type of grant, what goal, what level, all those kind of things. And then if it's a broader question, for example, other Department of Education programs, like the I-3 program, which is actually not out of IES, there, again, it depends on the nature of the type of grant you're going for, a development validation or scale up. So that's a good example if you have a very specific question about a specific grant program and WWC standards, then you should submit it to the help desk. There are grant programs, not just at the Department of Education, at other federal agencies that do point to the WWC standards as the standard for when a study has been constructed in a way that any evidence provides on the effectiveness of intervention is, you know, acceptable standard for a high-quality study, but it really depends on the program you're looking at.

Great. Thank you. For Roberto, another question on attrition. "If attrition can be explained, is it still considered in determining the rating of a study?"

Unfortunately, yes. The problem with attrition is that -- and I can draw my own experience here where I've had sites drop out. Sometimes they've literally told me the reason why they're dropping out. And while that's useful, what we don't know is how they may differ from other sites in unobservable ways, the classic selection bias problem. So if a site drops out or individuals drop out, it is considered attrition, and that's why the topic areas have adopted either a liberal or conservative boundary. They've adopted the extent to which they can tolerate attrition before the simulation work we've done has shown that it starts to bias results in a significant way. And so at the end of the day, it does ultimately count.

Okay, thank you. Jill, a more overarching question. "What is your advice to school districts that provide data to researchers and are trying to promote stronger study designs?"

Oh, that is a great question. Oh, from some of our school district attendees it must be. Okay, so I'm going to answer this with both hats on, my WWC one, but also my researcher. Yeah, so it's a lot of work, research in school districts. So I would say a couple of things. So the best things are always to plan ahead, if possible, for any sort of new intervention or any curriculum or policy or something that's going to go in place. If you can get out in front of it, then you have the best chance for a flexible sort of research design, including, so, for example, you want to offer something and you can't possibly roll it everywhere your first year, because you're just constrained resource-wise or infrastructure-wise, that's a great time. You can do a very nice randomized control trial approach without it being controversial, because if you can't roll it out everywhere, you could say, well let's pick some places and do it some places and not others since we don't have the resources to do it everywhere. And if you're forward thinking, you can get a very rigorous study design in place. If you're trying to do it kind of middle of the road and you're retroactively providing data to researchers, that can still work, since we have different kinds of standards. But you're already kind of starting a little bit with a less strong study design if you're unable to think ahead enough to do a randomized control trial, so I would say that's the main thing. Be aware of what kind of study designs might be possible, and to the extent that you can get out there ahead of when you've actually rolled out the intervention, do that. And then otherwise, I think it's just being mindful of some of these other things that, as Roberto has indicated, that kind of trip up perfectly well designed studies. So you had a great plan for studying something and then something else got introduced at the same time, or if you did, amazingly, manage to get a randomized control trial in place, you know, schools or students drop out or the things that pains us the most when people get very good study designs in place is they start moving people around in a not random fashion, so things like that. So it's a great, great question, but like I said, if at all possible, thinking ahead and thinking about a research design right before, the year before, the semester before you put it in place is your best approach to the most strong study.

Thanks, Jill. For Roberto, "How do I know what topic area my study will be reviewed under; that is, how do I plan for the best way to establish baseline equivalence, as it might differ across areas?"

So here what we would recommend you do is look at the different topic areas that are reviewed by the What Works Clearinghouse, and you can find that very easily on the homepage of the What Works Clearinghouse. And if it's clear to you where the intervention or program that you're going to be studying, if it's clear to you which topic area it will fall under, then the only thing you have to do is look at the review protocol that is used in that topic area. But if you're uncertain or if possibly your intervention might be reviewed by multiple topic areas, which can happen, I would suggest, although I'd

ask Neil and Jill to confirm this, to submit a question to the What Works Clearinghouse help desk just to be sure that you're thinking about the right topic areas in designing your study.

Great. Thanks.

Is that right, Neil, or Jill?

Uh-huh, yeah. Yeah, I think that's a great suggestion.

They can ask a question, yeah, okay.

Right.

And also protocol, you did mention that protocols are available online so --

They are, yeah.

Yeah.

So if you think you're thinking, gee, I think I have, you know, beginning reading or adolescent literacy type intervention, you can literally go look at the protocol online and see what it falls under, or what the specifications are. But if you don't think yours falls anywhere, then certainly submit a question.

Okay, great. Jill, "Why are different dosages of the same intervention not an eligible comparison for the WWC?"

Oh, that's an interesting question. Actually I would say my answer in that is a little bit it depends. It wouldn't be an eligible comparison, I would say, for what we call an "intervention report." So if you're trying to study a certain reading program and that intervention report is about the effectiveness of a reading program, then doing a study where you compare X number of hours, you know, for children, you know, two hours a week, compared to four hours a week isn't giving you a clean test of the overall effectiveness of that intervention. So that's the kind of thing, that wouldn't mix well or combine well with other studies that are really just testing the overall effectiveness of that intervention. Now it's possible -- again, this is one of those contextual sort of questions. We do sometimes report on single studies that have -- you know, they're just looking at one question and they're not trying to look at the overall effectiveness of an intervention or a program. And it may be -- again, it depends on the context and the importance of the study -- that if there's something about that study or that particular comparison that's very important or policy relevant and it was well executed, it is the kind of thing that might fall under our single study or quick review protocol that gets mentioned prominently. But the author would have to be very careful about the conclusion; right. It would have to be we're just showing this much of an intervention is, you know, better, or it doesn't make any difference, or whatever the finds are than this particular intervention. So that's all it would be. You wouldn't combine it in a report that's looking at the overall effectiveness, but it may be that there's some other way that information might be used. The other place where it could get used in what we call our "practice guides." That's a very good example, because practice guides are really based on, you know, what practices are effective in a classroom. And if there is a study that was tested in a way that it showed, you know what, more of this is better or more of this doesn't make a difference, and that could be good supporting material to one of our other products called the "Practice Guide." But that's correct; it would not fit into our intervention report framework, in general.

Okay, thanks. For Roberto, "In an RCT, if we find that there is a baseline difference between the two groups that's greater than .25, does that mean we should reassign the randomly assigned students?"

So I'm assuming here that the study has the baseline data in hand. They conduct their random assignment, and before, of course, they implement the program or intervention with the treatment group, they can already see that the treatment and comparison group differ along this baseline measure. And then I guess the question is, should we redo the random assignment since we know there's a baseline difference that's greater than .25. The trouble with doing random assignment again is that you did so because of something that you saw with the first results; that is, you know, there was a baseline difference that's greater than .25. And trying, then, to calculate impacts based on a re-randomized sample, that's a very complicated issue. And so this is really a good example of a way to avoid this problem by design that I mentioned earlier, and that is to do stratified random assignment. So if your study sample, let's say it's schools, and you need the schools across the treatment and comparison group to be comparable along baseline achievement, if you see in your study sample that baseline achievement varies widely. In that type of a sample, particularly if it's a small study sample -- I

talked earlier about a 20-school sample that would have 10 treatment, 10 comparison schools with a wide range of baseline achievement at the school level, if you randomly assigned 10 to treatment 10 to comparison, there's a good chance that they could differ, the treatment and comparison groups can differ along baseline achievement, perhaps by this .25 threshold. So a better way to avoid this problem is to solve it by design. Split your sample into low-achieving schools and high-achieving schools or even create -- or even separate them further. Depending on the number of schools you have, you can separate them into three or four groups, where each group contains schools that are comparable along baseline achievement, and then do the random assignment within each of those strata. That will give you -- that will help maximize your chance that you have a baseline equivalent sample. One thing to note is that if you do stratified random assignment you do have to account for the fact that you created strata when you were conducting the random assignment. You do have to account for that in your impact calculations, in your statistical tests of the impacts. Specifically, you lose degrees of freedom for having done the stratification, but there's work that's been done, I believe by Steve Raudenbush, that shows that even with relatively small sample sizes, the benefits of doing stratification far outweigh the statistical loss in precision in doing so.

Okay, thanks. For Jill, "Is there a minimum sample size for studies to be reviewed, or does the sample size affect the rating. And also, do groups have to be similar in size?"

Oh, good questions, and actually related a little bit to some of the things Roberto was just discussing. So the short answer is, no, there isn't a minimal sample size to be reviewed. We've actually looked at that a few times. There's not a great statistical reason for doing so. Although, the sample size does, indirectly, not so much directly, can, not always, but can affect the ratings, so we'll take small studies. Generally what could happen is what you'd expect in small studies. You don't have a lot of power or precision, or you could have a lot of noisiness in your measures that your findings tend not to be statistically significant. And statistical significance does add to your overall qualitative rating, in that we have some requirements for achieving statistical significance to get a reading of the highest rating of an intervention of positive effects. So it can affect your rating, although, like I said, not directly. It usually is because you won't get statistically significant findings. The groups don't have to be comparable in size. They could be very uneven in size. But there, again, if one group is much smaller than the other, either group is small, that will work against, a little bit, your statistical precision. And then the third rating where, actually, sample size is the one rating where it does fit in directly, which is what we call our "extensive evidence rating." That is our rating where we try to give some idea of how large or how many places the study was conducted in, and there, there are some minimum requirements for a number of studies and how large the schools had to be, so that's in what we call our "extensive evidence," which is really just either small or moderate to large.

Okay, thanks, Jill. And thank you both for all of your answers to the questions. And thank you all the participants for sending in your questions. We've tried to address as many as we could in the time that we have, and if there are others that you want to send into us, please feel free to send them either through the webinar or through the help desk, and we will try to address them. Just to remind you, the slides and the recording will be available, and the handbook is posted on the WWC website, along with some other resources if you need some more information.