Designing Quasi-Experiments: Meeting What Works Clearinghouse Standards Without Random Assignment

March 3, 2015

Hello, everyone, and thank you for attending today's webinar, Designing Quasi-Experiments, Meeting What Works Clearinghouse Standards Without Random Assignment.

Just some housekeeping before we get started.
You can make the slides bigger on your screen by clicking on the grey triangle in the bottom right corner of the slide's window and dragging it. We encourage you to submit questions throughout the webcast using the Q&A tool on the webinar software on your screen. You can ask a question when it comes to mind, you don't have to wait until the Question and Answer session. Because we're recording it, every member of the audience is in listen only mode. That improves the sound quality of the recording, but it also means that the only way to ask questions is through the question-and-answer tool. So please use that.

We've scheduled an hour for this webinar. If we do not make it through all the questions we've received, we will respond to your questions via mail after the event. The slide deck, and the recording, and transcript of the webinar will be available on the What Works Clearinghouse website for download. So with that introduction, let's get started.

I'd like to introduce Joy Lesnick, Associate Commissioner, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Joy, you now have the floor.

Great.

Thank you all for joining us. I'd like to welcome you to this What Works Clearinghouse webinar on Designing Quasi-Experimental Studies. We're pleased to have a growing set of archived webinars and other resources that we hope will be helpful to you in designing, carrying out, analyzing, and interpreting your results, and the results of other studies in education. The idea for this webinar came from some of you and aims to respond to many questions we have received through the What Works Clearinghouse Help Desk about our standards and procedures for reviewing studies, specifically those that do not use random assignment. And we hope that the information and additional resources that the presenters talk about today will help you in your own work, and we welcome your ideas and feedback for other ways we can better support your use of What Works Clearinghouse information.
And with that I'll turn it over to Neil and Jean.

Thanks, Joy.
And thank you all for joining us today. I'm Neil Seftor, a Senior Researcher in Mathematica Policy Research and the Project Director of Mathematica's work on WWC. My co-presenter, Jean Knab is also a Senior Researcher at Mathematica, who has overseen WWC topic areas and training for several years.

Before we get started, just a couple of quick reminders. We encourage you to submit questions whenever they come to mind using the Q&A tool on the webinar software on your screen. We've also made available some resources for you under the Resources widget, which is a green folder icon, to use during or after the webinar.

For about the next half hour, Jean and I have some prepared slides that we're going to talk through. First, I will provide an overview of the WWC and quasi-experimental designs. Next, Jean will go into some specifics of key features of quasi-experimental designs. Then I will summarize the key points, along with providing a set of resources you can use and ways you can stay connected to the WWC. And finally we will turn to your questions.

I'd like to start by briefly describing who we are, what we do and don't do, and why we exist.
By we and the WWC, I mean the Institute of Education Sciences, Mathematica, and our other partners who make up the WWC as a whole.

As one of IES's first investments, the WWC was established in 2002 to be a central and trusted source of scientific evidence for what works in education. The Clearinghouse aims to identify all relevant research studies on the topic, review those studies against design standards, and then synthesize findings from the high-quality research.

It's also important to clarify what the WWC does not do. We don't directly test or study interventions. We don't commission new research on interventions. And we don't recommend or approve specific interventions. We let decision makers use the information we have generated to decide for themselves.

The WWC's goal is to help decision makers efficiently make evidence-based decisions, informed by the most rigorous research. Therefore, the core of our work focuses on the assessment of individual studies against design standards.

As you probably know, not all research can reliably isolate the effect of an intervention. The WWC Standards are a set of criteria developed by panels of methodological experts to assess research quality. In particular, the standards focus on the causal validity of the study design and analysis. These standards allow us to be confident that any improvement in student outcomes is due to the intervention being studied and not some other difference between the treatment and control of districts, schools, teachers, or students. The standards are applied by trained and certified reviewers to each study, and ultimately are used to assign one of three ratings.

The highest rating, meets WWC group design standards without reservations, is reserved for studies with the highest degree of causal validity, such as randomized control trials with low attrition. The next rating, meets WWC group design standards with reservations, applies to studies that provide causal evidence but with a lower degree of confidence, such as a quasi-experimental design that has shown that the study groups are similar. The final rating, does not meet WWC group design standards, is given to studies that have not demonstrated a causal link between the intervention and the findings. More specifically, the group design standards focus on randomized control trials and quasi-experimental designs.

Before we talk about high quality quasi-experimental designs, it is useful to look at some differences in these two designs. Randomized control trials, or RCTs, are considered the gold standard in evaluation research. That's because when participants are randomly assigned to the treatment or comparison group, they should be similar on both observable and unobservable characteristics. In that case, as the only difference between the groups is the exposure to the intervention, any differences in outcomes can be attributed to the intervention. This type of study can receive the highest rating, meets WWC group design standards without reservations, since we have the highest level of confidence that the findings are caused by the intervention.

Quasi-experimental designs, or QEDs, are often used when random assignment is not feasible. Again, some participants receive the treatment, and some do not, forming a comparison group. Because they aren't assigned randomly, we can, at best, only show that the groups are similar on observable characteristics. There is no way that we can show the groups are similar on unobservable characteristics, such as motivation or interest. So the groups might differ due to something other than just the intervention, and that difference might affect the outcomes. Therefore the effect we might see might not only be due to the intervention, so the study can't receive the highest rating. However, a high-quality QED can receive the next rating, meets group design standards with reservations.

So what do we mean by the high-quality QED? While there are obviously a lot of things to think about when designing a study, we'd like to focus on four critical ones today. Making appropriate choices about these factors makes it much more likely that your study will be able to meet standards. This figure provides an example of a high-quality QED, and can be used to illustrate the four factors. First, there are

two groups. A group of third graders on the left, who will receive the reading supplement. And a group of third graders on the right, who will not. Second, there is a pretest given to both groups at the same time, the fall of 2014, prior to the intervention, which is used to demonstrate baseline equivalence. Third, the only difference in what happens to the two groups is the use of the reading supplement, and there aren't any other confounding factors. Finally, the study uses eligible outcomes after the intervention, spring 2015, to measure reading outcomes for both groups at the end of the year.
With that, I'll now turn the presentation over to Jean, who will go into more detail about each of these critical factors.

Thanks, Neil. And just a reminder, if you have a question at any time, you can submit that through the Q&A box.

As Neil mentioned, I'm going to talk about four key factors to keep in mind as you're designing your QED and analyzing your data.

The first factor I'll discuss is distinct groups. For a QED to meet standards, it needs to have at least two groups, one group that receives the intervention and another that serves as a comparison group. It's important for a QED to have a comparison group so that we can see what would have happened without the intervention. The groups that are compared must be distinct. And by distinct I mean that you cannot simply compare a treatment group to itself at an earlier point in time. Those simple pre-post, or before-and-after, studies cannot meet group design standards.

For example, a study might look at a change in test scores from fall to spring among students who received a new reading curriculum during the academic year and find large gains in reading. But we know that students generally improve in reading during the course of the school year. So without a comparison group that didn't receive the new curriculum, we don't know how much of the students' improvement is due to the new curriculum versus natural gains in reading skills. Therefore, studies without a distinct comparison group cannot meet group design standards.

Other than having distinct groups, there are no additional group formation requirements for QEDs. Group formation can be accomplished through various, non-random processes. Groups could be created largely for convenience, for instance, choosing comparison group schools that are near your intervention schools, or by comparing schools that volunteer to implement a new curriculum to schools that did not. Researchers may use propensity score methods to create a comparison group, matching intervention and comparison use on characteristics that are likely to be related to the outcome of interest. Finally, and this is a question that we get fairly often, researchers can use retrospective data, such as using a nationally-representative data set to compare the outcomes of youth who went to full day versus half-day kindergarten. So while all these methods of group formation may be allowable under the WWC standards, some are more likely than others to meet standards due to the next factor we'll discuss, which is establishing baseline equivalence of the groups.

To ensure baseline equivalence of the intervention and comparison groups, we strongly suggest using data to inform your group assignment. That will increase the likelihood your QED will meet group design standards.

So why do we care about establishing baseline equivalence? The key reason we care about establishing baseline equivalence is that if the intervention and comparison groups are different at the start of the study, those differences could be carried through to the outcomes and be mistaken for an effect of the intervention. For instance, if the intervention group's scores are substantially higher than the comparison group's scores on the pretest, and the intervention group also has higher test scores than the comparison group on the post-test, it's impossible to know whether the high spring test scores are the result of the intervention or the preexisting differences in the groups. It's possible that the intervention group is simply better at reading, regardless of the curriculum used. Therefore, the WWC requires QEDs to demonstrate that the groups were similar prior to the intervention, or at baseline, on key characteristics that are likely to be correlated with the outcomes. When groups are similar on

observed characteristics, like age and test scores, prior to the intervention, we can be more confident that any observed effect was caused by the intervention.

The specific characteristics that must be equated, and the rules for demonstrating equivalence, vary by review topic and are documented in the review protocols. In most cases, the WWC looks at equivalence on a pretest measure of the outcome, so you should collect pretest data whenever possible. However, some topic areas require equating on more than just a pretest. For topic areas that focus on specific subpopulations, such as youth with emotional disturbance, a study must also establish that the intervention and comparison groups are similar on behavior at baseline regardless of the outcomes that are examined. Finally, some topic areas, such as the dropout and post-secondary areas, focus on areas for which there are no natural pretests or prior measures. Their outcomes are measures like dropped out of school, or credits attained in college. In those cases, the protocols specify the proxy measures that must be equivalent, such as socioeconomic status. We suggest you review relevant protocols when you're designing your study to ensure you are collecting the data necessary to demonstrate the equivalence of the groups.

To determine whether the groups are equivalent at baseline, the WWC looks at the size of the difference between the groups on each key measure of interest. The difference is measured in effects-size units, and you can see the WWC Handbook for the specific calculation that we use. Using an effect size, rather than statistical significance, ensures a consistent metric to assess equivalence regardless of the study's ability to detect differences between the groups, which can vary with the sample size.

The chart on the bottom of this slide explains the two ways to demonstrate equivalence of the groups. First, groups are equivalent if there are very small differences between them. If there are effect-size differences of less than or equal to 0.05 standard deviations on a baseline measure, then groups are considered equivalent on that measure. You've demonstrated that the groups have nearly equivalent means before the intervention, so we can be confident that there weren't preexisting differences that carried through to the impact of the intervention. The second way to demonstrate equivalence is to have moderately sized differences, defined as a difference greater than 0.05 and less than or equal to 0.25 standard deviations on a baseline measure but statistically control for that difference in the analysis of outcomes. Specifically, when you are analyzing the outcome that has the baseline difference on the pretest, you need to use a regression model, such as an OLS, ANCOVA, or HLM that includes the baseline measure as a covariant. Oftentimes we see analyses that simply subtract the baseline mean from the outcome mean, for example using a gain score or a change score. But those methods are not sufficient for meeting WWC standards because they don't statistically adjust for the baseline level. So please be sure you use an acceptable method of statistical adjustment for your analyses.

Finally, and this is very important, if there are differences between the groups that are greater than 0.25 standard deviations, baseline equivalence cannot be established for that measure. More importantly, if there are differences greater than 0.25 standard deviations on a measure, the analyses of that outcome and all related measures cannot meet standards. The groups were shown to be too different prior to the intervention that even if we control for the differences, we can't be confident an observed effect on that measure, or any measures we expect to be highly correlated with that measure, was due to the intervention. See the relevant review protocols for how outcomes are classified within domains and how equivalence is assessed within those domains for more information on this because it's specific to the topic areas.

Another reason a study might not meet standards is that we did not have enough information to assess baseline equivalence. It's important that you provide enough information for the WWC to calculate the effect sizes and determine what co-variants are in the models estimating the outcomes. Not having that information is a common reason studies don't meet standards, particularly older studies. We will reach out to authors for that information, however, we don't always receive it. We strongly encourage you to review the WWC Reporting Guide for Study Authors for guidance on writing up your results. In addition, we also recommend you assess baseline equivalents for your own QEDs before finalizing your analysis. If you find that your groups are not equivalent, you could potentially derive a comparable sample using a subset of your data and matching techniques, such as propensity scores, for exact matching.

Next we want to make sure that there isn't a component or factor of the study that may influence the outcome and is completely aligned with only one study condition for a group. If there is a factor that is present in one condition, but not the other, it is impossible to separate the effect of the intervention from the effect of the confounding factor. Therefore we cannot attribute the observed impact solely to the intervention, and the study does not meet standards. The easiest way to understand confounds is to talk through some of them.

The most common confound we see is that there is only one unit in one or both conditions. For instance, a new behavioral intervention is implemented in third grade in one school, and outcomes are compared to the third grade in the nearest school. Unfortunately in this case we cannot isolate the effect of the intervention because the intervention and comparison schools may differ in ways related to behavioral outcomes. For instance, one school may have a school-wide positive behavior program or more teachers' aides, which could be influencing behavioral outcomes. Therefore, to decrease the likelihood that there is some factor confounding the results, both the intervention and comparison groups must contain at least two units, or schools in this particular example.

Another confounding factor we also see is that there is only one person providing the intervention, and that person only interacts with the intervention group. For instance, a single math teacher pulls students out of math class once per week for intensive support, and the comparison group remains in math class. In this case the math specialist may be an incredibly dynamic teacher, and her personality may affect student outcomes as much as the particular curriculum or strategy she is implementing. Therefore, this study cannot meet standards as a test of the curriculum or strategy because we cannot separate out how much of the observed effect was due to the intervention and how much was due to the particular math specialist. Again, to decrease the likelihood of the confounding factor influencing your results, you would need to have at least two interventionists or have an intervention in which the teacher is in both conditions to control for the teacher characteristics.

The third type of confound we often see is that the intervention of interest is bundled with some other services that are not intended to be part of the assessed intervention. For instance, in a study designed to test teacher professional development, a new curriculum is also provided to the intervention group. In this case, we cannot isolate the effect of the professional development because the curriculum changed at the same time. The WWC would treat this particular study as a study of the combined teacher professional development and curriculum, but the study could not be used to assess the effectiveness of teacher professional development alone due to the confounding factor.

Finally, the last confound to avoid is a history confound. A history confound occurs when the pre-tests and post-tests are not assessed that the same time for both groups, such as comparing the change in fall-to-spring test scores among this year's third graders to last year's third graders. When pre-tests and post-tests are not assessed at the same time for both groups, there could be something occurring at the same time as the intervention, such as a change in curriculum or policies, which could be causing the observed outcomes. Therefore, studies must collect data at the same time for both groups in order to meet standards.

The last aspect of QEDs I want to talk about is outcome measures. Outcomes are used to estimate the impact of the intervention, therefore the WWC wants to ensure that the outcome measures meet some basic criteria. First and outcome measure must have face validity. That is, it measures what it says it measures. Very few outcomes don't meet this bar. Second, an outcome must have sufficient reliability, which ensures it measures concepts consistently over time and across people. The specific reliability requirements vary across protocols and measures. For instance, there are typically higher reliability requirements for measures derived from observational data. So please become familiar with the review protocols in the areas you conduct research. Third, an outcome must not be over-aligned with the intervention.

Overalignment is a concept you may be less familiar with. Overalignment occurs when outcome measures are more closely aligned with one of the research groups, typically the intervention group,

than the other. When outcome measures are closely aligned with or tailored to the intervention, they may give an advantage to the intervention group. For example, an outcome measure derived from reviewing specific reading passages that were used as part of the intervention but not in the comparison classrooms would likely be considered overaligned. This type of outcome measure may be useful to a researcher as an interim outcome, but the broader question of interest to the WWC is not whether they could read that particular passage well, but whether they had improved their ability to read new passages. The decision about whether a measure is overaligned is specific to particular topic areas and is made by the review team leadership.

Just as outcome measurement should not be tailored toward one of the groups, data should be collected similarly for both groups. For instance, data that is collected using surveys for one group and administrative data for another would not be allowable. Similarly, using teacher-reported observations for the intervention group and researcher observations for the comparison group would not be allowable either.

Finally, the WWC does not allow imputed data for quasi-experimental design studies. All outcome data must be observed, and all youth with outcome data must have baseline data in order to establish baseline equivalence. So all QED analyses should be complete case analyses if you wish to meet WWC group design standards. So that summarizes the four key factors we assess for QEDs, and now I will send it back to Neil.

Thanks, Jean.

Before we move on, I'd just like to reinforce the key points to take away, to keep in mind, about high-quality QEDs. First, start with at least two well-matched groups with data collected before and after the intervention. Assess baseline equivalence before estimating impacts. And adjust for differences in baseline measures if necessary. Try to avoid confounding factors such as single units or bundled interventions. Use valid, reliable measures that are not over-aligned with the intervention. Clearly document your design, data collection and analytic procedures. And use WWC resources to help with measure selection, analysis and reporting.

Here are just a few of the resources available from the WWC to help you in designing and reporting on your study. Our Handbook provides a comprehensive description of all of our procedures and standards so that anyone can understand how we work and how decisions are made. The Reporting Guide for Study Authors provides some guidance on the types of information we need to be able to assess studies against standards. The Study Review Guide is the tool we use when assessing studies. It is available to the public, as well as the instructions for its use. It should be noted, however, that it is designed to be used by trained and certified reviewers. All of our review protocols are posted on the web to facilitate transparency and help when you're designing studies. And finally, if you're looking for whether we've reviewed a study or how it was rated, we provide a database of reviewed studies containing the more than 10,000 studies that have been identified for one of our reviews.

At the top of the WWC Home page, is a set of rotating images that highlight recent products or resources around a theme. These images are regularly updated. For example, we will soon have a research bulletin that will provide readers with links to new products highlighting our review process and standards. You can receive regular updates on new WWC products through social media by friending us on Facebook or following us on Twitter. You can also sign up to receive email news flashes. Click on the News & Events and select the news flash. When you enter your email address, you can click the plus next to NCEE and check the What Works Clearinghouse. Finally, you can also help us improve the Clearinghouse. If you have a suggestion or question, the best way to reach us is through our Help Desk. You can get to the Help Desk through our Contact Us page, which can be found under About Us.

And with that, we'll move on to your questions. For those of you who missed the beginning, if you want to submit questions, use the Q&A tool on your webinar software. To help us get through the questions, I'd like to introduce Russ Cole, a Senior Researcher at Mathematica, who currently oversees work on our review-based products.

Russ?

Thank you, Neil. We've had a number of questions that have come in already, and I'm going to try to sort those questions to Jean and Neil in turns.

So Jean, here's the first question for you. Can we use extant data to identify comparison groups for quasi-experiments?

Thanks, Russ. Yes, you can use extant data to identify a comparison group. The key is to come back to these basic principles. Can you establish baseline equivalence for these groups? And can you ensure that there are no confounds? We wouldn't want all of the data in one case to come from one school district and one for another.

Thank you, Jean.

So the next question that we received, we've actually had a number of questions about where the Handbook and review protocols are located. In addition we had a question about where we could see a list of To Dos that would help to prepare a successful submission to the Clearinghouse. So Neil, could you provide some guidance for where those resources can be obtained?

Sure. We provide the Handbook and the protocols in a number of places. If you go to the WWC Home page, you can easily get to them through the Publications and Reviews tab, if you look under that menu, under Reference Resources are links to the Handbook and the protocols. You can also find that information in the inside the WWC menu. There are links to the review protocols, the Handbook, and the Reporting Guide for Study Authors. That's the guide you can use that will give you the kinds of information that we need to have in order to assess your study. Important for you to provide, as Jean mentioned, is enough information that we can make sure that your groups are similar, that there is baseline equivalence. So it's very important to provide sample sizes, means, standard deviations, both before and after your intervention, so we can assess equivalence and look at the findings.

Thank you, Neil.

So now we have a question for Jean. Jean, what about regression discontinuity designs? Can these designs be reviewed by the What Work Clearinghouse as quasi-experiments?

So regression discontinuity studies are actually reviewed under a separate set of standards. There is a set of pilot standards that are also available on the What Works Clearinghouse website, and there's an alternate set of criterion that they need to meet as well, which was beyond the scope of this webinar. And because at this point we see so few studies, we would treat that as a separate design and discuss in a separate webinar.

That's great.

Neil, do you want to follow up with any of the key criterion?

I think that that's probably beyond this webinar, but our standards are available on the web. And in the future we may have a training for the RDD standards.

Great.

So this question is for Neil. So Neil, can you talk about the control condition as it relates to business as usual? Does the What Works Clearinghouse want us to control this condition in more ways than what is typically done in instruction?

That's a good question. There are many times in education research when we're looking at a particular intervention and all we can do is compare to what is normally going on in the schools. For example, in

reading and math, it's nearly impossible to look at the effects of an intervention compared to nothing. Most students are involved in some sort of reading or math curriculum. So the best we can do is – the best researchers can do is clearly document what's going on in the comparison condition. And we can try to provide that contextual information along with the findings in our reports. We describe each study, including the setting and the sample, of every study that meets standards in our reports so that people can try to understand what particular comparison was actually made and what the intervention was being compared to. Beyond that, business as usual or the standard practice is what it is. We don't expect researchers to try to go and change what that condition is.

Great.

So this next question is for Jean. Jean, is there any way to conduct a quasi-experimental design without a control group? For example, if I monitor students' growth over the years but I do not have a control group, can this type of design be reviewed as a quasi-experiment?

So that type of design cannot be reviewed as a quasi-experiment. That is essentially a pre-post study, in which case you have multiple pre-post studies it sounds like. You could potentially find an extant data source, as someone asked before, and try to find a group of youths that align in terms of baseline equivalence with the youth that you have at the start of the time point and then look at the change in their growth and see – use that as your benchmark. But right now, without a control group, that design could not meet standards. Some of you may have heard we also have another set of standards that are single case design standards which do use typically only one group and compare it to itself, but those have very specific requirements about repeated measurement and active manipulation of an intervention, so I don't think that type of design could meet under the single case design standards either.

Great.

So Neil, we've had a number of questions about confounding factors. How can you confirm in your study whether it has a confounding factor or not? Isn't having the single treatment and a single control group and N equals one confounding factor?

Hmm, interesting question. So having a group prevents the N equals one confound. What we're really worried about is something happens in one condition that's not at all observed in the other condition. So Jean gave a couple of examples. If all of the students receiving an intervention are in one school, and there are ten comparison schools, then we don't know whether an observed effect is due to the new treatment that was given in that one school or the fact that a new principal came in and changed everything around. So we really want to separate other things that could affect the outcome from the intervention that we're looking at. So you can have groups of students who are within the same teacher, within the same school. As long as the factor is in both groups, we're not worried about it in the analysis. It kind of gets eliminated. The real concern is if you have something going on that is only happening in one group, and so we really can't separate the effect of the intervention, which is what we want to look at, from this other factor that we don't.

Great.

So this next question is for Jean. Jean, can we establish that two groups are similar at baseline via surveys? So are surveys considered an observable characteristic on which we can establish baseline equivalence?

Yes, you can absolutely use surveys. By observed characteristics, really we mean measured characteristics, not characteristics that, you know, that maybe we can't observe like somebody's motivation or disposition. Basically we'll be looking at observed measured characteristics that are gotten through administrative data, survey data, or perhaps teacher or researcher classroom observations or something like that. But yes, absolutely, survey data is fine.

Great.

So now this next question is for Neil. So Neil, can dose response studies qualify as an acceptable quasi-experiment? For example, students who used a lot of something versus students who received very little of something, would that be considered a potential valid quasi-experiment under the Clearinghouse?

So no, that type of design wouldn't be eligible for review by the Clearinghouse. We are focused on studies where one group receives an intervention or a treatment and the other does not. In that type of study, we couldn't examine whether the treatment was having an effect. In that case you're really examining whether some amount of a treatment is more effective than another amount of a treatment, and we don't make those kinds of comparisons within a treatment. We're only comparing the effect of an intervention versus not having that intervention.

Great.

So Jean, here we've got a question about types of baseline equivalence. So can we use a state math test to establish baseline equivalence of groups that are receiving or not receiving a new eighth grade science curriculum on a high school biology assessment? So that's the first question. And then a follow-up question was, when should baseline equivalence be established?

Sure. So I'll take a stab at this one and Neil can also respond. So science is one of those subjects that you don't necessarily have as regular testing as in other subjects. So I believe, because math tests are so highly correlated with science that the science protocol allows math as a proxy for establishing baseline equivalence for science. Again, recognizing that you may not have a fall assessment just prior to the eighth grade science curriculum being implemented, you know, you could potentially use the spring seventh grade test to establish equivalence of your sample. Neil, is there anything you wanted to add to that one?

You're right, Jean. The science area review protocol says that when a pretest of a science achievement outcome measure is not available, any one of mathematics, reading achievement or literacy outcome measures can be used instead. And you're right, you would want to – the ideal time to establish equivalence is immediately before the intervention you're studying. If it's much earlier, there might be concerns that other things happened between the two groups that would be captured by your study.

So the next question that we have is for Neil. So Neil, are the QED standards limited to student-level interventions and student-level outcomes, or are school-level interventions and school-level outcomes also eligible? This person wanted to indicate that they don't have student-level data, they only have school-level data.

That's an excellent question. These standards apply actually to all levels. While much of the research we review is at the student level, there are interventions that are conducted at the teacher level or the school level, and these standards apply to those as well. You would want to establish that your groups, whether they are groups of teachers who receive a certain type of professional development, or groups of schools that are implementing some sort of school-wide program, you'd want to establish that the schools that were in your treatment group and the schools that were in your comparison group were similar before the intervention using that school-level data.

Great. Thank you, Neil.

So Jean, this next question is for you. Jean, can you talk a little bit more about statistical adjustments? So what might this look like when you're a study that requires a statistical adjustment in order to meet evidence standards?

Sure. So sometimes when people are estimating impacts, they simply just look at the unadjusted mean differences of the groups, so what is the average outcome score for the treatment group and what is the average outcome score for the comparison group, and then they subtract those two. That's okay if there

are very small differences as we discussed at the beginning prior to the intervention. If statistical adjustment is required, simply doing those mean differences is not sufficient. What you need to do is you need to calculate adjusted mean differences using a regression framework. So you might have an OLS model where your outcome is on the left-hand side, and then you have an indicator of the treatment, and then also the pre-test. So by doing that you're controlling for where each of the youths started from, their initial test score, when you're estimating the difference between two groups at the end. Because you know they started in different places, and so you want to account for that when you're estimating the impacts.

Great.

So we have a more general question about the Clearinghouse more generally than about quasi-experimental standards specifically, and this one is for Neil. So Neil, how does the Clearinghouse decide which studies to review?

That's a good question. So there are several different ways that we identify studies, and those ways depend on the type of product that we're talking about.

For our main products, our intervention reports, we identify interventions that we want to conduct a review on, and then we do a systematic, comprehensive review of all of the literature we can find on that intervention. So once an intervention is identified to be reviewed for a report, we try to find everything there is. Published or not, as long as it's publicly available. And then we review every single study of that intervention.

Practice guides have a slightly different approach in that they're focused on specific recommendations. A similar process occurs in which the panel decides the kind of recommendation they think is appropriate, and then we, again, try to go and find any piece of research that applies to that recommendation and we'll review it.

The two other products we have are quick reviews and single study reviews, and these are both reviews of individual studies. So quick reviews are triggered by some media attention that's given to a particular education study. Once there's enough media attention, the Clearinghouse will review that study so that we can provide our assessment of the quality of the research. So we monitor a variety of media sources to identify research that's being discussed by the media. For our single study reviews, those come from several sources. Many of them are quick reviews that we've done and then we do a more complete review in our single study review. But more recently we've been reviewing studies for other purposes, including IES-funded studies and studies that are submitted in support of grant applications. For both of those we receive lists of studies to review, and we review all of them and include them in our studies database. And sometimes when the research is important or interesting, we will do a single study review on those as well. So we have a variety of ways that studies are identified for our review.

Thank you.

So this next question is for Jean. Jean, when the intervention is a curriculum, the teacher is going to be delivering the intervention. So would the same teacher need to be using the treatment curricula and the comparison curricula, or is it sufficient if there's more than one teacher in each group? So essentially the question here is how can one get around the teacher delivery confound?

Um hmm. So they can do it either way. One way would be, as you mentioned, the teacher could provide both the intervention curriculum and also teach in the comparison condition, and that way you're controlling for the teacher effects. But no, absolutely, you can also just have more than one teacher delivering the intervention, and that would get around the confound. We recognize that there are going to be teacher effects embedded in what we're looking at. The rules that we set for confounds are to say in this case we really can't separate out the effects of this particular teacher from this curriculum versus sort of average teacher effects that we're a little more comfortable with as the sample size grows.

Great.

So this is a question for Neil about a slightly different topic. So Neil, will the SCD standards become official standards soon or as they going to remain as pilot standards?

Sure. We've had the standards for RCTs and QEDs have been around since the beginning of the Clearinghouse and have gone through some revisions over time. But more recently we had groups of methodologists create standards for assessing the quality of both regression discontinuity designs, as Jean talked about, and also single case designs. And both of those were pilot because they were very – the review of that kind of research was very new. What we've been using the single case design standards for reviewing studies for a while now, we've identified a number of things that we would like to change in them, some clarifications, some additions. That's why they're still pilot. Additionally, we're still figuring out the best way to report on single case design research. We have individual single case design studies that meet standards, but we have yet to have a full body of evidence for single case design that provides enough evidence that we would use it for rating an intervention. So there are still issues that we need to consider about the presentation of SCDs, and until we go through those and iron out some of the details, I believe the single case design standards will remain pilot.

Great.

This question is for Jean. So Jean, does a quasi-experimental design need to show baseline equivalence by subgroups if the impact analysis looks at subgroups?

Yes. So each WWC protocol talks about the subgroups that they will look at and report on in an intervention report appendix. And any of the analyses that we would present in a WWC report, we would assess equivalence separately for those subgroups. So if there was an analysis by gender, we would look to see the treatment and control groups within males looked similar before the intervention and the same for females. So, yes, we would encourage you to provide tables and data that would allow us to make those assessments for subgroups.

Great.

This next question is for Neil. So Neil, how can you determine how many and which factors need to be statistically equivalent prior to the intervention?

So for the specific factors and how many, you would want to refer to the review protocol for the area. As Jean mentioned, most of our review areas use some sort of pre-intervention achievement test that is related to the outcome to establish equivalence. There are some areas where those kinds of measures don't exist, so there are lists of variables that need to be established, you need to have equivalence established. But one clarification I want to make is that they don't have to be statistically equivalent. So we don't look at the statistical significance of the difference between the groups because that's heavily affected by the sample size, so, as Jean pointed out, we really want to look at the magnitude of the difference between the two groups on those key characteristics.

That's great. Thank you.

And I have a follow-up question that's somewhat similar, and this one is for Jean. Jean, when you assess baseline equivalence between participant and comparison groups, if you have a natural baseline variable, for example, a baseline math score when you're evaluating a program that's looking at a math intervention, and we see that the mean scores are equivalent, can we ignore other variables like demographics?

So, again, each protocol describes the variables that they will require in order for establishing equivalence, and some topic areas will require equivalence again because there may not be sort of a natural pretest. But even in the ones where there is a natural pretest, the protocols – it is within the review team leadership description to look at all of the demographic variables and see if they think that

the groups still look equivalent even across demographic variables. Sometimes you may have a pretest variable that looks, you know, somewhat equivalent, but depending on where it is in that scale of effect size, if it's toward the bottom or if it's close to not being equivalent, and then if you saw differences in demographic characteristics, that may cause them to deem the groups as not equivalent. So we think that you should look at those, and you should present those, and control for any of those that appear to be somewhat different in your analyses.

Great.

So we had a general question that came in about can we just offer a little bit more clarification about confounding factors. And so, Neil, would it be possible for you to just reiterate some of the thinking about how the Clearinghouse thinks about confounding factors?

Sure. It is a little bit of a confusing concept in that we're really trying to isolate the effects of the intervention. So we want to make sure that there's nothing else that could possibly have caused the impact that we saw. We've mentioned cases – the frequent cases we get are when there's one unit, such as one teacher, or one school in one of the conditions. So in those situations you can imagine a new reading program was given by one teacher, and the regular reading program was given by four other teachers. We can't say whether – if we saw an effect for the students who got the new reading program, we wouldn't be able to separate whether it was due to the new reading program or the very dynamic and engaging teacher. When there's more than one, we have some averaging of the characteristics. So if there were two intervention teachers and four comparison teachers, that would be okay. There wouldn't be just one teacher that was associated with the intervention and wouldn't be able to separate the effects. So more than one of the units deals with the confounding problem.

And that applies to both teachers and schools. If there's just one school giving an intervention, that can be problematic. We can't separate the effects of the intervention and the school. But if there's two, or three, or more schools, we're less worried that there's some factor just in that group that is aligned with the intervention and is causing all of the effects that we see.

That's great. Thank you.

So we're nearing the end of our time here. I've got one more question that I'm just going to ask to Jean. Jean, there's a question, can we impute outcomes or co-variants in a quasi-experimental design?

So at this time under the standards, QEDs cannot have any imputation. You cannot impute outcomes, and you cannot impute baseline data that is used as a co-variant. So no imputations and no imputation in QEDs.

Thanks, Jean.

So I think that will end the period of time that we've got for Q&A, and I'll turn this back over.

This concludes the webcast for today. Please submit feedback to our presentation team in your browser window when the event concludes. If you are unable to provide your feedback at this time, you can do the on-demand recording of the event and access the survey widget there. The on-demand recording will be available approximately one day after the webcast and can be accessed using the same audience link that was sent to you earlier. Alternatively you can submit feedback through the Contact Us form on our website, whatworks.ed.gov. Thanks and have a great day.