WWC Recertification for Standards Version 4.0
Office Hour #1 Transcript
January 22, 2018

Hello everyone and thank you for attending today's webinar: What Works Clearinghouse Recertification Training Office Hour #1. The webinar will begin with an introduction from Neil Seftor, Mathematica's Project Director for the What Works Clearinghouse. After that, the recertification team will respond to your questions about the recertification process and the new version 4.0 group design standards. I will be briefly going through housekeeping information before we get started. You can make the slides larger on your screen by clicking the bottom right corner of the slide window and dragging out. If you have accessed the audio for the webinar through the teleconference line, you may experience a slight delay. If possible, we encourage you to listen to the webinar through your computer or device speakers. We encourage you also to submit questions throughout the webinar using the Q&A tool on the webinar software on your screen. You can ask questions at any time throughout the webcast. Because we're recording this, 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 will try to answer as many questions as possible. A transcript of the webinar will be available on the WWC website for download. With that introduction, let's get started.

Neil you now have the floor.

Thanks, Brice. And thank you all for joining us today. As Brice said, today is the first of two office hours we are providing, following up the webinar on Friday, January 12th, that covered the major changes in standards of procedures as we move from version 3 to version 4 of the What Works Clearinghouse procedures and standards. The handbooks are online and the webinar can be accessed through the On24 link that you used to attend the webinar, and the test is already available so you can take it at any time. But today we'll be answering questions that were submitted during the webinar and since the webinar, and if you have questions that you want to submit today we'll try to address those too. Why don't we go ahead and get started.

We thought we would start with questions related to the cluster standards since that was not covered as part of the webinar but was covered in a separate video module. We will start with those and Dana Rotz will be providing most of our answers, but she will be helped along by Allison McKie and Elias Walsh.

The first cluster study question is: what denominator should be used to assess whether the analytic sample of individuals is representative of clusters in step five of the review process for cluster assignment studies?

The goal of Step 5 of the review process for cluster design studies is to assess whether the data used to estimate effects are representative of the clusters from which the data are drawn.

The denominator for the calculation should be all individuals present in the clusters that contribute data to the analytic sample around the time the outcome data was collected.

The numerator should be the number of individuals comprising the analytic sample. If cluster-level data is used to estimate impacts, the numerator should be the number of individuals contributing data to the cluster-level statistics.

Suppose we had 10 classrooms, each with 10 students usually enrolled in them. And we had outcome data for 90 of them. Then we would assess the overall representativeness in step 5 using the fraction 90 over 100. To compare this to the attrition thresholds, you would want to take 1 minus 90/100 to get 10 percent (because the attrition thresholds reference the sample lost, as opposed to the sample which is represented).

The next question is: how would the WWC classify this design? Classrooms in a school district are randomly assigned to be part of a study. Data are collected only for students in selected classrooms. Within those selected classrooms, students are assigned to treatment or control status.

In classifying the design, we focus on how sample members were assigned to intervention and comparison conditions.

The question first indicates that classrooms were randomly assigned to be a part of the study. That tells us something about how the study sample was selected – classrooms were randomly selected to participate – but it doesn't tell us anything about how sample members were assigned to the intervention and comparison groups, so it does not affect how we classify the design.

In this question, the key is how students in the selected classrooms were assigned to conditions, which the question did not specify. If students were randomly assigned to intervention and comparison conditions, then this study would be an individual-level randomized controlled trial. Otherwise, the study would be an individual-level quasi-experimental design, provided that there is no overlap between the intervention and comparison groups.

Also, please note that the WWC prefers the use of the language of "intervention" instead of "treatment" and "comparison" instead of "control."

The next question is: would an RCT with assignment at the individual level, but in which the intervention is delivered in groups, be considered a cluster RCT or an individual-level RCT.

As we talked about in the previous question, what matters is how sample members were assigned to the different study groups, being the intervention group and the comparison group. If students were randomly assigned as individuals, the study is an individual-level RCT. If students were randomly assigned as a group, the study is a cluster RCT.

Now a question about joiners: how should we calculate individual-level non-response when a cluster RCT has joiners?

There are a couple of issues here to bear in mind.

First, remember that we assess individual non-response only if any joiners included in the analytic sample do not provide a risk of bias. If, say, there were late joiners that were included in the analytic sample and late joiners pose a risk of bias, then we skip the step where we assess individual non-response and jump to the next step of the review process, which involves assessing whether the study establishes baseline equivalence for the groups in the analytic sample. Looking at non-response is really only relevant when joiners don't pose a risk of bias.

Let's suppose that any joiners that are included in the analytic sample do not pose a risk of bias. Then what we have to remember is that a correctly performed individual non-response calculation will always have a numerator that is a strict subset of the denominator. The numerator has to be smaller than the denominator. The numerator is the number of individuals included in the analytic sample. When there are joiners included in the analytic sample, the reference sample used as the denominator must be taken at a time that includes all those joiners.

For example, suppose the analytic sample includes early joiners, and the review protocol indicates that only late joiners pose a risk of bias – so those early joiners are okay. The reference sample must be taken at a time after all of the early joiners had entered the clusters. So, what we want to look for is essentially the earliest reference sample that fits that criteria. If more than one reference group is available that meets this requirement, we would calculate non-response using the reference sample from the earliest period.

The discussion of Step 3 in the Cluster-Level Assignment Module and in Section II.B of the Standards Handbook walks through how exactly the allowable reference samples for calculating individual non-response vary with the joiners associated with a risk of bias, which would be specified by the review protocol.

Thanks, Dana. Why don't we move now to some questions about missing data and how the WWC reviews studies that deal with missing data. The first question is: what do you do when there is a cluster RCT with missing data?

When we discussed missing data in the recertification webinar, we really stuck to examples with individual-level assignment studies. But the same principles can be applied to cluster-level assignment studies.

However, because neither the Standards Handbook nor the recertification webinar has missing data examples specific to cluster studies, we recommend that a reviewer who comes across a cluster study with missing or imputed data should consult review team leadership. The review team leadership can raise the issue with the STAT if necessary.

The next question is: do the missing data standards beyond step 2 only apply to studies using the same measure at baseline and outcome?

If you'll recall, step 1 of the review process essentially assessed whether the authors used an acceptable approach to handling missing data. And step 2 – I can pull this up for everyone to see.

[Reference Slide 33 from the January 12, 2018, recertification webinar]

Step 2 assessed whether the study is a low-attrition RCT. Step 3 looks at: does the study limit potential bias from imputed outcome data.

[Reference Slide 34 from the January 12, 2018, recertification webinar]

And step 4 looks at whether the study is a high attrition RCT that analyzes the full randomly assigned sample using imputed data, and step 5 calculates baseline equivalence. But none of these steps are actually restricted to situations where the baseline and the outcome measure are the same. All of the formulas that we use to assess potential bias or to assess baseline equivalence essentially include an adjustment for the correlation between the baseline and outcome measures. This allows us to use the formulas for any set of baseline and outcome measures, regardless of their correlation, so that the baseline and outcome measures do not have to be the same.

Thanks for that clarification. The next question is: what about using the dummy-variable method to handle missing data for characteristics that are not needed to assess baseline equivalence?

You can use the dummy variable method for a QED to impute missing data if the measure is not required to satisfy baseline equivalence.

The WWC considers the dummy variable method to be acceptable to account for missing data for QEDs, high and low attrition RCTs, or RCTs with compromised random assignment for variables in general. The restriction of the dummy-variable method to RCTs with either low or high attrition is only for measures required by the review protocol to satisfy baseline equivalence. So, if we saw a study that used the dummy-variable method to handle missing data, and, regardless of the design of that study, if the characteristic that the dummy-variable method was applied to was not required to satisfy baseline equivalence, we'd consider that an acceptable approach.

Right. Here is a situation that reviewers are likely to encounter: what should reviewers do if they need a correlation to review a study but the authors do not provide one?

If the information is needed in order to rate the study, the reviewers should conduct an author query, just like they do when other information which is needed to rate a study is missing.

Okay. The next question is: why do you say multiple imputation is a type of regression imputation? I don't think that's correct. Can you use multiple imputation with other imputation methods?

Right, that's a really good point. This is correct – at one point in the slide, we did say that multiple imputation was a form of regression imputation. In that particular case, it was just that multiple imputation was being used with regression imputation. Multiple imputation, the concept, can be used with other forms of imputation. We were implicitly assuming it was multiple imputation by regression in the case that we were looking at. There's nothing that necessarily links multiple imputation with regression imputation – it can be applied elsewhere.

But, there is an extra thing with regression imputation that's associated with multiple imputation, and that's that when authors account for missing outcome data using regression imputation, the WWC will only report the p-value associated with the estimate if their method of estimating the p-value, or the standard errors that are used to calculate the p-value, accounts for the missing data in some way. For example, using multiple imputation, but also potentially using methods like a bootstrap.

The next question is: does case-wise deletion count as attrition?

When an author of an RCT deletes cases because of missing outcome data, this sample loss is considered attrition, and we can just treat case-wise deletion, for this reason, as attrition.

Now, some specifics about some of the formulas. Do the formulas for baseline equivalence with missing data imply that high correlations between the pretest and the posttest make it easier to satisfy baseline equivalence?

What a higher correlation means is that a known outcome mean will probably be more informative about an unknown baseline mean. So when the correlation between the outcome and the baseline is high, the difference in outcome means between the analytic sample with observed baseline data and the full analytic sample will tell us more about the difference in baseline means for those samples. When the difference in outcome means is small, we can then be more confident that the corresponding difference in baseline means is also small, and that's essentially how the formulas are using the correlation.

Now I have a question about some of the notation in the handbook. In the Standards Handbook, in the Appendix C on missing data, what does the notation xbar_i~y mean?

Here we're getting a little bit into the weeds but let me just talk a little bit about the notation that we used for the handbooks that will hopefully help clarify some questions that we've had about it.

We're using x here to denote the baseline, and y here to denote the outcome, and then the tilde sign we're really thinking about as meaning "not". So tilde y, essentially, is being used to mean that y was not observed, or the outcome variable was not observed, which, in the case of the particular formulas that we're looking at, would mean that the observations for y were imputed as opposed to being observed. So xbar_i~y is the mean of the baseline for the intervention group, within the sample with observed baseline data but imputed outcome data.

Thanks. Another question from the handbook: Table II.6 of the Standards Handbook lists different requirements for approaches for handling missing data to use in order for the WWC to report the statistical significance of the estimated results. For example, if authors use regression imputation, the table indicates that they must use a method that reflects the missing information, such as a bootstrap method, or multiple imputation. What happens if these requirements are not met?

The adjustment of the standard errors, and going along with that, the p values, will not impact the rating received by the study. The p-values and the associated statistical significance are simply not reported in WWC products. So, those additional requirements are just requirements for the reporting of those p-values and significance levels. You should also note here that in complex cases where multiple methods are used to handle missing data (for example, if you used a combination of case-wise deletion and regression imputation), review team leadership should be consulted to determine whether p-values and significance can reasonably be reported by a WWC product, based on exactly what was imputed and what method was used in order to correct the standard errors for the missing data.

Thanks, Dana. Let's turn to some questions related to the WWC's new standards for baseline equivalence. One of the questions is: do parallel forms count as the same test?

We're going to talk about a couple of cases here, but essentially what it boils down to is that if the test has documentation that indicates that parallel forms can be directly compared, then these can be considered to be the same test for the sake of looking at whether baseline equivalence can be satisfied with a baseline difference in the adjustment region if the authors use a difference in difference approach, gain scores, or fixed effects. If you're uncertain about whether the parallel forms can be directly compared, you should consult with review team leadership, and the team's content expert may be asked to weigh in on a decision based on the exact nature of the test.

Okay. The next question is: Are standardized assessments that are conducted in different grade levels always considered to be different assessments, even if they are designed on a vertical scale or otherwise aligned?

Right. So, like I said before, these two questions are related, and the answer is similar to that of the question that Neil asked about parallel forms. If the test has documentation that indicates the vertical scaling allows direct comparisons between scores on tests in different grades, then we can consider these to be the same test. Again, if you're not sure, consult with review team leadership and the team's content expert will be asked to weigh in if needed.

The next question relates to slide 23 of the recertification webinar slides.

[Reference Slide 23 from the January 12, 2018, recertification webinar]

This slide indicates that difference-in-differences, gain scores, and fixed effects are additional adjustments that can be used to satisfy the baseline equivalence requirement when the baseline difference falls in the adjustment required range, the pretest and posttest measures are the same unit, and the correlation between the tests is 0.60 or higher.

Does "additional" mean difference-in-differences, gain scores, and fixed effects alone are sufficient for satisfying the baseline requirement, or do those adjustments have to be done in addition to regression covariate adjustment or ANCOVA"?

When the conditions are met – that is, the pretest and posttest measure are the same unit, and the correlation between the tests is 0.6 or higher, and baseline difference falls in the adjustment range—difference-in-differences, gain scores, or fixed effects alone are sufficient to satisfy the baseline equivalence requirement. You could also do a regression covariate adjustment or ANCOVA, but you could also simply do difference-in-differences, or gain scores, or fixed effects.

The next question has a slightly more complicated study. Suppose a study is an RCT with blocked random assignment, and the probabilities of assignment to condition varied across blocks. The analysis accounted for the different probabilities using an acceptable method. Should the baseline means be calculated using the same method?

This question involves two important issues. The first is accounting for different assignment probabilities and the second is assessing baseline equivalence.

Let's break this down. For clarity, let's remind ourselves of the key points about the first issue. A study that uses different probabilities for random assignment must account for those different probabilities in the analysis to preserve the integrity of the random assignment. The authors can do this in a couple of different ways. First, the authors could apply inverse probability weights, using weighting methods to re-balance the intervention and comparison groups. They could also include dummy variables that differentiate the blocks in their regression model. Or, they could estimate effects separately by block and then average the estimates across blocks to produce a single impact estimate.

Now let's think about baseline equivalence. If the study uses inverse probability weights in the impact analysis, the baseline means must also be calculated using the same weights. The same thing is true if separate estimates are calculated by strata, and then those estimates are combined to form a single impact estimate.

A little bit differently, if the study includes dummy variables for the different blocks in the impact analysis, the baseline means may also be adjusted using the same dummy variables, but the WWC will also allow unadjusted means to be used in order to assess baseline equivalence.

We have one question that was submitted already about a procedural question and then we will go back and address some of the questions that have been submitted during the webinar. The procedures question is about defining a study and the question is: wouldn’t a pure replication of a study (for example, the same design, data collection strategy, and study team) be considered the same study as the original study?

So the new rules for defining a study are far more prescribed than the previous ones and we designed these rules to be easier for review teams to apply. This additional specificity does require drawing some bright lines, so we would really approach a replication study the way same way we would approach any other study to determine whether it would be part of the same study as the original study. We consider four key characteristics: the sample members, the assignment or selection process used to create the intervention and comparison groups, the procedures used for data connection and analysis, and the research team.

[Reference Slide 10 from the January 12, 2018, recertification webinar]

The question indicated that that the original and replication analyses were conducted by the same study team and had the same design, and presumably they also used the same assignment process for both of the analyses so that’s two of our four characteristics. The data collection strategy is also the same but the question doesn't say anything about the analytic procedures. So if the analytic procedures are also the same between the original and replication study – that would make three out of four – and the WWC would consider the findings from the original and the replication analyses to be the same study even if they used distinct samples. So it all just boils down to looking at these four different characteristics of a study and seeing if three of them overlap. And if that is the case, then we consider two findings to be part of the same study.

Okay great. So we’ve got a set of questions that actually cover a whole range of areas, so why don’t we go back to cluster standards and we have either a three-part question or three questions that are related. The questions are related to demonstrating cluster level effects in cluster level RCTs. The first question is: to demonstrate cluster level effectiveness is there any minimum number of clusters required? For example, if I have two interventions and two comparison classrooms, is this eligible for the “with Reservations rating,” as long as the other requirements are met (for example, steps 1-6 of slide 14 of the cluster module).

So long as there is not a confounding factor, which you would have if you only had one intervention or one comparison cluster, there is no minimum number of clusters. Another way of putting that is you need two clusters in the intervention group and two in the comparison group to avoid an n=1 confound in general, but we don't have any further requirements on the number of clusters.

The second question is: is baseline equivalence assessed the same way for cluster effects as for individual effects? For example, if baseline effect sizes are in the adjustment range do you make statistical adjustments in the same way to the cluster effects calculation?

I believe what this question is getting at is “are the baseline equivalence requirements the same for clusters as for individuals?” The answer to that is yes, with one caveat, that to satisfy baseline equivalence based on clusters we would need to have the sample of individuals that contribute the baseline data to be representative of the clusters.

Dana let me jump in here as well. I wonder if the question is also about satisfying the statistical adjustment requirement. If the baseline difference is in the adjustment range then to satisfy equivalence of clusters the adjustment for the baseline measure must be a cluster level measure as well. To step back, to establish equivalence of individuals when it is in the adjustment range, the baseline data must be adjusted at the individual level and to do it for clusters it has to be adjusted at the cluster level.

Great point.

Great, thanks Elias. That leads into the next question about whether a study can demonstrate cluster representativeness if it doesn't report actual sample sizes, but the overall and differential attrition rates can be assessed, or does the sample size absolutely have to be reported?

So long as we can understand the share of the individuals in the clusters that are represented in the analytic sample, that is sufficient in order to demonstrate cluster representativeness. So instead of knowing that 99 out of 100 students in classrooms contributed baseline data or contributed outcome data, if we just knew 99 percent contributed outcome data, that would be sufficient to demonstrate representativeness. The number of individuals that the sample size is needed to estimate the effect size in certain cases, either the baseline effect size or the outcome effect size, because the WWC does a small sample adjustment, so the exact sample numbers might be required for that. But in terms of just understanding cluster representativeness it should be sufficient to know the percent of the cluster which is represented in the analytic sample.

And Dana let me just clarify percent across all of the clusters in the study not individual cluster by cluster. The overall sample.

That is right. And we would need it for both the intervention and the comparison groups in order to assess overall and differential.

Great. The next set of questions are related to missing data and they’re about some additional explanations of some of the handbook text or the slides. The questions are about slide 53 from the original webinar so possibly could go to that, Dana. The questions are about studies with imputed baseline data.

[Reference Slide 53 from the January 12, 2018, recertification webinar]

This person found the handbook text difficult to parse on the topic especially for step 5B: does the study satisfy baseline equivalence using the largest baseline difference accounting for missing or imputed baseline data? “If you could go through slide 53 from the January 12th webinar, the summary chart, and the associated handbook text, that would be helpful. Specifically, for each column in the chart on slide

53, can you confirm that each of the marked items must be provided except that we can substitute option 2 if option 1 is not available?”

Sure. I can go through a couple of these cases. You all should be able to see the slide 53 that I have pulled up. We can look at the most straightforward case which is where the outcome data in the analytic sample is always observed and the baseline data in the analytic sample is always observed. This is our normal case in that this is the most common situation we will be in. In this case, the baseline standard deviations and samples sizes for the sample with observed baseline data would be needed, as well as the baseline means for the sample with observed baseline data. That's what we need to assess baseline equivalence – our means, our standard deviations and our sample sizes. It’s just framed a little bit differently here.

We can look at another case, which is where the outcome data are always observed and the baseline data are sometimes missing. That was a specific case that we looked at when we were looking in the slides at the second most common case that people will see when they are reviewing studies, you are missing the pretest for instance for a couple of individuals that you have outcome data for. In that case you need the baseline standard deviations and sample sizes for the sample with observed baseline data. You would need that in order to calculate the effect size within the observed data. You need the correlation between the outcome and the baseline measures, the observed standard deviations, and sample sizes for the sample with observed outcome data.

And you would also need the baseline means for the sample with observed baseline data, and the outcome means for the sample with observed outcome data, which is the full sample, as well as the sample with observed baseline and outcome data. So again, here, some things we’re referring to a little bit differently, and the sample with observed outcome data for instance in this case is actually the full sample, but we’re going to need those means. I am not sure if that addresses the question, or if anyone else wanted to chime in on that, if that addresses the concern.

Yes, I think also the clarification that if there is an X in that box for that column it is a required piece of information for assessing baseline equivalence.

Yes, that’s right, thank you for clarifying.

I have another question here based on this slide. The question is: can you talk a bit about how to use the table on slide number 53? In the third column of this table, where outcome data is always observed and baseline data is sometimes imputed, the correlation between the outcome and baseline measures row is not checked but the standards indicate that you do need the correlation to assess step three: does the study limit potential bias from an imputed outcome data. Is the table on slide 53 only for information needed to assess step 5?

Exactly, so this table was formulated explicitly to show the data needed to assess baseline equivalence. So you’re right that when the outcome data is sometimes imputed and the baseline data is always observed, you do need the correlation between the outcome and the baseline measures in order to look at step 3, whether the potential bias is limited from imputed outcome data. You just don't need it to assess baseline equivalence. Because in this case your baseline data is always observed, you can use the observed baseline data in the same way that you do normally to assess baseline equivalence.

Okay great. Another follow-up related to the table. The question is: does the last line of the table mean that both observed baseline and outcome data are needed?

Yes, thanks for asking for clarity there. We need the outcome mean for the sample with observed baseline and outcome data so this is specifically the outcome mean restricted to the sample in which both the baseline and the outcome are observed.

Okay. Thank you Dana.

We have a couple of follow-ups on the cluster question. One, and I think Elias, you can address this, someone noted that the handbook says that individual means and standard deviation can be used to establish cluster level equivalence. I think you said it had to be cluster level?

To satisfy the adjustment requirement for cluster equivalence, the handbook says any required statistical adjustments must be made using data at the same level as those used to assess baseline equivalence. So, yes, I believe my statement before was too strong because you can satisfy equivalence of clusters using individual level data, and if you do that, if the study uses individual level data to establish equivalence, then you can use that same individual level data to satisfy the adjustment requirement. So that is correct, I misspoke before. My statement was too strong before. But the data should align – the data used to actually measure the baseline difference and the data used to conduct the statistical adjustment should align.

Thanks for that clarification. Another question about representativeness, the question is: is it true that representativeness must be determined for both baseline and follow-up analytic samples?

We do different things when we look at representativeness for the baseline and follow-up analytic samples. So in certain cases you actually would not be required to satisfy both. Suppose you have a low attrition RCT that is not eligible for a *Meets Standards without Reservations* rating. In that case, the low attrition RCT – I should be more specific there – it's an RCT with low cluster level attrition that for one reason or another (maybe joiners or representativeness) cannot meet standards without reservations. In that case you would actually only need to satisfy representativeness for the follow-up analytic sample because a cluster RCT with low cluster level attrition can be rated meets standards with reservations if it shows representativeness at follow up.

If you have a study that needs to satisfy baseline equivalence of clusters, it does need to satisfy representative of clusters as part of satisfying baseline equivalence. In addition any study that would be required to satisfy baseline equivalence of clusters would also need to satisfy representativeness of the follow-up analytic sample. So I believe to just summarize, in cases where you have to show representativeness at baseline you also have to show representativeness at follow-up, but there are some cases where you might have to show representativeness at follow-up but not baseline depending on the precise nature of the design.

Okay, thank you.

We have another follow-up, possibly related to the slide you have up, but going back to step 5b and dealing with missing data, and that step, again, is: does the study satisfy baseline equivalence using the largest baseline difference accounting for missing or imputed baseline data? So the question notes that the handbook says "when the outcome measure is observed for all subjects in the analytic sample the

WWC requires the following data from authors." One of the pieces, listed as "D" in the handbook, is "an estimate of the baseline differences based on study data. If the study did impute baseline data, then the WWC would include the imputed data when calculating the means, but use standard deviations based only on observed data". For this requirement can we use means calculated on observed data to demonstrate baseline equivalence, or do we have to use means calculated on imputed data if we know imputation happened for this purpose?

Okay I think this is going to boil down to the option 1/option 2. And essentially, the way that these were developed is we have to have some sort of measure of the baseline difference based on study data so if baseline data are sometimes imputed that's either going to be based on the sample including only observed baseline data or the sample which combines both observed and imputed baseline data. So we've chosen as option 1 that we would typically use the sample with observed or imputed baseline data, however you can use the sample with only observed data. So either option is all right, and the WWC has formulas to deal with both. And those formulas are I believe all in appendix C of the Standards Handbook.

Okay, thanks, Dana.

We have a couple of quick procedure questions that I will try to answer. One is: where can we find the updated study review guide? The second is: when will we be able to see it or when will it be available for use?

The new study review guide that incorporates version 4 standards is a web based study review guide, it's going through its final testing now and when it becomes available hopefully in February you will be able to access a link to it on the WWC reviewer page. Where you now see a link to download the Excel file and you will have a link to log into the system. That's for reviewers. There will also be a public version that will not be connected to all of the data that are part of WWC reviews. You will also be able to access that version on that page as well. So hopefully these will both be available in February.

That is all the questions that we have right now. Are there any other last questions that people might have? We'll wait just a minute, but let me quickly note while we are waiting that the second office hours covering recertification information on the change standards and procedures for version 4 will be this Friday at 1:00. Also, we will be making a recording of this webinar available on the WWC site at some point but you can always access it through the link that you use today to get to the webinar.

I don't see any more questions so I will now turn it over to Brice to sum up for us.

Oh, hang on one second. We have one more question. We have a question about slide 57 – would you mind going to that, Dana? The question is: is the term x bar_i~y reported on slide 57 or was it accidentally left up?

[Reference Slide 57 from the January 12, 2018, recertification webinar]

x bar_i ~y – that's the mean of the pretest for those with observed pretest data and imputed post test data – that is not included on slide 57 that I am seeing right now. There are two possibilities here.

[Reference Slide 53 from the January 12, 2018, recertification webinar]

If we look back at slide 53, we are in the case where both outcome and baseline data are sometimes imputed so, x bar_i~y is the observed baseline data imputed outcome data. That is in option 1. I'm thinking out loud here. That's option 1, so possibly option 2 was included which is observed or imputed baseline data for the baseline sample? Yes, that is included there.

[Reference Slide 57 from the January 12, 2018, recertification webinar]

So we don't really need that x bar_i~y because it is one of the options. So here we have the data from option 2 as opposed to the data from option 1. So it is both not included and not an error.

Great, thank you, and thanks for solving that on the fly.

The next question is more of a general question and I assume it can apply in a number of situations. The question is: how do you calculate attrition in an RCT where the author did random subsampling?

Okay. So in this case it's going to depend on the point of the random subsampling, but the idea here is we don't want to count individuals who were randomly subsampled out of the follow-up as attrition because that's calling something that's random, nonrandom. So essentially, and it's going to depend on the precise nature of the subsampling and when it occurs and how it was done, but typically what you would look at as kind of the denominator for the attrition calculation would be anybody that was included in the random subsample and the numerator would be the size of the analytic sample. It's hard to say more about that—maybe Elias can elaborate—without knowing precisely when and how the random subsampling was done.

So first of all this topic is discussed in one of the online training modules, the online training module on attrition. There are a couple of examples related to this question about random subsampling and the basic take away is that if the study randomly subsamples based on the originally randomized sample, the sample that was originally assigned to conditions (that full sample), then that is treated as an exogenous subsample. The random assignment is preserved and analysis based on that subsample can be treated as an RCT with no attrition. But if the subsampling is done based on a sample that has lost sample members – a common example of this might be you have a survey with multiple waves and after wave one the researcher doesn't have enough money to sample everybody in wave two so they randomly select a group of respondents from the wave one sample. That is problematic because the wave one sample had some attrition already baked into it, and randomly subsampling from that selected sample, the WWC would count that as attrition in that case. So again the key is just, is the subsample taken from the full original sample or is it from some selected sample. And if it is from a selected sample, such as after a wave one survey, then all of the samples loss is counted as attrition. And again, the online training module on attrition has some more specific examples on this.

Thanks Elias. To put you on the spot again Dana, we have a follow-up on the missing data example that you have up there. The questioner is trying to understand how we are using some of the formulas that appear in appendix C, under option 2. So the question is: is option 2 used somehow to inform option 1?

[Reference Slide 53 from the January 12, 2018, recertification webinar]

So I think the trick here is that you can essentially do the calculations when there is imputed baseline data in two ways. You can build them off of assuming that we have some imputed data or you can build them off of treating all of the imputed date as missing. So when you are trying to assess baseline equivalence you either start with the effect size within the complete observed data or you start with the observed data combined with the imputed data. So essentially what we are doing by having those option one versus option 2, we are allowing one set of formulas to be used versus another set of formulas. I think there’s a C1 and a D1, a C1 star and a D1 star. I am not 100 percent positive because I don't have it in front of me, but I think it is allowing you to use the C1 formula as opposed to the C1 star formulas. So hopefully that clarifies the issue. It’s not using the observed data to infer something about the imputed data, it’s just using a different starting point and adjusting the starting point differently.

Thanks, Dana. If there are any follow-ups to that or if, after you think about this a little more, you have more questions, please feel free to send questions to our recertification email address. And with that I will turn it over to Brice and not interrupt him again.

Thanks Neil. This actually does conclude the web cast for today 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. The transcript will also be available on the WWC website later this month. You can submit any feedback to the team through the contact us form on our website whatworks.ed.gov. Thank you and have a great day.