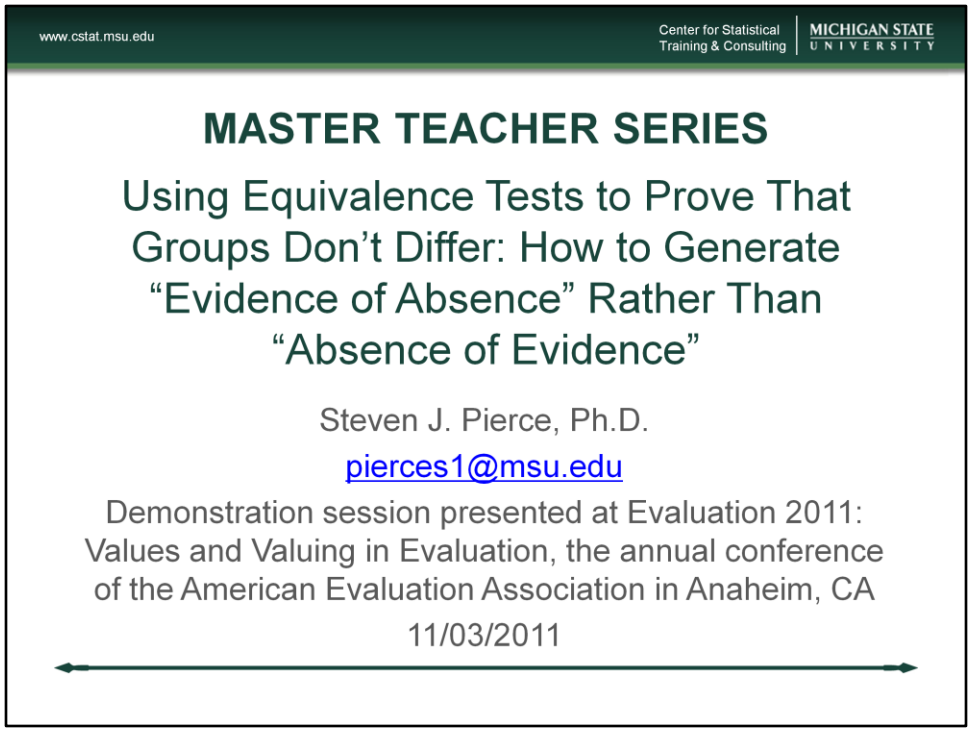


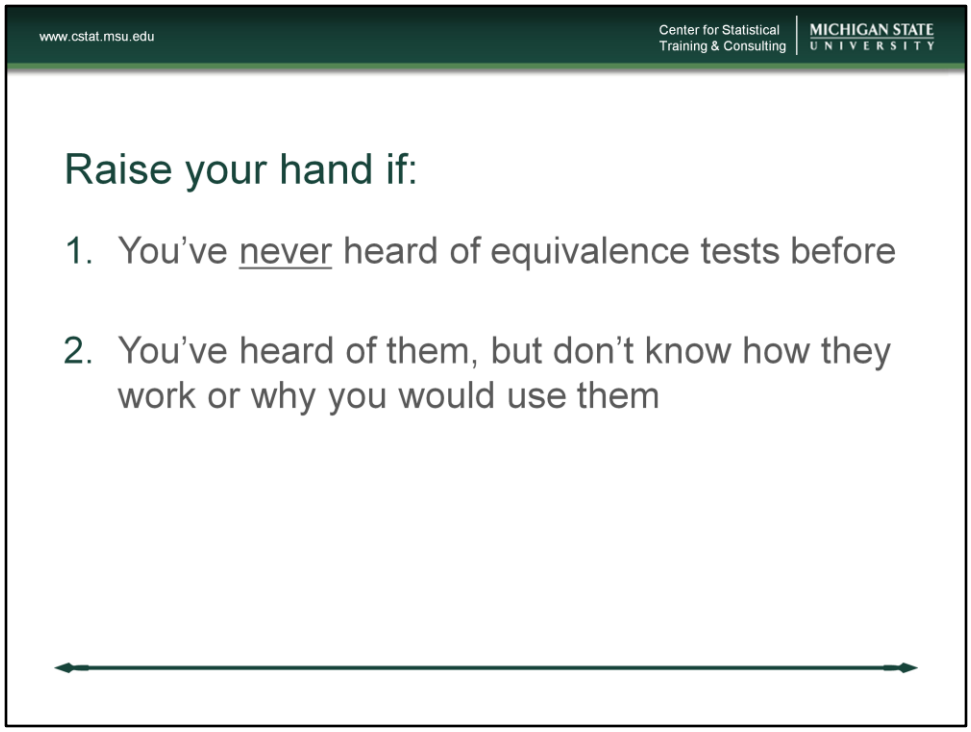
Good afternoon. I'm Steve Pierce and this demonstration session is entitled *Using Equivalence Tests to Prove That Groups Don't Differ: How to Generate "Evidence of Absence" Rather Than "Absence of Evidence".* I've put a stack of handouts at the back of the room – that's a list of resources, including a note on how to get copies of the slides after the talk.

Demonstration Session #409. Limit ≤ 90 min (75 min talk + 15 min Q&A)

Citation

Pierce, S. J. (2011, November). *Master Teacher Series: Using equivalence tests to prove that groups don't differ: How to generate "evidence of absence" rather than "absence of evidence".* Demonstration session accepted for presentation at Evaluation 2011: Values and Valuing in Evaluation, the annual conference of the American Evaluation Association, Anaheim, CA.

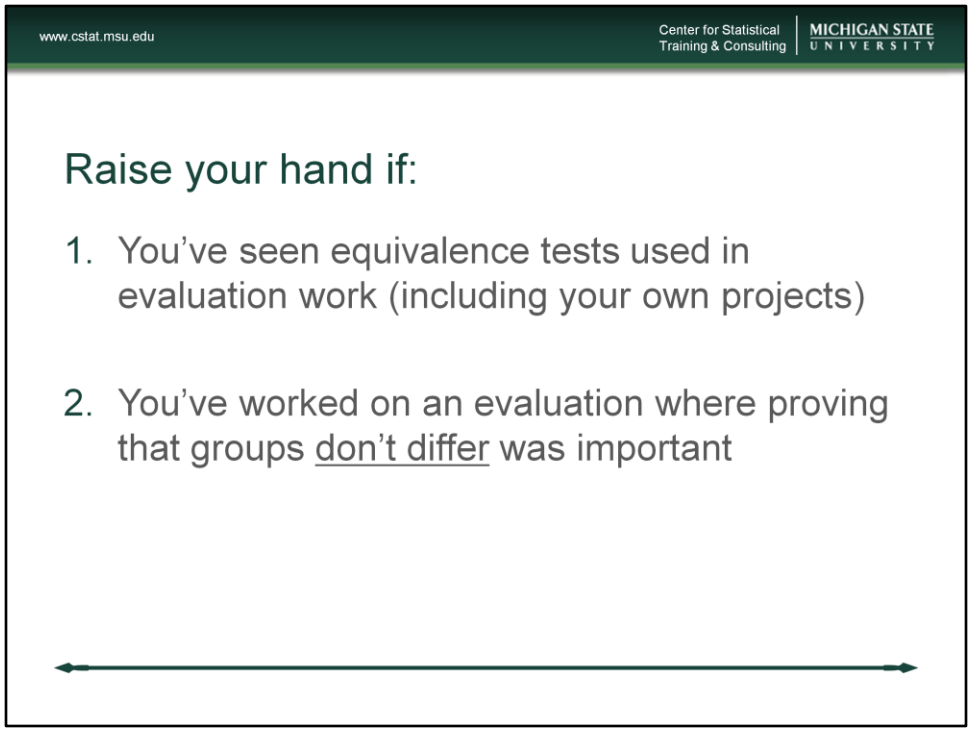
Abstract: This intermediate session will demonstrate the use of equivalence tests, which are statistical methods designed to use the evidence in the data to explicitly prove that two groups are actually equal on some outcome measure. This departs from the goal of classical statistical methods (e.g., t-tests, chi-square tests, etc.), which aim to use the data to prove that two groups actually differ from one another. The session will explain what kinds of evaluation questions equivalence tests can answer, how these tests work, and why they provide more credible evidence of equivalence than offered by simply finding a non-significant effect with a classical method. The session will offer practical advice on how and when to use equivalence tests in your evaluation work. Because equivalence tests are rarely covered in basic graduate level statistics courses, this session aims to provide the audience with a user-friendly introduction to new statistical methods.



I want to just poll the audience briefly. It's useful to see where the audience is at in terms of prior exposure to my topic.

First, please raise your hand if you have never heard of equivalence tests before seeing this talk listed in the program.

Second, please raise your hand if you've heard of equivalence tests before, but don't know how they work or why you would use them in an evaluation.



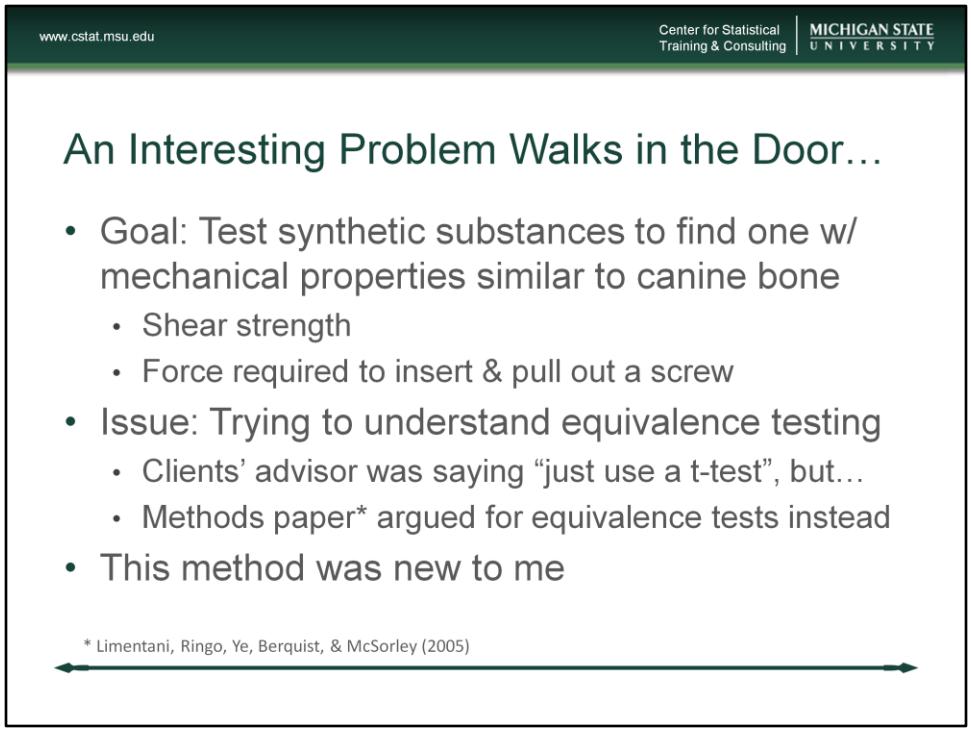
Ok. Just a couple more questions:

Please raise your hand if you've actually seen equivalence tests used in evaluation work (including on your own projects).

Finally, raise your hand if you've worked on an evaluation where proving that groups actually don't differ on some measure was important.

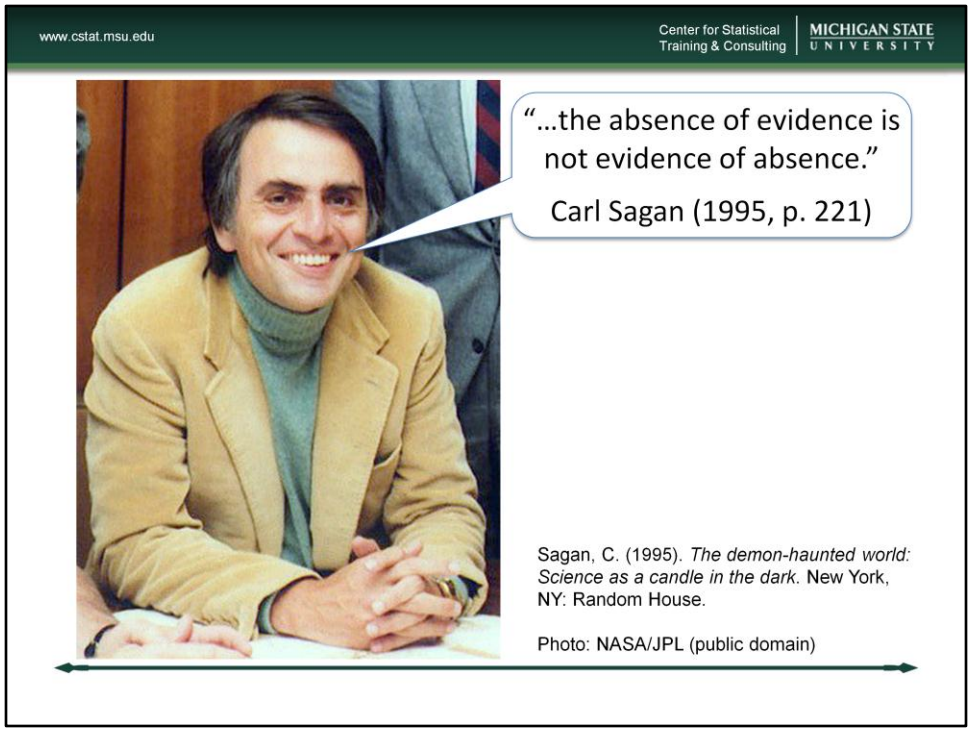
Thanks.

OK. I'd like to just share a bit of background on how I got interested in offering a session on equivalence tests. I've spent the last couple years working at Michigan State University, which is a pretty large institution. It currently has about 4900 faculty and academic staff and about 11,000 graduate and professional students. The Center for Statistical Training & Consulting where I work provides statistical consulting services to the faculty, research staff, and graduate students on campus. We serve all disciplines and departments on campus, so as you can imagine we got lots of clients. In fact, we served about 491 clients last fiscal year. Because MSU is a large research institution, we get clients coming in to ask for help on an incredibly diverse set of research problems. So, working at CSTAT gives me a lot of opportunities to learn new methods and see how tools from one discipline might be applied in other disciplines.



The idea to do a talk about equivalence tests here came up after a client came in for help on a really interesting problem. He was a veterinary medicine student doing research aimed at finding a synthetic substance (such as certain kinds of plastics) that had mechanical properties very similar to canine bone. The idea was to find something that they could use as a substitute for real bone when doing certain kinds of research, to both reduce the cost of their work and to reduce the number of dogs they would have to use as subjects. He was looking at things like the shear strength of the material and the amount of force required to insert and remove a screw from the substance. It was critical to find a substance that behaves just like actual bone.

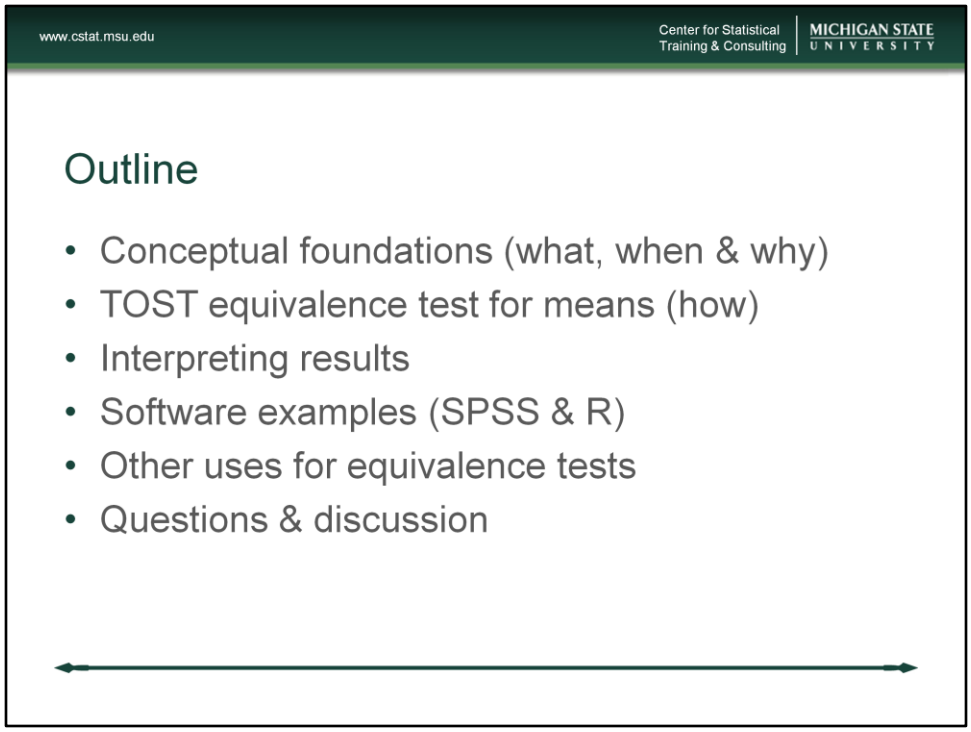
The issue was that while his advisor was saying he should just use t-tests and look for the ones with non-significant results, he had found a paper in the literature arguing that equivalence tests are more appropriate. He was having trouble understanding the paper and deciding whether or not to just do what his advisor was suggesting, so he'd come to ask us for advice. Despite having had fairly broad exposure to statistical methods, this was the first time I'd run equivalence tests. I was able to rapidly pick up the essentials though, so after the call for abstracts came out I started thinking about how these methods could be used in evaluation.



While trying to compose a catchy title to lure people to a statistical methods talk, I borrowed from this famous quote by Carl Sagan: "the absence of evidence is not evidence of absence". I gather from your presence that this tactic worked. I'll have to try that trick again next year.

Beyond it's value for luring in an audience, this quote really is deeply connected to my topic today. Think about how it might apply to statistical comparisons of different groups in an evaluation. Many classical statistical methods (e.g., t-tests, chi-square tests, etc.) are designed to test whether groups differ significantly. This quote suggests that failing to find a significant difference between groups is not the same thing as proving that they do not differ. Indeed it concisely conveys a message many statisticians and methodologists have offered, which is that failure to reject the standard null hypothesis of "no difference" can't be interpreted as support for the assertion that they do not differ. A non-significant finding from a t-test yields only an "absence of evidence" with respect to the hypothesis that two groups are equivalent.

So how do we prove that groups don't differ? That is, how do we generate compelling and credible "evidence of absence" with respect to group differences? That's what equivalence tests are for. My goal today is to introduce you to such tests and demonstrate how they may be useful to you in an evaluation. I'm guessing a fair number of you are wondering why you've never heard of equivalence tests before. I suspect part of the answer is that they're not usually covered in basic statistics courses and they're rarely mentioned in the evaluation literature.

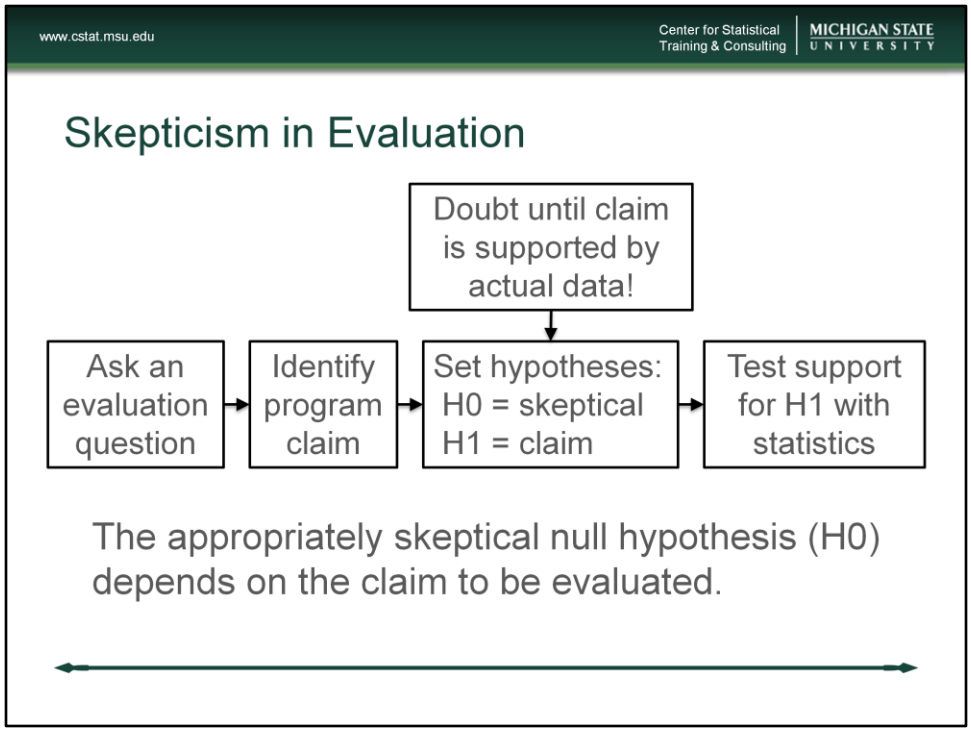


Here's a high-level outline of what I'll be talking about today. My most important goal is to lay the conceptual foundations for what equivalence tests are, when and why they might be appropriate tools for an evaluator to use. So, I'll talk a bit about the most common situations where I might point someone toward these methods.

Then, I'll focus on explaining how one of the simplest equivalence tests works and how to interpret the results. I'll demonstrate how to do this test using two different software packages (SPSS and R).

After that, I'll very briefly point out some other potential uses for equivalence tests that I've located in the literature. I'll wrap up by taking questions and discussing issues or applications raised by the audience.

OK. Let's get started!

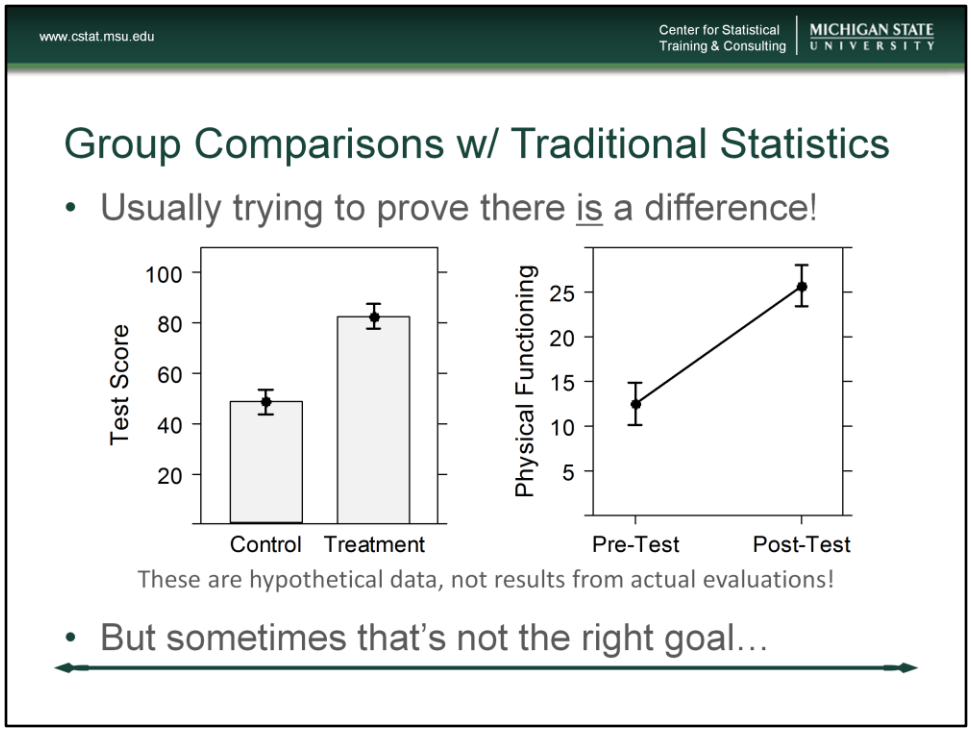


A healthy dose of skepticism plays an important role in evaluation, just as it does in other branches of science. When we're asked to answer an evaluation question like "Does this program improve outcomes for participants?", we are in a sense tasked with testing whether there is evidence supporting a particular claim that the program developers might like to make. Namely, that the program is effective at improving outcomes.

Adopting a skeptical position means that we will require compelling proof before we declare support for that claim, rather than just assuming it is true. So, we want to doubt the claim until the data convince us that it is plausible. That doubting position helps us identify how to translate the claim into a formal statistical hypothesis.

In the statistical hypothesis testing framework, we have to choose both a null hypothesis (H0) and an alternative hypothesis (H1). To declare support for the alternative hypothesis, we must first find evidence in the data that warrant rejecting the null hypothesis. So, the claim we have identified is really the alternative hypothesis and the null hypothesis represents the skeptical position wherein that claim is not tenable.

So, how does this relate to the equivalence testing? I want to emphasize that the appropriately skeptical null hypothesis we should be adopting really depends on the specific claim we need to evaluate. Statistical analysis is just a method for extracting evidence from data, but it's one that depends crucially on properly aligning the hypotheses with the evaluation goals. With that in mind, let's now look at how we use group comparisons in evaluation work.

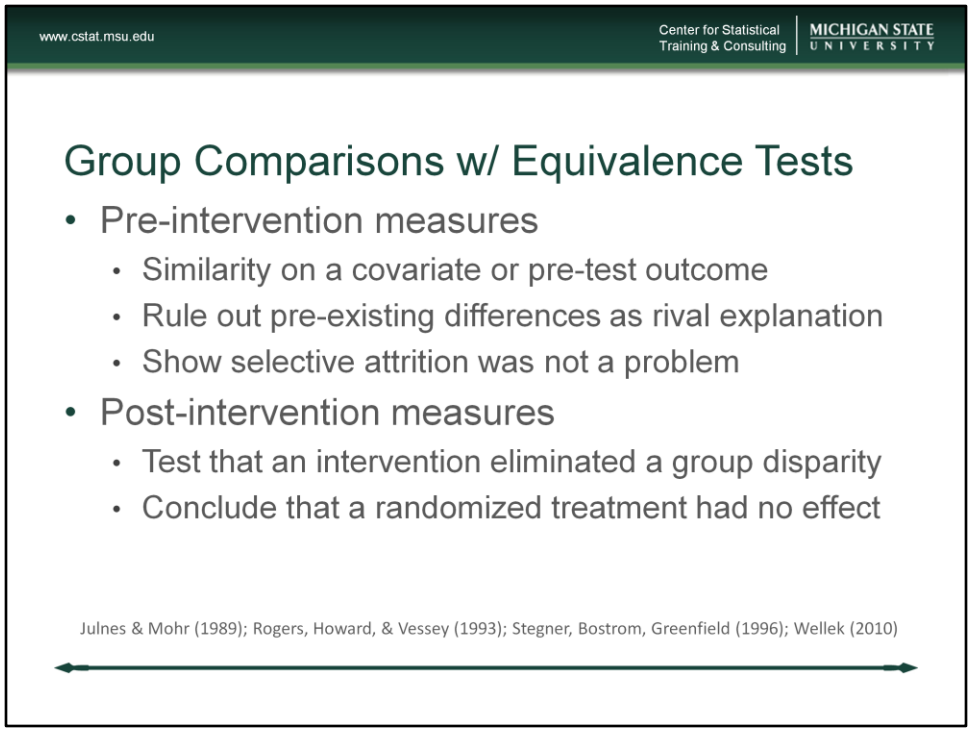


Clients often want to prove that there is a difference in outcomes between groups. For example, we may need to compare participants in an educational intervention to a control group because our clients want to show that the treatment group has higher test scores than the control group, thereby justifying continuation of the program. Since the test scores are a continuous outcome variable, we could use an independent t-test or an ANOVA to test that claim. If the focus was instead on a binary outcome, we could use a chi-square test or a z-test on proportions.

Alternatively, we could need to test whether outcomes are better at post-test than they were at pre-test. Here we could imagine using a paired t-test or repeated measures ANOVA to examine whether patients have higher levels of physical functioning after receiving a new form of physical therapy. If we had a binary outcome instead, such as whether or not patients can climb a flight of stairs, we could use McNemar's test, which is analogous to a paired t-test for proportions.

These common statistical methods adopt a null hypothesis that assumes that there's no difference between the groups because the claim they are trying to prove is that the groups really do differ. Rejecting the null hypothesis means that the data provide credible affirmative evidence supporting that claim. So, when the claim to be tested is that there is indeed a difference, traditional statistical methods are well aligned with the evaluation goal because they rely on an appropriately skeptical null hypothesis.

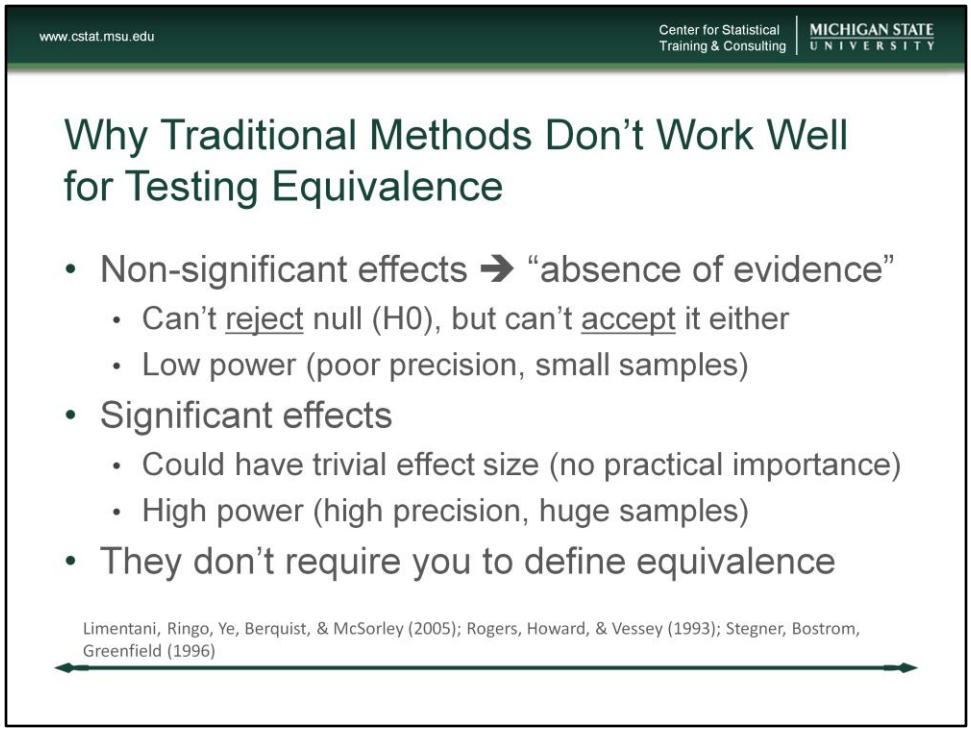
But sometimes, proving that groups differ is really not the right goal because what we need to do is prove that they are actually equivalent and don't differ in any meaningful way.



Here are some scenarios where we might need to test claims asserting there is no difference between groups, either on pre-intervention measures or on post-intervention measures. First, we might want to prove that two groups are equivalent with respect to a either covariate or a pre-test outcome measure. This could be especially important when group membership is not-randomly assigned because it helps you rule out pre-existing differences between groups as a rival explanation for group differences in outcomes identified by subsequent analyses. However, it's also useful as a way to show that random-assignment was not compromised in some way. We could also test for group equivalence to show that selective attrition was not a major problem with a study.

Second, we might also want to prove that groups don't differ on some outcome after the intervention. For example, imagine evaluating a compensatory education program whose goal has been to eliminate a performance gap between "at risk" students and regular students who are not at risk. To prove that, we need to show that the performance of "at risk" students is equivalent to that of the other students. The principal claim is that there is no difference in performance after the program, so that's the alternative hypothesis, not the null. A good skeptical null hypothesis would be that there is still a disparity in performance. An equivalence test is an excellent way to test such a claim because it provides "evidence of absence" rather than an "absence of evidence" by more properly aligning the hypotheses with the goals of the evaluation than would classical statistical methods.

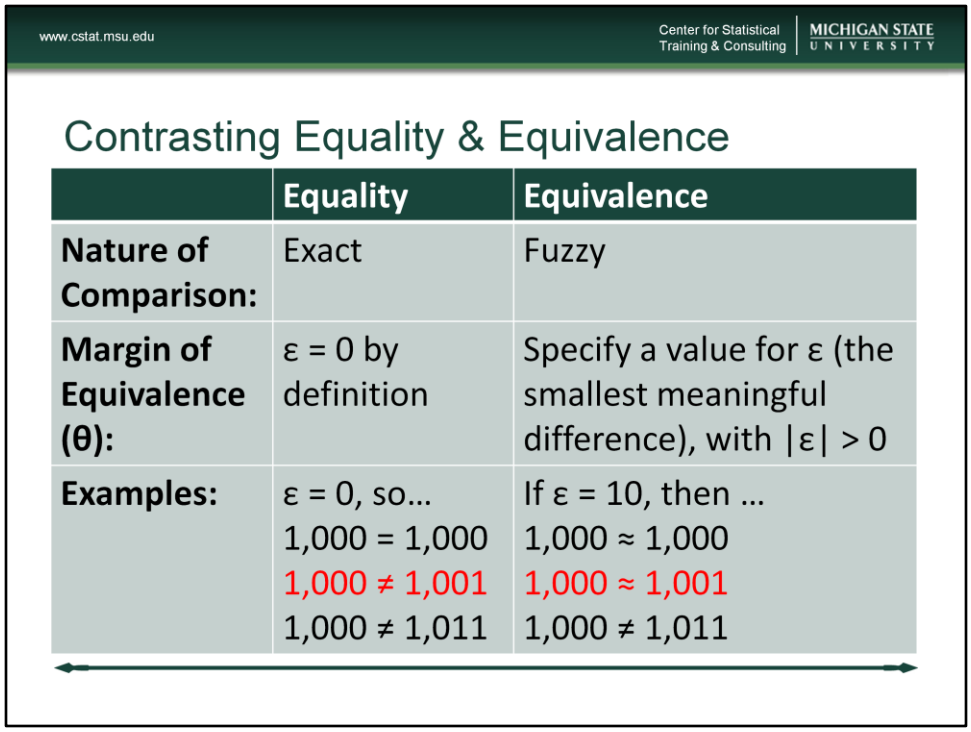
We could also use equivalence tests to bolster a conclusion that the treatment in a randomized experiment truly had no effect. That could be useful when comparing a an existing program to a new program that is less costly to run, but hopefully achieves equivalent outcomes, thereby making it more cost effective.



So, let's examine why traditional statistical methods don't really answer questions about equivalence very well. While it is tempting to interpret a non-significant a t-test or ANOVA as evidence that the groups are equivalent, but that's not what it means. The non-significant effect only means you have no compelling evidence for a difference and therefore can't reject the null hypothesis that the groups have equal means. However, you can't actually accept the null hypothesis either on the basis of that result. The "absence of evidence" for a difference could simply mean you ran a study with very low power, perhaps because your measurements have very poor precision or your sample was too small . We should not draw important evaluative conclusions on that basis.

In addition, a significant effect from a t-test could actually be observed even when the groups are functionally equivalent. That sounds odd, but consider the case where you detect an effect that may be statistically significant, but the difference between the group means is so trivially small that it has no practical importance. In that case, the trivial effect size implies that the groups are functionally equivalent. That can happen when you have really high power, perhaps because of very high precision or when you're using really large samples.

Finally, classical statistical methods are designed to test for differences, not for equivalence, so they don't force you to actually define equivalence. Using traditional methods to try to prove groups don't differ would mean your analysis methods are poorly aligned with your evaluation goal. We can do better than that by using equivalence tests because they're designed to provide affirmative evidence to support concluding that groups are equivalent.

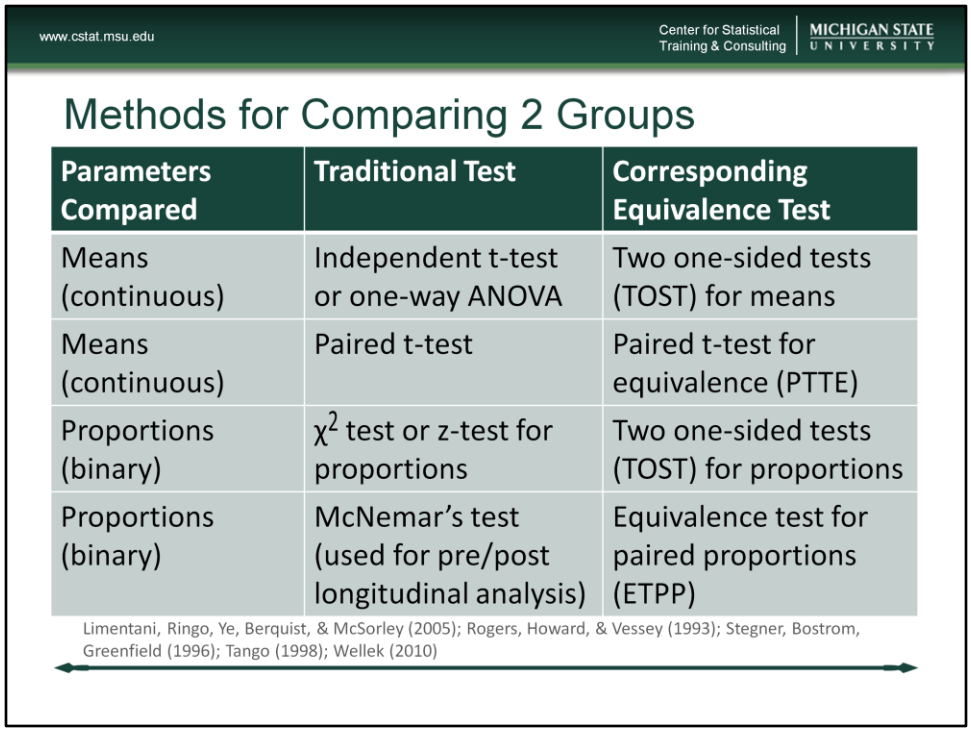


I mentioned a moment ago that traditional statistical methods don't require defining what counts as equivalence. There's a crucial conceptual difference between equality and equivalence. Traditional methods build hypotheses around the concept of equality, which implies that we are making an exact comparison between two estimates. By exact, I mean that the margin of equivalence for the difference between two numbers is exactly zero: Only identical values are considered equal.

In contrast, equivalence tests build hypotheses around fuzzy comparisons where you actually have to define the smallest meaningful difference between the two groups. Only differences smaller than that margin of equivalence indicate the groups estimates are close enough to each other to consider the groups equivalent. The contrast between hypotheses based on equality and equivalence is important because they can lead to different conclusions given the same data.

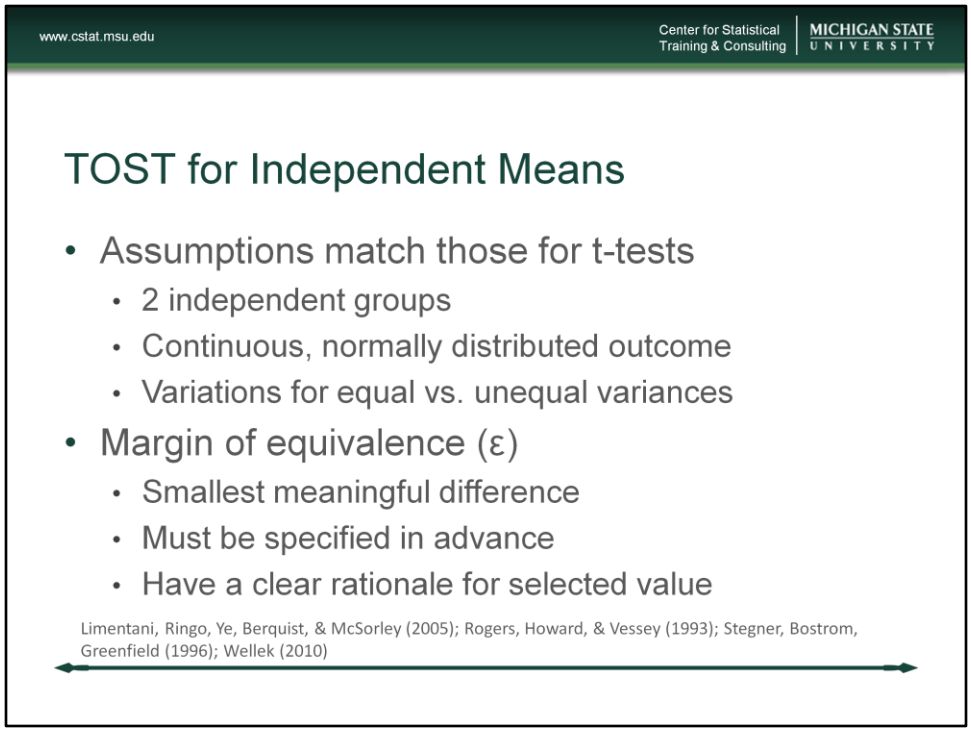
Here are some simple numerical examples. If we have two groups that have population means of 1,000, then we can easily see that they are both equal and equivalent. However, if the group means differ by even 1 point (1,000 vs. 1,001), they are no longer equal. If we adopt say a 10-point margin of equivalence, groups with means of 1,000 and 1,001 are equivalent, but not equal; group means of 1,000 and 1,011 would be both unequal and not equivalent.

So, one advantage of equivalence tests is that they require you to be clear about what counts as equivalence.



Many of the classical statistical methods that you've learned in statistics courses have corresponding equivalence tests. Here's a small menu of choices related to simple methods for group comparisons. The first two rows are for tests based on comparing group means, while the last two are for tests based on comparing percentages or proportions when you have a binary dependent variable. The first row in each pair lists tests for independent groups, while the second lists tests you would use for longitudinal analyses.

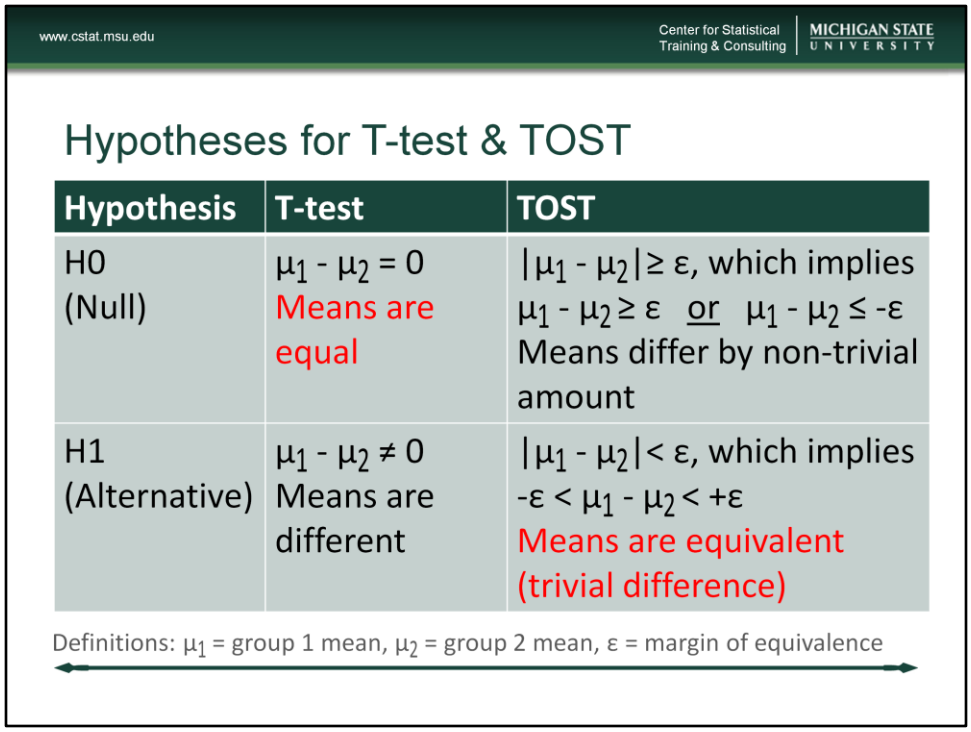
The basic approach for independent groups actually relies on using two-one sided tests (TOST) for either means or proportions. I'm going to elaborate on that approach next.



The TOST approach relying on two-one sided tests for independent means is one of the simplest equivalence tests around. It's the counterpart to the traditional t-test, so many of the assumptions for the TOST match those required for a t-test.

For example, you're assuming that you have two independent groups with random assignment , and a continuous outcome that follows a normal distribution. There are variations of the formulas for situations where you can assume the variances for the two groups are equal, and for when the variances are unequal.

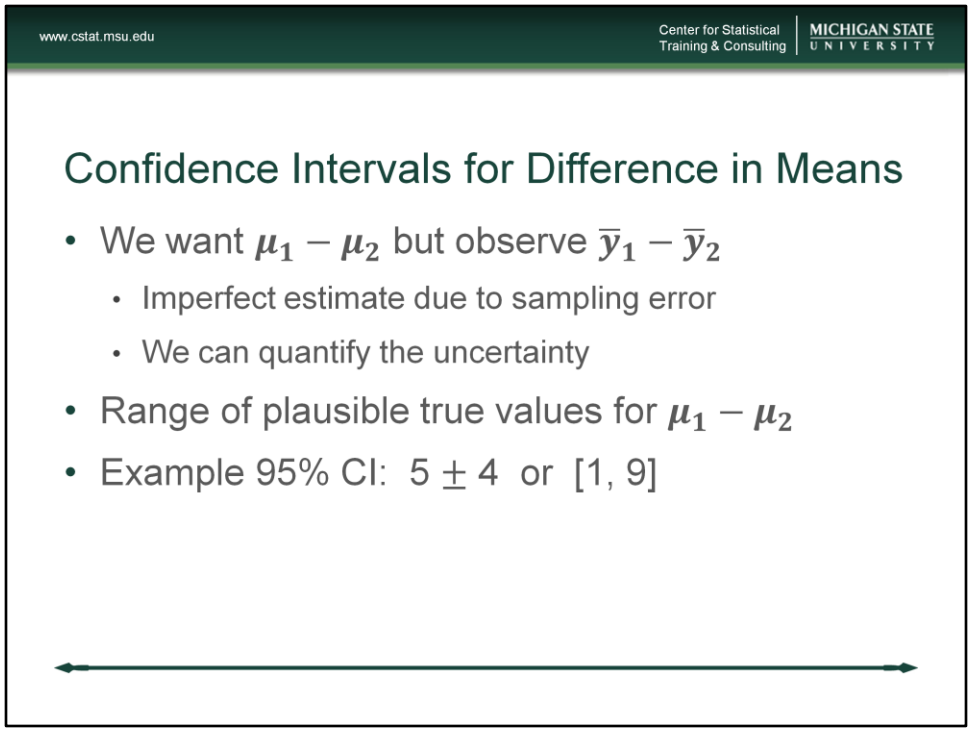
The TOST procedure requires that you define a margin of equivalence representing the smallest meaningful difference between the means that would make you conclude that they're not equivalent. I'm using the Greek symbol epsilon (ε) for that margin. You need to specify that margin in advance to run the equivalence test. It's very important that you think carefully about the value you set for this margin: you need a clear rationale for why that value is appropriate.



This table illustrates how the hypotheses for a traditional t-test compare to those from the TOST procedure. TOST essentially reverses the hypotheses so that the null (H0) hypothesis now corresponds to assuming there really is a non-trivial difference between the groups until proven otherwise.

By asking whether the absolute value of the difference is larger than the margin of equivalence, we are really examining two different one-sided hypotheses. The first one says that the size of a positive difference between the means is equal to or larger than the selected margin, while the second one says that the size of a negative difference is equal to or smaller than the negative value of the margin. In either case, these hypotheses are saying the difference is farther from zero than the margin of equivalence. The burden is on the analyst to use the data to prove that the two groups really are equivalent by showing that the actual difference lays within the equivalence interval that runs from –ε to +ε.

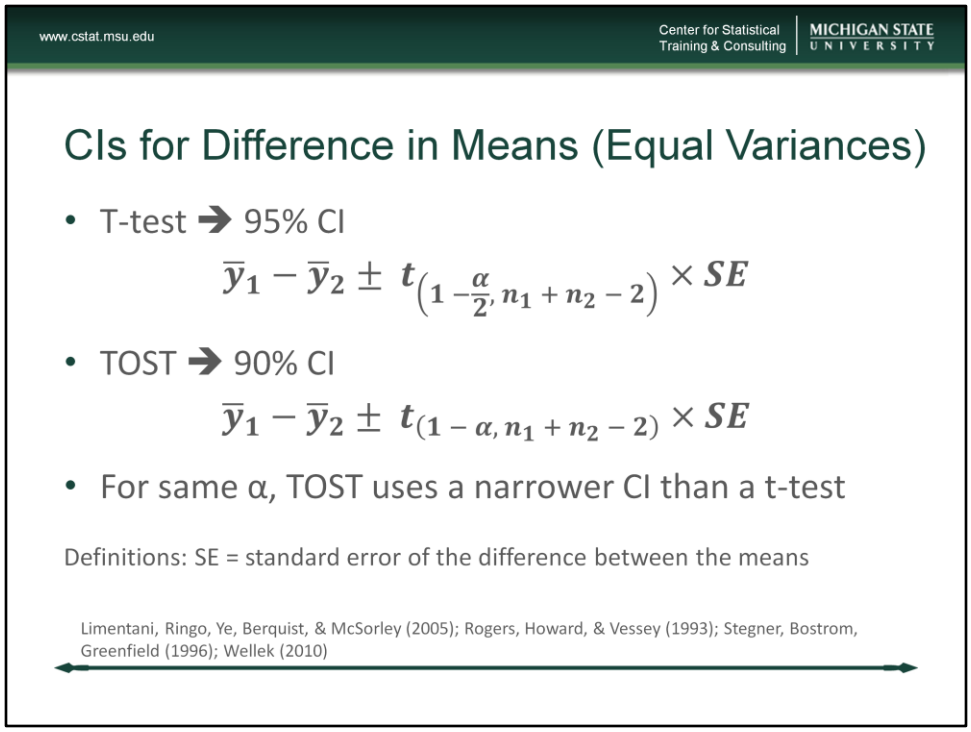
This helps to generate “evidence of absence” by not rewarding you for using measures with poor precision or small sample sizes. Both of those factors could inappropriately stack the deck toward a non-significant difference in a traditional t-test, but would not help you in an equivalence test.



There is more than one way to conduct a TOST equivalence test for comparing group means. I'm showing the one that I think is easiest to explain and understand. It focuses on estimating and interpreting the confidence interval for the difference between the means. Our goal is to draw conclusions about whether or not the groups are equivalent at the population level, but since it's usually not feasible to obtain population level data, we're using statistics to make an inference from the sample data. Although our best single estimate for that difference in the population is the difference between the sample means for the two groups, we know that's an imperfect estimate because of sampling error.

Fortunately, as long as we know the sample standard deviations and sample sizes for each group, we can quantify how much sampling error there is surrounding that estimate and use it to calculate a CI that tells us how large or small that true difference might really be. So, you can think of the CI as the range of plausible true values for the difference between the group means.

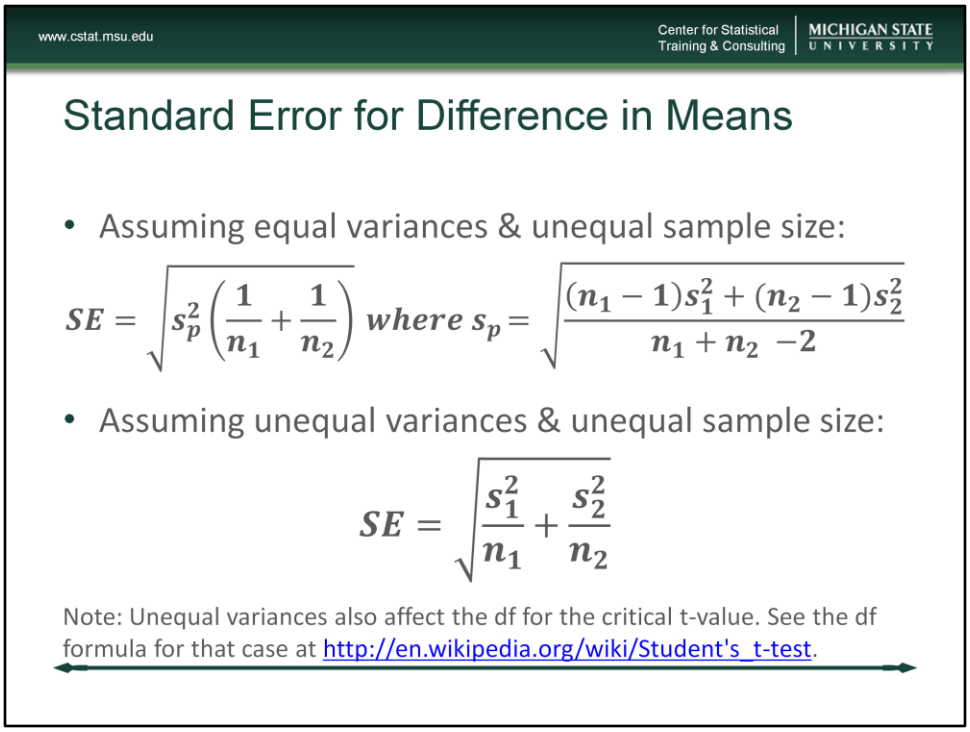
We usually write CIs in one of two ways, as the mean plus or minus an amount of error, or we can actually do the addition and subtraction to get the upper and lower limits then report those two endpoints inside a set of square brackets. For example, if the mean is a 5-point difference between groups, plus or minus 4 points due to sampling error, then the interval runs from 1 on the low end to 9 on the high end.



Here are the CI formulas for both a traditional t-test and the TOST equivalence test. Note how similar they are: They both use the difference between the sample means as the center point of the interval, then calculate upper and lower bounds around the center by multiplying a critical value from the t-distribution by the standard error of the difference between the means. The only real difference between these two formulas is the critical t-value used.

For the same Type I error rate (e.g., α = .05), the critical value of the t-statistic used to construct the CI for the TOST analysis is smaller because we are really doing 2 one-tailed tests instead of 1 two-tailed test. That means the TOST CI is narrower than the corresponding t-test CI. Where we look at the 95% CI for the t-test, we would use the 90% CI for the TOST analysis. To get the critical t-value, we just need to know the Type 1 error rate α and the degrees of freedom (which are based on sample size).

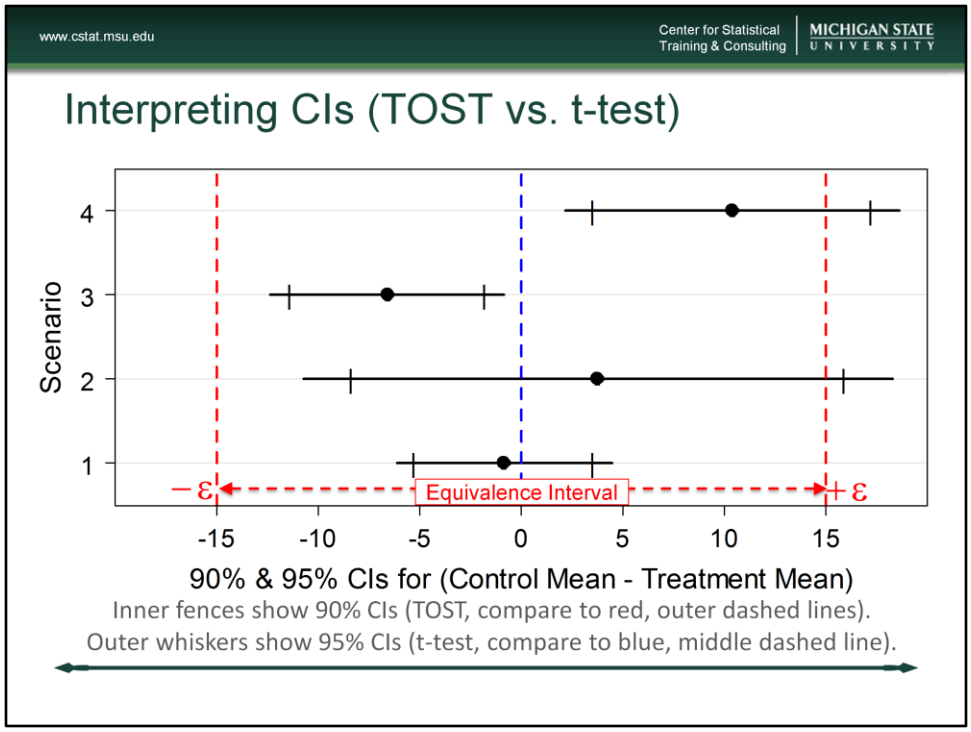
We use the same methods for calculating the standard errors for these two formulas. However, the exact formula depends on whether we can assume equal variances across the two groups, or we need to assume unequal variances. Let's look at that formula next.



We can use the exact same formula for the standard error of the difference between the means in a TOST equivalence test as we would have used for the corresponding t-test. At the top here I'm showing the variation you would use if you are willing to assume that the variances for the two groups are equal.

This is an important assumption: If it is not reasonable, using this version of the formula could yield inaccurate inferences and you would be better off using the variation for unequal variances instead. I've actually seen that happen with real data. One of my clients was comparing the amount of force required to insert a screw into bone as opposed to a synthetic material under consideration as a bone substitute. The force required with synthetic material was much more consistent than the force required with real bone (probably as a result of quality control during manufacturing) and that made a big difference in the conclusions once we switched him over to using the correct formula.

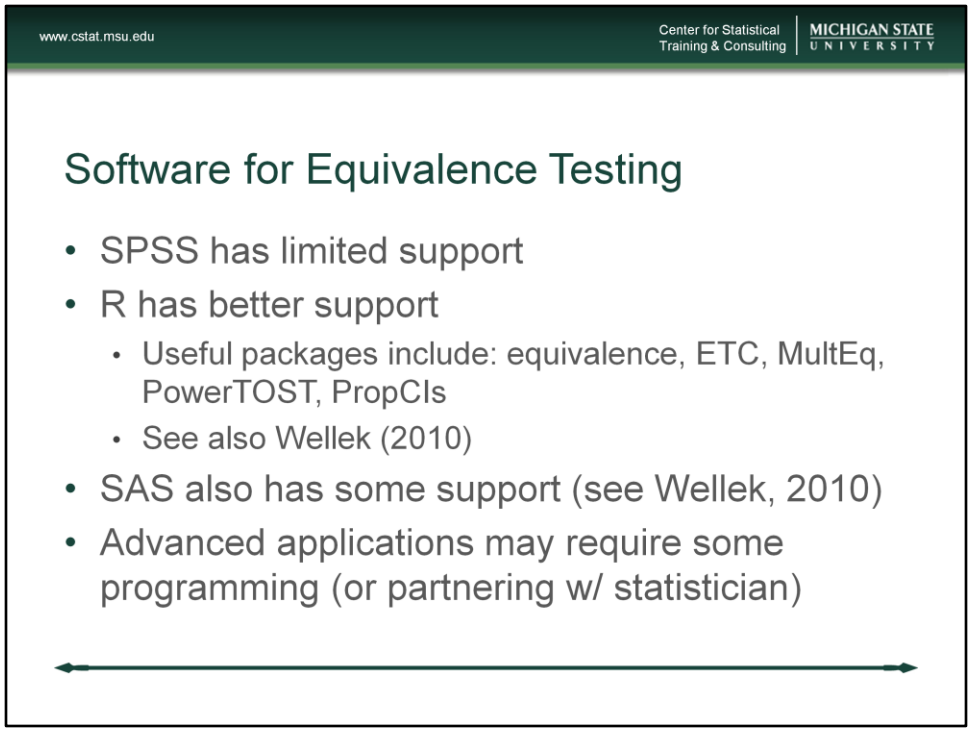
Regardless of whether you are assuming equal or unequal variance, you just need to know the sample sizes and estimated SDs for the two groups to calculate the SE. Although these formulas may look imposing to some of you, rest assured that statistical software will usually do these calculations for you automatically.



This graph illustrates 4 different possible results of comparing the answers you might get from a TOST equivalence test and a t-test on the same data. I've used simulated data here, assuming in all 4 scenarios that the margin of equivalence is 15. For each scenario, I've graphed both the 90 and 95% confidence intervals for the difference between the means of a control group and a treatment group. The inner bars on each interval show the upper and lower limits of the 90% CI, which corresponds to doing a TOST equivalence test. If the entire span of that 90% CI falls inside the equivalence interval marked by the red, outer dashed lines, then we can reject the null hypothesis that the two groups differ and conclude that they are equivalent on this outcome. Meanwhile the outer whiskers of each interval show the upper and lower 95% CIs corresponding to a traditional t-test. So, if that wider interval overlaps the blue, dashed line at zero, then your t-test cannot reject the null hypothesis that the means are equal.

Scenario 1 shows a case where the groups are equal (because the interval includes zero as a plausible value) and equivalent because both ends of the TOST interval fall well within the margin of equivalence. Scenario 2 is slight different because the t-test would conclude the groups are equal, but the TOST equivalence test would conclude that they are not equivalent because the end of the 90% CI sticks out beyond the margin of equivalence. That means it is plausible for the difference to be slightly larger than 15.

In Scenario 3, we see an example where you would get a significant t-test, indicating that the groups means are different. However, the TOST test would conclude that while the difference is probably not zero, it is small enough to be considered trivial, so the groups are in fact equivalent. Finally, in Scenario 4, we see a situation where the t-test would be significant (indicating a difference between the means) and the TOST equivalence test would conclude that the groups are not equivalent.

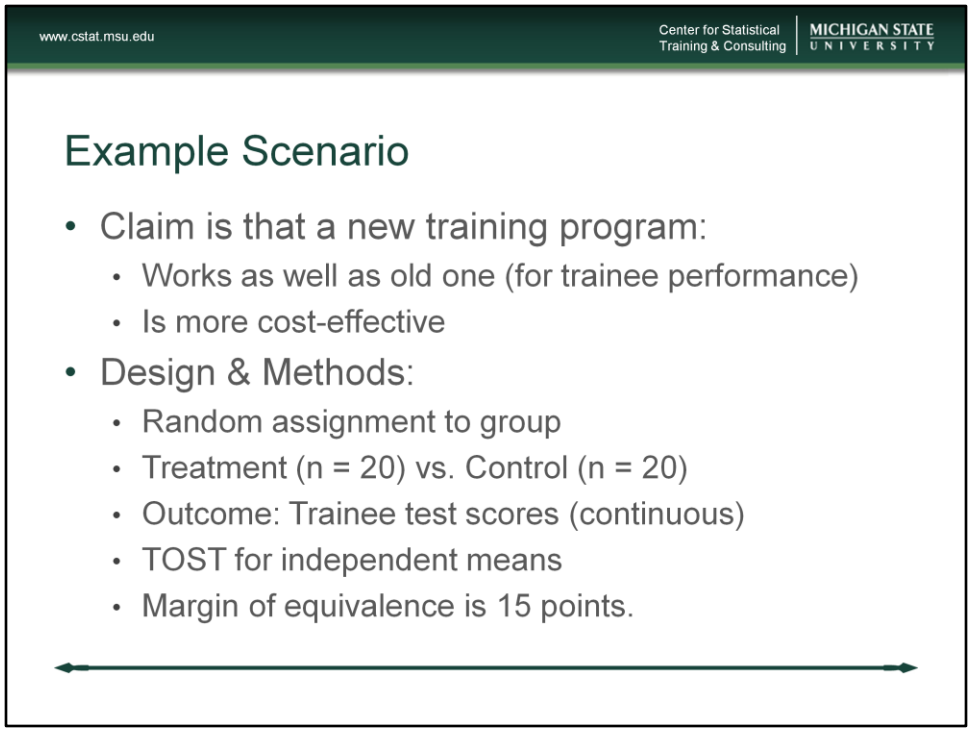


I'm sure most of you plan to use software to run your statistical tests. From what I've seen so far, SPSS can handle some kinds of equivalence testing, but its support is limited and not well documented. Still, I know it's a popular software package, so I'll show you how to use it for a simple equivalence test.

My own preferred stats software tool is R. It seems to have better support for equivalence tests, but mostly through some add-on packages that you can download.

I'm sure that SAS is quite capable of doing equivalence tests, but using it is not my forte so I'm not going to demonstrate using it for that purpose.

Finally, I suspect that more advanced applications of equivalence testing may require some custom programming, or perhaps partnering with a local statistician. Still, the simple stuff is something you can easily do on your own.

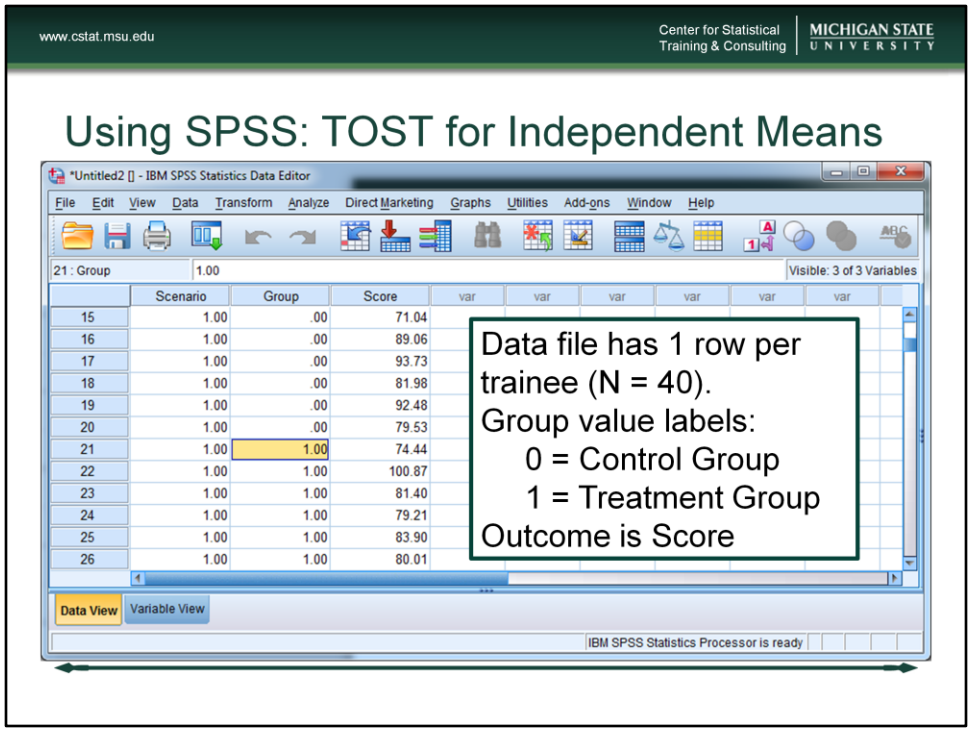


Before I start showing you how to use SPSS and R to run a simple equivalence test, let's describe a scenario where an equivalence test would be warranted. Let's say your evaluation client has been considering replacing an existing training program with a new one that it believes will be more cost-effective. However, the program director wants to know that the trainees' performance is as good under the new program as it was under the old one before she commits to fully replacing the old program: She doesn't want to sacrifice too much effectiveness to get the cost-savings. While traditional methods might be good for testing the cost-effectiveness claims, an equivalence test is a better way to examine the effectiveness claim.

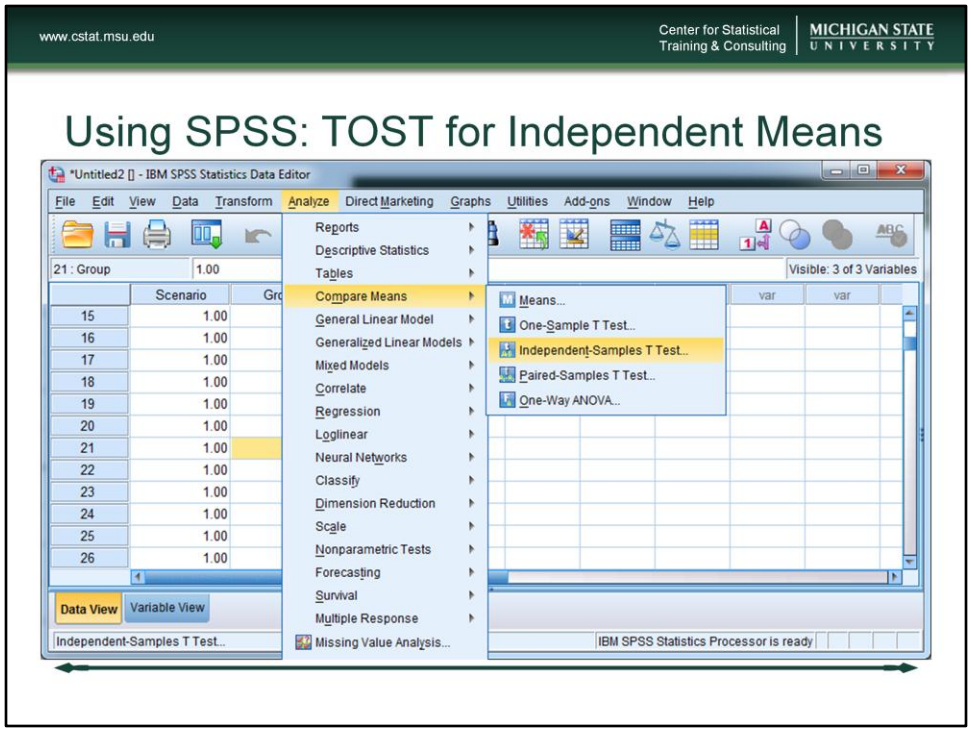
The study will be simple. 20 randomly assigned people will be trained under the new program as the Treatment group; another 20 randomly assigned people will be trained under the old program as the Control group. The outcome will be scores on a certification test taken at the end of the training.

Discussions with the program director reveal that she will be satisfied that the programs are sufficiently equivalent with respect to trainee performance if the means differ by less than 15 points on the test they administer to trainees.

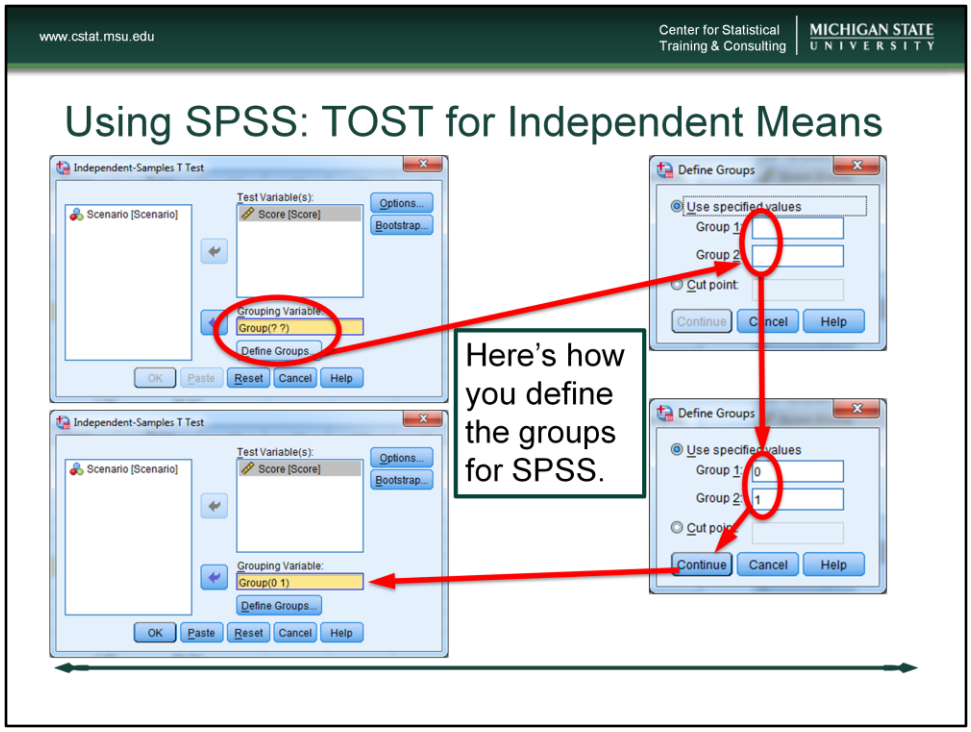
I'm using simulated data for this demonstration, but you would of course collect real data. The files you would need to replicate my demonstration here will be bundled with the slides I post online after the presentation. So, now I'll demonstrate how to analyze the data with both SPSS and R.



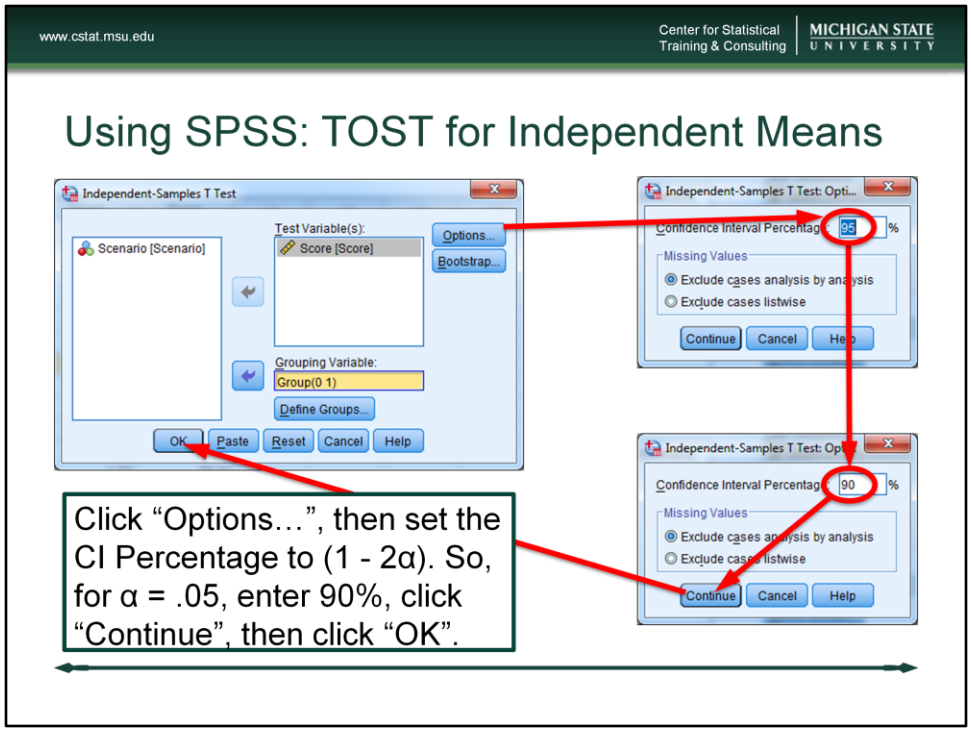
You would set up the data file in SPSS using the same structure you would use for a traditional t-test. So, it would need to have one row per trainee, and at a minimum, you would need two variables. One variable would record the Group assignments, and one would record the trainee's test scores. Here I've coded group as 0 for control and 1 for treatment.



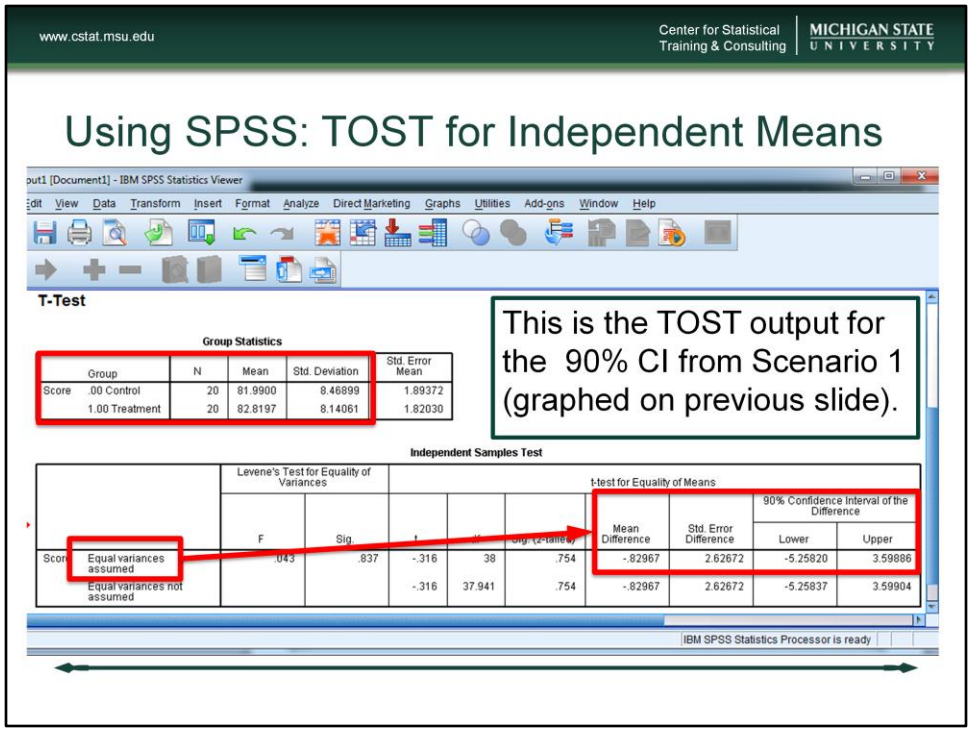
To actually run the TOST analysis, you can actually use the t-test procedure, because it provides an option to get the confidence interval for the difference between the means for the two groups. So, you would click the Analysis menu, select Compare means, and then click on Independent samples t-test.



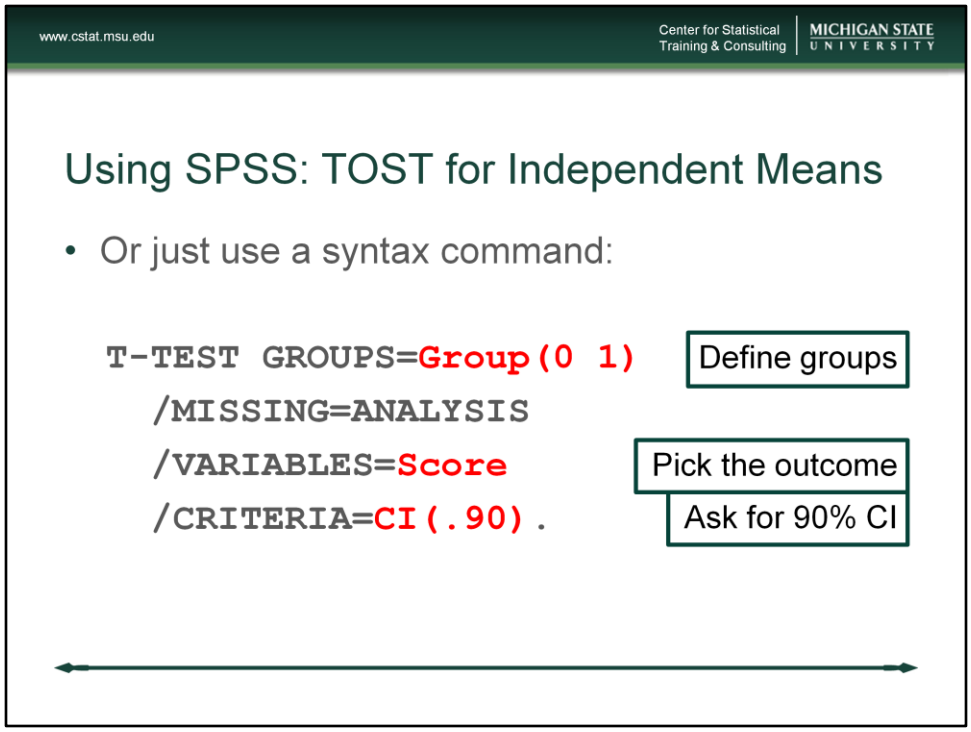
This is the dialog box you get when you do that. Next, you need to define the groups, so after you move the Scores variable to the "Test variables" box and the Group variable over to the "Grouping variable" box, you need to click on the Define Groups button, fill in the values that identify the groups you want to compare, then click Continue. Notice that the Grouping variable box now shows the group numbers.



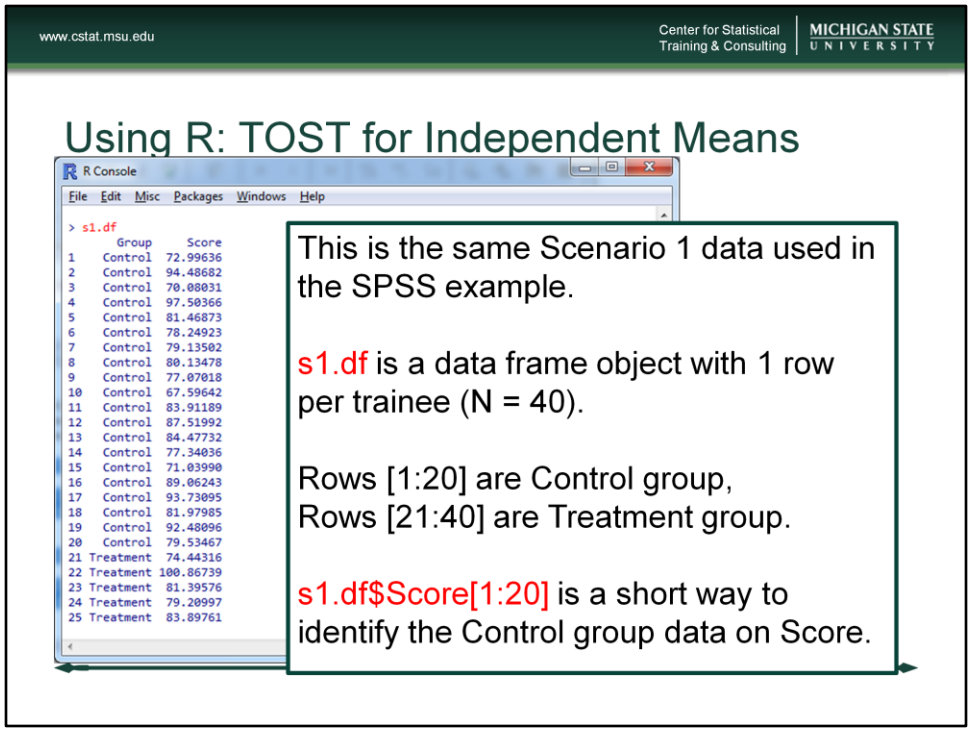
Next, you need to use the Options button. Click that, then change the value in the Confidence Interval Percentage box from the default of 95% to 90%. That's how we get the 90% CI we need for a TOST equivalence test that uses two one-sided tests, each at a Type 1 error rate of 5%. Click Continue to get back to the main dialog box, then click OK to run the analysis.



Here's the output you would get from the Scenario 1 data I've created. In the upper left corner, notice that you get the group means and standard deviations. Since the standard deviations are very similar to each other, you want to focus on the row labeled "Equal variances assumed" in the bottom table. On the right side of that lower table, you can see that we now have the estimated difference between the means, its standard error, and, most important of all, the lower and upper limits of the 90% CI. To finish the TOST analysis, you just need to see if this entire CI falls within the equivalence interval that runs from -15 to +15. It does in this case, so you could tell your client that the new training program is equivalent to the old one with respect to trainee test scores.



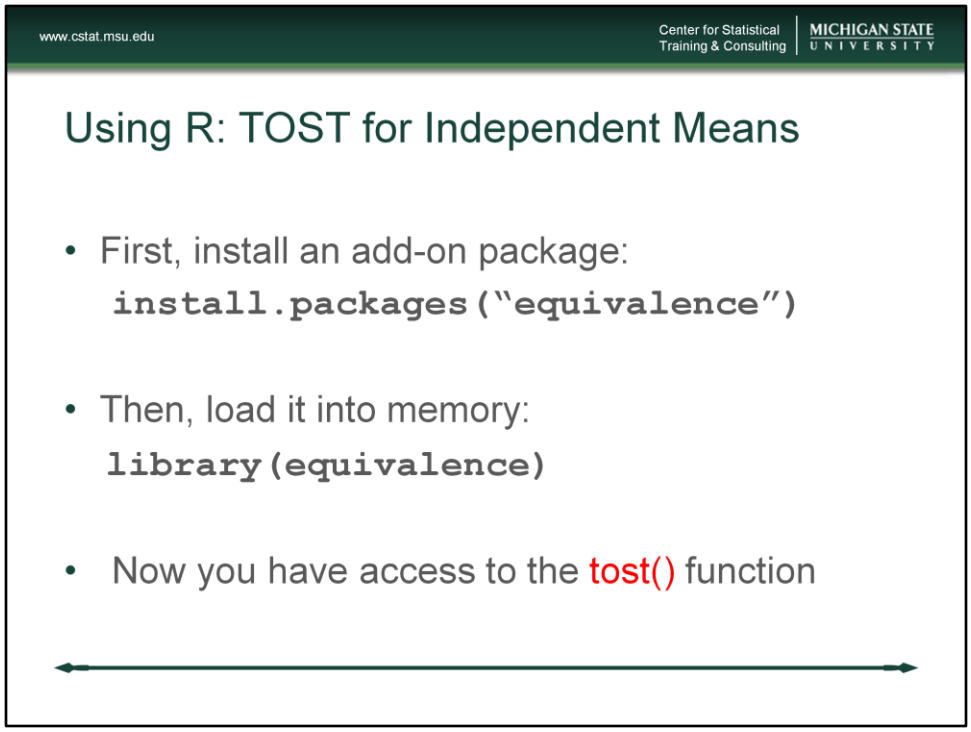
Although I just demonstrated how to do the TOST analysis with the menus and dialog boxes in SPSS, I prefer to just use the syntax language to run analyses. Here's how you would do the analysis that way. On the first line, you're telling SPSS to use the t-test procedure to compare the two groups of scores associated with values of zero and one in the Group variable. In the 3rd line, you're telling SPSS what outcome variable to use. Finally, the crucial piece for getting the right CI is on the 4th line, where you tell it you want the 90% CI instead of the default 95% CI.



Now let's look at how you would do the same analysis in R, which is high-quality, free, open-source statistical computing software that is rapidly becoming very popular. It's a package that largely requires you to write scripts or syntax files to do your work, but I have found it to be an excellent tool.

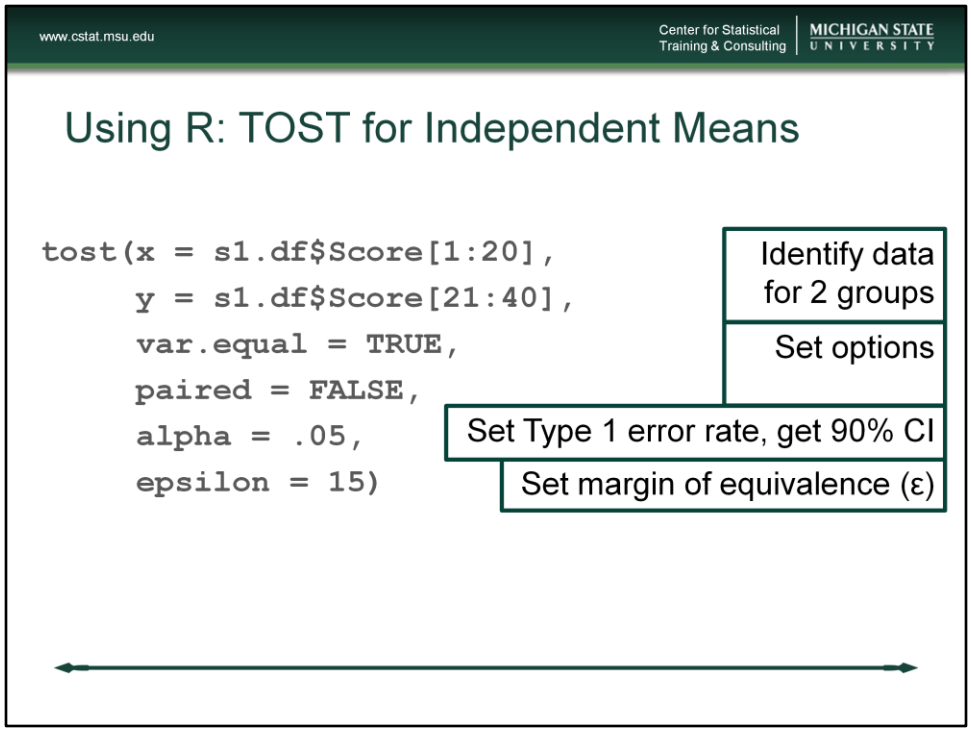
I'm going to assume that if you're trying to use R, you know at least the basics of that package. There are lots of good intro tutorials about using R, so, I'm skipping some steps here and assuming that you already have your data entered into what R calls a data frame. Here, I called the data frame s1.df and have displayed the first 25 records or so on the screen by typing the name of the data frame into the R console. You can see that it has just 2 variables in it: Group and Score. It also has one row per trainee, with the first 20 rows being the Control group and the last 20 rows being the Treatment group. This is the same data I was using in SPSS.

It helps to know that in R, you can select a subset of values for a particular variable by writing the data frame name and a dollar sign, followed by the variable name. You can then identify the rows you want within a set of square brackets. Here I've demonstrated how to select the Control group data for the Score variable.

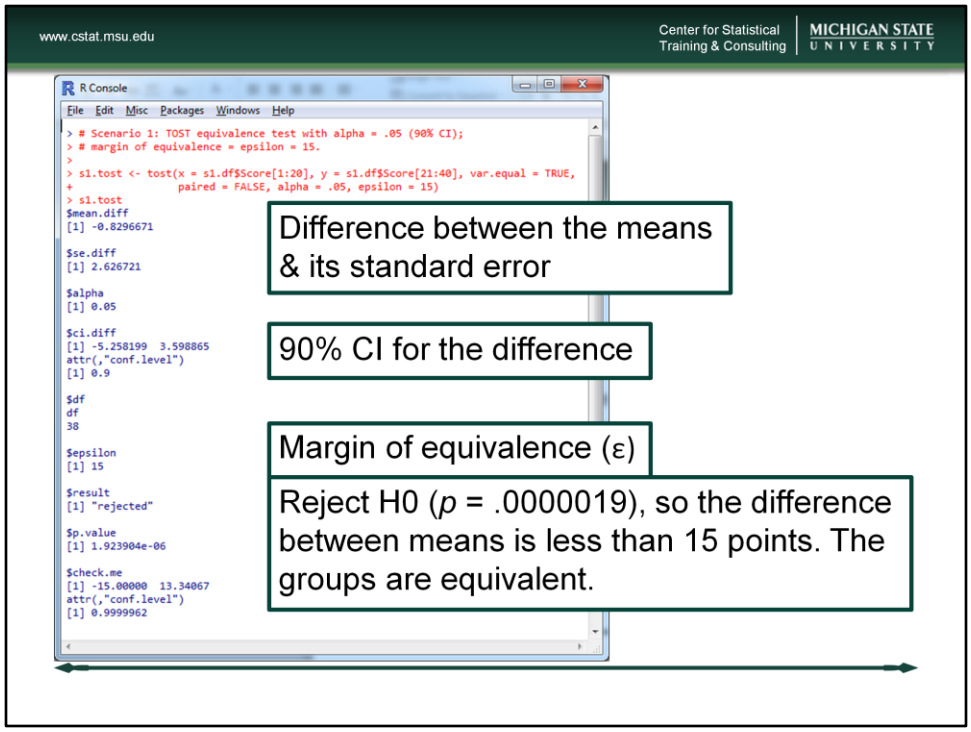


The base version of R doesn't do equivalence testing, so you will need to install an add-on package of additional commands before you can use the analysis method I'm about to demonstrate. Fortunately R makes it very easy to install any of the thousands of additional packages people have created, all of which are also free, open-source software. That package I'm using is conveniently called "equivalence". To install it, you use the install.packages command and put the name of the desired package inside the parentheses, surrounded by quotes. You should only need to install the package once for any given version of R.

After you've installed the package , you need to tell R to load it into memory. You can do that with the library function. Just put the name of the package – without quotes – inside the parentheses. You would need to do this once per R session. Once it's loaded, this package provides a library of new functions, including the tost() function that I'm about to demonstrate.

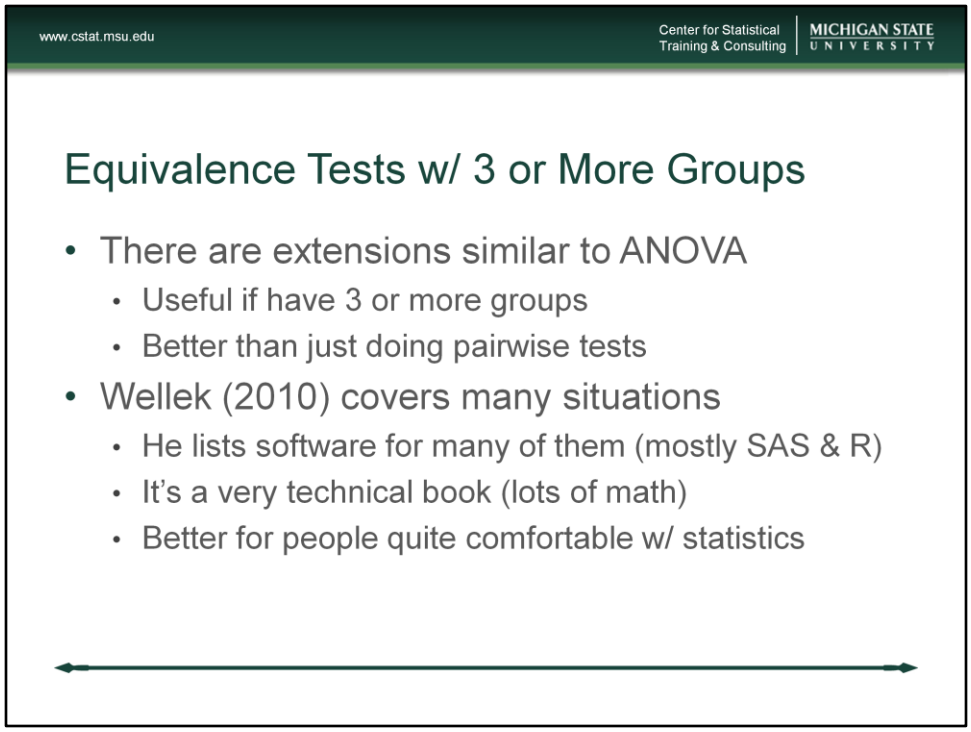


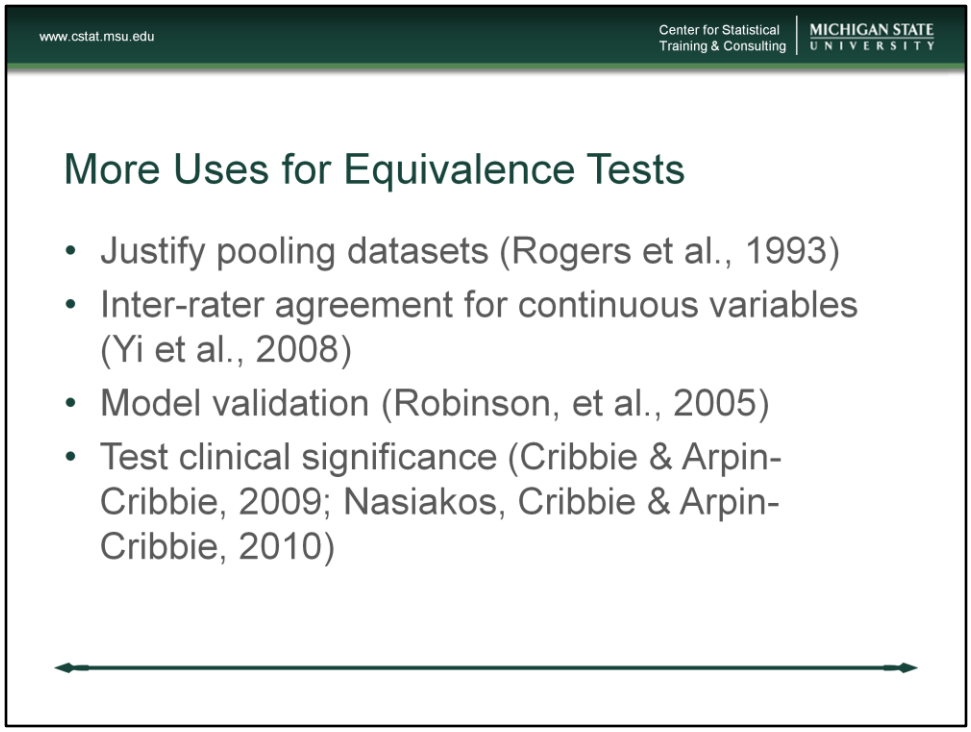
Once you have your data in a data frame and ther package loaded, you are ready to actually do the TOST analysis. You do that by calling the tost function. Inside the function, the inputs labeled x and y are used to identify the subset of data belonging to the two groups. Here x refers to the Control group Scores, and y refers to the Treatment group Scores. The var.equal line tells the R to assume that the groups have equal variances, while the paired line tells R that these are not paired groups (so it should treat them as independent). The alpha line sets the Type 1 error rate for each of the two one-sided tests, thereby ensuring that you get the 90% CI on the difference between the means. Finally, the epsilon line sets the margin of equivalence.



The output is just a list of different pieces of information. At the top, you see the estimated difference between the group means and it's standard error. After that the output displays the alpha level you selected, then shows you the 90% CI for the difference between the means, followed by the degrees of freedom and the margin of equivalence you had selected.

It also shows you the statistical conclusion with respect to the null hypothesis that the group means differ by a non-trivial amount, and a corresponding p-value. In this case H0 is rejected, because the p-value is very, very small. You can therefore conclude that the groups are equivalent with respect to test scores.


www.cstat.msu.edu
Center for Statistical Training & Consulting
MICHIGAN STATE UNIVERSITY
Equivalence Tests w/ 3 or More Groups
There are extensions similar to ANOVA
Useful if have 3 or more groups
Better than just doing pairwise tests
Wellek (2010) covers many situations
He lists software for many of them (mostly SAS & R)
It's a very technical book (lots of math)
Better for people quite comfortable w/ statistics



To give you a quick preview of other uses that people have identified for equivalence tests, I put together this list. These papers are listed in the resources handout I'm providing, along with some papers and books focused on the basic applications and issues.

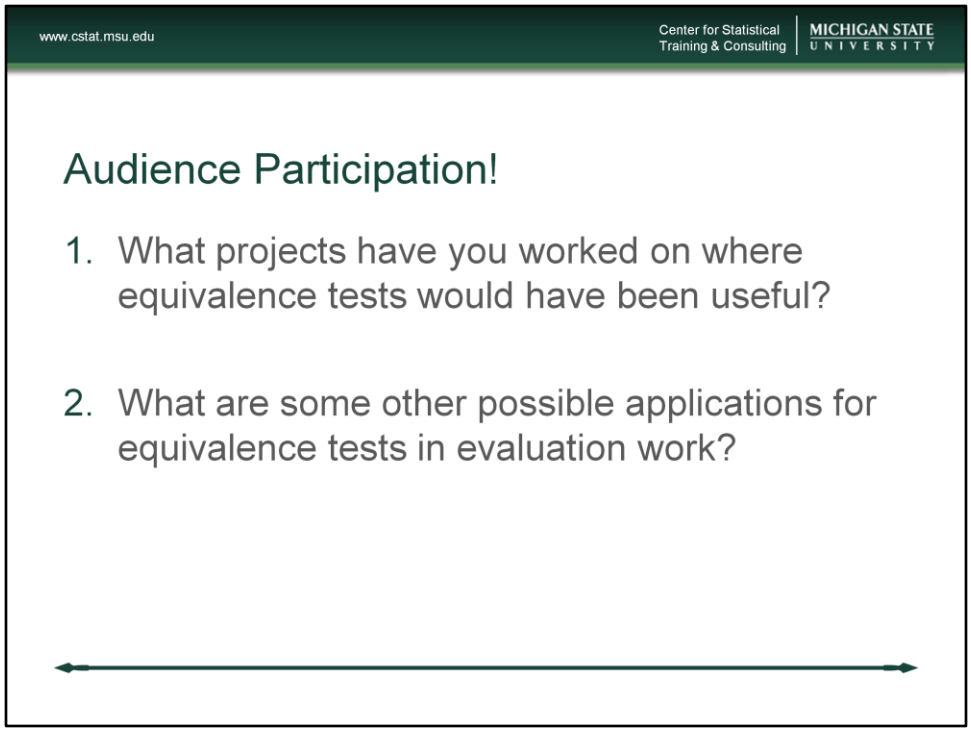
Rogers and colleagues suggested you could use equivalence testing to justify pooling different datasets.

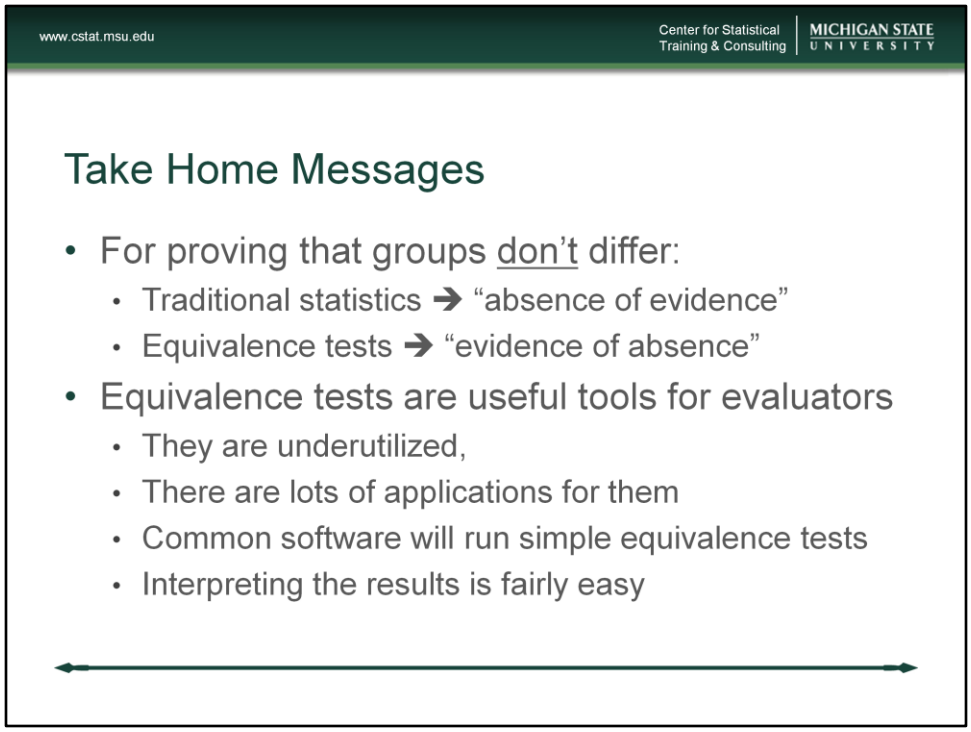
YI and colleagues describe a way to use equivalence tests to assess interrater agreement on continuous measurements, which is not quite the same thing as inter-rater reliability.

Robinson and colleagues talk about how you can use equivalence tests to validate the fit of regression models.

Finally, Cribbie and colleagues talk about assessing the clinical significance of psychotherapy by using equivalence tests to test whether patients are functioning equivalently to normal persons after the conclusion of the therapy intervention.

Those are just the things I found so far. There are probably many other uses as well.

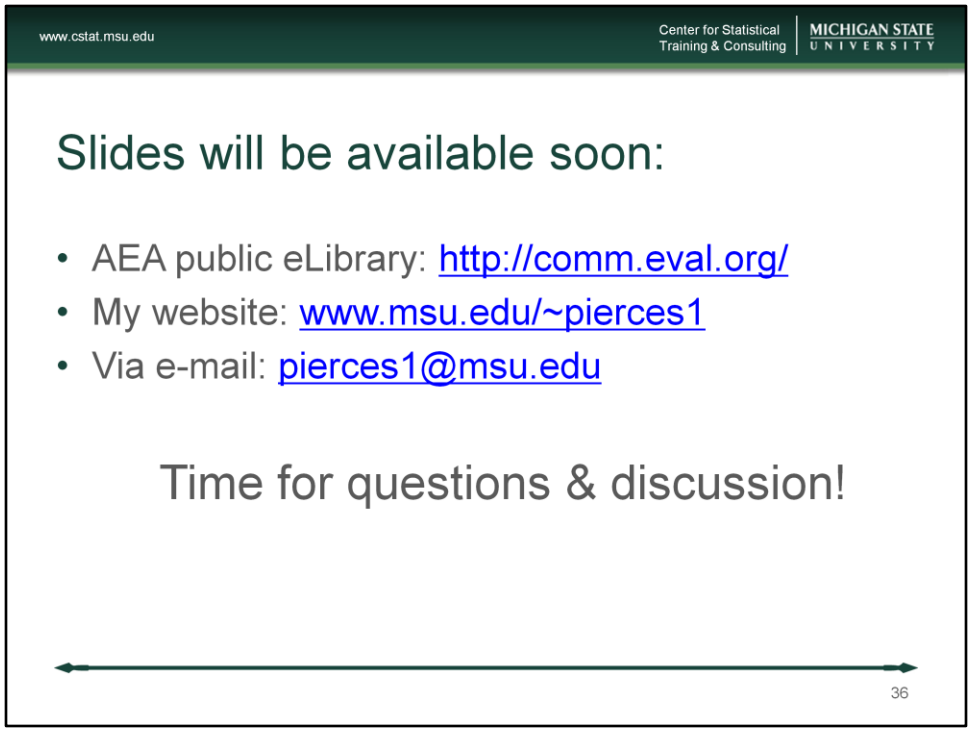

www.cstat.msu.edu
Center for Statistical Training & Consulting
MICHIGAN STATE UNIVERSITY
Audience Participation!
1. What projects have you worked on where equivalence tests would have been useful?
2. What are some other possible applications for equivalence tests in evaluation work?



So here are the simple take-home messages from this talk. For the basic purpose of proving that two groups don't differ, traditional statistics fail because they produce only an "absence of evidence", but equivalence tests can solve that problem because they produce "evidence of absence" with respect to group differences.

Broadly speaking, I think equivalence tests are useful tools for evaluators, but that they are underutilized so far. There appear to be lots of applications for them, and common software will run simple equivalence tests quite readily. The results are not difficult to interpret, so I encourage you all to consider how you can integrate them into your own work.

Thanks you!


www.cstat.msu.edu
Center for Statistical Training & Consulting
MICHIGAN STATE UNIVERSITY
Slides will be available soon:
• AEA public eLibrary: http://comm.eval.org/
• My website: www.msu.edu/~pierces1
• Via e-mail: pierces1@msu.edu
Time for questions & discussion!
36