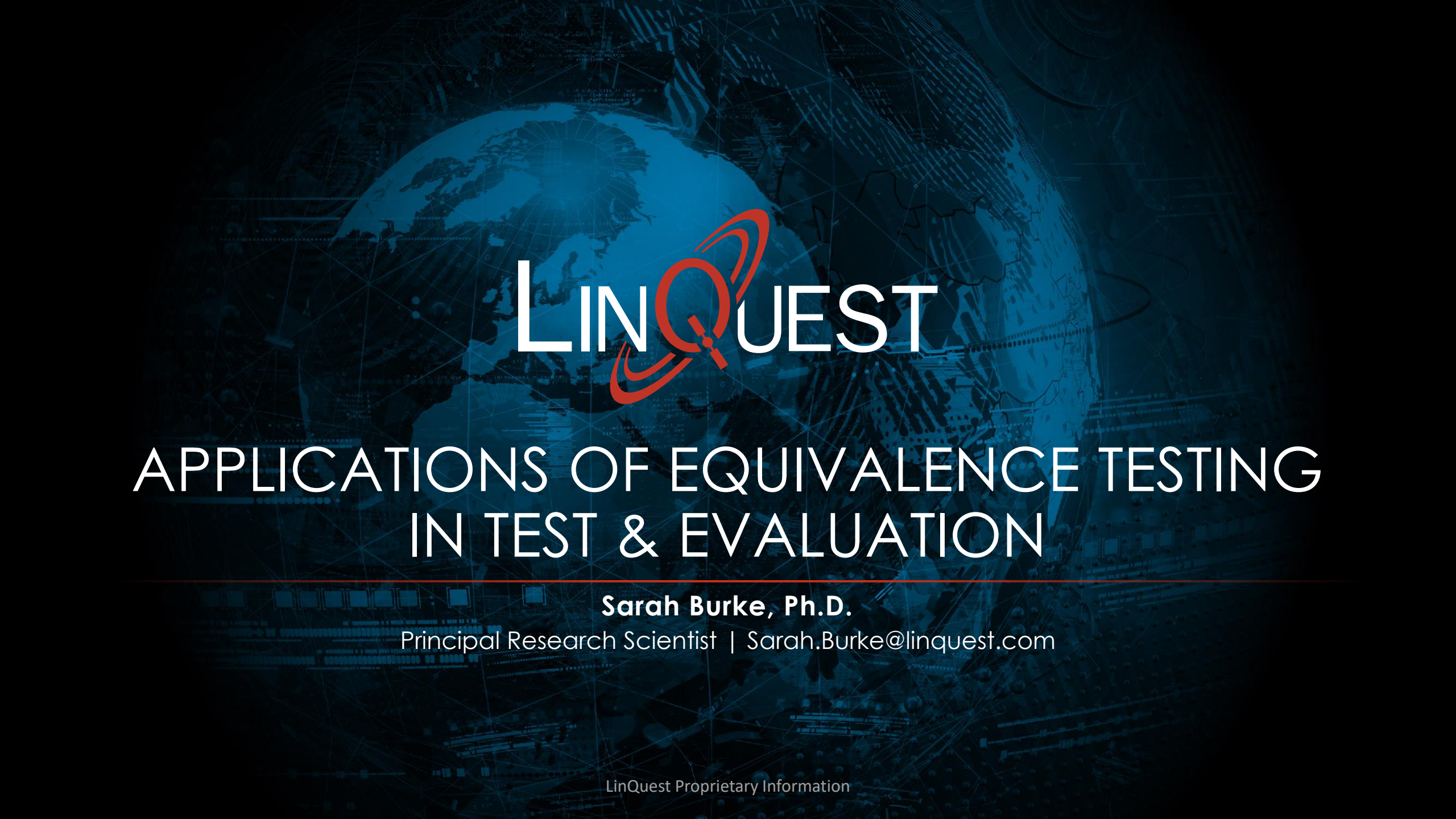


APPLICATIONS OF EQUIVALENCE TESTING IN TEST & EVALUATION

Sarah Burke, Ph.D.
Principal Research Scientist | Sarah.Burke@linquest.com

OUTLINE

- Traditional hypothesis testing vs equivalence testing
  - Mechanics of equivalence testing
  - Relationship between hypothesis tests and equivalence tests
- Examples
  - M&S Validation
  - Training Comparison
- Considerations
  - Equivalence Criterion
  - Power
  - Noninferiority and Superiority

MOTIVATION

- Confirm that results from a model or simulation are similar to live test data
- Prove that an updated training program has not affected performance of trainees
- Verify that a change in parts has not affected the performance of the system
- Show that the efficacy of a generic drug is equivalent to a brand-name drug

In each case, the goal is prove no change/difference

TRADITIONAL HYPOTHESIS TESTING

- Traditional hypothesis test for two population means:

  H_0: Group 1 mean and Group 2 mean are the same

  H_1: Group 1 mean and Group 2 mean are different

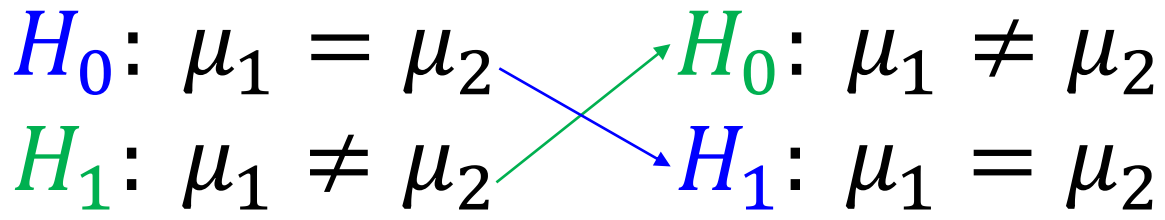
- Possible decisions:
  - p-value $\leq \alpha$: Reject H_0; Conclude the two means are different
  - p-value $> \alpha$: Fail to Reject H_0; Conclude there's not sufficient evidence that the means are different

What conclusion do you really want to make?

Absence of evidence does not imply evidence of absence

EQUIVALENCE TESTING

- If you want to prove that two means are similar, you cannot fail to reject the null hypothesis that they are equal and conclude the means are equal
- So reverse the hypotheses
- Assume that the means are different and gather evidence to prove they are the same



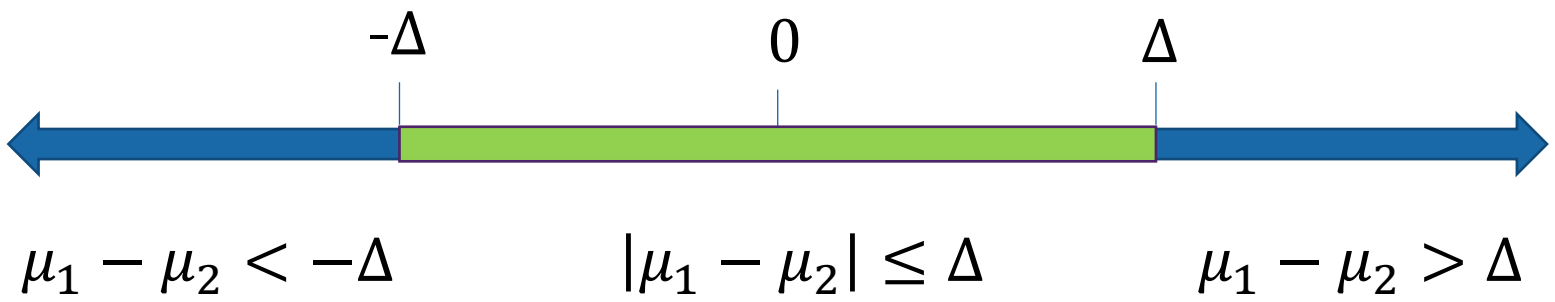
Traditional Hypothesis Test Equivalence Test

EQUIVALENCE TESTING

- How close is close enough to be considered the same?
- If we assume that the two means are different, then the difference of means that will allow you to reject the null hypothesis is Δ

$$H_0: |\mu_1 - \mu_2| > \Delta$$

$$H_1: |\mu_1 - \mu_2| \leq \Delta$$



MECHANICS OF EQUIVALENCE TESTING

- Two One-Sided Tests (TOST)

$$H_0: \mu_1 - \mu_2 \leq -\Delta$$

$$H_1: \mu_1 - \mu_2 > -\Delta$$

$$\text{AND}$$

$$H_0: \mu_1 - \mu_2 \geq \Delta$$

$$H_1: \mu_1 - \mu_2 < \Delta$$

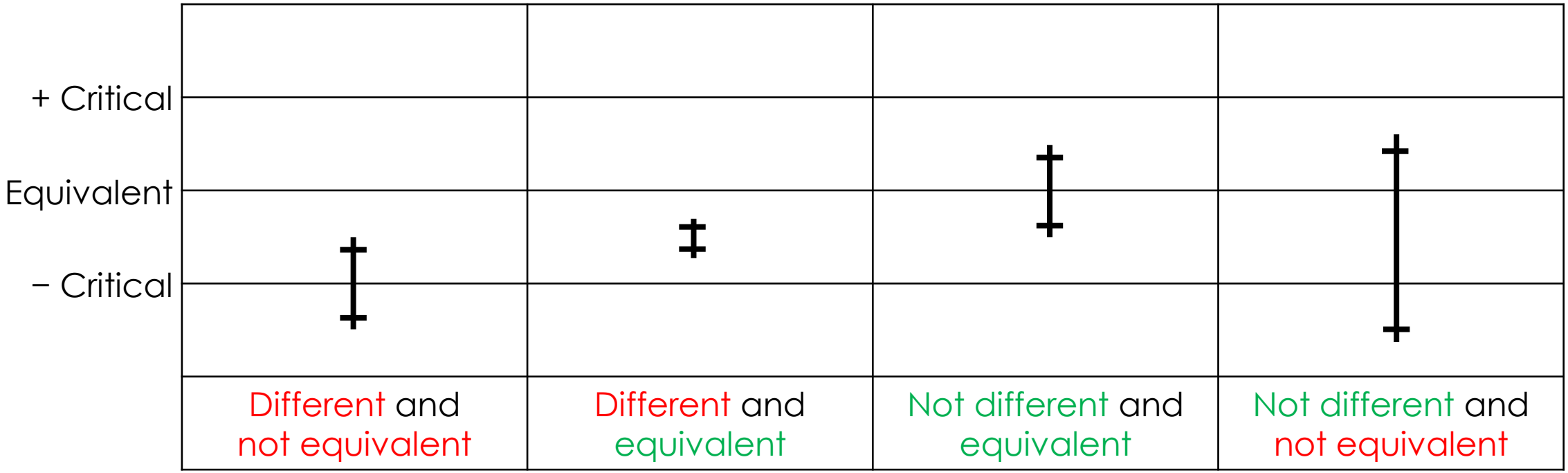
- If both null hypotheses are rejected, population means are practically equivalent
- If one or none are rejected, the two means may not be equivalent

MECHANICS OF EQUIVALENCE TESTING

- The two group means are equivalent if a 100(1-2α)% confidence interval of $\mu_1 - \mu_2$ lies completely between $[-\Delta, \Delta]$
- If both one-sided tests are rejected, the CI will be contained in the $\pm\Delta$ interval, and we can conclude equivalence

- Note: the CI yields an α-size equivalence test, not a 2α level test because we do two α-level one-tailed tests

RELATIONSHIP BETWEEN HYPOTHESIS TESTS AND EQUIVALENCE TESTS



High power test
(large n or small σ)

Low power test
(small n or large σ)

Figure adapted from Allen and Seaman, *Different, Equivalent or Both?* Quality Progress, 2006

Examples

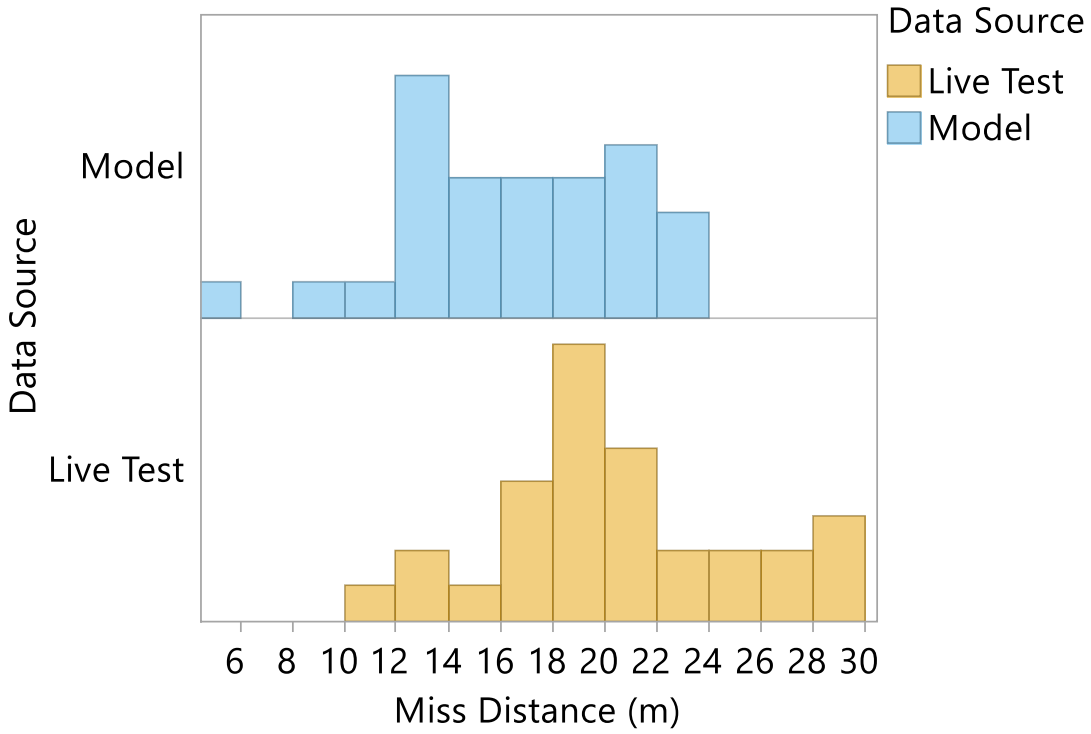
EXAMPLE: MODEL VALIDATION

- A weapon system program developed a model to estimate miss distance (in meters)
- Program then executed live testing and compared to the model results
- SMEs agree that a difference of 5 meters would not have an operational impact

EXAMPLE: MODEL VALIDATION

$$H_0: \mu_{model} - \mu_{live} < -5 \text{ OR } \mu_{model} - \mu_{live} > +5$$

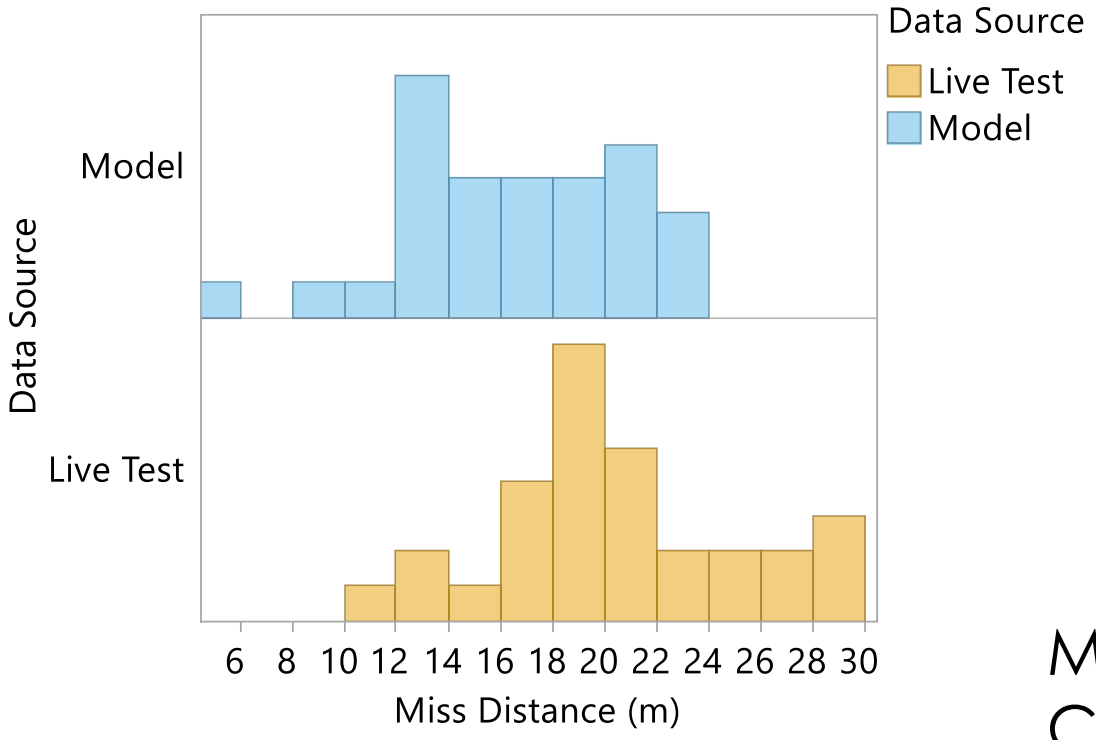
$$H_1: -5 \leq \mu_{model} - \mu_{live} \leq +5$$



All data are notional

EXAMPLE: MODEL VALIDATION

Observed Difference: 4.38 meters



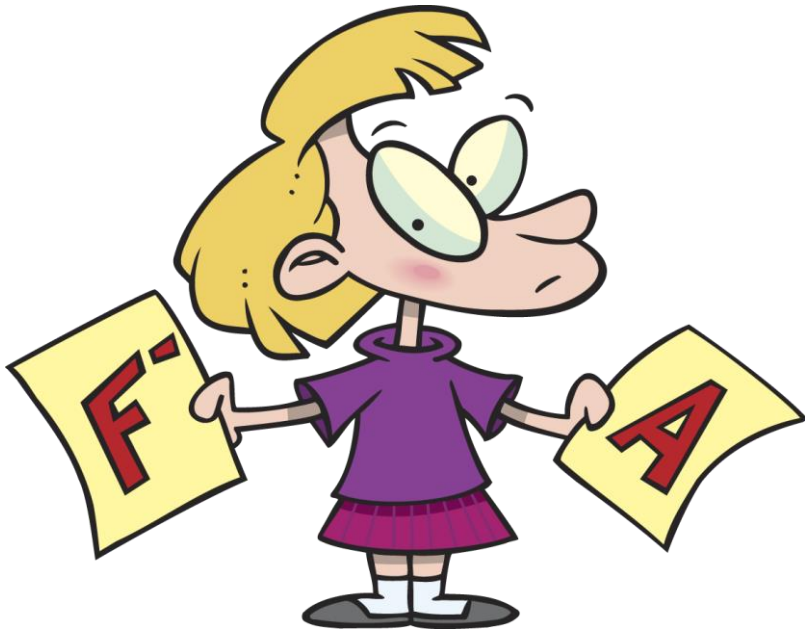
Lower Bound p-value	<0.0001
Upper Bound p-value	0.2959
90% Confidence Bounds	[2.44, 6.31]

Max p-value > 0.05
Cannot conclude the model and live test results are practically equivalent

All data are notional

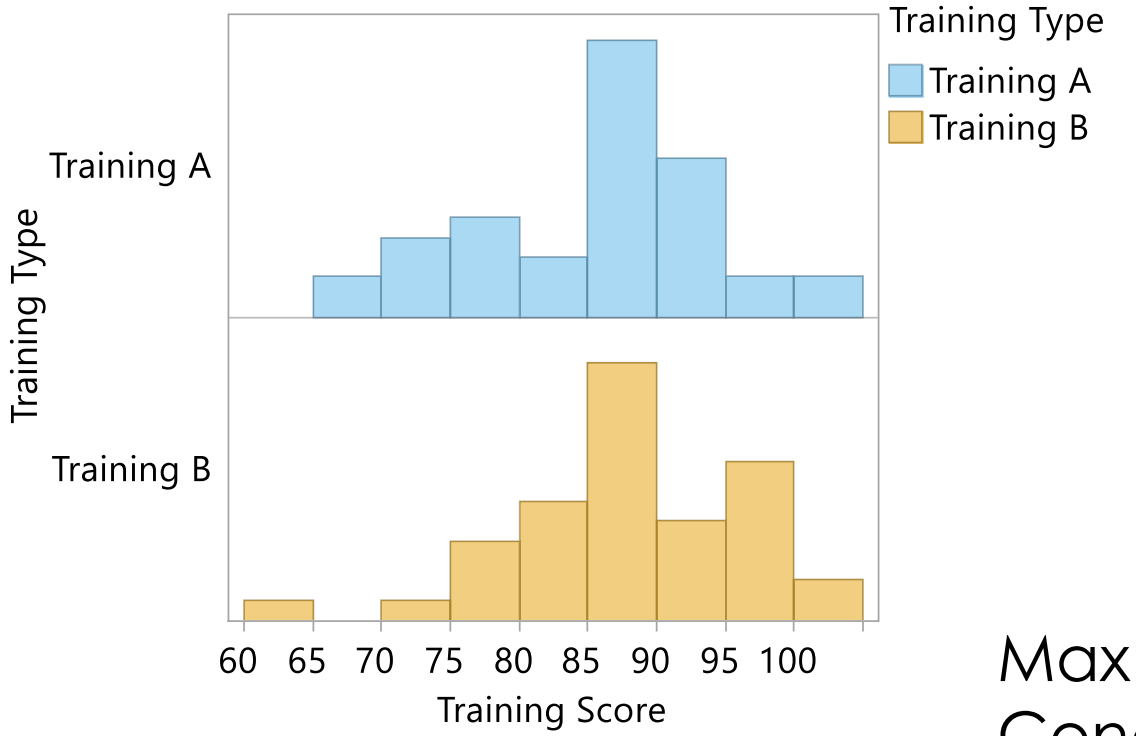
EXAMPLE: TRAINING

- A program has implemented a new training syllabus and must confirm the new syllabus has not negatively affected student performance
- Student scores are measured on a scale from 0 to 100
- Instructors determine that scores within 10 are considered equivalent



EXAMPLE: TRAINING

Observed Difference: 2.03 points



Lower Bound p-value	<0.0001
Upper Bound p-value	<0.0001
90% Confidence Bounds	[-1.14, 5.19]

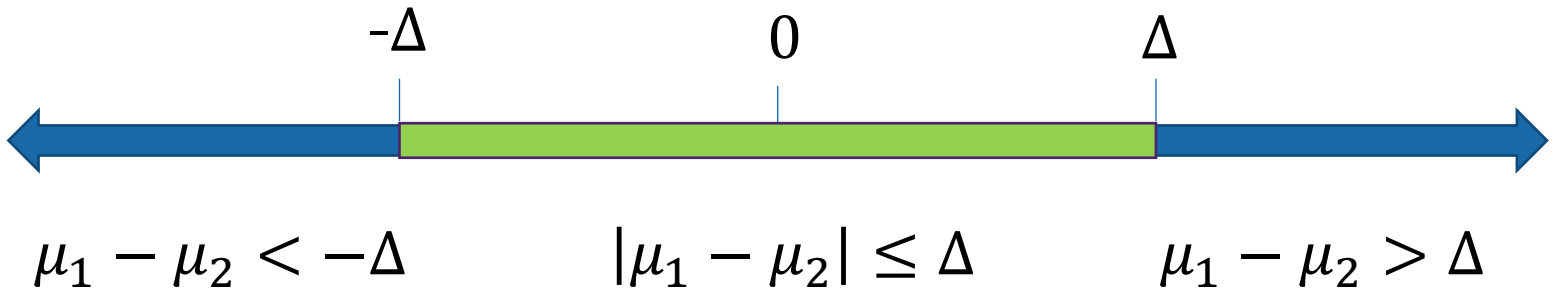
Max p-value < 0.05
Conclude the mean training scores are practically the same for both syllabi

All data are notional

Considerations

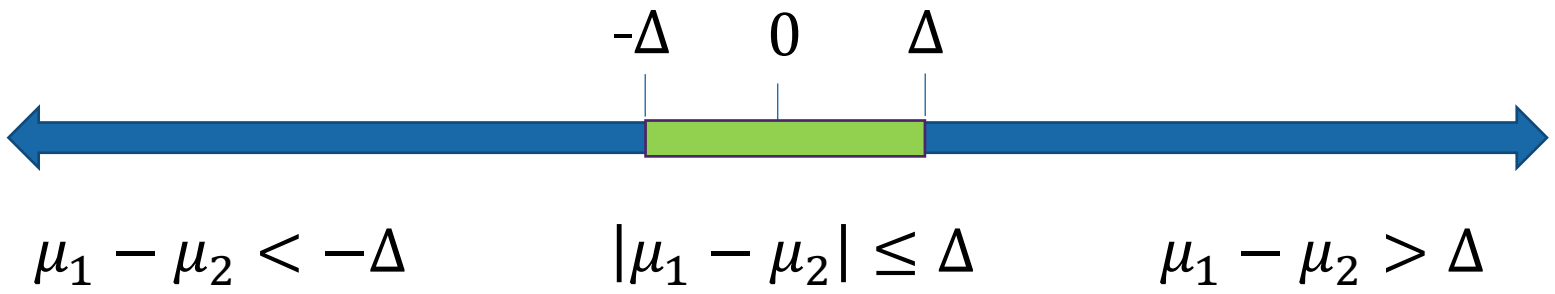
HOW TO DETERMINE Δ

- Δ is the "equivalence acceptance criterion"
  - Level at which the population means are close enough to be acceptable
  - If limited information, could use % difference from control group (e.g., 5-20%)
  - May be set by regulatory guidelines or requirements



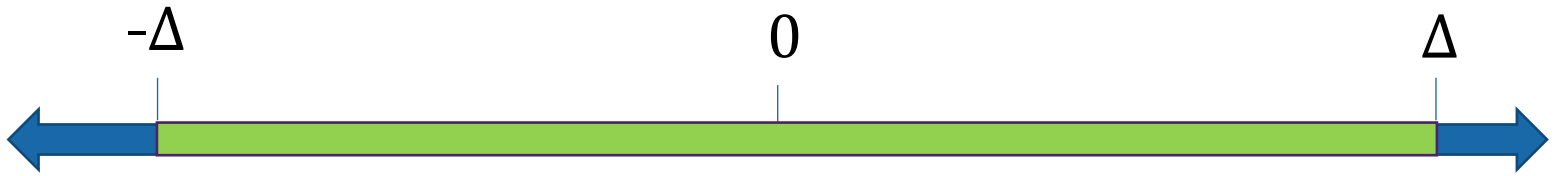
HOW TO DETERMINE Δ

- What does "close enough" mean for your application?
- Should be agreed upon by (PRIOR to data analysis):
  - Subject matter experts
  - Decision-makers
  - Operators



HOW TO DETERMINE Δ

- What does "close enough" mean for your application?
- Should be agreed upon by (PRIOR to data analysis):
  - Subject matter experts
  - Decision-makers
  - Operators

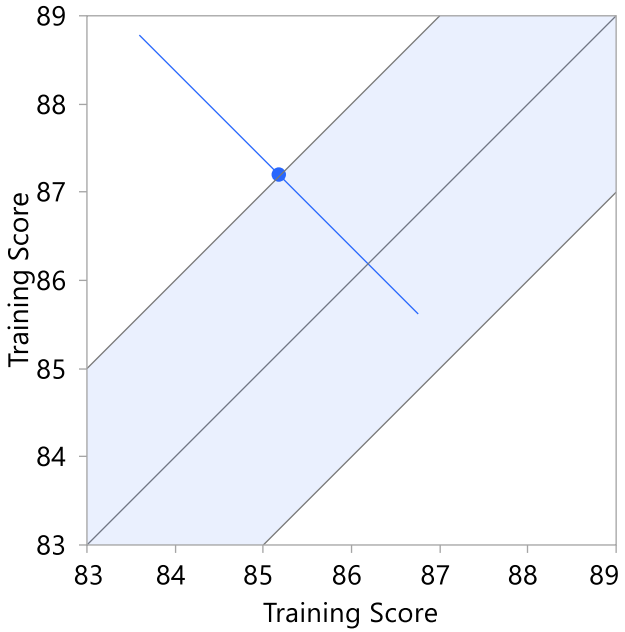


$\mu_1 - \mu_2 < -\Delta$ $\quad |\mu_1 - \mu_2| \leq \Delta$ $\quad \mu_1 - \mu_2 > \Delta$

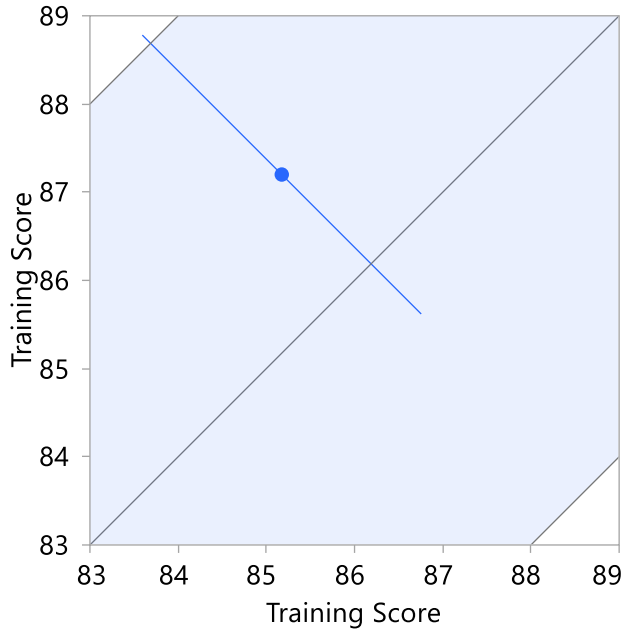
EFFECT OF Δ ON RESULTS

- How would our decision on the training results change for different values of Δ?

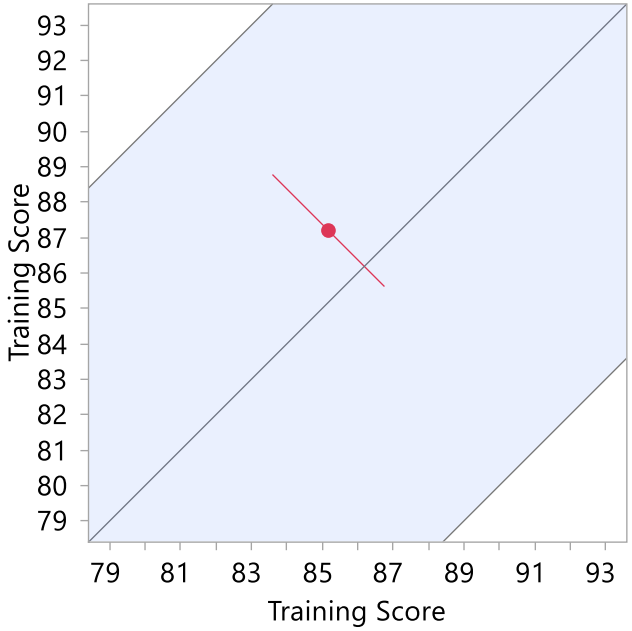
$\Delta = 2$



$\Delta = 5$



$\Delta = 10$

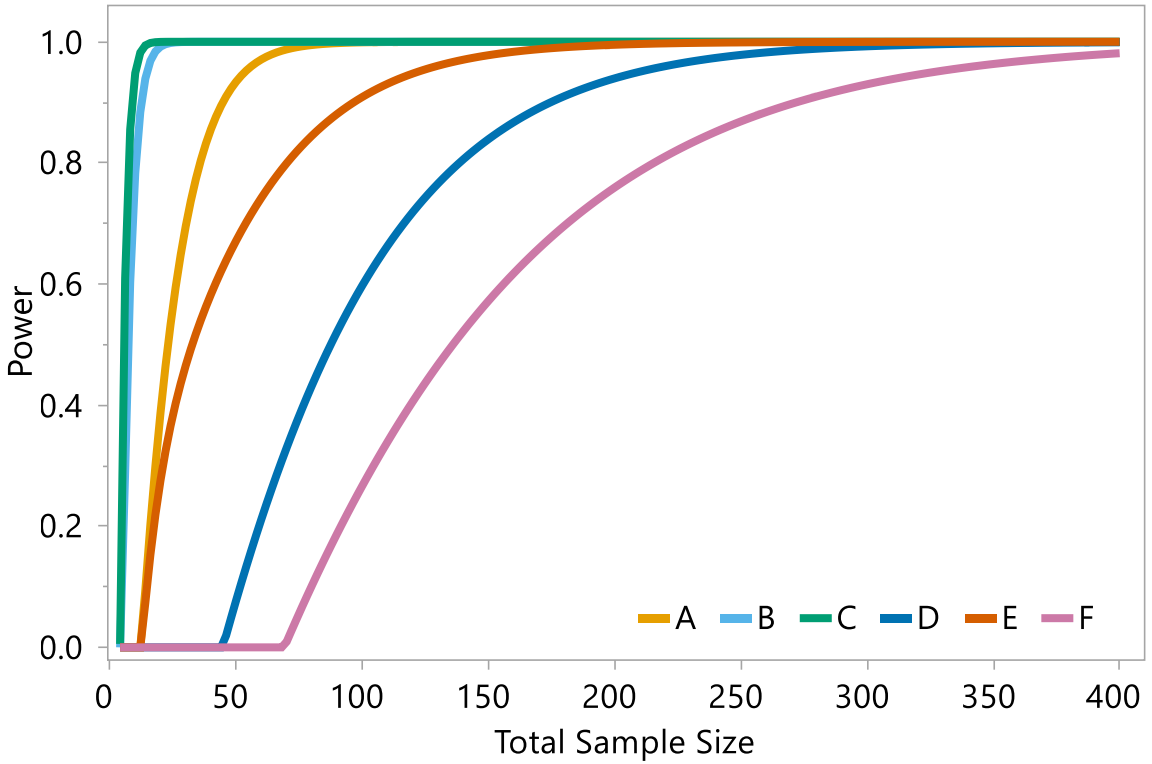


Select equivalence criterion PRIOR to data collection & analysis

SAMPLE SIZE AND POWER

- Just like traditional hypothesis testing and/or DOE, you should perform power analysis to estimate required sample size
- Power of equivalence testing based on non-central t-distribution and requires estimates of:
  - Equivalence criterion
  - Actual difference
  - Group 1 standard deviation (s1)
  - Group 2 standard deviation (s2)

SAMPLE SIZE AND POWER



Label	s1	s2	Delta	Observed Difference
A	5	5	5	0
B	5	5	10	0
C	2	2	5	0
D	10	10	5	0
E	5	5	5	2
F	5	5	2	0

- Explore sensitivity of power estimates across values of standard deviation, equivalence criterion, and true difference
- Easy to determine equivalence when noise is low and/or equivalence criterion is large

EQUIVALENCE, SUPERIORITY, & NONINFERIORITY

- Noninferiority
  - Results are not practically **worse**
- Superiority
  - Results are practically **better**
- Equivalence
  - Results are within practical range

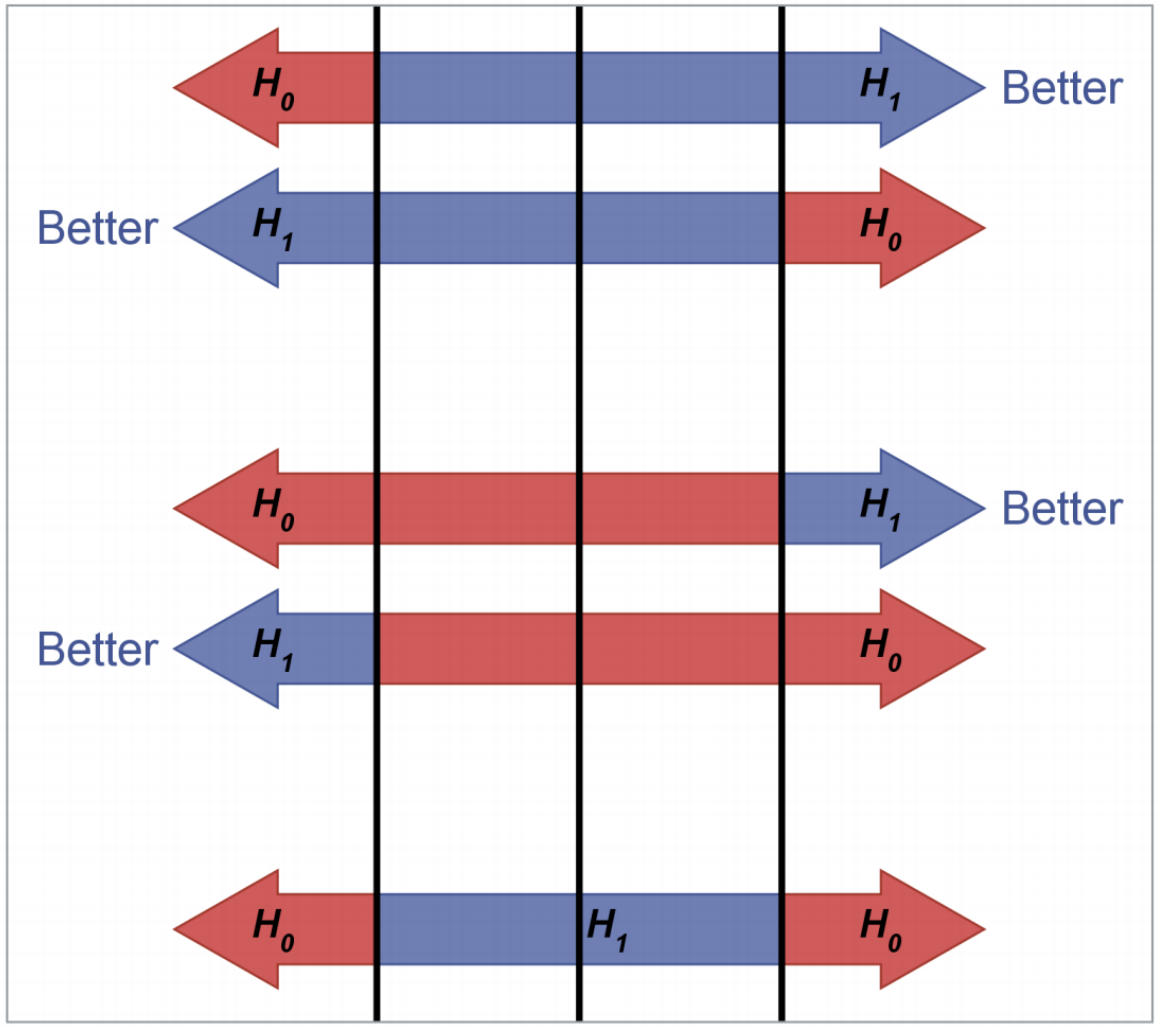


Figure Source: https://support.sas.com/resources/papers/proceedings15/SAS1911-2015.pdf

NONINFERIORITY TESTS

- Suppose larger values are "better"
- Define $\delta > 0$ as the largest acceptable difference (noninferiority margin)
- Hypotheses:

$$H_0: \mu_1 - \mu_2 \leq \theta_0 - \delta$$
$$H_1: \mu_1 - \mu_2 > \theta_0 - \delta$$

- θ_0 represents a requirement, standard, or threshold
- δ is the "close enough" fudge factor

Figure Source: https://support.sas.com/resources/papers/proceedings15/SAS1911-2015.pdf

VARIATIONS

- Focused on comparing two independent means here…
- Other applications:
  - Single mean against a standard
  - Standard deviation against a standard
  - Two independent standard deviations
  - Linear regression model coefficients

CONCLUSIONS

- There are many applications of equivalence testing in T&E
- Consider what decision you need to make:
  - Group means are different?
  - Group means are similar?
- Consider the equivalence criterion and sample size during test planning

Questions?