# Assessing the Correspondence of Results in Replication Studies

***Peter M. Steiner***
University of Wisconsin-Madison

***Vivian C. Wong***
University of Virginia, Charlottesville

*Open Science, Trier, March 12-14, 2019*

Supported by NSF grant #2015-0285-00

# Motivating Examples

# Example 1

## Study 1

Effect estimate: 12 pts
s.e.: 4 pts
→ sig. positive effect

## Study 2

Effect estimate: 11 pts
s.e.: 7 pts
→ insig. effect

| Correspondence measure | | Results replicate? |
|---|---|---|
| Significance pattern: | different | → no |
| Difference test: | insignificant | → yes |
| Equivalence test: (threshold of 3 pts) | insignificant | → no |

# Example 2

| Study 1 | | Study 2 | |
|---|---|---|---|
| Effect estimate: | 12 pts | Effect estimate: | 34 pts |
| s.e.: | 4 pts | s.e.: | 7 pts |
| → sig. positive effect | | → sig. positive effect | |

| Correspondence measure | | Results replicate? |
|---|---|---|
| Significance pattern: | identical | → yes |
| Difference test: | significant | → no |
| Equivalence test: (threshold of 3 pts) | insignificant | → no |

# Choice of Correspondence Measures

- Conclusions strongly depend on the choice of a correspondence measure
- Practical issue: ad hoc choices without careful considerations of the replication goal (Anderson & Maxwell, 2016a)
- Only discuss correspondence measures that
  - treat the results of both studies as stochastic (i.e., sampling uncertainty) and
  - use unbiased test

# Correspondence Measures

- Conclusion-based measures
  Would researchers and policy makers draw the *same conclusion* from each of the two studies? That is, has the intervention a meaningful effect or not?
- Distance-based measures
  Is the *difference in effect estimates* small / large enough to claim a significant equivalence / difference?

Anderson & Maxwell (2016a, 2016b), Gilbert et al. (2016), Steiner & Wong (2018), Tryon (2001), Valentine et al. (2011)

# Conclusion-based Correspondence Measures

# Conclusion-based Correspondence

- Direction of effects

  Is the sign of estimates identical?

$$C_C^D = 1[\text{sgn}(\hat{\tau}_I) = \text{sgn}(\hat{\tau}_{II})]$$

- Magnitude of effects

  Do the estimates exceed a certain magnitude?

$$C_C^M = 1[(\hat{\tau}_I \geq \lambda \;\&\; \hat{\tau}_{II} \geq \lambda) \text{ or } (\hat{\tau}_I < \lambda \;\&\; \hat{\tau}_{II} < \lambda)]$$

- Statistical significance pattern

  Is the significance of estimates identical?

$$C_C^S = 1\left[\{\text{sgn}(\hat{\tau}_I) = \text{sgn}(\hat{\tau}_{II}) \;\&\; p_I \leq \alpha \;\&\; p_{II} \leq \alpha\} \text{ or } (p_I > \alpha \;\&\; p_{II} > \alpha)\right]$$

# Conclusion-based Correspondence

Correspondence of results depends on

- Power of study 1 & 2 to detect the true but unknown effect
  - Magnitude of true effect
  - Sample size
  - Error variance
- Direction and magnitude of bias (if any) in study 1 & 2 (e.g., due to attrition, variation in treatment fidelity and instrumentation)

Note: the minimum detectable effect size (MDES) of study 1 & 2 only reflects sample size and error variance, but is unrelated to the unknown true effect!

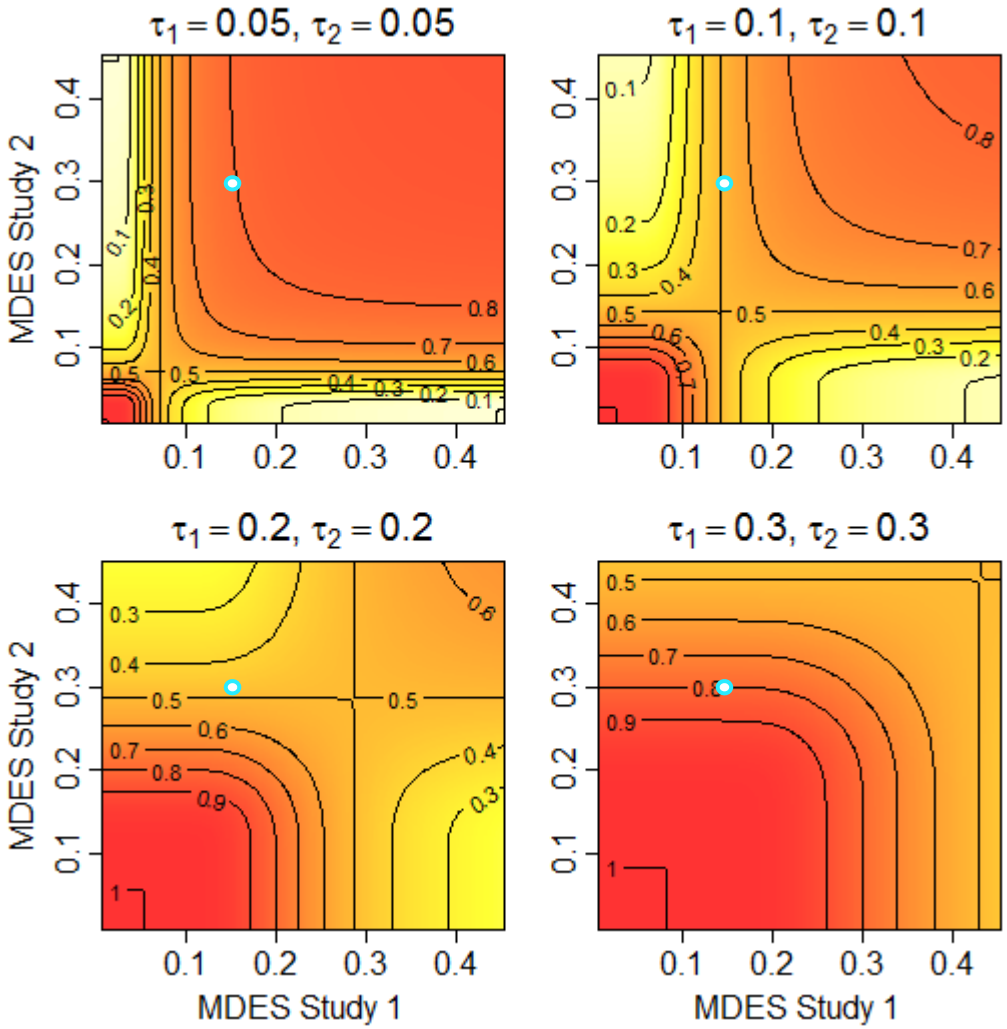
# Probability of Identical Significance Patterns

*True effects are* ***identical*** *across studies*

$$\tau_1 = \tau_2$$

.05, .1, .2, .3 *SD*

- MDES Study 1 = .15
- MDES Study 2 = .30



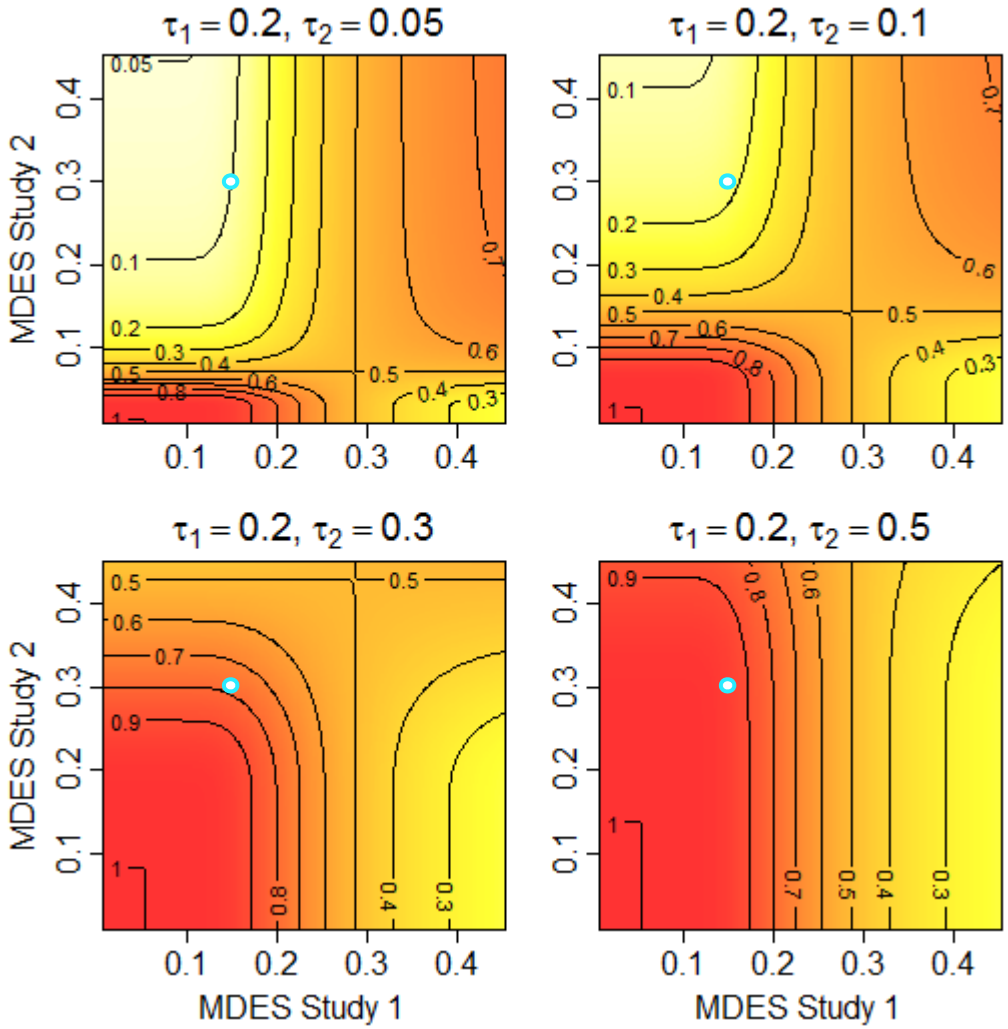
# Probability of Identical Significance Patterns

*True effects* ***differ*** *across studies*

$\tau_1 \neq \tau_2$

$\tau_1$= .2 *SD*

$\tau_2$=.05, .1, .3, & .5 *SD*



# Conclusion-based Correspondence

Since we don't know the true effects,

- it is impossible to know which correspondence probabilities to expect
- even if MDES of both studies would be identical

Thus, conclusion-based measures are often not very informative!

# Distance-based Correspondence Measures

# Distance-based Correspondence

Investigates the estimated effect difference $\hat{\tau}_1 - \hat{\tau}_2$

- Difference test
  Is the difference in estimates insignificant?
  (two-sample *t*-test)
- Equivalence test
  Is the equivalence in estimates significant (with respect to a given equivalence threshold)?
  (Tryon, 2001; Tryon & Lewis, 2008)
- Correspondence test
  Combines the difference and equivalence test
  (Tryon & Lewis, 2008; Steiner & Wong, 2018)

# Difference Test

- Standard null-hypothesis significance test (NHST) -> two-sample *t*-test
- Equivalence of effects is formulated as null hypothesis $H_0$: $\tau_1 - \tau_2 = 0$
- Correspondence requires insignificant *t*-test

$$C_D^D(\alpha) = 1[p_D > \alpha]$$

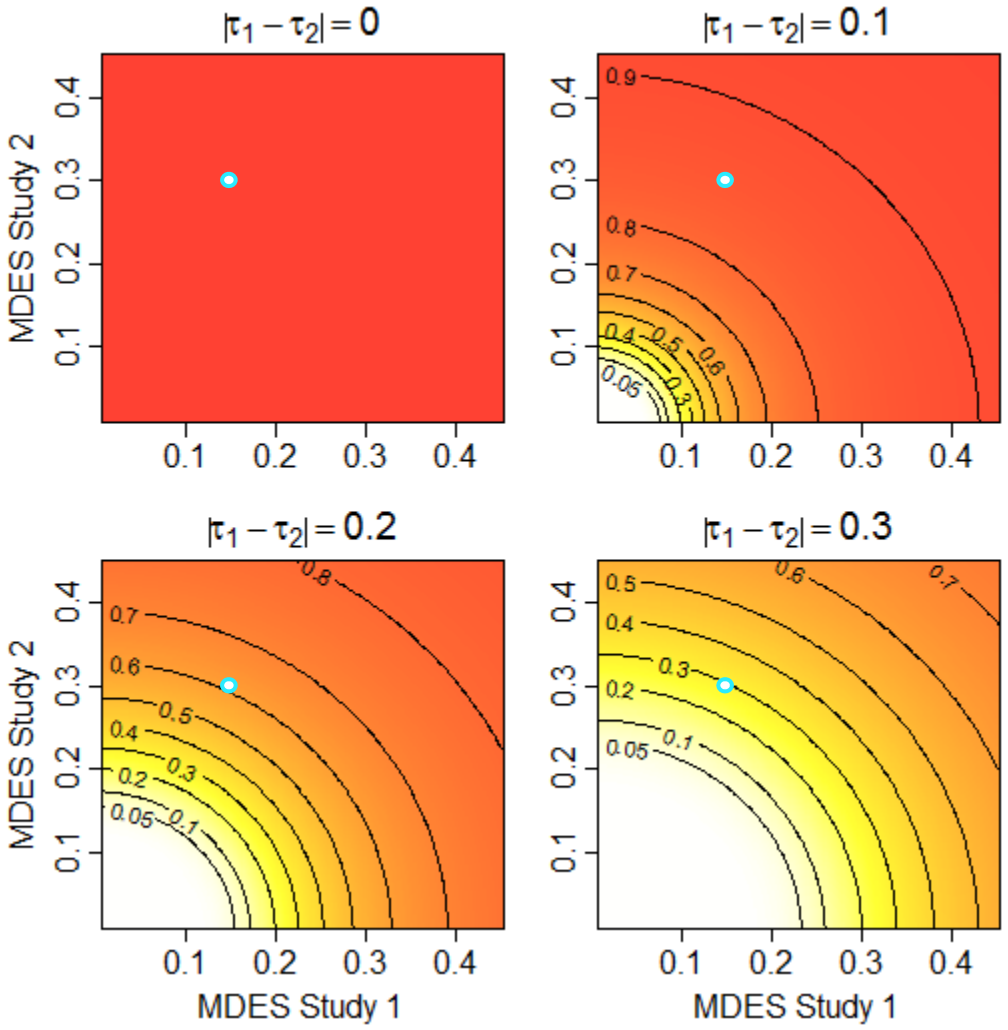
Issues

- Incorrect interpretation of failure to reject the null
- Lack of power may result in failure to reject the false null hypothesis (no difference in effects)
  - occurs even if both studies are sufficiently powered for a given MDES

# Corresp. Probabilities – Difference Test

*True effect difference:*

$|\tau_1 - \tau_2| =$
0, .1, .2, .3 *SD*



# Equivalence Test

- Uses NHST, but tries to overcome the difference test's weakness
- Equivalence is formulated as alternative hypothesis with respect to an equivalence threshold $\delta_E$

$$H_0: |\tau_1 - \tau_2| \geq \delta_E \qquad H_1: |\tau_1 - \tau_2| < \delta_E$$

- The composite null hypothesis can be reformulated as two one-sided hypotheses

$$H_{01}: \tau_1 - \tau_2 \geq \delta_E$$
$$H_{02}: \tau_1 - \tau_2 \leq -\delta_E$$

- Equivalence can be tested with two one-sided *t*-tests
- Correspondence requires two significant t-tests

$$C_D^E(\delta_E, \alpha) = 1[p_{E1}(\delta_E, \alpha) \leq \alpha \;\&\; p_{E2}(\delta_E, \alpha) \leq \alpha]$$

# Equivalence Test

Issues

- Determination of equivalence threshold ($\delta_E$)
  - Effect difference that is substantively inconsequential or trivial
  - Small thresholds (.1 SD or smaller) require large sample sizes in each study
- Lack of power may result in a failure to reject the false null hypothesis (difference)

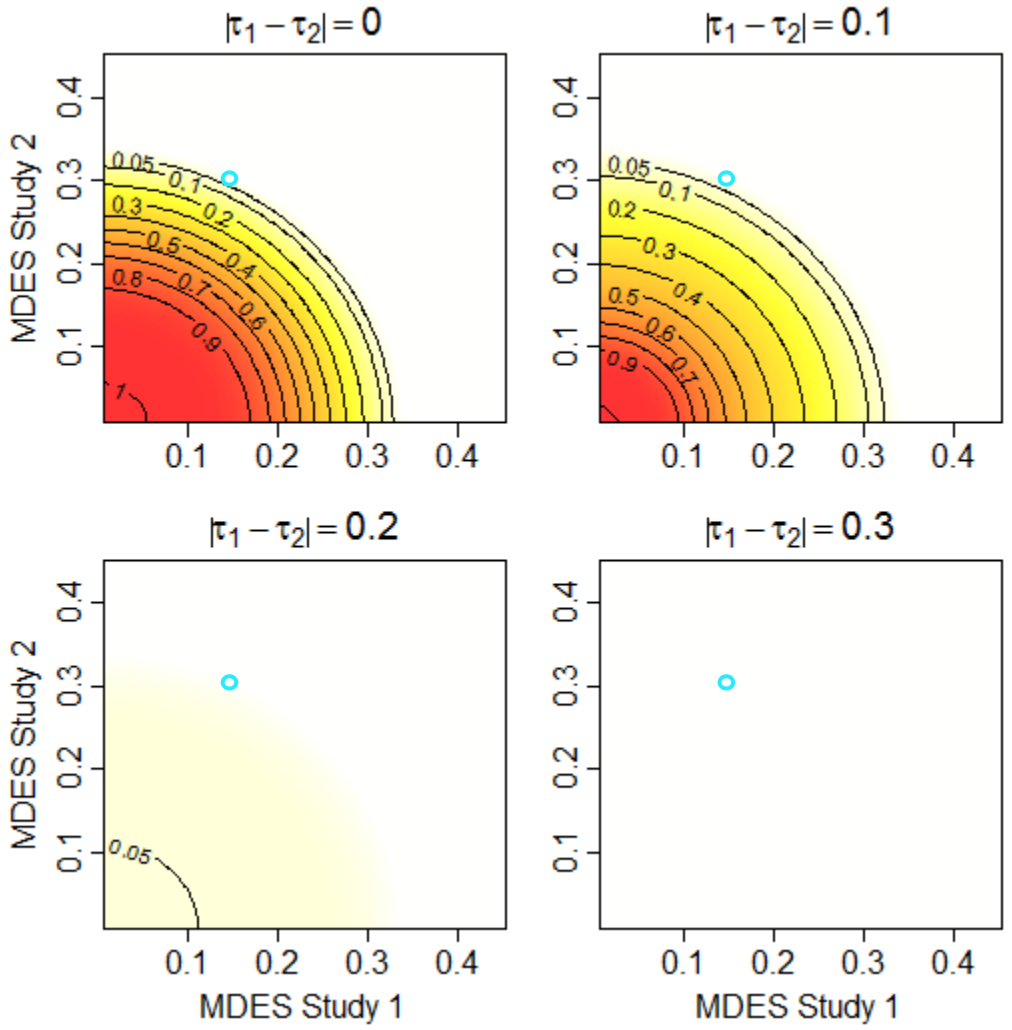
# Corresp. Probabilities – Equivalence Test

*True effect difference*

$|\tau_1 - \tau_2| =$
0, .1, .2, .3 *SD*

*Equiv. threshold*

$\delta_E = .2\ SD$



# Correspondence Test

- Combines the difference and equivalence test into a single test with four possible outcomes

| **Difference ($C^D$)** | **Equivalence ($C^E$)** | |
|---|---|---|
| | $C^E = 0$ insig. equivalence | $C^E = 1$ sig. equivalence |
| $C^D = 0$ sig. difference | **Difference** | **Trivial Difference** |
| $C^D = 1$ insig. difference | **Indeterminacy** | **Equivalence** |

# Correspondence Test – Scenario I

- True effects are identical:

$$|\tau_1 - \tau_2| = 0$$

- Equivalence thresholds

$$\delta_E = 0.1, 0.2, \& \; 0.3 \; SD$$

- Correspondence test should ideally indicate equivalance but not a difference

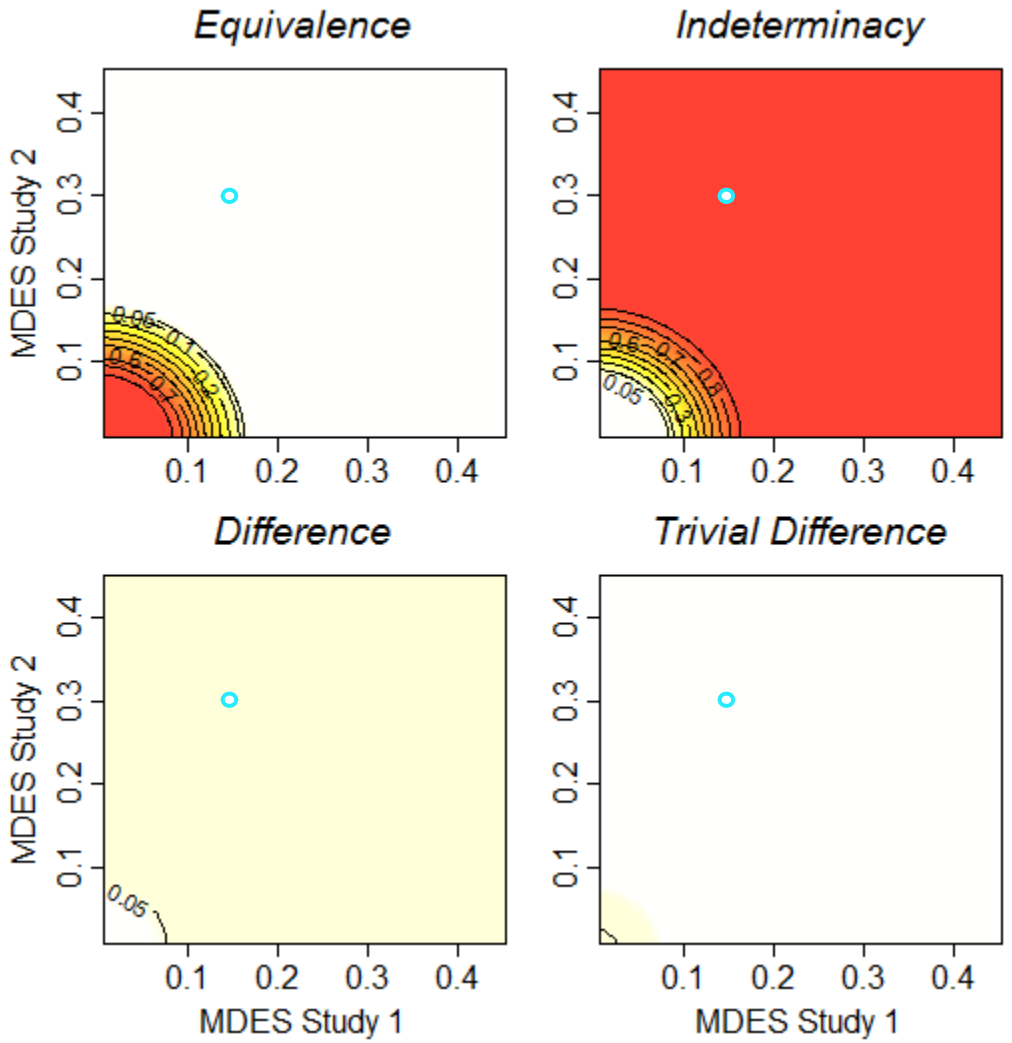
# Outcome Prob. – Correspondence Test

*True effect diff.*

$|\tau_1 - \tau_2| = 0\ SD$

*Equiv. threshold*

$\boldsymbol{\delta_E = .1\ SD}$



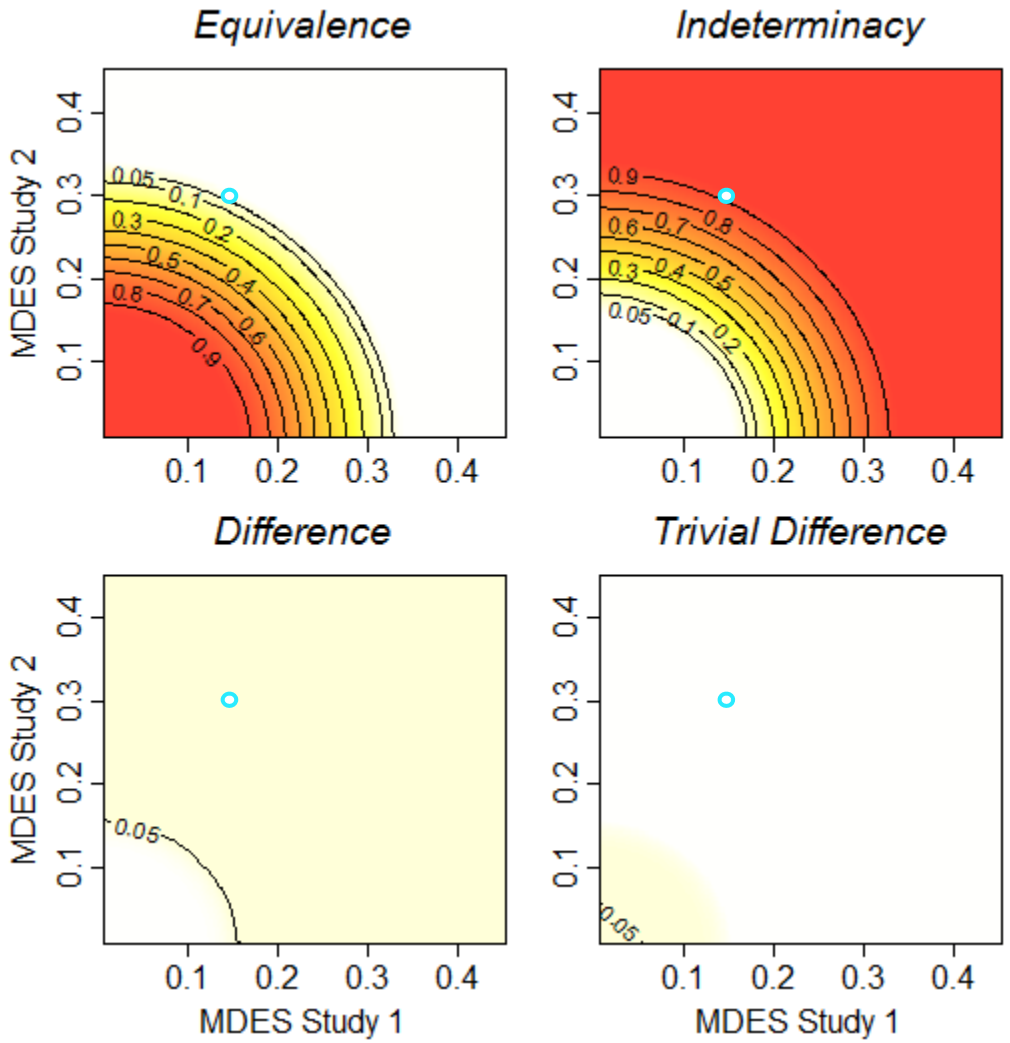
# Outcome Prob. – Correspondence Test

*True effect diff.*

$|\tau_1 - \tau_2| = 0$ *SD*

*Equiv. threshold*

$\boldsymbol{\delta_E = .2\ SD}$



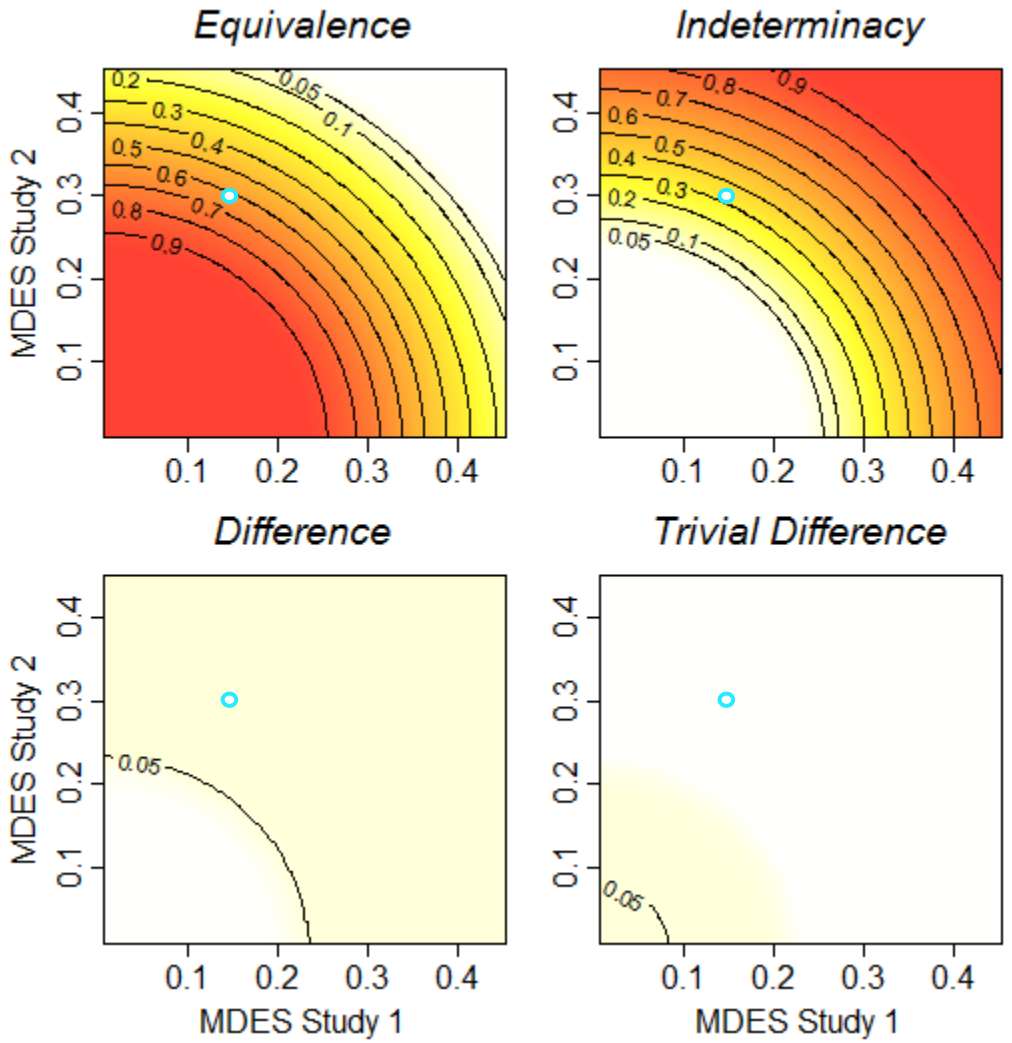
# Outcome Prob. – Correspondence Test

*True effect diff.*

$|\tau_1 - \tau_2|$ = 0 *SD*

*Equiv. threshold*

$\boldsymbol{\delta_E = .3\ SD}$



# Correspondence Test – Scenario II

- True effects differ:

$$|\tau_1 - \tau_2| = 0.2\ SD$$

- Equivalence thresholds

$$\delta_E = 0.1, 0.2, 0.3, \& \ 0.4\ SD$$

- Correspondence test should ideally indicate a difference but not equivalence as long as the threshold is less than .2 SD

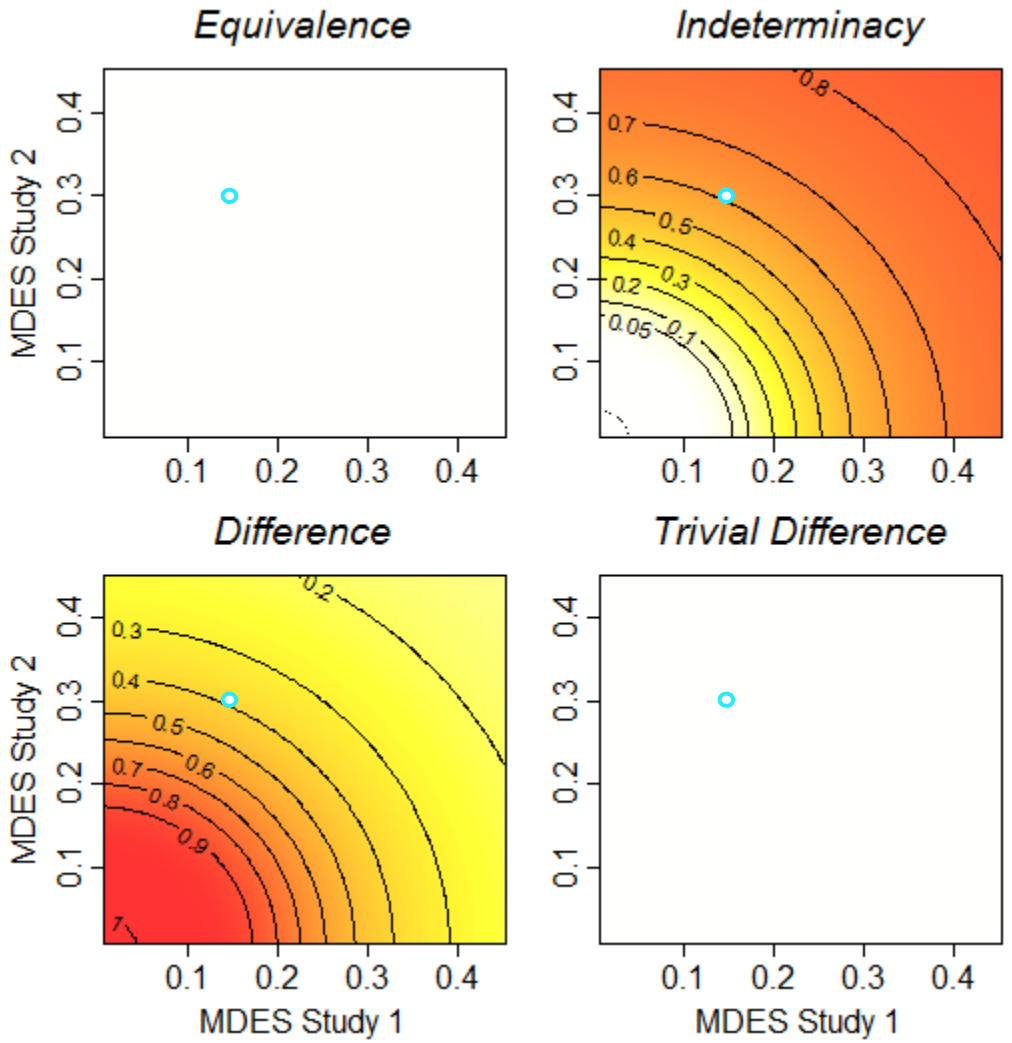
# Outcome Prob. – Correspondence Test

*True effect diff.*

$|\tau_1 - \tau_2|$ = .2 *SD*

*Equiv. threshold*

$\boldsymbol{\delta_E = .1\ SD}$



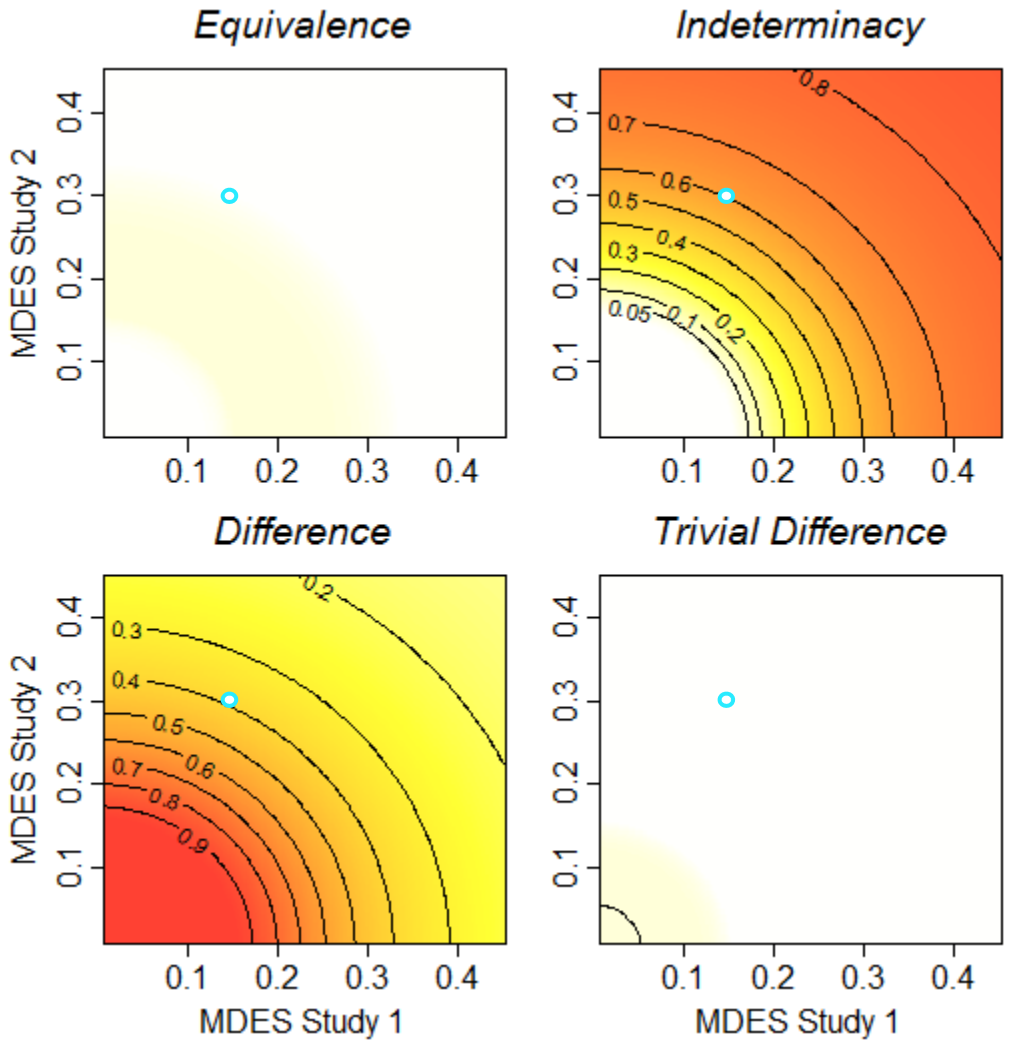
# Outcome Prob. – Correspondence Test

*True effect diff.*

$|\tau_1 - \tau_2|$ = .2 *SD*

*Equiv. threshold*

$\boldsymbol{\delta_E = .2\ SD}$



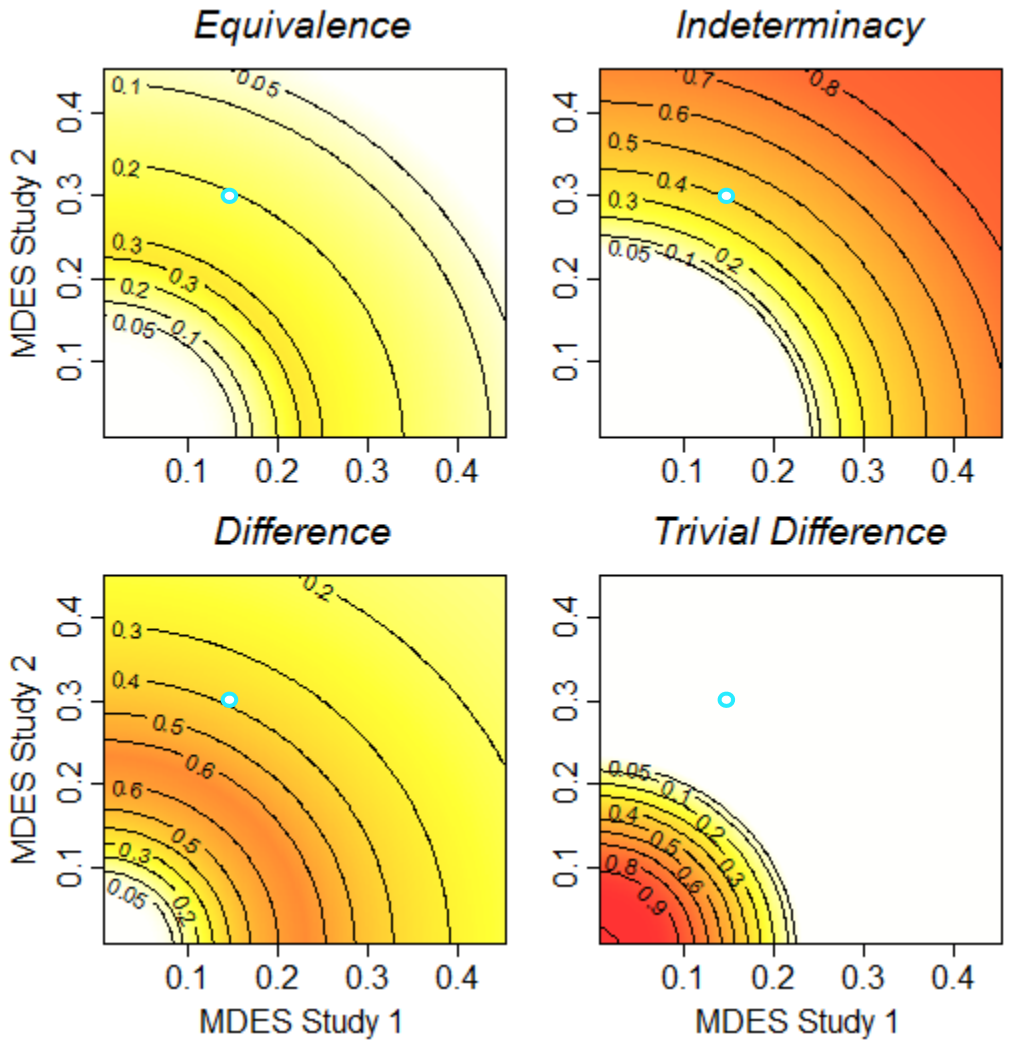
# Outcome Prob. – Correspondence Test

*True effect diff.*

$|\tau_1 - \tau_2|$ = .2 *SD*

*Equiv. threshold*

$\boldsymbol{\delta_E = .3\ SD}$



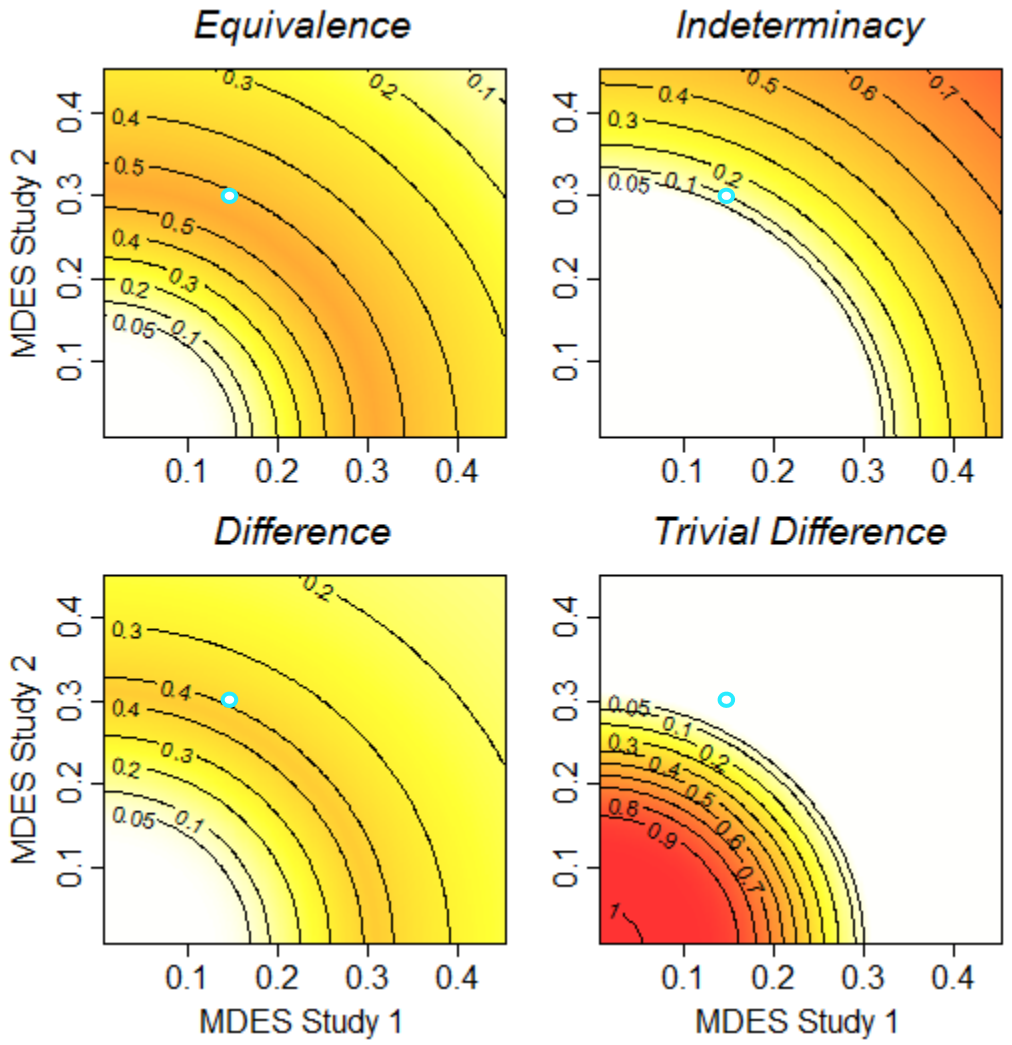
# Outcome Prob. – Correspondence Test

*True effect diff.*

$|\tau_1 - \tau_2|$ = .2 *SD*

*Equiv. threshold*

$\boldsymbol{\delta_E = .4\ SD}$



# Correspondence Test

- Explicitly deals with lack of power
- Correctly indicates equivalence or difference with high probability if both studies are sufficiently powered (MDES < .1 SD)
- With insufficiently powered studies, indeterminacy is the most likely outcome
- Choice of equivalence threshold is crucial (affects power and test outcome)

# Conclusions

# Conclusions

- Choice of correspondence measure should depend on replication question and be chosen prior to conducting the replication studies
- Conclusion-based measures: directly relevant for policy decisions, but depend on the magnitude of the unknown true effects
- Distance-based measures: establishing significant equivalence/difference is hard – need two highly powered studies
  (ideally with MDES < .1 SD)

# Conclusions

- Post hoc replication efforts often start with an insufficiently powered orginal study (for replication purposes)
- Need to prospectively plan replication studies and determine required sample sizes for both studies (power calculations)
- Alternative approach: Meta-analysis with multiple studies
  (e.g., Hedges, 1987; Hedges & Schauer, 2018)

# Thank you!

psteiner@wisc.edu