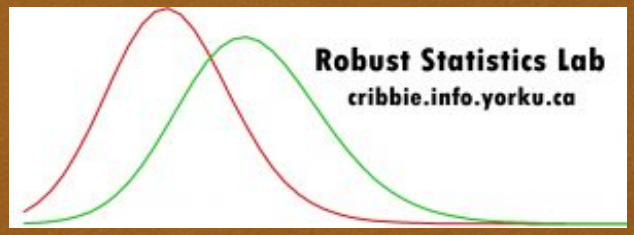
Equivalence Testing: A *Soon-To-Be* Household Name in Psychology?

Rob Cribbie



Collaborators

- I have been blessed with fantastic undergraduate and graduate students who deserve much of the credit for the work I will be presenting:
  - Teresa Allan (PhD Student, Psychology, University of Ottawa)
  - Nataly Beribisky (MA Student, QM Program, York University)
  - Alyssa Counsell (Assistant Professor, Psychology, Ryerson University)
  - Heather Davidson (Biostatistician, Sunnybrook)
  - Linda Farmus (MA Student, QM Program, York University)
  - Jason Goertzen (Senior Consultant - Research, Alberta Health Service)
  - Naomi Gutierrez (Undergraduate Student, Psychology, York)
  - Joseph Hoyda (Undergraduate Student, Psychology, York)
  - Jamie Kim (Independent Researcher)
  - Constance Mara (Assistant Professor, Division of Behavioral Medicine and Clinical Psychology, Cincinnatti Children's Hospital)
  - Victoria Ng (Data Analyst, Mental Health Systems)

Equivalence Tests

- Many empirical questions in psychological research involve a lack of relationship among variables
  - For example, a researcher may be interested in demonstrating that a clinical group subjected to a therapeutic intervention will score *equivalent* to a normal comparison group following the treatment
    - Often called 'normative comparisons'
  - Or, a researcher may hypothesize that caffeine intake is *not* related to levels of depression

Goal of Equivalence Tests

- The goal of equivalence tests is not to show that there is no relationship among variables (e.g., $\mu_1 = \mu_2$), only that any relationship that exists is too small to be considered meaningful (e.g., $-\delta < \mu_1 - \mu_2 < \delta$)
- The 'bounds' that define the upper and lower limits for an inconsequential relationship are termed the equivalence interval $(-\delta, \delta)$
  - In some cases a one-tailed equivalence test is conducted (e.g., inferiority tests), in which case there would only be an upper or lower bound, not an interval

Frequency of Equivalence-Based Hypothesis Tests

- Our lab conducted a review of the 2009 editions of three psychology journals to ascertain how frequent "equivalence based" research questions were being evaluated
  - Journal of Consulting and Clinical Psychology
  - Journal of Abnormal Psychology
  - Journal of Experimental Psychology-General
- For all three journals more than 50% of articles contained at least one equivalence-based hypothesis
  - The most frequent types of hypotheses were:
    - Group equivalence on demographics/other confounds
    - Primary Hypothesis(es)

Example: Equivalence-based Hypothesis

Personality and Individual Differences

Volume 74, February 2015, Pages 122-126

Comparing the psychosocial health of tattooed and non-tattooed women

Kathleen Thompson

Hypotheses

That tattooed women will be as psychosocially healthy as non-tattooed women as measured by scores on the Loyola Generativity Scale (LGS).

Frequency of Equivalence-Based Hypothesis Tests

- A more specific investigation of clinical studies was recently conducted

- Studies that compared psychological treatments and were published between 2000 and 2010 were included
  - 270 studies that compared two psychological treatments, psychological treatments to drug treatments, etc. were found
  - Of these studies, 154 specified no specific direction of effects, 91 hypothesized a difference between treatments, and 25 hypothesized that the treatments would be equivalent

Frequency of Equivalence-Based Hypothesis Tests

- Of the 25 studies that hypothesized equivalence, all used a difference based test for the comparison
  - Interestingly two studies used equivalence tests, but both incorrectly used them to investigate differences

- Further, approximately half of the studies that found no significant difference between treatments used "equivalence-based" language to summarize the findings
  - E.g.., "equivalence", "comparable", "equally effective"

What about at the U of M?

Reductions in Goal-Directed Cognition as a Consequence of Being the Target of Empathy

Jacquie D. Vorauer[1], Matthew Quesnel[1], and Sara L. St. Germain[1]

Personality and Social Psychology Bulletin
2016, Vol. 42(1) 130–141

DOI: 10.1177/0146167215617704
pspb.sagepub.com
SAGE

As preliminary analyses across the three studies revealed no evidence that sex moderated any effects of the mind-set manipulation, and sex did not vary according to experimental conditions ($Fs < 1$), it is not discussed further.

What about at the U of M?

Eur J Psychol Educ (2014) 29:175–194
DOI 10.1007/s10212-013-0193-2

The longitudinal effects of achievement goals and perceived control on university student achievement

Lia M. Daniels · Raymond P. Perry · Robert H. Stupnisky · Tara L. Stewart · Nancy E. G. Newall · Rodney A. Clifton

et al. 2008). In regards to achievement, we hypothesized that because this is a college sample, performance goals and primary control would positively and directly predict achievement, whereas mastery goals would have a non-significant direct effect. The existing research linking

What about at the U of M?

Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale
2016, Vol. 70, No. 2, 93–98

1196-1961/16/$12.00 http://dx.doi.org/10.1037/cep0000082

The Production Effect in Recognition Memory: Weakening Strength Can Strengthen Distinctiveness

Glen E. Bodner
University of Calgary

Randall K. Jamieson and David T. Cormack
University of Manitoba

Dawn-Leah McDonald and Daniel M. Bernstein
Kwantlen Polytechnic University

implemented by Merritt, Cook, and Wang (2014). Three 2 (production: unproduced vs. produced) × 2 (design: mixed vs. pure) ANOVAs showed that the pure-list production effect was similar in size to the mixed-list production effect in each of the 20%, 50%, and 80% groups, $F(1, 50) = 1.62$, $p > .20$, $F(1, 50) = 2.77$, $p > .10$, and $F(1, 50) = 0.21$, $p > .60$. In summary, the pure- and mixed-list production effects were similar and no evidence for a distinctiveness influence was obtained.

Traditional and Equivalence-based Hypotheses

- Two Populations, Mean Difference
  - Traditional Nondirectional Null & Alternate Hypotheses
    - $H_o: \mu_1 = \mu_2$, $H_a: \mu_1 \neq \mu_2$
  - Equivalence Null & Alternate Hypotheses
    - Two One-sided Testing Procedure (TOST)
      - $H_{o1}: \mu_1 - \mu_2 \geq \delta$; $H_{o2}: \mu_1 - \mu_2 \leq -\delta$
      - $H_{a1}: \mu_1 - \mu_2 < \delta$; $H_{a2}: \mu_1 - \mu_2 > -\delta$
      - Rejection of H_{o1} implies that $\mu_1 - \mu_2 < \delta$, and rejection of H_{o2} implies that $\mu_1 - \mu_2 > -\delta$.
      - Thus, rejection of both null hypotheses implies that $\mu_1 - \mu_2$ falls within the bounds of $(-\delta, \delta)$ and the means are deemed equivalent

Can't a Traditional *t*-test be Used to Evaluate Equivalence?

- You cannot use non-rejection of the null hypothesis of a traditional difference-based *t*-test to evaluate equivalence because:
  - Theoretically, non-rejection of the null hypothesis does not prove the null to be true
  - Power is backward
    - Power increases as sample sizes decrease and error variances increase

Two One-sided Testing (TOST) for $\mu_1 - \mu_2$

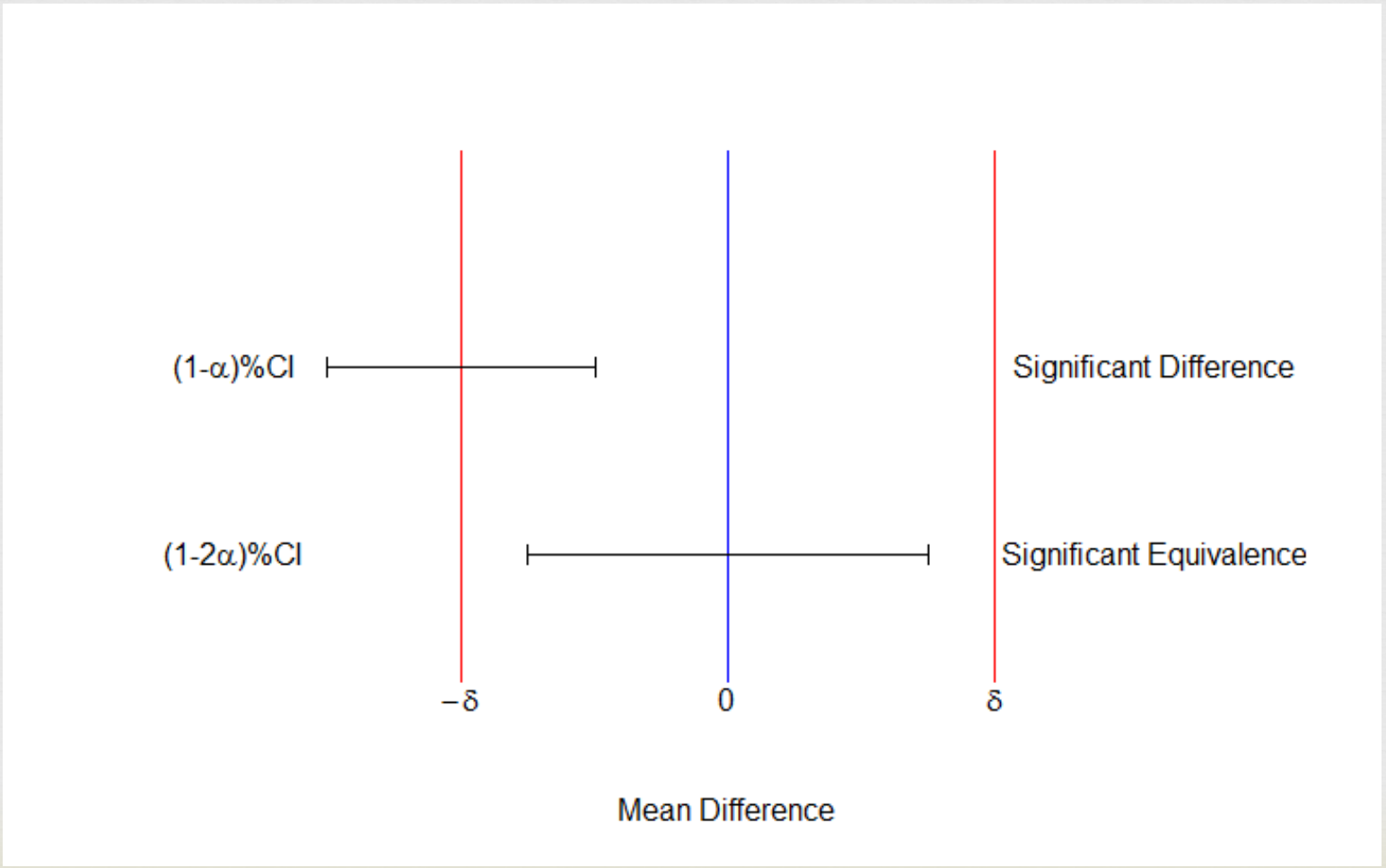
- $H_{o1}: \mu_1 - \mu_2 \leq -\delta$
  - H_{o1} is rejected if $t_1 \leq t_{\alpha,\, df=n_1+n_2-2}$
- $H_{o2}: \mu_1 - \mu_2 \geq \delta$
  - H_{o2} is rejected if $t_2 \geq t_{1-\alpha,\, df=n_1+n_2-2}$

$$t_1 = \frac{(\bar{X}_1 - \bar{X}_2) - \delta}{\sqrt{\frac{(n_1+n_2)[(n_1-1)s_1^2 + (n_2-1)s_2^2]}{n_1 n_2 (n_1+n_2-2)}}}$$

$$t_2 = \frac{(\bar{X}_1 - \bar{X}_2) - (-\delta)}{\sqrt{\frac{(n_1+n_2)[(n_1-1)s_1^2 + (n_2-1)s_2^2]}{n_1 n_2 (n_1+n_2-2)}}}$$

- Both H_{o1} and H_{o2} must be rejected in order to conclude population mean equivalence
- This is equivalent to testing if the (1-2α) CI falls within (-δ, δ)

Difference vs Equivalence: Confidence Intervals



Extensions of Equivalence Testing: Lack of Association

- One extension of equivalence testing is to the problem of demonstrating that two variables are *minimally related*

- For the same reason that a traditional *t* test cannot be used to evaluate the equivalence of two independent groups, a traditional correlation (or regression) statistic cannot be used to demonstrate a lack of association
  - Recall that the hypotheses are backward, and power would be maximized by decreasing N and increasing error variance

Lack of Association Tests

- The goal of a lack of association test is to demonstrate that any relationship between the variables is too small to be considered meaningful

- ρ^* is used to represent the smallest correlation between the variables that would be considered meaningful
  - $(-\rho^*, \rho^*)$ forms the equivalence interval

Lack of Association Test

- Following the logic of the TOST procedure, the composite null hypotheses, H_{o1}: $\rho \geq \rho^*$ and H_{o2}: $\rho \leq -\rho^*$, are rejected if $t_1 \leq t_{\alpha, N-2}$ and $t_2 \geq t_{1-\alpha, N-2}$, respectively, where:

$$t_1 = \frac{r - \rho^*}{\sqrt{\frac{1-r^2}{N-2}}} \qquad t_2 = \frac{r - (-\rho^*)}{\sqrt{\frac{1-r^2}{N-2}}}$$

We have also investigated versions of this test based on: 1) Fisher's *z* transformation, and 2) Resampling

Power Rates of Lack of Association Tests

	$N = 50$			$N = 100$			$N = 500$			$N = 1{,}000$		
ρ^*	eq_r	eq_fz	eq_rs	eq_r	eq_fz	eq_rs	eq_r	eq_fz	eq_rs	eq_r	eq_fz	eq_rs
$\rho = 0$												
.05	0	0	0	0	0	0	0	0	0	0	0	.003
.10	0	0	0	0	0	0	.425	.432	.434	.874	.874	.875
.15	0	0	0	0	0	0	.914	.917	.917	.998	.998	.999
.20	0	0	.008	.237	.262	.279	.995	.995	.995	1	1	1
.25	.044	.085	.137	.591	.621	.622	1	1	1	1	1	1
.3	.315	.365	.372	.820	.847	.849	1	1	1	1	1	1

TOST

Fisher's *z*

Resampling

Lack of Association Tests

- Why is it hard to find a 'lack of association' with small n/ρ*?

ρ = 0

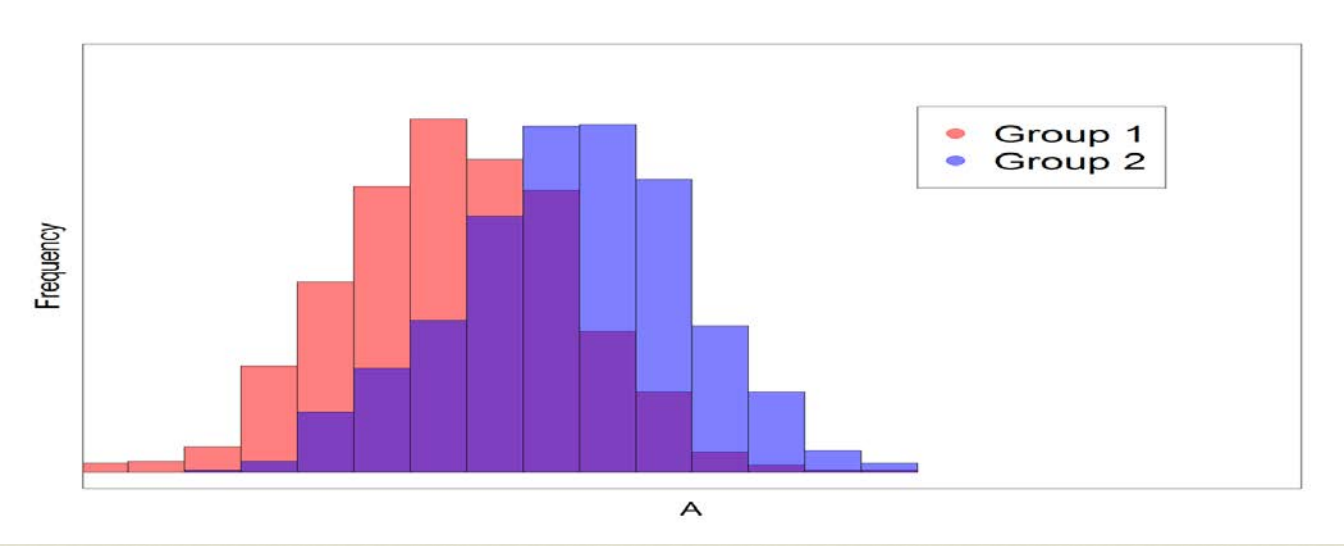
	Sample correlation (r) magnitude (absolute value)									
N	>.9	>.8	>.7	>.6	>.5	>.4	>.3	>.2	>.1	>0
10	.001	.007	.025	.067	.140	.252	.398	.578	.780	1
15	0	<.001	.005	.019	.060	.139	.273	.465	.720	1
20	0	0	.001	.005	.025	.082	.199	.400	.687	1
25	0	0	<.001	.002	.012	.046	.145	.334	.628	1
50	0	0	0	0	<.001	.004	.036	.163	.492	1
100	0	0	0	0	0	0	.003	.048	.323	1
200	0	0	0	0	0	0	0	.005	.159	1

What is an appropriate equivalence interval?

- The "Elephant in the Room" when it comes to many discussions of equivalence testing
  - Essentially, a researcher is asked to determine *the smallest effect that is of practical importance* within the nature of the study
  - Our lab has started investigating what represents the smallest meaningful relationship among variables in common research settings (almost no prior research on the topic)

Smallest Meaningful Difference in Central Tendencies

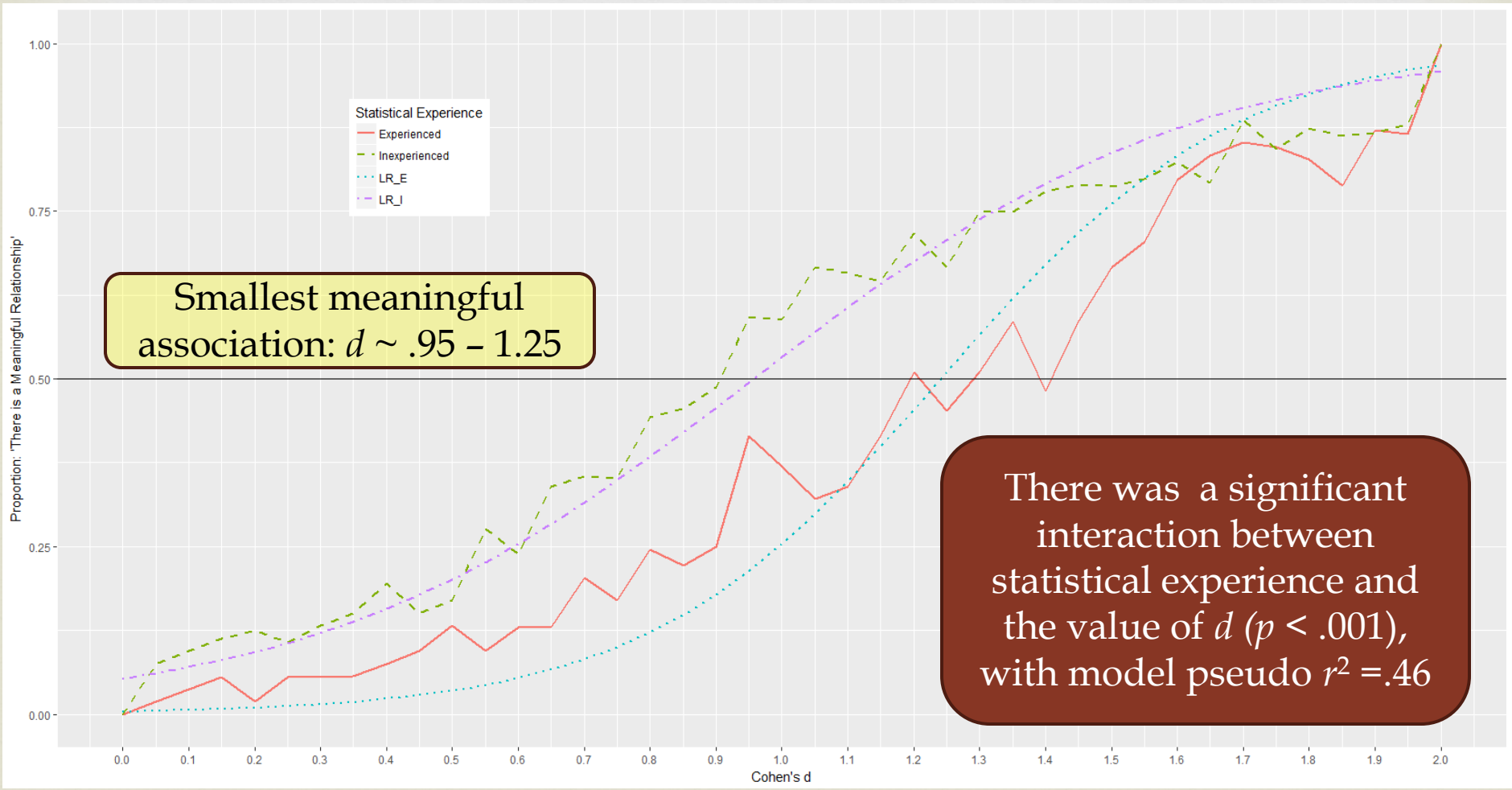
- Overlapping histograms displayed two distributions that were separated by a population Cohen's *d* ranging from 0.00 to 2.00, in .05 increments
  - For each value of Cohen's *d*, five plots were generated and subjects saw only one randomly chosen plot



Smallest Meaningful Difference in Central Tendencies

- Participants were asked to indicate whether the difference in the distributions was meaningful
- Training for interpreting the plots was provided
- Participants were separated by level of statistical experience (< 3 semester courses in university statistics, 3+ semester courses in university statistics)
- Based on past research, the value of Cohen's *d* where 50% of participants assert that the relationship is meaningful was used to represent the smallest meaningful association

Logistic Regression Results – Mean Difference

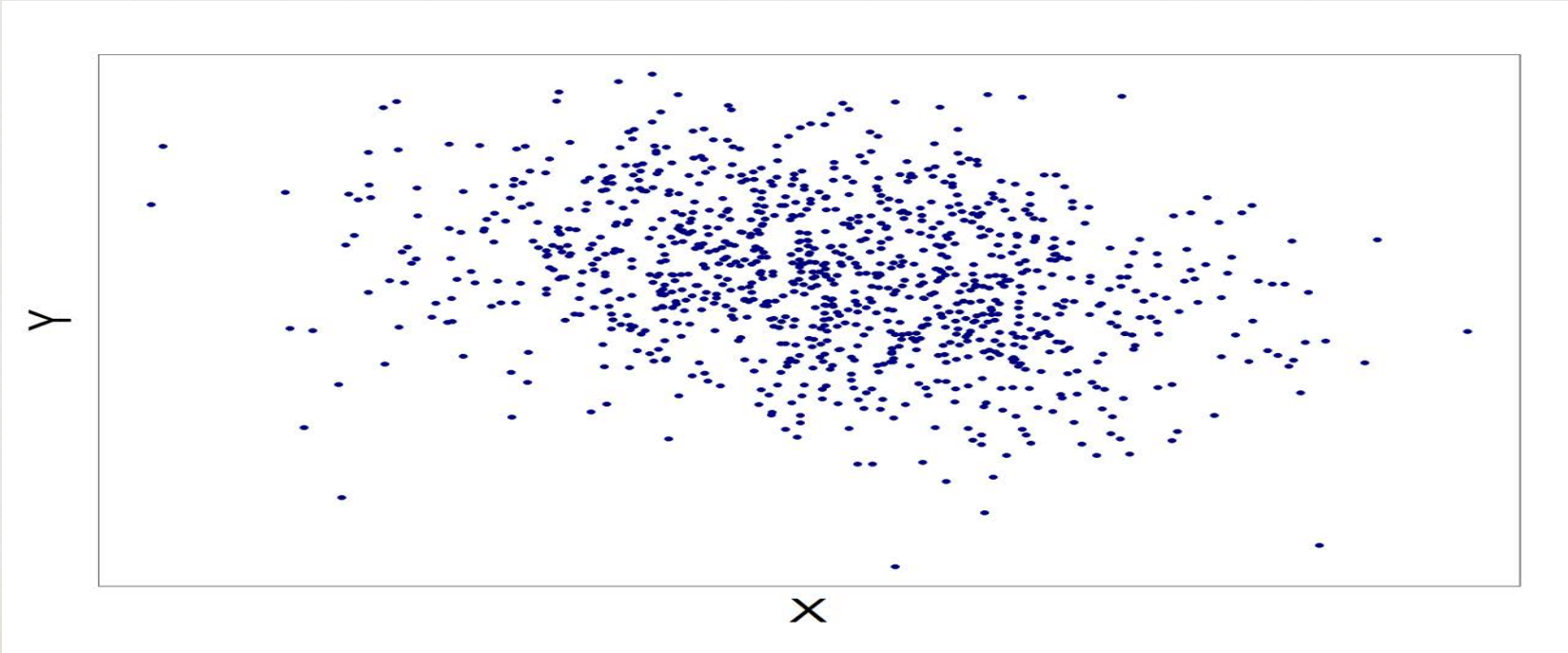


Smallest meaningful association: $d \sim .95 - 1.25$

There was a significant interaction between statistical experience and the value of d ($p < .001$), with model pseudo $r^2 = .46$

Smallest Meaningful Association among Variables

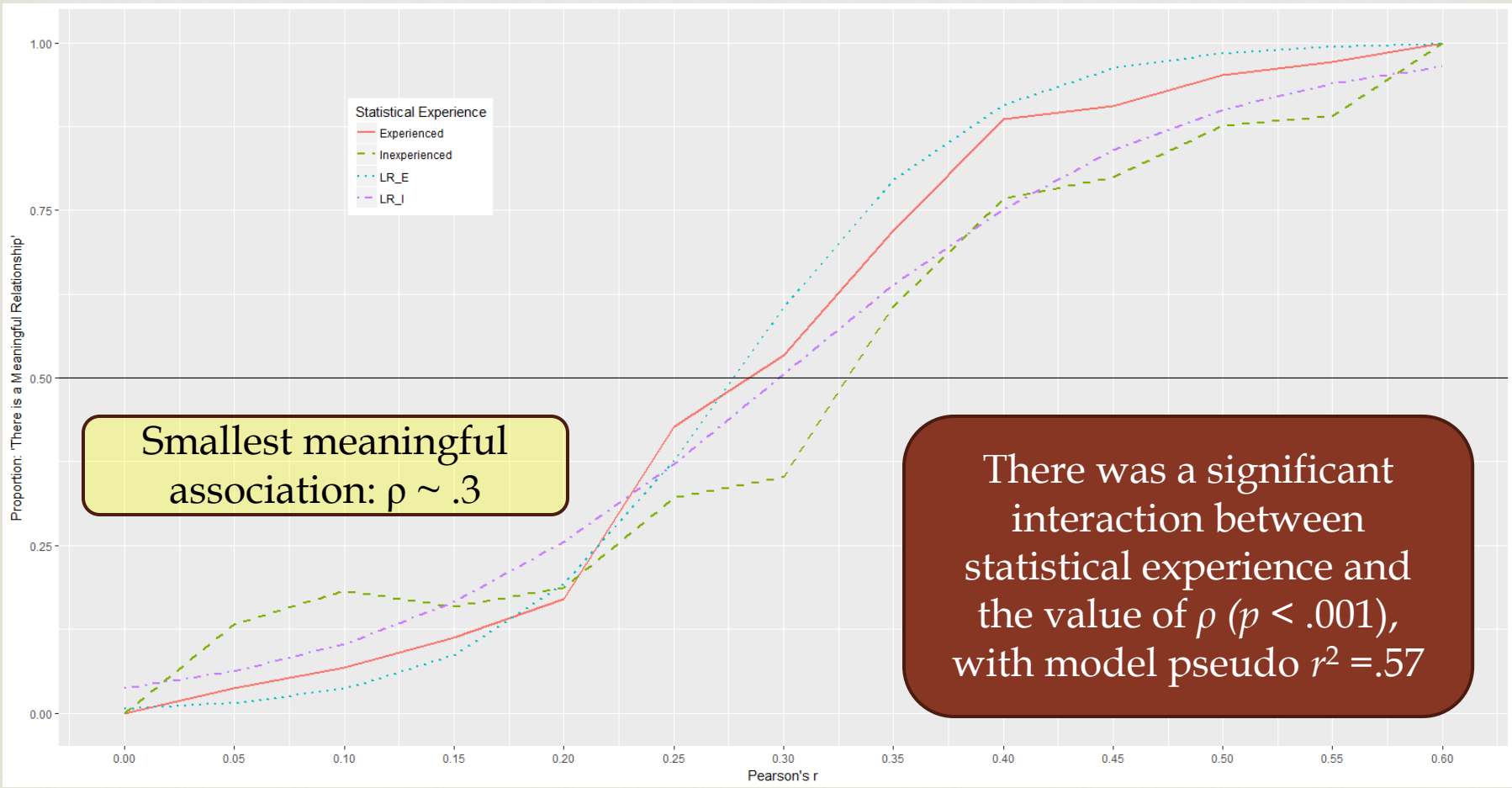
- Scatterplots displayed population correlations ranging from ρ = - .60 to ρ = .60 in .05 increments
  - For each value of ρ, five plots were generated and subjects saw only one randomly chosen plot



Smallest Meaningful Association among Variables

- Participants were asked to indicate whether the association was meaningful
- Training for interpreting the plots was provided
- Participants were separated by level of statistical experience (< 3 semester courses in university statistics, 3+ semester courses in university statistics)
- Again, the value of ρ where 50% of participants assert that the relationship is meaningful was used to represents the smallest meaningful association among the variables

Logistic Regression Results – Correlation



Smallest meaningful association: ρ ~ .3

There was a significant interaction between statistical experience and the value of ρ ($p < .001$), with model pseudo r^2 =.57

Conclusions – Smallest Meaningful Association

- The minimally important relationship in a correlation setting (ρ = .30 to ρ = .35) was moderately higher than what Cohen suggested as the lower bound for a "small" correlation (ρ = 0.10)
- The minimum meaningful difference in central tendencies (d = .95 to d = 1.20) was much larger than what Cohen suggested as the lower bound for a "small" standardized mean difference (d = 0.20)
- We re-ran the study asking if "the groups were equivalent" or "there was a negligible association" and found essentially the same results

Conclusions – Smallest Meaningful Association

- It is hoped that this research will help researchers better interpret the magnitude of effect sizes and set appropriate boundaries in equivalence testing
  - Current research is exploring what aspects of the figures participants focus on when making decisions regarding the meaningfulness of the association
    - Eye Tracking
  - Another current study allows participants to manipulate the visualizations until the smallest meaningful association is represented
    - Will this produce different results?

Extensions of Equivalence Testing

- Our lab has explored many proposed extensions of equivalence testing, including:
  - Tests of Substantial Mediation (Mara)
  - Tests of the Equivalence of Correlation or Regression Coefficients across Groups (Counsell)
  - Negligible Interaction Tests (Counsell, Jabbari)
  - Multiplicity Issues in Equivalence Testing (Davidson)
  - Equivalence Tests for Longitudinal Data (Ng, Mara)
  - Equivalence Tests for Categorical Data (Shiskina)
  - Bayesian approaches to Equivalence Testing (Counsell, Hoyda)
  - Equivalence-based Homogeneity of Variance Test (Mara, Kim)
  - Equivalence Tests for Measurement Invariance (Counsell)

Current Approaches to Testing for Variance Equality

- Difference-based tests
  - E.g., Levene's (1960) test for homogeneity of variances (HOV)
    - An ANOVA is conducted on the absolute value of the deviations from the group means:
      - $Z_{ij} = |X_{ij} - \bar{X}_j|$
- Brown-Forsythe (1974): $Z_{ij} = |X_{ij} - Mdn_j|$

Levene's HOV Test

- $H_o: \sigma_1^2 = \sigma_2^2$
  - HOV is concluded when H_o is NOT rejected
- Since the research hypothesis deals with variance equality, H_a, not H_o, should be aligned with the research hypothesis

Applying Equivalence Testing to Homogeneity of Variance Tests

- Wellek's (2003) one-way equivalence test is used for detecting the equivalence of multiple population means
- Borrowing logic from Wellek's one-way equivalence test, and Levene's HOV test:

$$H_0 : \Psi^{2^*} \geq \varepsilon^2$$

- Ψ^{2*} is an estimate of variance inequality (i.e., a modified ANOVA on the $Z_{ij} = |X_{ij} - \bar{X}_j|$)

$$H_a : \Psi^{2^*} < \varepsilon^2$$

- ε^2 represents the smallest difference in the variances that is meaningful (same metric as Ψ^{2*})

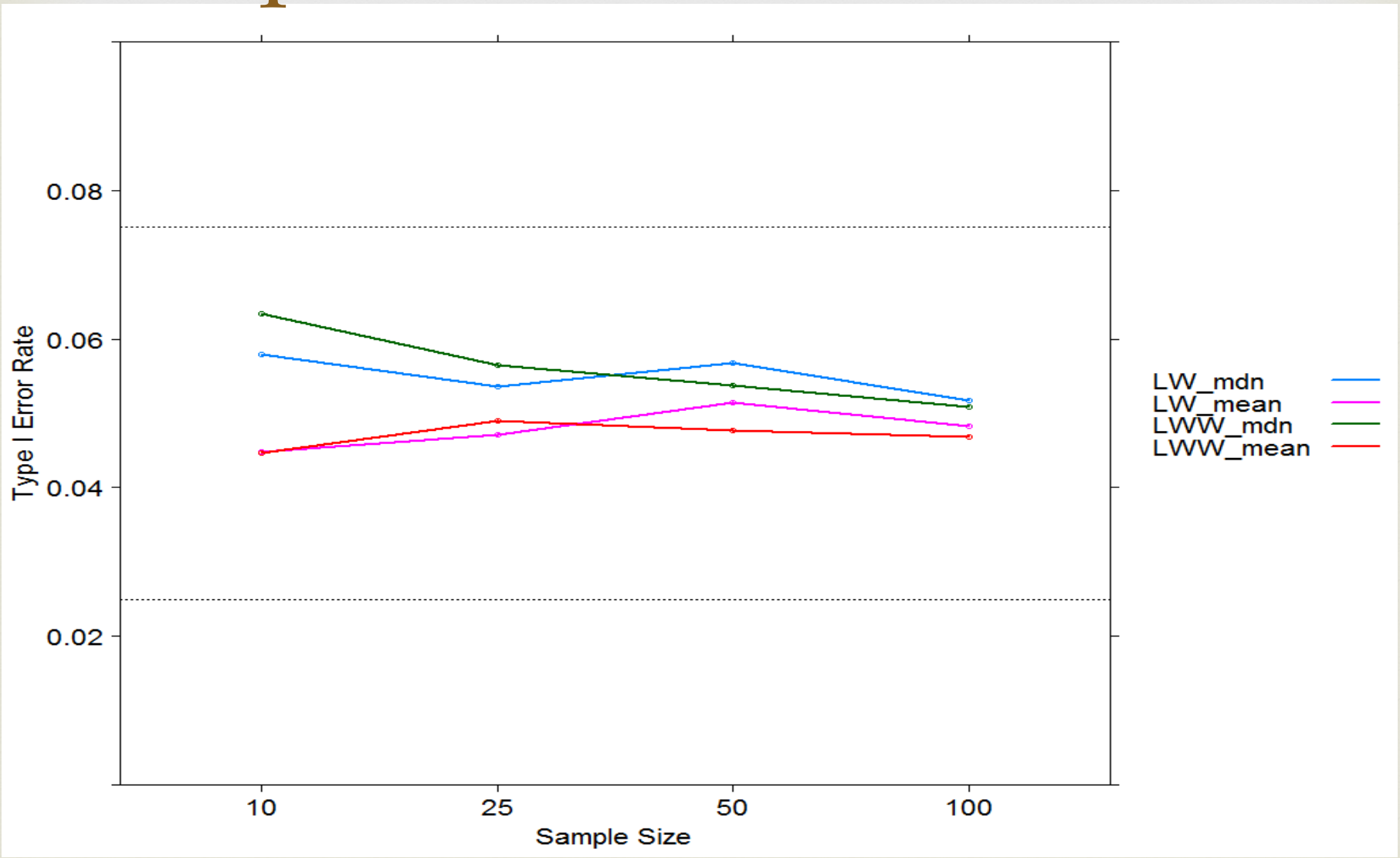
Monte Carlo Study

- 4 equivalence tests:
  - Levene-Wellek using $Z_{ij} = \left|X_{ij} - \bar{X}_j\right|$ (LW_mean)
  - Levene-Wellek using $Z_{ij} = \left|X_{ij} - Mdn_j\right|$ (LW_mdn)
  - Welch-adjusted versions of each of these tests (LWW_mean, LWW_mdn)
- Compared to 4 non-equivalence versions:
  - Original Levene (Lev_mean)
  - Brown-Forsythe (Lev_mdn)
  - Welch-adjusted versions of each of these tests (LevW_mean, LevW_mdn)

Monte Carlo Study

- Outcomes:
  - Type I error control
  - Probability of detecting equivalence
    - Equivalence Test: Reject H_0
    - Traditional Levene Test: Don't Reject H_0
- Manipulated variables:
  - Group sample sizes
  - Level of variance equality
  - # of groups = 2 & 4
  - Equivalence Interval (ε) =.50 & .25
  - α = .05

Mean Type I Error Rates: Equivalence-based Tests



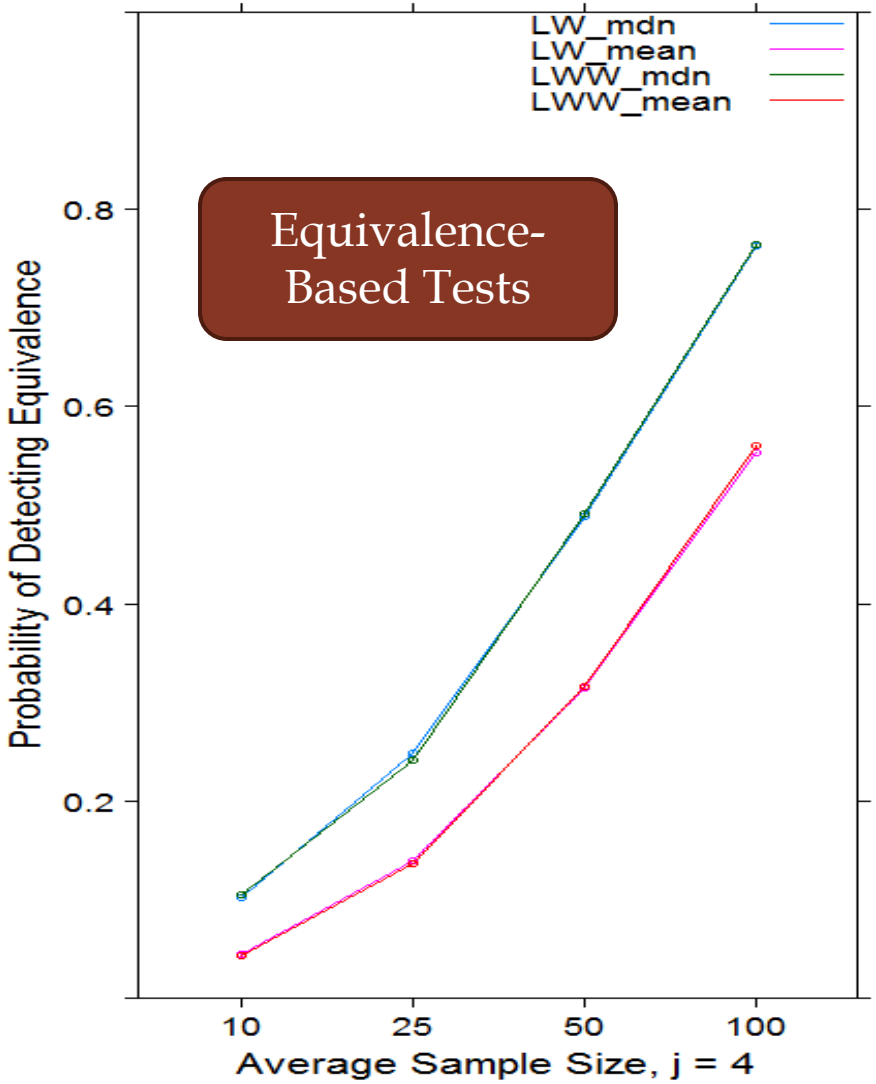
Type I Error Rates

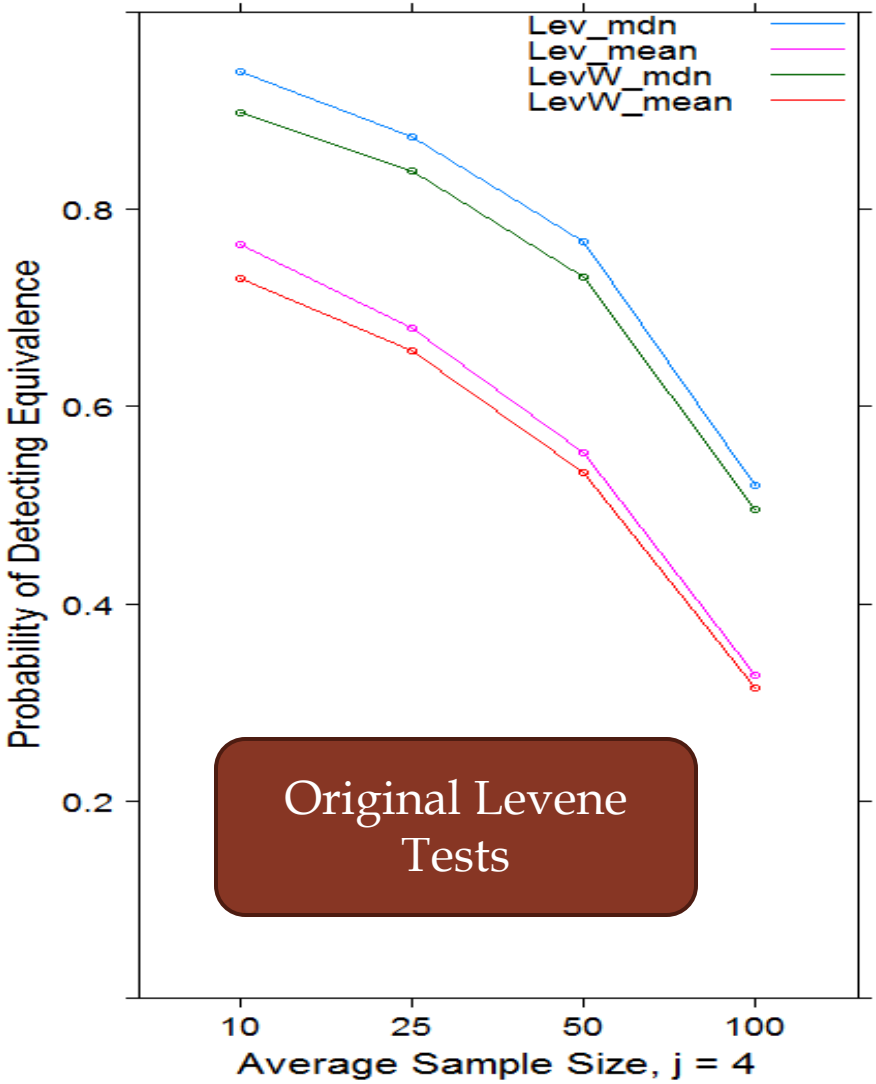
Test	Min. Empirical Type I Error Rate	Max. Empirical Type I Error Rate	# of Times Type I Error Rate Exceeded the Bounds of .025-.075
Levene-Wellek mean	.0200	.1113	12
Levene-Wellek median	.0223	.0886	6
LWW mean	.0182	.0859	9
LWW median	.0237	.1014	2

Issues in only 2/96 conditions

Probability of Declaring Equivalence (4 Groups)

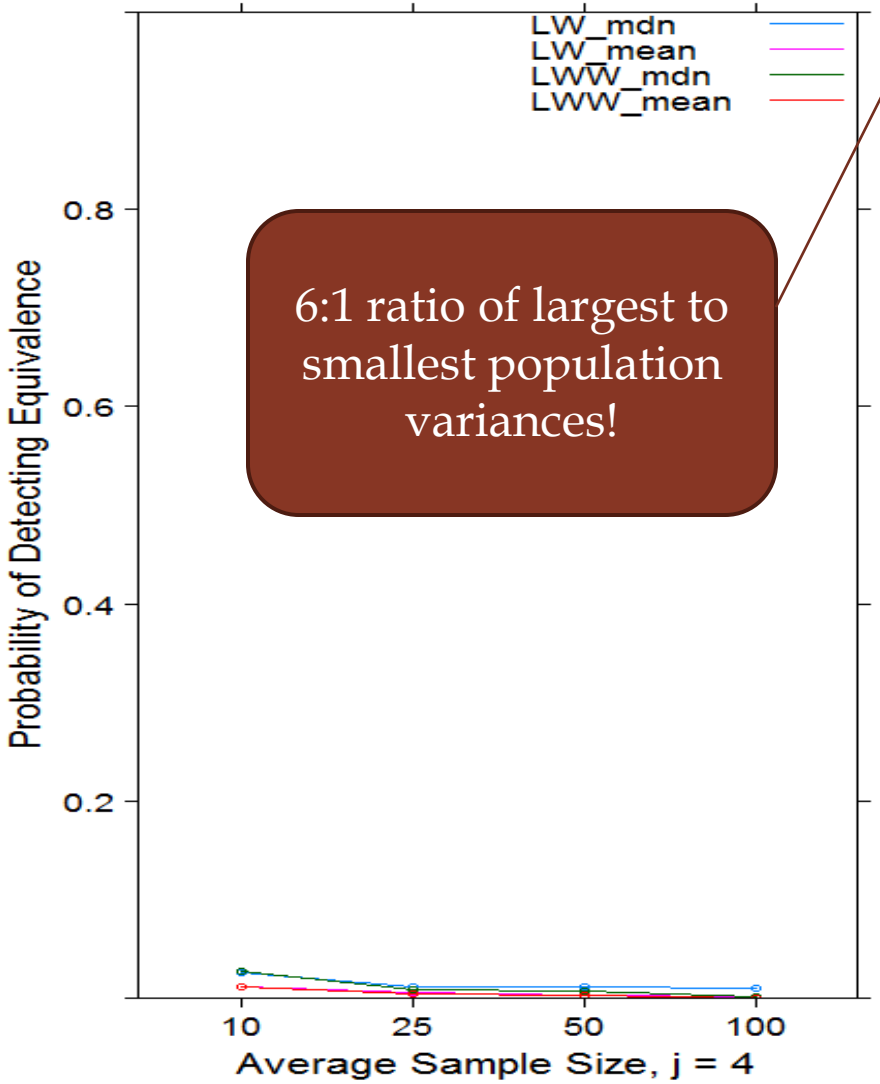
$\varepsilon = .50$, $\sigma^2 = 1, 1.33, 1.66, 2$ $(\Psi^2 < \varepsilon^2)$



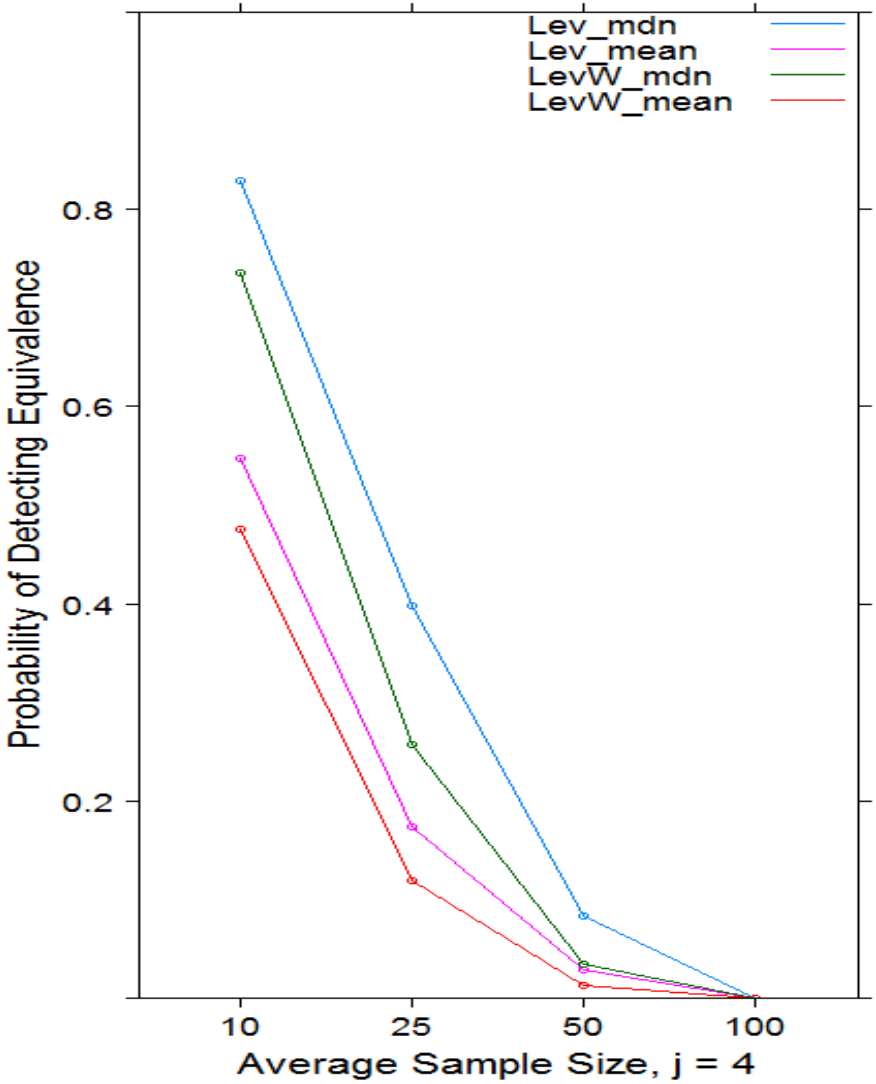


Average Probability of Declaring Equivalence

$\varepsilon = .50$, $\sigma^2 = 1, 3, 4, 6$ $(\Psi^2 > \varepsilon^2)$



6:1 ratio of largest to smallest population variances!



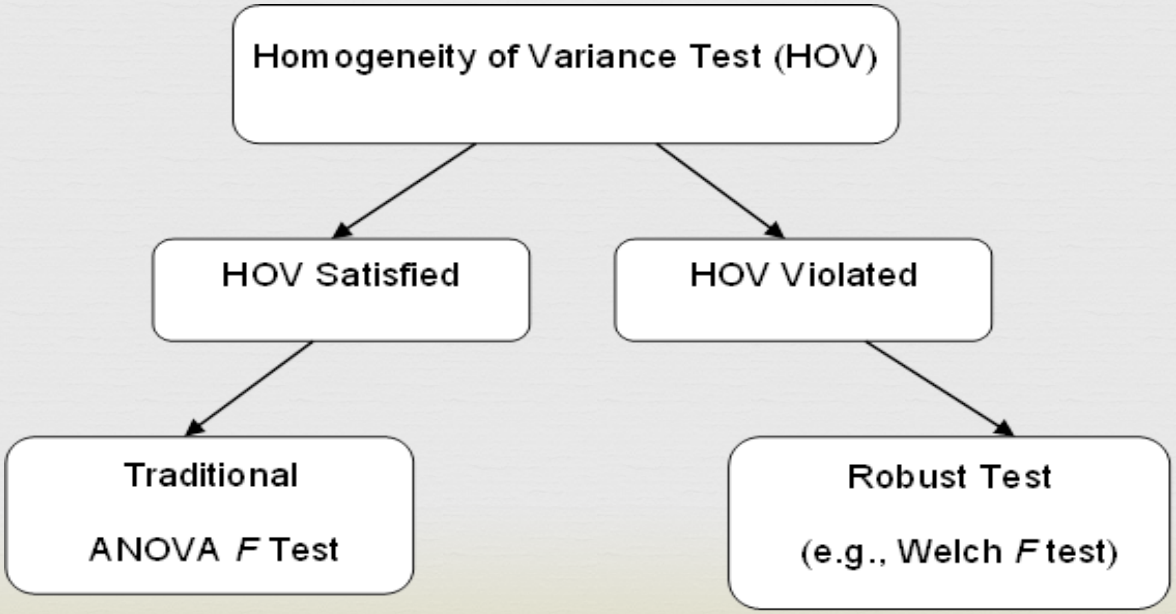
Conclusion - HOV

- Traditional HOV tests address the problem from the wrong perspective
  - Levene difference-based tests attempt to NOT REJECT H_0: $\sigma_1^2 = \sigma_2^2$

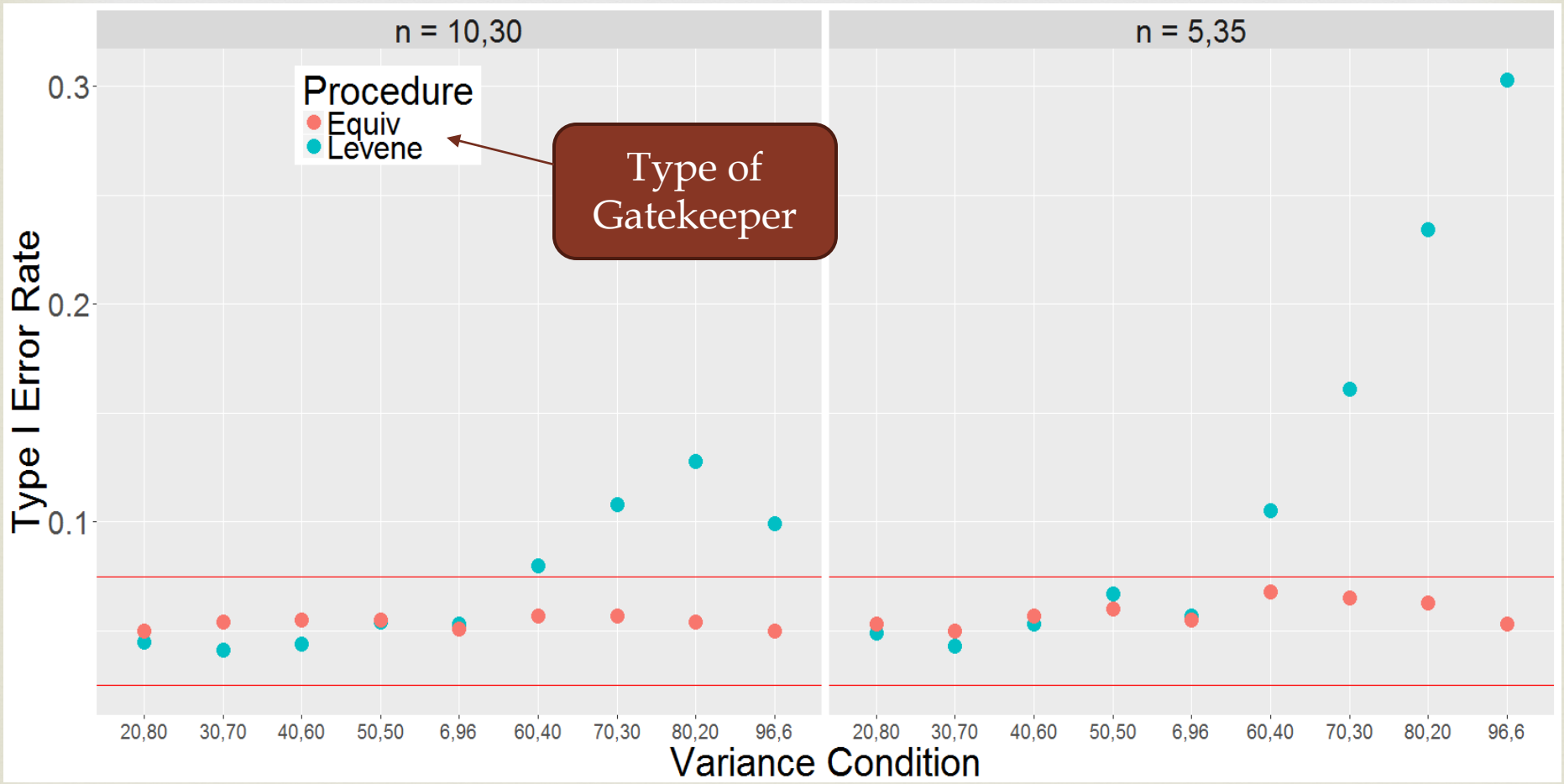
- The proposed equivalence-based tests correctly address the research question: "Are the population variances equivalent?"
  - Power increases with sample size and large differences in variances are not declared equivalent with small N

Extension: Equivalence-based HOV Tests

- Previous research has found that Levene-type tests are not effective gatekeepers for deciding between traditional and robust tests
- But what about an equivalence-based HOV test?



Type I error Rates With Gatekeepers

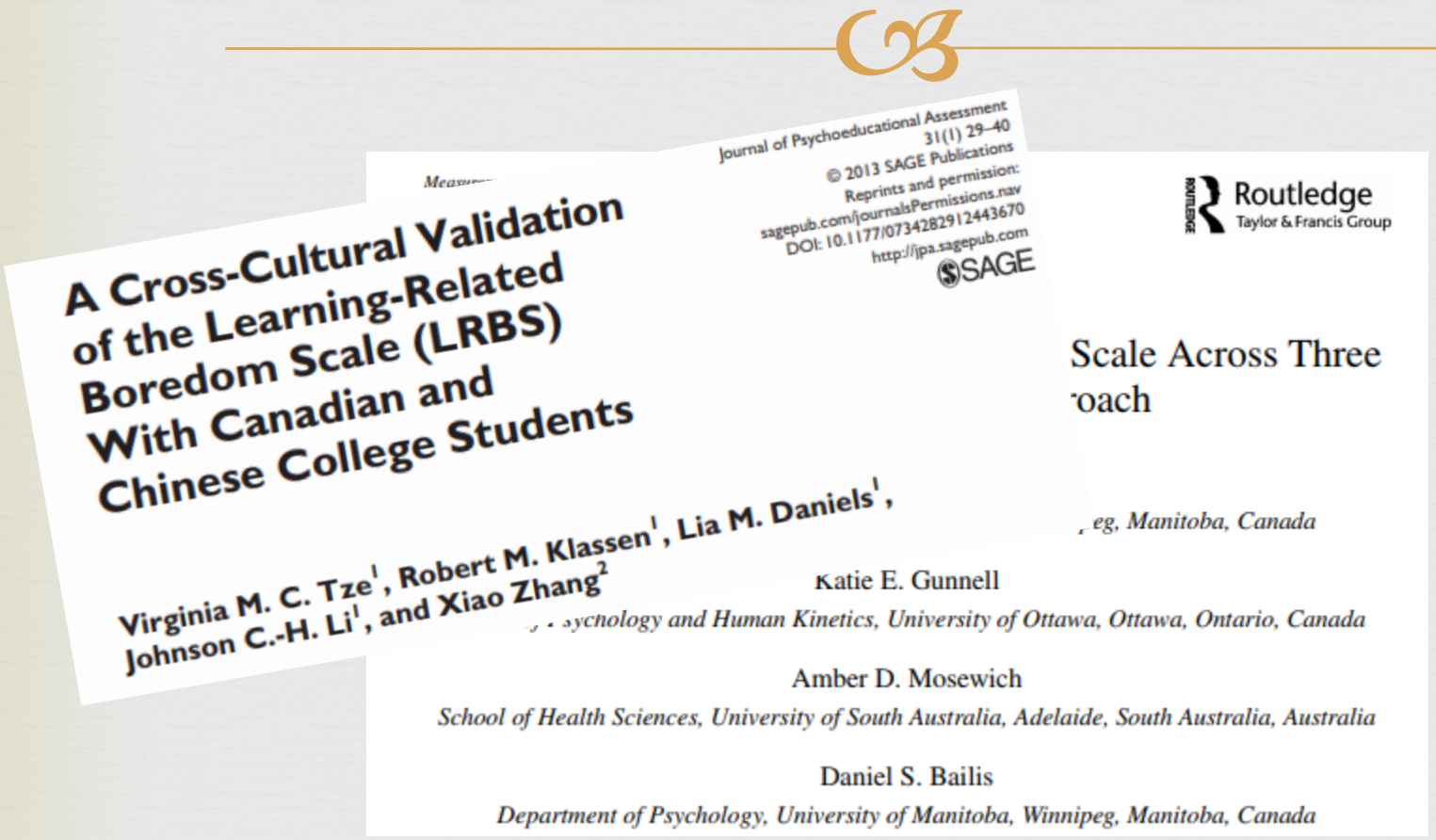


Measurement Invariance (MI)

- Researchers often seek to compare independent groups on a construct of interest
  - For example, do males and females score similarly/differently on depression?

- When discussing group differences on a construct, it is important to ensure that these differences are indeed a function of the group membership, rather than the manner in which the construct is measured
  - For example, maybe the items on a depression scale are interpreted differently by males and females

MI at the U of M



Journal of Psychoeducational Assessment
31(1) 29–40
© 2013 SAGE Publications
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0734282912443670
http://jpa.sagepub.com
SAGE

A Cross-Cultural Validation of the Learning-Related Boredom Scale (LRBS) With Canadian and Chinese College Students

Virginia M. C. Tze[1], Robert M. Klassen[1], Lia M. Daniels[1], Johnson C.-H. Li[1], and Xiao Zhang[2]

Routledge
Taylor & Francis Group

Scale Across Three
oach

eg, Manitoba, Canada

Katie E. Gunnell

ychology and Human Kinetics, University of Ottawa, Ottawa, Ontario, Canada

Amber D. Mosewich

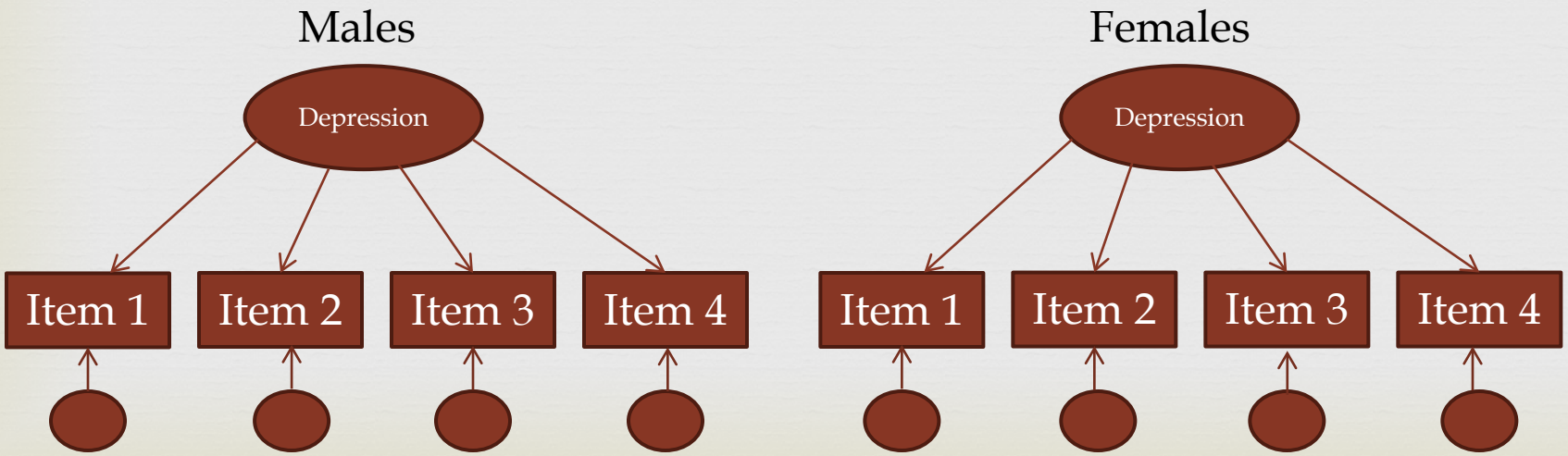
School of Health Sciences, University of South Australia, Adelaide, South Australia, Australia

Daniel S. Bailis

Department of Psychology, University of Manitoba, Winnipeg, Manitoba, Canada

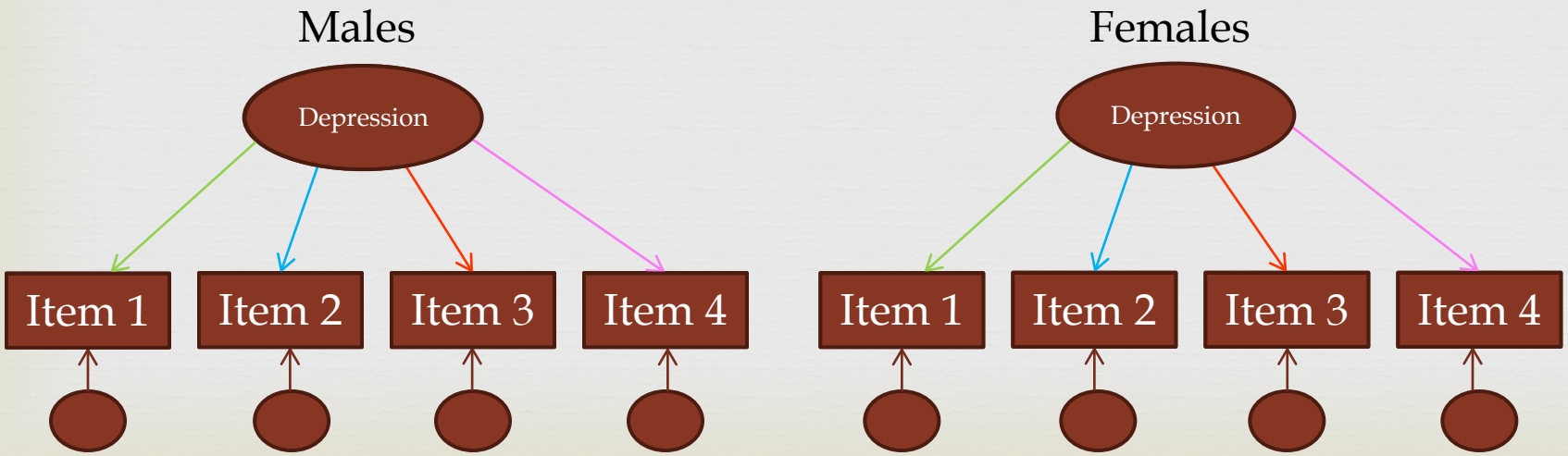
Assessing MI

- There are multiple levels of MI
  - **Configural Invariance: same factor structure**
  - Metric (Weak) Invariance: factor loadings are equal
  - Scalar (Strong) Invariance: intercepts are equal
  - Strict Invariance: error variances are equal



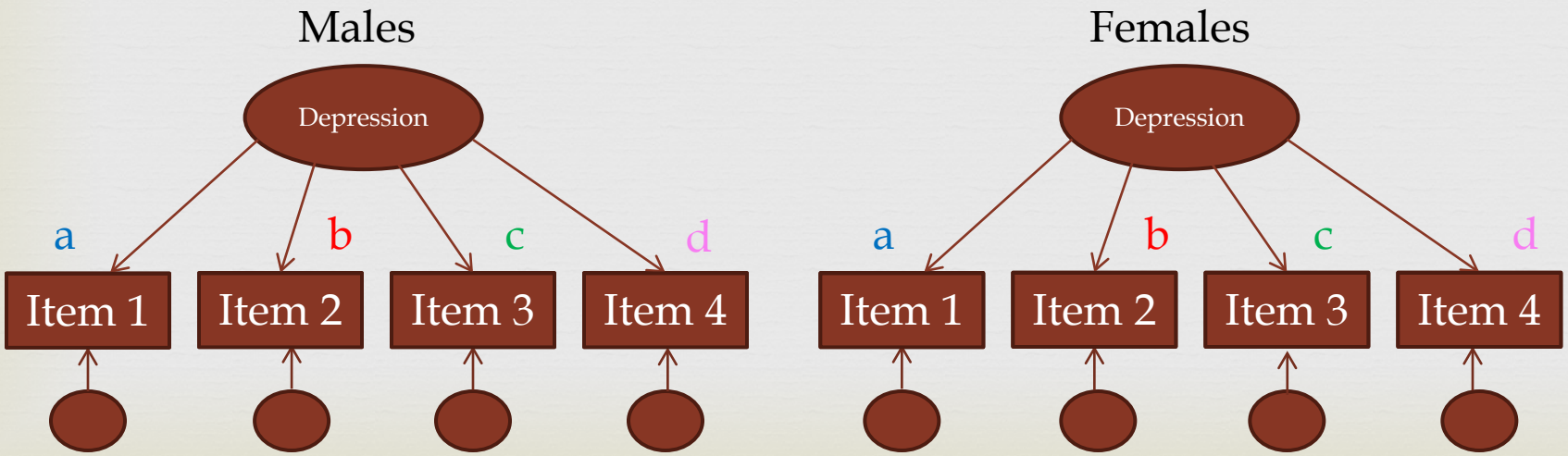
Assessing MI

- There are multiple levels of MI
  - Configural Invariance: same factor structure
  - **Metric (Weak) Invariance: factor loadings are equal**
  - Scalar (Strong) Invariance: intercepts are equal
  - Strict Invariance: error variances are equal



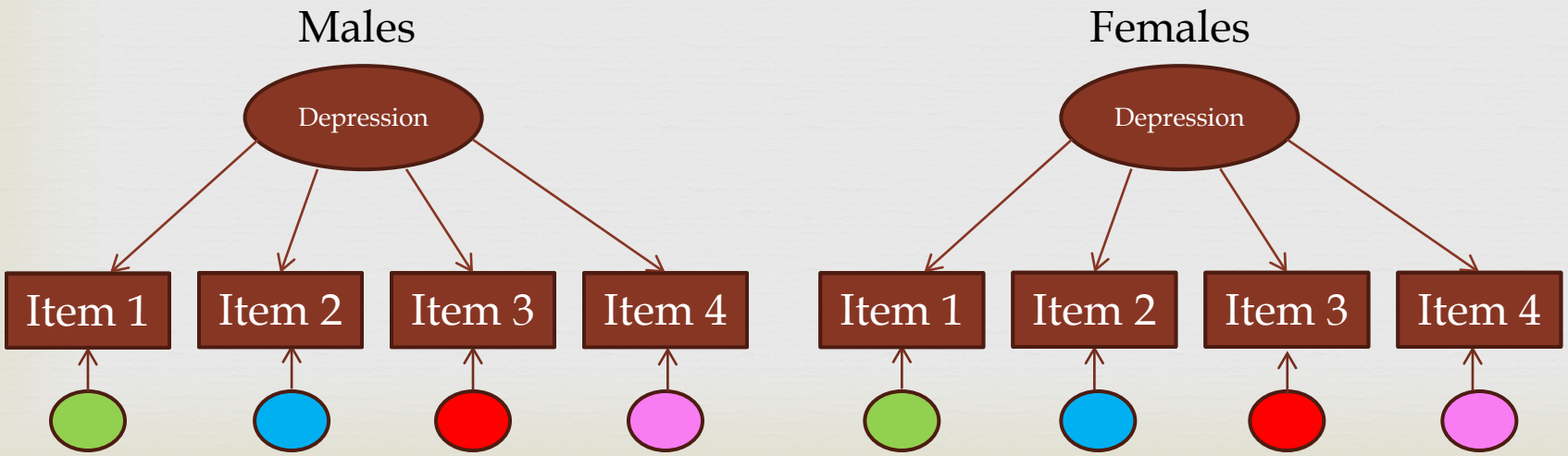
Assessing MI

- There are multiple levels of MI
  - Configural Invariance: same factor structure
  - Metric (Weak) Invariance: factor loadings are equal
  - **Scalar (Strong) Invariance: intercepts are equal**
  - Strict Invariance: error variances are equal



Assessing MI

- There are multiple levels of MI
  - Configural Invariance: same factor structure
  - Metric (Weak) Invariance: factor loadings are equal
  - Scalar (Strong) Invariance: intercepts are equal
  - **Strict Invariance: error variances are equal**



Traditional Approaches to MI

- Nonsignificant χ^2 difference test
  - Nested model comparisons ($\chi^2_{\text{more constrained}} - \chi^2_{\text{less constrained}}$)
  - Example (metric invariance): Is the χ^2 fit statistic significantly smaller if we fix the factor loadings to be equal in males and females than if we let males and females have unique loadings?

- Using change in fit indices (ΔGOF)
  - ΔCFI (Comparative Fit Index)
  - ΔMNCI (Macdonald's Noncentrality Index)
  - ΔRMSEA (Root Mean Square Error of Approximation)

Issues with Traditional Approaches to MI

- The χ^2 difference test approach has several limitations:
  - "Accepting" the null hypothesis
  - Unrealistic to expect *zero* difference in any parameters between groups
  - Power to find invariance is highest when sample sizes are small
- Fit indices are descriptive in nature and there is much debate about appropriate ΔGOF cut-offs

Equivalence Tests for MI

- Yuan and Chan (2016) proposed using equivalence testing principles to evaluate MI
- Equivalence testing null hypotheses:
  - $F_{ml0} > \varepsilon$ at the configural stage
  - $F_{bc0} - F_{b0} > \varepsilon$ for all subsequent stages

where F_{ml0} is the population fit function, and $F_{bc0} - F_{b0}$ is the difference in fit functions of two nested models where *b* indexes the baseline model and *bc* indexes the baseline model with constraints, and ε is the largest tolerable amount of model misspecification

What is ε?

- As discussed earlier, one of the biggest challenges with equivalence testing is setting an appropriate equivalence interval
- Yuan and Chan (2016) relate ε to the RMSEA

$$\varepsilon = \frac{df(RMSEA^2)}{K}$$

- Where *df* is the model degrees of freedom at the configural stage or difference in *df* when comparing nested models, *K* is the number of groups
- This value is rescaled into a noncentrality parameter (ncp) for use in calculating a noncentral χ^2 statistic

$$\delta = (N-K)\varepsilon$$

Yuan and Chan's Recommended Adjustment

- Despite outlining this test statistic with conventional RMSEA values used for calculating ε, Yuan and Chan argue that power based on these values is too low

- They provide functions for an adjusted ε/RMSEA but provide little theoretical or empirical justification for the adjusted test's performance

Simulation Study

- We evaluated the power and Type I error control of Yuan and Chan's outlined equivalence testing method (EQ) and their recommended modification using adjusted RMSEA values (EQ-A)

- The equivalence tests' performance was compared to using a nonsignificant χ^2 difference test and ΔGOF
  - For simplicity the results for the ΔGOF method are not presented, but a brief conclusion regarding the method is provided

Conditions

- Measurement Model
  - 2 factors with either 4 or 8 indicators each
- Sample Size
  - 100, 250, 500, 1000, or 2000 per group
- Equivalence Bound (ε)
  - Based on RMSEAs of .05, .08, or .10
- Factor Loadings
  - .5, .7, .9

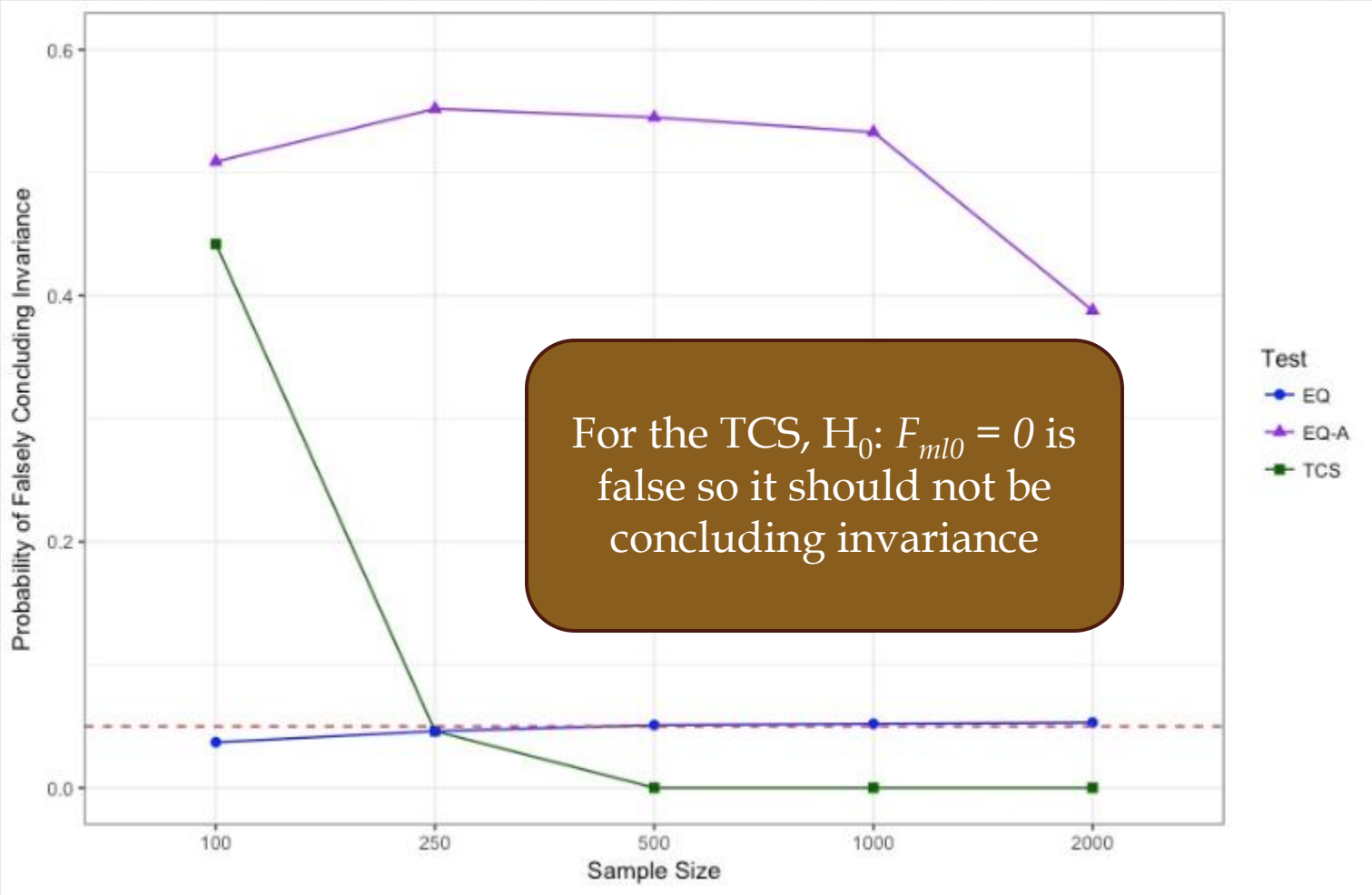
Type I Error and Power

- To evaluate Type I error rates, the population model misspecification was created such that $F_{ml0} = \varepsilon$ in each group at the configural stage and $F_{bc0} - F_{b0} = \varepsilon$ at all other stages
  - Error covariances were added to differing observed variables in each group at the configural stage to invoke lack of fit
  - At later stages, either one parameter differed or 25% of parameters differed (e.g., loading, intercept, error variance)

- To evaluate power we tested a condition where the groups' population models were identical and one where there were differences smaller than ε

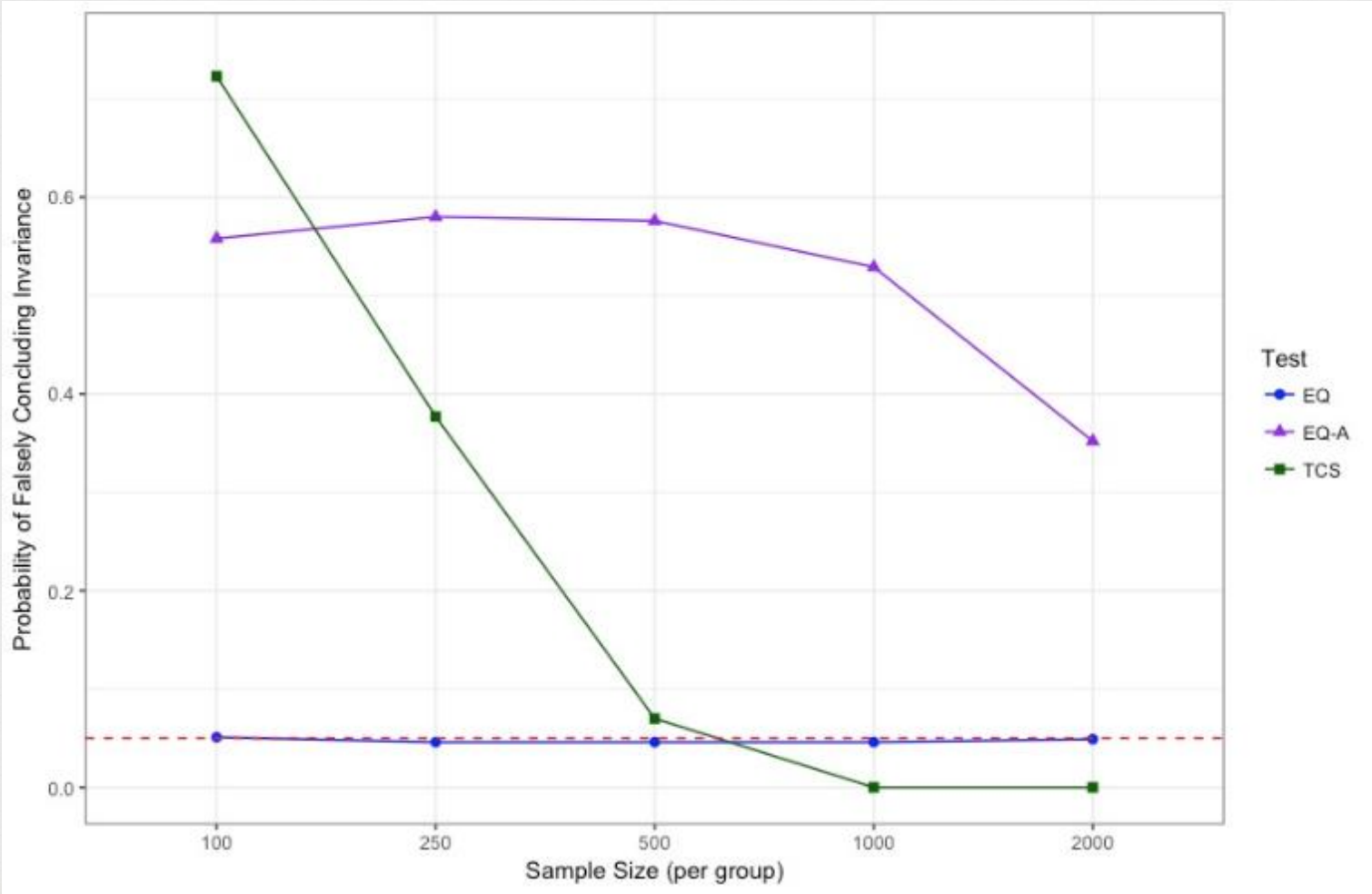
Type I Error Results

Configural Invariance: RMSEA = .08, 4 indicator Model



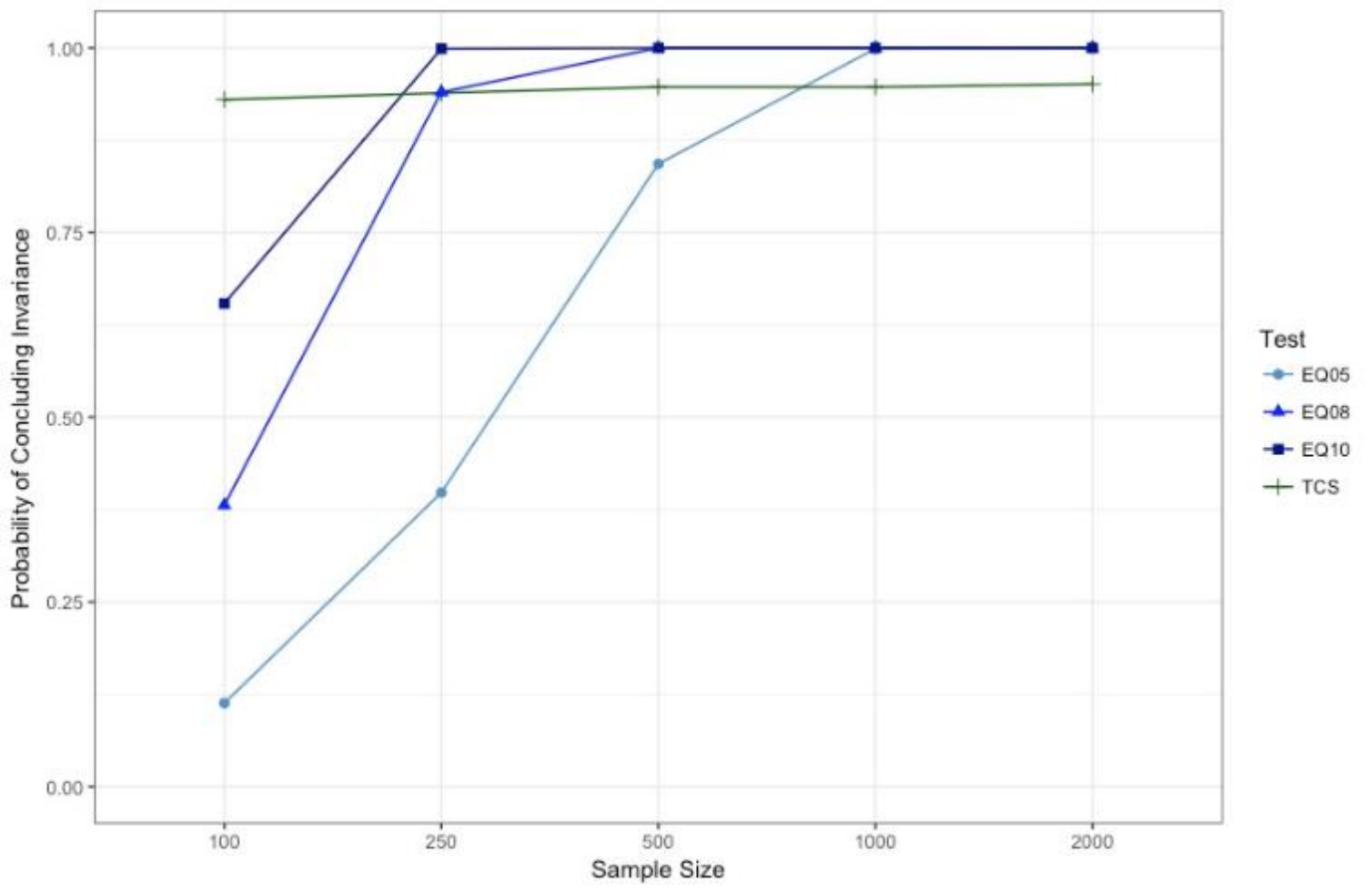
Type I Error Results

Metric Invariance: RMSEA = .08, 4 indicator Model



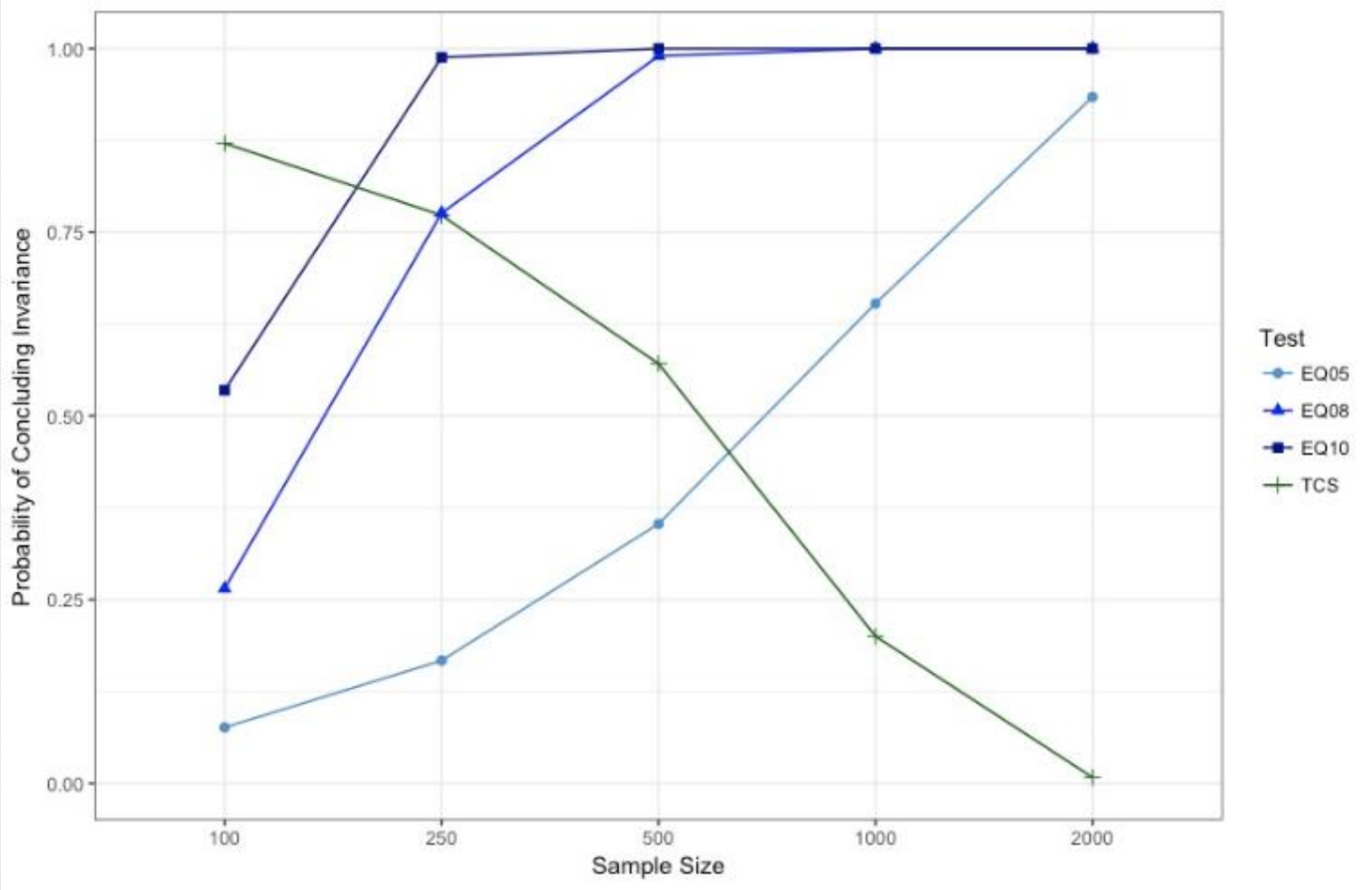
Power Results

Configural Invariance: 4 indicators, equal models



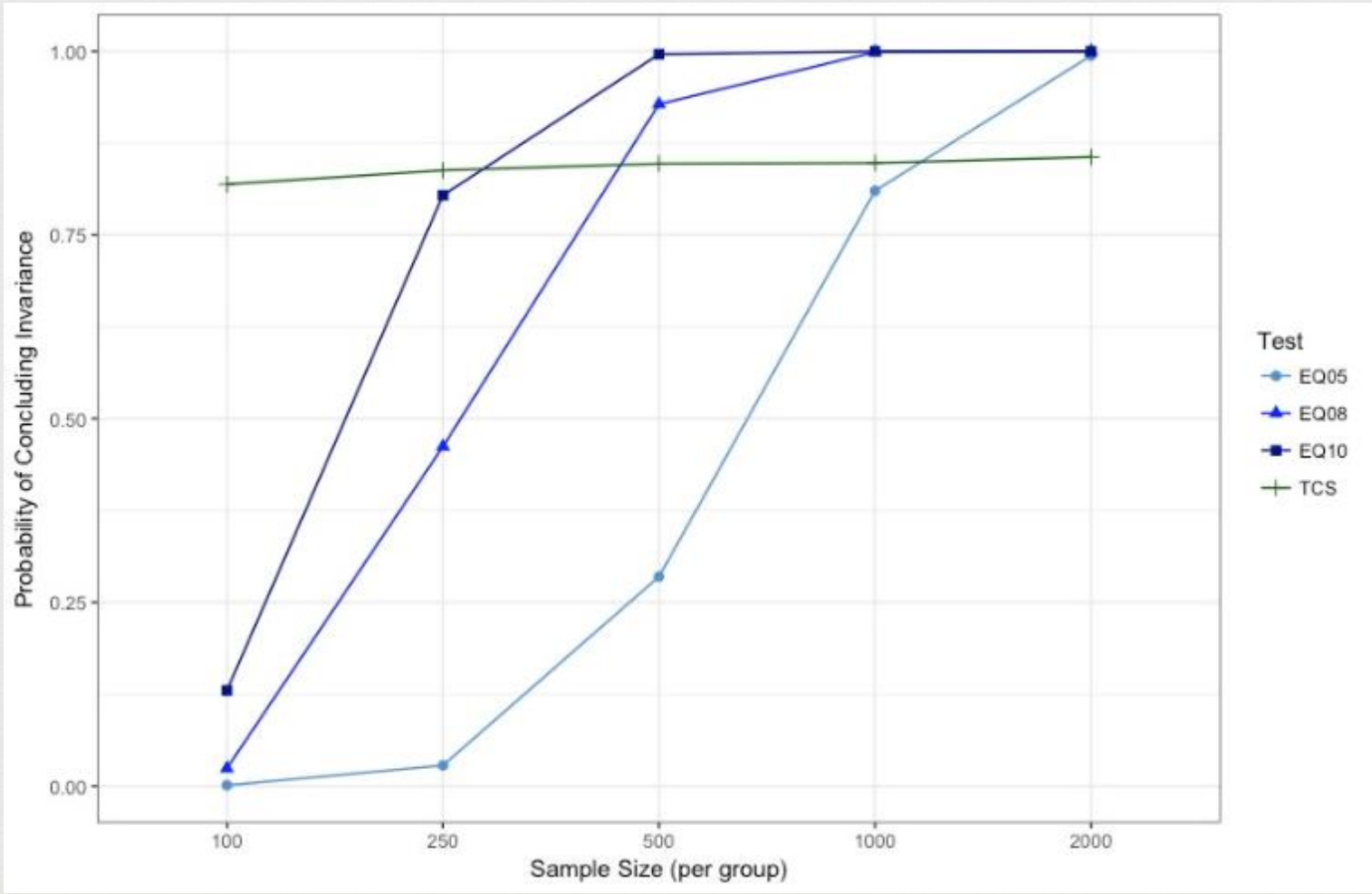
Power Results

Configural Invariance: 4 indicators, slightly different models



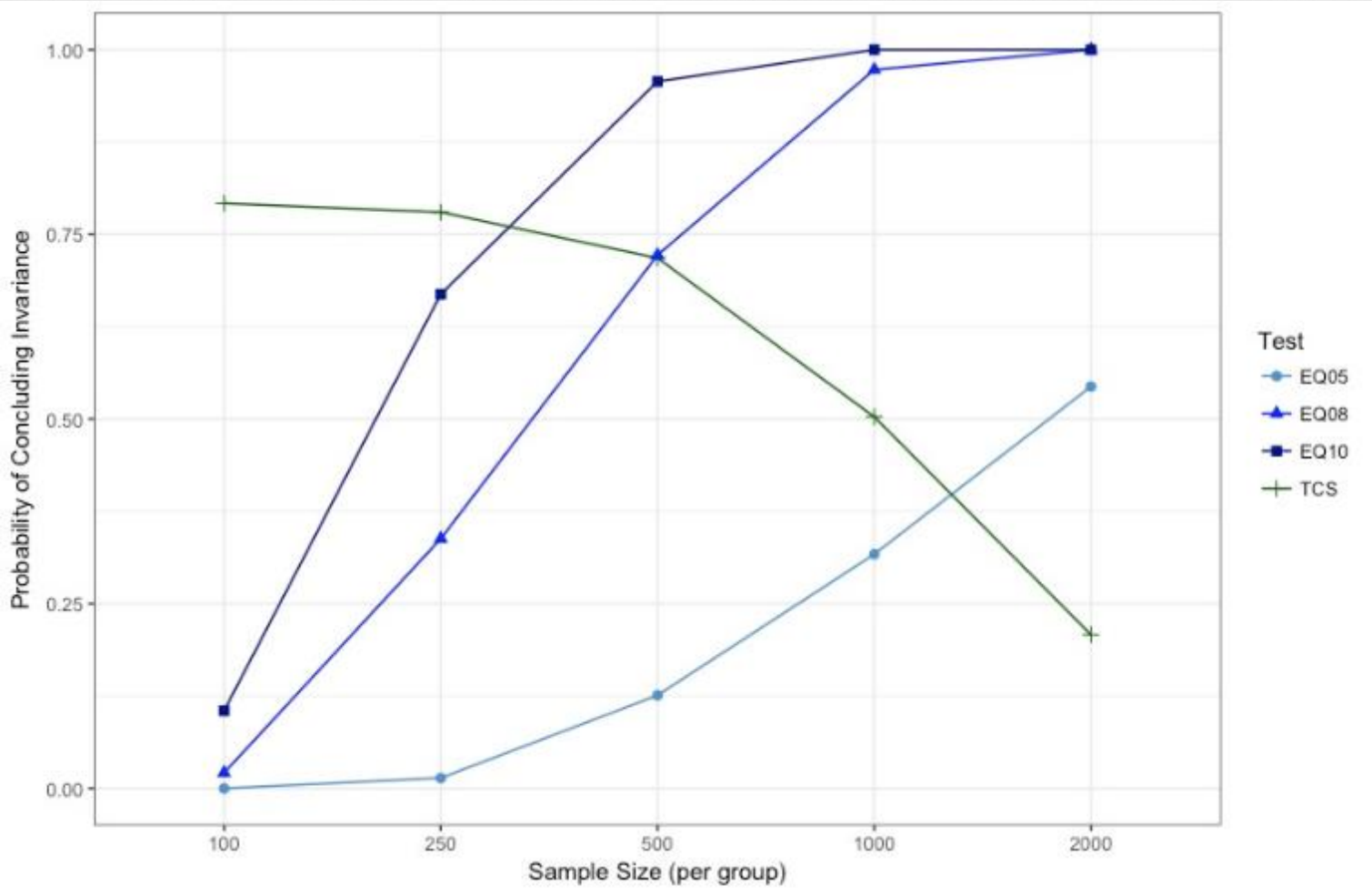
Power Results

Metric Invariance: 4 indicators, equal group models



Power Results

Metric Invariance: 4 indicators, slightly different models



Results Summary

- As expected, the χ^2 difference test results in illogical properties such as backwards power for finding invariance
- Yuan and Chan's EQ-A approach increased power at the expense of Type I error control
- The EQ method demonstrated good statistical properties, although power is low with very small sample sizes

- Power for finding invariance using fit indices depended on the degree of population misspecification
  - Performance of CFI cut-off strongly depended on condition

Conclusion

- The χ^2 difference test is not appropriate for establishing MI
- ΔGOF using a rigid adoption of common cut-offs is not recommended
- Equivalence testing is the logical statistical tool for testing MI
  - Yuan and Chan's (2016) adjusted RMSEA method is not recommended
    - No theoretical justification and very liberal Type I error rates
  - Some caution is required for the EQ approach with larger *ncp*s and small sample sizes

General Conclusion

- Researchers in Psychology frequently explore equivalence-based hypotheses
- Equivalence tests are rarely adopted because of unfamiliarity with the methods and lack of availability of software
- It is hoped that Psychology researchers will begin to recognize situations in which an equivalence test would be appropriate and that the software for conducting these tests becomes more available and user-friendly

Thanks!

cribbie@yorku.ca
cribbie.info.yorku.ca