Equivalence Testing Approaches to Measurement Invariance

Alyssa Counsell & Robert Cribbie
Department of Psychology, York University, Toronto, ON

INTRODUCTION

- Measurement Invariance (MI) is important for demonstrating lack of measurement bias across groups
- There are four main stages of MI: Configural, Metric (Weak), Scalar (Strong) and Strict
- The most common testing approach is using a non-significant χ^2 (difference) statistic
- This approach has important limitations:
  - Power to find invariance does not function in a logical way (i.e., increasing *N* decreases a researcher's power to find support for the research hypothesis that MI has been met)
  - It is unrealistic to expect identical parameters across groups

Equivalence Testing

- Equivalence testing addresses the limitations above
  - The research hypothesis is aligned with the alternative hypothesis as opposed to the null hypothesis
- It involves specifying an a priori interval that represents the smallest effect size that a researcher considers practically meaningful to conclude a difference

Equivalence Tests for Measurement Invariance

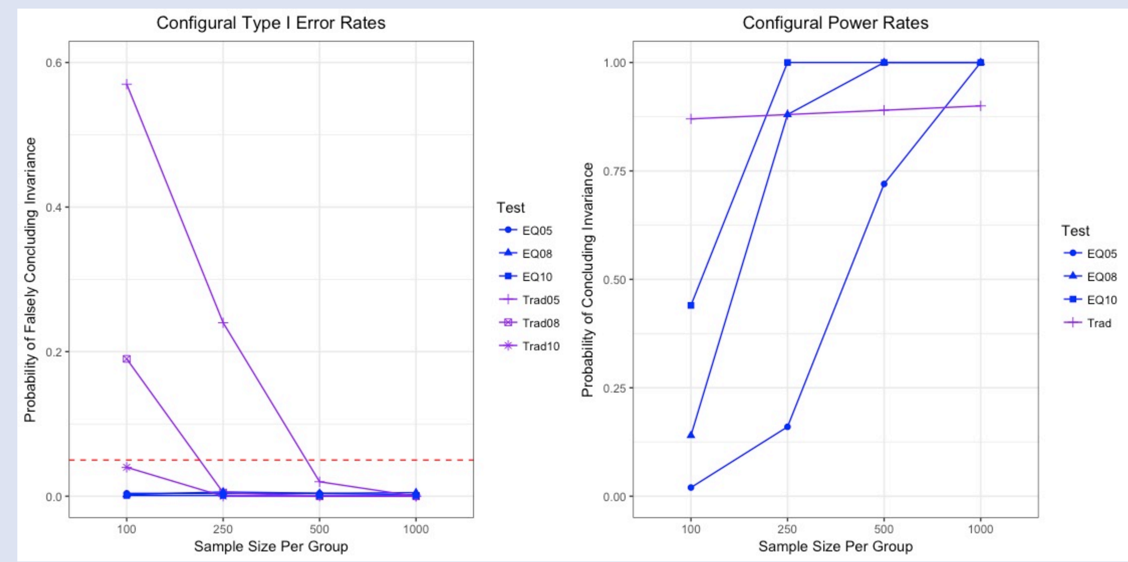
- Yuan and Chan (2016) proposed equivalence test versions of the χ^2 and χ^2 difference tests for SEM models
- For configural invariance, model fit is assessed separately in each group using the χ^2 statistic with the following null and alternative hypotheses: H_0: $F_{ml0} \geq \varepsilon_0$ and H_1: $F_{ml0} < \varepsilon_0$
  - ε_0 is the researcher specified equivalence bounds calculated as $\varepsilon_0 = df\ (RMSEA^2)/K$ and K is the number of groups
- The value of ε_0 is used to calculate a population noncentrality parameter, $\delta_0 = (N\text{-}K)\ \varepsilon_0$
- With $c_\alpha(\varepsilon_0)$ as the left-tail critical value of the noncentral $\chi^2(\delta_0)$ at probability α, one rejects the null hypothesis when $\chi^2(\delta_0) \leq c_\alpha(\varepsilon_0)$
- Similarly, one can apply the same logic to an equivalence based version of the χ^2 difference test
- Its null and alternative hypotheses are: H_0: $F_{bc0} - F_{b0} \geq \varepsilon_0$ and H_1: $F_{bc0} - F_{b0} < \varepsilon_0$
  where the *b* subscript refers to the baseline model and the *bc* refers to the baseline model with additional parameter constraints
- The rest of the details of the test remain the same as described above except the *df* refer to the difference in *df* between the models
- This test is used to establish metric, scalar, and strict invariance
- Rejection of the null hypothesis means that the added parameter constraints do not significantly worsen model fit and therefore the corresponding MI stage would be concluded

METHOD

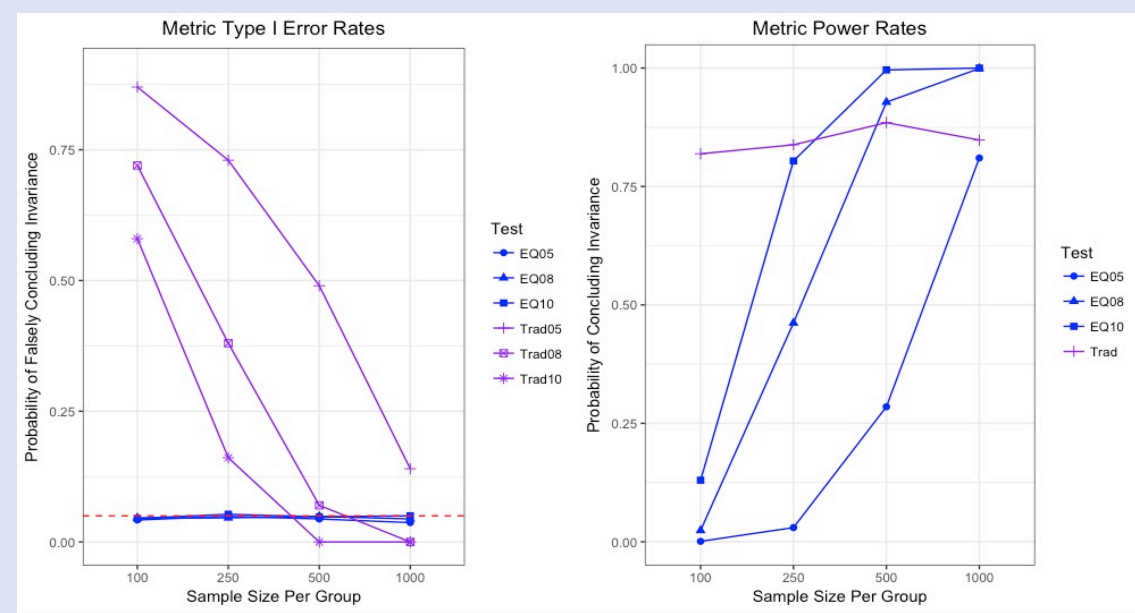
- A simulation study was conducted in R to compare the power and Type I error rates of Yuan and Chan's proposed equivalence test for MI to using a traditional nonsignificant χ^2 result for MI
- Several different conditions were examined:
  - Measurement Model: 4 indicators or 8 indicators per factor (both with 2 factors)
  - Different reliability scores for factor loadings of .5, .7, and .9
  - Differing δ_0 values based on RMSEA values of .05, .08, and .10
  - Two groups with sample sizes of 100, 250, 500, or 1000 per group
- The empirical power and Type I error rates of the MI tests were obtained by dividing the number of true (power) or false (Type I error) rejections by the number of replications (5000).
- To investigate Type I error rates, data were generated from a population model where the misspecification based on the χ^2 statistic was equal to the noncentrality parameter, δ_0 (i.e., $F_{ml0} = \varepsilon_0$ or $F_{bc0} - F_{b0} = \varepsilon_0$)
- To investigate power rates, data were generated from identical population models
- In the results, "Trad" refers to the traditional χ^2 test and "EQ", the equivalence test, whereas the 05, 08 and 10 refer to the values of the RMSEA used for calculating ε_0

RESULTS

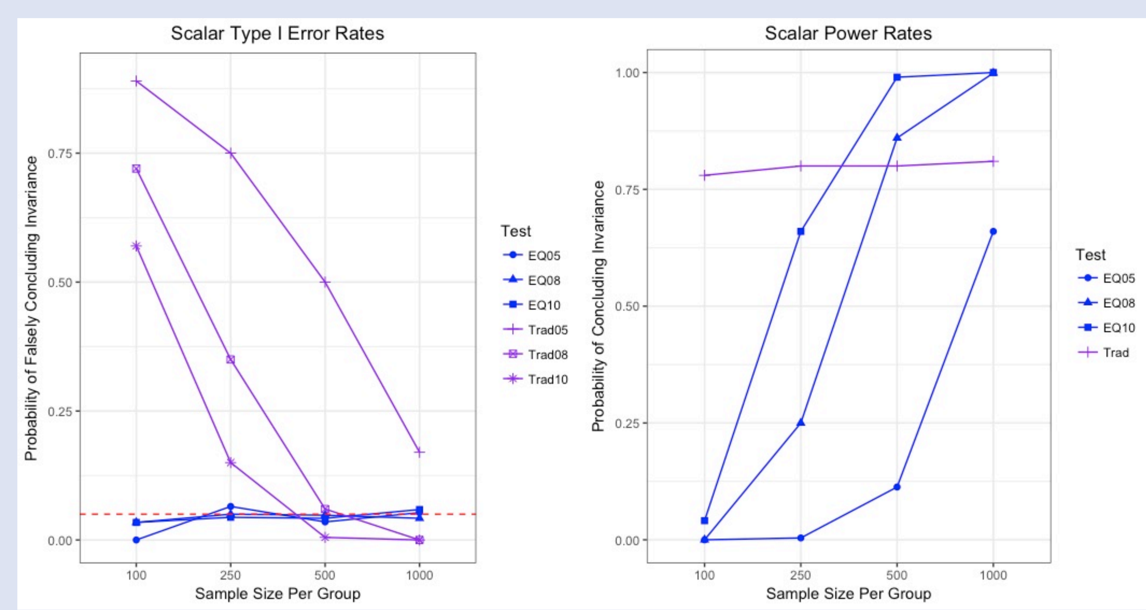
CONFIGURAL INVARIANCE



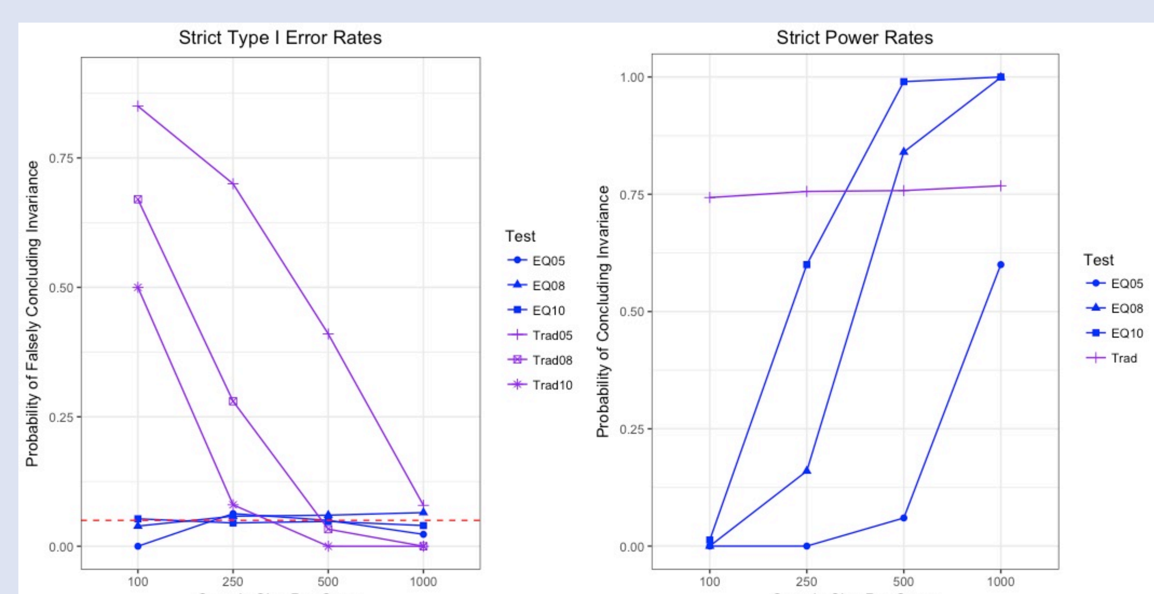
METRIC INVARIANCE



SCALAR INVARIANCE



STRICT INVARIANCE



CONCLUSION

- Yuan and Chan's equivalence test for MI provided accurate Type I error rates across all of the conditions
- Due to the stepwise nature of MI testing, the equivalence test's power to reach the final stage of strict invariance requires relatively large sample sizes or a larger equivalence interval (e.g., using RMSEA =.10)
- As such, conventional cut-off values for RMSEA (e.g., .05,.08) may be too conservative for MI testing
- Using the traditional χ^2/χ^2 difference statistic is statistically inappropriate for invariance testing because its ability to detect invariance when there is no model misspecification from the population model does not change across different sample sizes
- When small amounts of model misspecification are present, the ability to conclude invariance actually decreases with larger sample sizes when using a non-significant χ^2 result

Future Directions

- Further research into empirically justified adjustments or decisions regarding the equivalence interval is warranted
- Extending the conditions to examine the impacts of partial invariance, non-normal data/including robust estimators is worthwhile

DEMONSTRATION

- To demonstrate the differences in conclusions and procedure using the traditional methods for MI vs. the equivalence test approach, I analyzed data on the Generic Conspiracy Beliefs Scale to test for MI between males and females
- The measurement model had 5 factors with 3 indicators each
- There were no cross-loadings, but each factor had a covariance with each of the other factors
- $N = 2359$ ($n_{\text{male}} = 1222$, and $n_{\text{female}} = 1137$)
- To calculate the equivalence interval (ε_0) and noncentrality parameter (δ_0), I used an RMSEA of .08 as the smallest amount of model misspecification that would be considered practically important

MI level	χ^2	*df*	δ_0	Trad *p*	EQ *p*
Configural: Male	408.79	80	625.15	< .001*	< .001
Configural: Female	364.92	80	581.63	< .001*	< .001
Metric	17.44	10	75.42	.065	< .001
Scalar	43.60	10	75.42	< .001	.004
Strict	153.84	15	113.14	N/A	.88

*configural invariance was concluded based on adequate fit indices such as CFI, RMSEA
Note: For concluding invariance, a researcher would look for $p > \alpha$ using the traditional χ^2 and $p < \alpha$ using the equivalence test method

- Using the traditional χ^2 tests, one would conclude metric invariance holds
  - The factor loadings are considered invariant for males and females
- Using the equivalence test, one would conclude scalar invariance holds
  - The factor loadings and intercepts are considered invariant for males and females

REFERENCES

Yuan, K. H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological methods*, *21*, 405-426.