A Note on Effect Size for Measurement Invariance

Sunthud Pornprasertmanit

March 12, 2025

This article aims to show the mathematical reasoning behind all effect sizes used in the `partialInvariance` and `partialInvarianceCat` functions in `semTools` package. In the functions, the following statistics are compared across groups: factor loadings, item intercepts (for continuous items), item thresholds (for categorical items), measurement error variances, and factor means.

The comparison can be compared between two groups (e.g., Cohen's d) or multiple groups (e.g., R^2). This note provides the details of the effect sizes in comparing two groups only. The comparison between multiple groups can be done by picking the reference group and compare the other groups with the reference group in the similar fashion to dummy variables. For example, the comparison between four groups would create three effect size values (i.e., Group 1 vs. Reference, Group 2 vs. Reference, and Group 3 vs. Reference). Alternatively, for the measurement invariance, the change in comparative fit index (CFI) can be used as the measure of effect size. In the measurement invariance literature [Cheung and Rensvold, 2002, Meade et al., 2008], the change in CFI is used to test the equality constraints for multiple items simultaneously. The functions in `semTools` will show the change in CFI for each individual item. That is, if an item were to allow to have different statistics (e.g., loading), how large the CFI would drop from the original model. Please note that more research is needed in finding the appropriate cutoffs for the change in CFI for individual items. Are the cutoffs of .002 or .01 appropriate for this context?

In creating effect size, a target statistic needs to be standardized. Sample variances are used in the standardization formula. If researchers can assume that target variances across groups are equal in population, then pooled variances can be used in the standardization. The pooled variance s_P^2 can be computed as follows:

$$s_P^2 = \frac{\sum_{g=1}^{G}(n_g - 1)s_g^2}{\sum_{g=1}^{G}(n_g - 1)},$$

where g represents the index of groups, G is the number of groups, s_g^2 represents the variance of Group g, and n_g is the Group g size. If the variances are not assumed to be equal across groups, I recommend to pick a reference (baseline) group for the standardization.

In the following sections, I will show how effect sizes are defined in each type of partial invariance testing.

1 Factor Loading

Let λ_{ijg} be the unstandardized factor loading of Item i from Factor j in Group g. A standardized factor loading λ_{ijg}^* can be computed [Muthén, 1998–2004]:

$$\lambda_{ijg}^* = \lambda_{ijg} \cdot \frac{\psi_{jg}}{\sigma_{ig}},$$

where ψ_{jg} is the standard deviation of Factor j from Group g and σ_{ig} is the total standard deviation of Item i from Group g. To quantify the difference in factor loadings between groups in standardized scale, the standard deviation in the standardization formula needs to be the same across groups. If Group A and Group B are compared, the standardized difference in factor loading is defined:

$$\Delta\lambda_{ij}^* = (\lambda_{ijA} - \lambda_{ijB}) \cdot \frac{\psi_{jP}}{\sigma_{iP}},$$

where ψ_{jP} is the pooled standard deviation of Factor j and σ_{iP} is the pooled total standard deviation of Item i. If Group A is the reference group, ψ_{jA} and σ_{iA} can substitute ψ_{jP} and σ_{iP}. Assume that standardized factor loadings are from congeneric measurement model, standardized factor loadings represent the correlation between items and factors. Cohen [1992] provide a guideline for interpreting the magnitude of the difference in correlations for independent groups. The correlations are transformed to Fisher's z transformation:

$$q = \arctan\left(\lambda_{ijA} \cdot \frac{\psi_{jP}}{\sigma_{iP}}\right) - \arctan\left(\lambda_{ijB} \cdot \frac{\psi_{jP}}{\sigma_{iP}}\right)$$

Then, the q values of .1, .3, and .5 are interpreted as small, medium, and large effect sizes.

For continuous outcomes, the amount of mean differences implied by the factor loading difference given a factor score can be used as an effect size [Millsap and Olivera-Aguilar, 2012]. Let X_ijg be the observed score of Item i loaded on Factor j from Group g and W_j represents the score of Factor j. The expected value of the observed score differences between Group A and Group B is calculated as follows:

$$E\left(X_{iA} - X_iB|W_j\right) = (\nu_{iA} - \nu_{iB}) + (\lambda_{ijA} - \lambda_{ijB}) \times W_j,$$

where ν_{ig} represents the intercept of Item i in Group g. Let the values between W_{jl} and W_{jh} be the values of interest. We can find the expected difference in the observed scores under this range of the factor scores. Millsap and Olivera-Aguilar [2012] proposed that, if the size of the expected difference is over the value of meaningful differences, the loading difference is not negligible. See their article for the discussion of the meaningful difference.

Note that, in the `partialInvariance` function, W_{jl} is calculated by (a) finding the factor scores representing a low z-score (e.g., -2) from all groups and (b) selecting the lowest factor score across all groups. W_{jh} is calculated by (a) finding the factor scores representing a high z-score (e.g., 2) from all groups and (b) selecting the highest factor score across all groups.

2 Item Intercepts

Let ν_{ig} be the intercept of Item i in Group g. A standardized intercept ν^*_{ig} is defined as follows [Muthén, 1998–2004]:

$$\nu^*_{ig} = \nu_{ig}/\sigma_{ig}.$$

Thus, the standardized difference between Groups A and B in item intercepts is defined:

$$\Delta\nu^*_i = (\nu_{iA} - \nu_{iB})/\sigma_{iP}.$$

Note that σ_{iA} can substitute σ_{iP} if Group A is the reference group. By using this scale, .2, .5, and .8 can be interpreted as small, medium, and large effect sizes according to Cohen [1992].

The proportion of the intercept difference over the observed score difference may be used as an effect size [Millsap and Olivera-Aguilar, 2012]:

$$(\nu_{iA} - \nu_{iB})/(M_{iA} - M_{iB}),$$

where M_{ig} represents the observed mean of Item i in Group g. Millsap and Olivera-Aguilar [2012] noted that a relatively small proportion (e.g., less than 20%) is ignorable. If the sign is negative or the value is over 1, the interpretation is doubtful.

3 Item Thresholds

Let τ_{cig} be the threshold categorizing between category c and $c+1$ for Item i in Group g. Note that the maximum number of c is the number of categories minus 1. Because thresholds are the location of the distribution underlying ordered categorical items (usually normal distribution), the location statistic can be standardized by dividing it by the standard deviation of the underlying distribution. The standardized threshold τ^*_{cig} is defined as follows:

$$\tau^*_{cig} = \tau_{cig}/\sigma^u_{ig},$$

where σ^u_{ig} is the standard deviation of the distribution underlying the categorical data for Item i in Group g. In theta parameterization of categorical confirmatory factor analysis, σ^u_{ig} may not be equal across groups. The standardized difference in thresholds between Group A and B needs the pooled standard deviation. The standardized difference in thresholds is defined:

$$\Delta\tau^*_{ci} = (\tau_{ciA} - \tau_{ciB})/\sigma^u_{iP}.$$

Note that σ^u_{iA} can substitute σ^u_{iP} if Group A is the reference group. By using this scale, .2, .5, and .8 can be interpreted as small, medium, and large effect sizes according to Cohen [1992].

4 Measurement Error Variances

Let θ_{ig} be the measurement error variance of Item i in Group g. A standardized measurement error variance θ^*_{ig} is defined as follows [Muthén, 1998–2004]:

$$\theta^*_{ig} = \theta_{ig}/\sigma_{ig},$$

Thus, the standardized difference between Groups A and B in measurement error variances could be defined:

$$\Delta\theta^*_i = (\theta_{iA} - \theta_{iB})/\sigma_{iP}.$$

Note that σ_{iA} can substitute σ_{iP} if Group A is the reference group. However, there is no direct guideline to interpret the magnitude of the difference in measurement error variances according to Cohen (1992). A new standardized difference in measurement error variances is needed.

Assume that σ_{iP} is always greater than θ_{iA} and θ_{iB}, which is usually correct, then θ_{iA}/σ_{iP} and θ_{iB}/σ_{iP} ranges between 0 and 1 similar to a proportion statistic. Cohen [1992] provided a guideline in interpreting the magnitude of the difference in proportions using arcsine transformation. The new index (h) is defined as follows:

$$h = \sin^{-1}\sqrt{\frac{\theta_{iA}}{\sigma_{iP}}} - \sin^{-1}\sqrt{\frac{\theta_{iB}}{\sigma_{iP}}}.$$

Then, the h values of .2, .5, and .8 are interpreted as small, medium, and large effect sizes.

If items are continuous, the proportion of the error variance difference over the observed variance difference may be used as an effect size [Millsap and Olivera-Aguilar, 2012]:

$$(\theta_{iA} - \theta_{iB})/(\sigma_{iA} - \sigma_{iB}).$$

If the sign is negative or the value is over 1, the interpretation is doubtful.

5 Factor Means

Let α_{jg} be the mean of Factor j in Group g. A standardized factor mean α^*_{jg} is defined as follows [Muthén, 1998–2004]:

$$\alpha^*_{jg} = \alpha_{jg}/\psi_{jg}$$

Thus, the standardized difference between Groups A and B in factor means is defined:

$$\Delta\alpha^*_j = (\alpha_{jA} - \alpha_{jB})/\psi_{jP}.$$

Note that ψ_{jA} can substitute ψ_{jP} if Group A is the reference group. By using this scale, .2, .5, and .8 can be interpreted as small, medium, and large effect sizes according to Cohen [1992].

References

Gordon W Cheung and Roger B Rensvold. Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling*, 9(2): 233–255, 2002.

Jacob Cohen. A power primer. *Psychological bulletin*, 112(1):155–159, 1992.

Adam W Meade, Emily C Johnson, and Phillip W Braddy. Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, 93(3):568, 2008.

Roger E Millsap and Margarita Olivera-Aguilar. Investigating measurement invariance using confirmatory factor analysis. In Rick H Hoyle, editor, *Handbook of structural equation modeling*, pages 380–392. Guilford, New York, 2012.

Bengt O Muthén. *Mplus technical appendices.* Muthén & Muthén, Los Angeles, CA, 1998–2004.