Appendix C: Effect size calculation

Appendix C: Effect size calculation

Calculation using d_{MACS}: lack of statistics

Lack of unstandardized statistics is a usual challenge in MI studies because it is suggested that standardized loadings and intercepts for the SEM framework are adequate to report (Gunn et al., 2020). Researchers may need a formula that converts standardized statistics to unstandardized ones under a following scenario: Given $\hat{\mu}_\xi$ and $\hat{\psi}$, it is not complicated to compute the unstandardized loadings of each indicator using the formula $b_{item} = \frac{\beta_{item} SD_{item}}{\sqrt{\hat{\psi}}}$ (Brown, 2006), where b_{item} is unstandardized loading, β_{item} is the unstandardized loading, SD_{item} is the item standard deviation, $\sqrt{\hat{\psi}}$ is the factor standard deviation. Note that SD_{item} for a particular sample group may also not be available. Researchers need to either borrow information from other primary studies in which descriptive characteristics of samples are as similar as possible to the target study, or estimate it using alternative method.

Note that for categorical variables, the variance of a single group can be estimated using the formula (e.g., the reference group):

$$\theta_{iR} = \sum((j - \hat{X}_{iR})^2 P_{ijk}) = \sum j^2 P_{ijk} - \hat{X}_{iR}^2$$

. Consequently, the standard deviation of that group should be: $SD_{iR} = \sqrt{\theta_{iR}}$. As long as the sample sizes for comparing groups are reported, we can estimate the pooled standard deviation SD_{ip} and then calculate d_{MACS}.

Calculation using MIMIC model

MIMIC model is an important model in covariance structure analysis and can be thought of a special case of CFA (Bohrnstedt, 1977). It is characterized as at least one latent variable with both multiple indicators (i.e., item questions as independent variables) and multiple causes (i.e., latent variable as the independent variable). One study included in our systematic review indicated that MIMIC model is a preferable model to detect DIF compared to other methods using IRT because it can generate proportional odds ratio expressing DIF (Yang & Jones, 2007). Scott et al. (2010) proposed that odds ratio can indicate the magnitude of DIF (i.e., a larger odds ratio indicates a larger magnitude of DIF). Therefore, proportional odds ratio reported by the article using MIMIC model or similar method (e.g., Mantel-Haenszel method) can be directly used as the effect size measure of measurement non-invariance.

Differential item functioning using item response theory.

Ordinary logistic regression method (OLR). Zumbo (1999) summarized that logistic regression is one of the most productive and recommended methods in the DIF analysis for binary or ordinal scored items. In the OLR method, the item responses are deemed as the criterion variable, which could be binary or ordinal. The independent variables include the grouping variable (e.g., dummy coded as 0 = reference group, 1 = focal group), total scale score for each observation (e.g., coded as TOT), and possible moderators (i.e., variables that contribute to the interaction effects). This method provides a DIF testing by a linear regression of a group of predictor variables on a latent trait variable. The formula of OLR is: $y^* = b_0 + b_1 TOT + b2GROUP + b3TOT * GROUP + \epsilon_i$, where $y^* = ln[\frac{p_i}{(1-p_i)}]$ (i.e., a natural log of the odds ratio; pi represents the proportion of endorsing the item underlying the latent trait).

As Zumbo (1990) and McKelvey & Zavoina (1975) suggested, the effect size measuring of DIF follows the same procedures as the statistical hypothesis testing for DIF in which researchers only need to work on the R-squared values. The procedures of DIF modeling imply a hierarchy of entering variables into the model: (1) First, entering the conditioning variable (i.e., total score); (2) entering the group variable; (3) entering the interacting variables. At each step, researchers are able to conduct chi-squared test and compute the model statistics of statistical tests for DIF (e.g., χ^2 value, R^2). To sum up, the R^2 measures can be used to measure the effect size in the hierarchical sequential modeling process for ordinal logistic regression models.

Graded response model. Graded response model (GRM) is another IRT model in which an item has ordered response categories and used to estimate location and information parameters for each item given the raw data (Samejima, 1997). Some studies used GRM to conduct DIF analysis as the authors consider item questions in the CES-D are ordinal instead of continuous. The GRM is extended and specified by Cohen et al. (1993): $P_j(\theta) = [1 + e^{-a_j(\theta - b_j)}]^{-1}$, where a_j is the discrimination parameter for item j; b_j is the difficulty parameter for item j; θ is the latent trait parameter.

The idea of calculating the effect size of DIF in GRM is comparable to that in categorical CFA. Auné et al. (2019) provided a practical instruction of determining the probability that an observation endorses a certain category in each item. Since the literature using GRM in our study included a specialized D parameter with a certain value (i.e., $D = 1.7$) in the model, the formula will be modified to: $P^*_m(\theta) = [1 + e^{-1.7a_j(\theta - b_j)}]^{-1}$. Here m is a certain category in an item, and hence $P^*_m(\theta)$ represents the probability that an observation with a latent trait score endorses a category m or greater in an item. When $m = 1$, $P^*_m(\theta) = 1$, because the cumulative probability of responding to the lowest category is a certain event (i.e., the observation must choose at least a category to give a response to the item); when $m = M + 1$ (M is the total number of categories), $P^*_m(\theta) = 0$ because it is impossible to choose a nonexistent category. Thus, the probability of a given category is defined as: $P_m(\theta) = P^*_m(\theta) - P^*_{m-1}(\theta)$. The idea of determining the category probabilities by Samejima (1969) and Auné et al. (2019) further helps calculate the expected score of a certain group for a given item using the expected mean and variance formula demonstrated in Wackerly et al. (2008). As a result, d_{MACS} can be computed given the available information.

Example of GRM:

```
es_grm <- function(d, a_ref, a_foc, b_1_ref, b_2_ref, b_3_ref, b_1_foc, b_2_foc, b_3_foc

  integrand <- function(theta) {

    # Reference Group (Female)

  # calculate cumulative probability using the GRM function
  prob_0_ref_temp <- 1
  prob_1_ref_temp <- 1/(1 + exp(-d*a_ref*(theta + b_1_ref)))
  prob_2_ref_temp <- 1/(1 + exp(-d*a_ref*(theta + b_2_ref)))
  prob_3_ref_temp <- 1/(1 + exp(-d*a_ref*(theta + b_3_ref)))
  prob_4_ref_temp <- 0

  # calculate the probability of each category
  prob_0_ref <- prob_0_ref_temp - prob_1_ref_temp
  prob_1_ref <- prob_1_ref_temp - prob_2_ref_temp
  prob_2_ref <- prob_2_ref_temp - prob_3_ref_temp
  prob_3_ref <- prob_3_ref_temp - prob_4_ref_temp

  # calculate the expected value
  exp_val_ref <- 0*prob_0_ref + 1*prob_1_ref + 2*prob_2_ref + 3*prob_3_ref

# Focal Group (Male)

  # calculate cumulative probability using the GRM function
  prob_0_foc_temp <- 1
  prob_1_foc_temp <- 1/(1 + exp(-d*a_foc*(theta + b_1_foc)))
  prob_2_foc_temp <- 1/(1 + exp(-d*a_foc*(theta + b_2_foc)))
  prob_3_foc_temp <- 1/(1 + exp(-d*a_foc*(theta + b_3_foc)))
  prob_4_foc_temp <- 0

  # calculate the probability of each category
  prob_0_foc <- prob_0_foc_temp - prob_1_foc_temp
  prob_1_foc <- prob_1_foc_temp - prob_2_foc_temp
  prob_2_foc <- prob_2_foc_temp - prob_3_foc_temp
  prob_3_foc <- prob_3_foc_temp - prob_4_foc_temp

  # calculate the expected value
  exp_val_foc <- 0*prob_0_foc + 1*prob_1_foc + 2*prob_2_foc + 3*prob_3_foc

# Expected value difference
  exp_val_diff <- exp_val_ref - exp_val_foc

# determine the distribution of the latent variable of the focal group
  exp_val_diff^2 * dnorm(theta, MeanF, sqrt(VarF))

  }

  density_value <- integrate(integrand, -Inf, Inf)
  ef_value <- sqrt(abs(unlist(density_value)$value))/sd_i
  ef_value

}
```

```
# Test the function
 es_grm(1.7, 1.36, 1.18, -0.69, 0.88, 2.18, -0.73, 1.04, 2.34, -0.15, 1.20, 0.78)
```

```
## [1] 0.08399943
```

Calculation using Item Characteristic Graphs (demonstration using Rasch model). Rasch model is a one-parameter logistic latent trait model proposed by Rasch (1980). Like logistic regression model used to model the latent variable, it shows the probability of an individual endorsing a response on a test item given their abilities on the latent trait. Generally, Rasch analysis generates a plot showing the person-item threshold distribution for the original set of items, and a set of item characteristic curves for target items. Under a scenario of lacking key information (e.g., λ_i for both reference and focal groups), we can use the idea of Riemann sum to approximate the area between the item characteristic curves between two groups. This area should be theoretically equivalent the squared difference between expected value of the reference and focal group at a given item (i.e., $(\xi)^2$. The definition of Riemann sum is: $\sum_{k-1}^{n} f(c_k)\Delta_{x_k} \approx \int_a^b f(c_k)$. Equivalently speaking, if f is defined on a closed interval $[a, b]$ and c_k can be any points in $[x_{k-1}, x_k]$, then the rectangle with width of Δ_{x_k} and height of $f(c_k)$ will be an area approximating the density. As the width of such a rectangle becomes smaller and smaller until infinitesimal, the sum of all the area (i.e., the Riemann sum) will be mostly close to the density value of $f(c_k)$ under the interval $[a, b]$. Once the approximated area is calculated, we can plug in the number into the d_{MACS} formula and calculate the effect size of non-invariance.

Example of Item Characteristic Graphs:

```
#CES_D_27 Item_17


# Using logistic distribution
 #proportion of participants in each bin; total number
```

```
  n_sample = 360
  mean_fac = -1.789
  sd_fac = 1.672
  sd_p = 0.716

  #also represents how many % of people in each bin
  n_bin1 = plogis(-3, -1.789, 1.672) - plogis(-4, -1.789, 1.672)
  n_bin2 = plogis(-2, -1.789, 1.672) - plogis(-3, -1.789, 1.672)
  n_bin3 = plogis(-1, -1.789, 1.672) - plogis(-2, -1.789, 1.672)
  n_bin4 = plogis(0, -1.789, 1.672) - plogis(-1, -1.789, 1.672)
  n_bin5 = plogis(1, -1.789, 1.672) - plogis(0, -1.789, 1.672)
  n_bin6 = plogis(2, -1.789, 1.672) - plogis(1, -1.789, 1.672)

  #total people between [-4, 3]
  n_total = (n_bin1 + n_bin2 + n_bin3 + n_bin4 + n_bin5 + n_bin6)*n_sample

# Using the histogram

  # bin(-4, -3)
  n_bin_1 = 30 + 1 + 41 + 1
  # bin(-3, -2)
  n_bin_2 = 34 + 27 + 22
  # bin(-2, -1)
  n_bin_3 = 1 + 19 + 1 + 20 + 27
  # bin(-1, 0)
  n_bin_4 = 27 + 10 + 8
  # bin(0, 1)
  n_bin_5 = 7 + 16 + 3 + 1
  # bin(1, 2)
  n_bin_6 = 6 + 3 + 4 + 3

  n_total_h = n_bin_1 + n_bin_2 + n_bin_3 + n_bin_4 + n_bin_5 + n_bin_6

# calculate the rectangle areas

  #area between [-4, -3]; in the form of (diff in expected value between male and fema
  area_1 <- (0.1 - 0)^2*(-3 - (-4))
  area_2 <- (0 - 0.2)^2*(-2 - (-3))
  area_3 <- (0.2 - 0.4)^2*(-1 - (-2))
  area_4 <- (0.1 - 0.4)^2*(0 - (-1))
  area_5 <- (0.5 - 1)^2*(1 - 0)
  area_6 <- (1 - 1.5)^2*(2 - 1)

  area_total <- area_1  + area_2  + area_3  + area_4  + area_5  + area_6

  #logistic distribution effect size
  es_ld = sqrt(area_1*n_bin1 + area_2*n_bin2 + area_3*n_bin3 + area_4*n_bin4 + area_5*n_

  #histogram effect size
  es_hs = sqrt((area_1*n_bin_1 + area_2*n_bin_2 + area_3*n_bin_3 + area_4*n_bin_4 + area

  #Take the average
  es_item17 = mean(c(es_ld, es_hs))
```