

## Model Evaluation with JRule The detection of misspecifications

William M. van der Veld - Radboud University, Nijmegen  
Willem E. Saris - ESADE, Barcelona  
Albert Satorra - Universitat Pompeu Fabre, Barcelona

1

---

---

---

---

---

---

---

---

## Overview

- What is the problem in model testing?
- An alternative to model testing
- Judgment Rule Aid (JRule)

2

---

---

---

---

---

---

---

---

## What is the problem in model testing?

- Model evaluation is based upon the (weighted) residuals:  $S \cdot \Sigma(p)$
- Residuals can be nonzero because of:
  - Misspecifications (incorrect theory)
  - Sampling fluctuations
- So, when should a model be rejected?
  
- When residuals deviate more from zero than can be expected by chance alone.
- This requires the specification of 'by chance alone'.
- Normally this is accomplished by selecting a probability level ( $\alpha$ ) for a test-statistic with a known probability distribution.
  
- Traditionally the test statistic  $T$  is used in SEM, more specifically,  $T_{ML}$ .  
$$T_{ML} = (N - 1) F(S, \Sigma(\hat{p})) = -2 \ln \left[ \frac{L(\hat{p}_{M_s})}{L(\hat{p}_{M_p})} \right]$$
- If the observed variables have a **multivariate normal distribution** and the **model is correct**.
- Then the test statistic  $T$  has a **chi-square distribution**.
- For a given  $\alpha$ -level, a model is rejected if  $T_{ML} > C_\alpha$

3

---

---

---

---

---

---

---

---



## What is the problem in model testing?

- What if a model is rejected according to this procedure?
- All estimates of the model parameters are based on the assumption that the model is correct and the data have a multivariate normal distribution.
- If the model is rejected we cannot trust the estimates!
- What now?
  - There are several possible causes for this rejection:
    - Misspecifications (model is wrong)
    - Violation of assumptions (e.g. non-normal distribution of the variables)
    - **Sample size & model complexity**
      - Size of incidental parameters
- The sensitivity of the CHI2-test to sample size and model complexity has lead to the development of alternative procedures.
- For example: GFI, AGFI, RMR, SRMR, CFI, RMSEA, and MI
- Are the alternative, so-called fit indices, performing any better than CHI2?

4

---

---

---

---

---

---

---

---

---

---



## What is the problem in model testing?

- Corten, Saris, Satorra (forthcoming) have shown that the behavior of many of the fit indices is very poor, using Hu & Bentler's (1999) suggestions.
- This is also illustrated in other articles, e.g. Beauducel & Wittmann (2005); Fan & Sivo (2005); Marsh, Hau, & Wen (2004); Yuan (2005), Saris & Satorra (in progress). See also Barret (2006) in the special issue on testing in PAID.
- What is the problem with the fit indices, but also with the CHI2?
  - Rejection of the model can be due to very small misspecifications for which the test is very sensitive (i.e. high power).
  - Acceptance of the model can happen despite very large misspecifications, if the test is insensitive to them (i.e. low power .....
- The bitter conclusion is that the standard model evaluation procedures do not satisfy the requirements that we wish them to satisfy, which is:

*"... if the model is truly a good model in terms of its fit in the population, we wish to avoid concluding that the model is a bad one.  
Alternatively, if the model is truly a bad one, we wish to avoid concluding that it is a good one."*

MacCallum et al. (1996)

5

---

---

---

---

---

---

---

---

---

---



## What is the problem in model testing?

- A possible solution has been suggested by Saris & Satorra with other authors in different publications in the eighties. They suggest to take the power of the test for the complete model into account.
  - Rejection of the model can be due to very small misspecifications for which the test is very sensitive (i.e. high power).
  - Acceptance of the model can happen despite to very large misspecifications, if the test is insensitive to them (i.e. low power .....

|            |                          |                       |
|------------|--------------------------|-----------------------|
|            | $T_{ML} \leq C_{\alpha}$ | $T_{ML} > C_{\alpha}$ |
| High Power |                          |                       |
| Low Power  |                          |                       |

- However, this approach is rather complex, and not very practical.
- So, a different perspective on model evaluation should be taken.

6

---

---

---

---

---

---

---

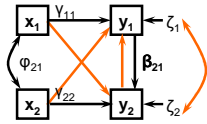
---

---

---

## An alternative to model testing

- Saris & Satorra (forthcoming) have suggested to switch from the so called model tests, which do not work as required, to testing for single misspecifications in the model. Previously this was performed with:
  - Estimates of the modification index (MI),
  - The standardized expected parameter change (EPC).



- The EPC is the expected value of the parameter if it is estimated.

- The significance of the EPC's are evaluated with the MI.
- The MI is a test for a single parameter, which is CHI2-distributed with 1 df.
- Unfortunately this test (MI) has the same deficiencies as the other tests/F!!
- That is why the power of the MI-test should be taken into account. How?

7

---

---

---

---

---

---

---

---

---

---

## An alternative to model testing

| power (1-β) | MI     | Conclusion               | JR |
|-------------|--------|--------------------------|----|
| ≥ 0.75 (HP) | ≤ 3.84 | No misspecification      | 1  |
| < 0.75 (LP) | > 3.84 | Misspecification present | 2  |
| ≥ 0.75 (HP) | > 3.84 | Use EPC                  | 3  |
| < 0.75 (LP) | ≤ 3.84 | No decision              | 4  |

- How to estimate the power?
- If a model is misspecified, the CHI2 distribution shifts to the right.
- The extent of this shift is indicated by the noncentrality parameter (ncp).
- The size of the ncp = (MI/EPC)<sup>2</sup>δ<sup>2</sup>
- Which is dependent on δ, the size of the misspecification one likes to detect.
- This ncp can be used to determine the power of the MI-test for a misspecification of δ for any value of the significance level α of the test, and for any restricted parameter.

8

---

---

---

---

---

---

---

---

---

---

## An alternative to model testing

- Saris & Satorra suggest to use the following δ's as important enough to detect:
  - For standardized structural parameter and for a correlated errors: 0.1
  - For standardized factor loadings the standard approach: 0.4.
- I want to add to this
  - For standardized equality constraints: 0.1
  - (It is sometimes also suggested to not accept cross-loadings: > 0.2
- These values are suggestions!
- Now we can compute: ncp = (MI/EPC)<sup>2</sup>δ<sup>2</sup> and look up the power in a table.



- A lot of work, even for a single parameter. That's why I developed JRRule.
- JRRule = Judgment Rule Aid, i.e. an aid to judge whether one or more zero-assumed parameters in your model are misspecified, taking into account the power.
- A illustration with a very simple model.

9

---

---

---

---

---

---

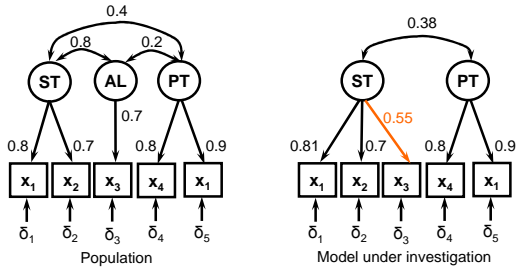
---

---

---

---

### Judgment Rule Aid



- Won't detect this misspecification easily, especially if no test due to  $df=0$ .
- Convergent validity and Discriminant validity.
- <sup>10</sup> ■ Note, I have selected this model with a reason.

---

---

---

---

---

---

---

---

### Judgment Rule Aid

Degrees of Freedom = 4  
 Minimum Fit Function Chi-Square = 10.33 (P = 0.035)  
 Normal Theory Weighted Least Squares Chi-Square = 10.15 (P = 0.038)  
 Estimated Non-centrality Parameter (NCP) = 6.15  
 90 Percent Confidence Interval for NCP = (0.28 ; 19.60)

Minimum Fit Function Value = 0.0080  
 Population Discrepancy Function Value (F0) = 0.0047  
 90 Percent Confidence Interval for F0 = (0.0022 ; 0.015)  
 Root Mean Square Error of Approximation (RMSEA) = 0.034  
 90 Percent Confidence Interval for RMSEA = (0.0074 ; 0.061)  
 P-Value for Test of Close Fit (RMSEA < 0.05) = 0.81

Normed Fit Index (NFI) = 0.99  
 Non-Normed Fit Index (NNFI) = 0.99  
 Comparative Fit Index (CFI) = 1.00

Root Mean Square Residual (RMR) = 0.022  
 Standardized RMR = 0.022  
 Goodness of Fit Index (GFI) = 1.00  
 Adjusted Goodness of Fit Index (AGFI) = 0.99

■ 1a MisFit.out



11

---

---

---

---

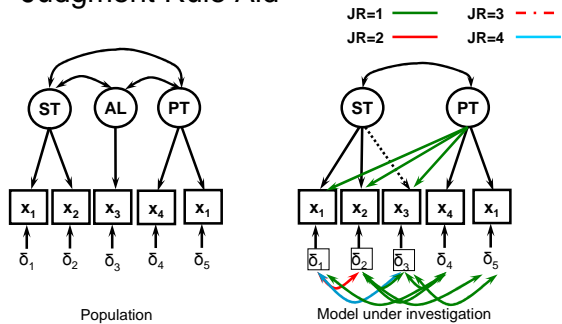
---

---

---

---

### Judgment Rule Aid



12

---

---

---

---

---

---

---

---

## Summary

- Model evaluation is hampered by the sensitivity to sample size, model complexity, and the size of incidental parameters, which affect the power to detect misspecifications.
- As a result standard model evaluation procedures do a poor job.
- We switch to parameter evaluation, not model evaluation.
- This is usually done with the modification indices.
- Those indices are suffering the same problems with sample size, etc.
- So we have taken into account the power of the test in our judgment of whether a parameter is misspecified or not: JR=1, JR=2, JR=3, JR=4.
- Because the procedure takes a lot of time a (free) software package has been developed called JRULE.

---

---

---

---

---

---

---

---