

Statistical Power Analysis for Construct Validity

Wei Chen^{1,2}, Jingfu Zhang²

¹School of Mathematics and Statistics, Southwest University,

²The Faculty of Psychology, Southwest University,
Chongqing, China

Corresponding author: Jingfu Zhang, e-mail: zhangjf@swu.edu.cn

Abstract

Confirmatory factor analysis method is usually used in psychological surveys to evaluate the construct validity. But most of the researchers have been wasting their resources (e.g. effort, time, money) without power analysis at the beginning of their studies. The main purposes of this study are to analyze the statistical power, to estimate sample size for the confirmatory factor analysis model and to maximize the utility of the sources owned by the researchers. Monte Carlo Simulation is adopted to explore the power of the five-factor model. According to the result, the power is changed under different conditions. We found that the model has low power when the RMSEA value is closed to the alternative hypothesis' RMSEA. That is, the lower RMSEA of model does not necessarily have high power. The change of power is very complex, because it is subject to a number of factors that are different sample size, different alternative hypothesis, different estimator and different model.

Key Words: Power, Confirmatory Factor Analysis, Sample Size, Monte Carlo, Construct Validity,

1. Introduction

Construct validity has been as one of the important evaluations for potential structure of the scale. It is commonly examined with confirmatory factor analysis that is the measurement part of structural equation model (SEM). If the assumed model can fit the data, in theory, the latent factor structure of the scale can represent the psychological traits that the subjects. That is, the scale has good construct validity(Wang & Wang, 2012). Because there are many goodness of fit statistics to determine that the assumed model is good or bad, actually this method has a high degree of subjectivity. In addition, even if the model fit is good, it does not represent the real situation because this test may have only the low power(Miles, 2003). In null hypothesis significance test (NHST), the null hypothesis of no effect is often rejected when the sample size is very large, and researchers consider that the alternative hypotheses had a significant effect. The fit indexes of SEM may also like this. No matter the p value how small, we must make a type II error (β). Statistical power is the real description of the quality of the statistical test (Cohen & Lea, 2004). The researchers can calculate the sample size appropriately for the study hypothesis with the special statistical power, they can save their resources (e.g. time, energy and money) and ensure that research resources are fully utilized(Myers, Ahn, & Jin, 2011).

Power in *NHST*

In the behavioral and social sciences, the null hypothesis (H_0) often refers to the concerning issues that they may have no effect, but the alternative hypothesis (H_1) is that the effect size (ES) is really existed.

Statistical principles are as follows:

$$H_0 : ES = 0 ; \quad H_1 : ES \neq 0$$

Comparing $p(\text{data} | H_0 \text{ is true})$ with $\alpha = p(\text{reject } H_0 | H_0 \text{ is true})$
(type I error);

$$\text{test statistics} = f(N) \times ES ;$$

$$\beta = p(\text{not reject } H_0 | H_0 \text{ is false}), (\text{this is type II error}) ;$$

$$1 - \beta = \text{power} = (\text{reject } H_0 | H_0 \text{ is false}) ;$$

$f(N)$ is a statistic related to sample size (N). ES shows the disagree between the data and null hypothesis (H_0). The following table displays the logic of NHST (Cohen & Lea, 2004).

Table 1 The logic of NHST

	What is true in the population	
	H_0 is really true	H_0 is really false
Results declare not significant (Don't reject H_0)	Correct Conclusion ($1 - \alpha$)	Type II error (β)
(Reject H_0)	Type I error (α)	Correct Conclusion ($\text{power} = 1 - \beta$)

Power in SEM

In the framework for SEM, as long as there is not enough evidence to reject the null model, it will consider accepting the assumed model due to the well goodness-fit-index. If our sample data is poor and may reject the null hypothesis, then the probability that we might make Type I error will increase. In contrast, we will get a high likelihood of type II error (Kim, 2005). The concept of statistical power is defined as the probability of rejecting the null hypothesis given that the null hypothesis is false (Cohen & Lea, 2004).

The null hypothesis in the context of SEM is defined by the specification of fixed and free elements in relevant parameter matrices of the model equations. This researchers' hypothesis may reflect the putative direct and/or indirect effects among the latent variables. The null hypothesis is assessed by forming a discrepancy function ($F = f(S - \Sigma)$) between the model-implied set of matrix (Σ) and the sample matrix (S). The parameters of the proposed model are estimated by minimizing the discrepancy between the S and the Σ implied by the model. Various discrepancy functions can be formed depending on the particular minimization algorithm being used (e.g. generalized least squares), but the goal remains the same—namely to derive a test statistic that has a known distribution, and then compare the obtained value of the test statistic against tabled values in order to render a decision vis-a-vis the null hypothesis (Kaplan, 1996).

The power of SEM is different from the traditional NHST. That is very complicated. Because the assessment of model and parameters refer to many fit indexes. Each fixed parameter in the model is potentially false and each can take on, in principle, an infinite number of alternative values. Thus, each fixed parameter needs to be evaluated one at a time. That is extremely complex and unrealistic.

There are two main types of power in SEM. One is based on the model parameters (MacCallum, Browne, & Sugawara, 1996; Muthén & Muthén, 2002). The

other is based on the goodness-fit-index(Kim, 2005; MacCallum et al., 1996; Wang & Wang, 2012). Generally speaking, the construct validity is described by the degree that the data fits our interesting model, so we choose the second method of power analysis using the goodness-fit-index that is the root-mean-square-error-of-approximation (RMSEA). According to Kim and M.B.S. method, we have chosen four values as the null hypothesis for observing the change of power. That is, the H_0 may be 0.00 (perfect fit), 0.03, 0.05 (close fit) or 0.08 (moderate fit).

In SEM, a not significant test statistic of overall fit is desired because the researcher typically does not want to reject a hypothesized model. However, this result can be due to lack of power. For example, a small sample size can guarantee low power. A not significant result in SEM will lead to an acceptance of a null hypothesis and may lead to publication. In other statistical methods a lack of power will result in demonstrating no effect and therefore it will not lead to publication. Therefore, power is an even more important issue in SEM than in other statistical methods.

This paper examines the relation among fit indexes, power, discrepancy function and sample size in SEM for evaluating the quality of construct validity. The two main existing methods of computing power have estimated by specifying an alternative hypothesis or alternative fit, but we chose the Kim and M.B.S method. These methods cannot be implemented easily and reliably. Our purpose is only the variety of power with considering different factors. At the same time, we can compute the sample size for our investment on a certain level of power. The different factors on power and sample size estimates are discussed.

2. Method

Monte Carlo simulations are growing in popularity, also known as statistical simulation and random sampling. It is a stochastic simulation method, based on mathematical statistics and probability theory and implemented in the computer. Monte Carlo method can generate random numbers (pseudo-random number) to address the research questions associated with the probability model and obtain the approximate solution. This is the idea of Monte Carlo analysis(Robert & Casella, 2009).

To the Paxton et al.'s point of Monte Carlo experiments, we adopted a nine-step procedure. The steps were conceptualized as occurring in three stages: research design, simulation implement, summaries(Paxton, Curran, Bollen, Kirby, & Chen, 2001).

The First Stage: Research Design

In this stage, five steps had been created.

Step (1): Research questions.

We want to know the power how to change under the different conditions (e.g. different sample sizes, estimators, observed items and alternative hypothesis).

Step (2): The selection of representative model.

The chosen model is depicted in Figure 1. The observed items can be varied each factor. Generally, the psychometric researchers expected that the low correlation among the latent factors, high loadings between factors and the observed variables. The measurement model has five factors, each of which may have three, five, eight or ten continuous factor indicators.

Step (3): Design Experiment.

We just want to the change of power with following factors. The latent factors is continuous data and obeyed the normal distribution, no missing. The sample size is from 100 to 500 at intervals of 100. The estimator are weighted least squares (WLS, this is asymptotically distribution free discrepancy functions), generalized least squares (GLS) and maximum likelihood (ML)(Browne, 1984; Jöreskog, 1967; Wang

& Wang, 2012).

Table 2 The comparison of four discrepancy functions

Discrepancy function	W derived as	Equation
WLS	Asymptotic covariance matrix	$F_{WLS} = \frac{1}{2} tr \left\{ \left[(S - \Sigma) W^{-1} \right]^2 \right\}$
GLS	Function of elements of S	$F_{GLS} = \frac{1}{2} tr \left[(I - S^{-1} \Sigma)^2 \right]$
ML	Function of elements of Σ	$F_{ML} = \log \Sigma - \log S \Sigma^{-1} + tr(S \Sigma^{-1}) - p$

The number of observed variables are three, five, eight or ten, respectively. The power of model is calculated by the method of Kim and M.B.S using the *RMSEA* index. This index is treated as the alternative hypothesis and acts as a important role of *ES*. However, according to the method of M.B.S, the null hypothesis has different values and we have chosen four numbers. Respectively, those are 0.00, 0.03, 0.05 and 0.08. The Alpha is fixed at 0.05. In summary, we consider the power with the 180 (5*3*4*3) different factors.

Step (4): The Set of Parameters.

Data are generated using the following parameter values. The factor loadings are 0.8. The residual variances of the factor indicators are 0.36. Factor variances are fixed to one to set the metric of the factors. The factor correlation is 0.25. All factor loadings are free. These population values are chosen so that the variances of the factor indicators are one, which makes the parameter values more easily interpretable. That is advised by Muthén and Muthén(Muthén & Muthén, 2002).

Step (5): Choosing Software.

Monte Carlo simulation is performed with the ML estimator in Mplus7.0. The CFA with GLS and WLS estimator is implemented in the same program (Muthén & Muthén, 2002; Wang & Wang, 2012). The power is computed in the R 2.15.2 and the code is provided by Preacher et al. (Preacher & Coffman, 2006).

The Second Stage: Simulation Implement.

Step (6): Executing the simulations.

Now we can get the raw data that was generated in Mplus with the target model designed and the values of the parameters determined. Note that the change to the code needed for every sample and estimator.

Step (7): File Storage.

At last there were 20 (5*4, equal to the different types sample size times the total number of model) raw data files that would be analyzed in CFA with the different estimators. The data of model with different items was stored in a new folder.

Step (8): Verification.

In every step of the programming, we focus on the change of the code, such as the new sample size, the new estimator of CFA and the new data filename. Specifically, despite some data may be not converged using different experiment conditions, we still kept them in the analysis and might be find something that were usually ignored. This is our purpose that the stability of estimator is found by mutual comparison under different conditions.

3. Results

The Third Stage: Summarizing Results. This is also Step (9).

Completion of the above-mentioned 8 steps, it indicates that Monte Carlo simulation have implemented thoroughly. We can calculate the power each *RMSEA*

using the Preacher's R-code. At last, we have summarized those power values in the following table 2.

First, we found that there are many spaces in the WLS column. This is not what we intentionally do not fill, but there is no value to write. Because the model can not converge with the WLS estimator in the same condition. This phenomenon shows that the WLS estimator may require a higher quality data compared to the ML and GLS methods in the same model. In addition, the stability of the WLS' convergence is getting worse with more and more complex model. There is a little difference in power using GLS and ML estimators under the same conditions. With the increase in the sample size and the complexity of the model, the power will be more close to 1.

Second, because of the different test reference standards (null hypothesis), the change of the power shows that it is low in the center and high at each end. This is very interesting. It is very different from our usual understanding. Not the model fitting better, the higher power. Because its trend is not monotonic.

At last, we have found that the power is stabilized and little change. Therefore, when we make a scale for psychological trials, the number of observed variables for each latent factor should not be too many or too few. This result indicates that about eight items for every factor is appropriate relatively.

Table 3 Power in different condition

Number of observed variables		3 items			5 items			8 items			10 items		
Sample Size (N)	RMSEA(H_0)	WLS	GLS	ML	WLS	GLS	ML	WLS	GLS	ML	WLS	GLS	ML
100	0.00	----	0.134	0.114	----	0.308	0.308	----	1.000	0.969	----	1.000	1.000
	0.03	----	0.053	0.064	----	0.245	0.065	----	0.487	0.598	----	0.656	0.969
	0.05	----	0.217	0.247	----	0.817	0.413	----	0.996	0.182	----	1.000	0.074
	0.08	----	0.851	0.873	----	1.000	0.999	----	1.000	1.000	----	1.000	1.000
200	0.00	0.999	0.211	0.088	1.000	0.591	0.131	----	1.000	0.437	----	1.000	0.785
	0.03	0.985	0.079	0.180	1.000	0.591	0.392	----	0.935	0.514	----	0.991	0.466
	0.05	0.731	0.552	0.746	0.992	0.999	0.996	----	1.000	1.000	----	1.000	1.000
	0.08	0.180	0.999	1.000	0.380	1.000	1.000	----	1.000	1.000	----	1.000	1.000
300	0.00	0.848	0.138	0.087	----	0.078	0.098	----	1.000	0.247	----	1.000	0.400
	0.03	0.307	0.245	0.331	----	0.813	0.776	----	0.999	0.978	----	1.000	0.998
	0.05	0.265	0.934	0.963	----	1.000	1.000	----	1.000	1.000	----	1.000	1.000
	0.08	1.000	1.000	1.000	----	1.000	1.000	----	1.000	1.000	----	1.000	1.000
400	0.00	0.801	0.255	0.077	1.000	0.975	0.090	----	1.000	0.153	----	1.000	0.246
	0.03	0.125	0.257	0.531	1.000	0.975	0.951	----	1.000	1.000	----	1.000	1.000
	0.05	0.698	0.981	0.998	1.000	1.000	1.000	----	1.000	1.000	----	1.000	1.000
	0.08	1.000	1.000	1.000	1.000	1.000	1.000	----	1.000	1.000	----	1.000	1.000
500	0.00	0.331	0.059	0.762	1.000	0.088	0.762	----	1.000	0.121	----	1.000	0.191
	0.03	0.328	0.738	0.694	0.997	0.993	0.995	----	1.000	1.000	----	1.000	1.000
	0.05	0.997	1.000	1.000	0.331	1.000	1.000	----	1.000	1.000	----	1.000	1.000
	0.08	1.000	1.000	1.000	1.000	1.000	1.000	----	1.000	1.000	----	1.000	1.000

4. Conclusion

Monte Carlo method has been used to explore the assumption of statistical theory. There have been many recent calls to use Monte Carlo method as a tool to improve applications of quantitative methods in substantive research(Maxwell, Kelley, & Rausch, 2008; Muthén & Muthén, 2002; Paxton et al., 2001; Thoemmes,

MacKinnon, & Reiser, 2010). The main purpose of this study is to demonstrate how Monte Carlo method can be used to estimate the power of the SEM within different conditions. This can be used to evaluate the construct validity of the questionnaire. However, the power of SEM is extremely complex as a function of number of observed variables, degrees of freedom, the RMSEA of null hypothesis, different estimators and sample size. According to the property of Monte Carlo method, the RMSEA statistics behave differently (Hu & Bentler, 1999). The power would be affected greatly these factors of our experiments. We recommend that the scales should include each factor of eight items in order to permit a high statistical power of the SEM. The construct validity may be good.

Primary limitation of this study is to consider only a class method of calculating the power of SEM. That is Kim and M.B.S. method using the RMSEA. We do not consider another goodness-fit-index (e.g. CFI, Steiger's γ and McDonald's fit index) (Kim, 2005). Finally, in this study our main concern is only a particular SEM with five latent continuous factors, normal distribution and no missing data, which the breadth of research results subjects to certain limitations.

References

- Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance-structures. *British Journal of Mathematical & Statistical Psychology*, 37(May), 62-83.
- Cohen, B. H., & Lea, R. B. (2004). *Essentials of statistics for the social and behavioral sciences*: John Wiley & Sons, Inc.
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling- a Multidisciplinary Journal*, 6(1), 1-55.
- Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32(4), 443-482.
- Kaplan, D. (1996). Statistical power in structural equation models. <http://www2.gsu.edu/~mkteer/power.html>.
- Kim, K. H. (2005). The relation among fit indexes, power, and sample size in structural equation modeling. *Structural Equation Modeling*, 12(3), 368-390.
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychol Methods*, 1(2), 130-149.
- Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annu Rev Psychol*, 59, 537-563.
- Miles, J. (2003). A framework for power analysis using a structural equation modelling procedure. *BMC Medical Research Methodology*, 3(1), 1-11.
- Muthén, L. K., & Muthén, B. Q. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, 9(4), 599-620.
- Myers, N. D., Ahn, S., & Jin, Y. (2011). Sample size and power estimates for a confirmatory factor analytic model in exercise and sport: A Monte Carlo approach. *Research Quarterly for Exercise and Sport*, 82(3), 412-423.
- Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. N. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, 8(2), 287-312.
- Preacher, K. J., & Coffman, D. L. (2006). Computing power and minimum sample size for RMSEA [Computer software]. Available from <http://quantpsy.org/>.
- Robert, C. P., & Casella, G. (2009). *Introducing Monte Carlo methods with R*: Springer.
- Thoemmes, F., MacKinnon, D., & Reiser, M. R. (2010). Power analysis for complex mediational designs using Monte Carlo methods. *Structural Equation Modeling* 17(3), 510-534.
- Wang, J. C., & Wang, X. Q. (2012). *Structural equation modeling: Application using Mplus*: Higher Education Press.