

SAMPLE SIZE IN MULTILEVEL STRUCTURAL EQUATION MODELING – THE MONTE CARLO APPROACH

Adam Sagan

Cracow University of Economics, Cracow, Poland
e-mail: sagana@uek.krakow.pl

ORCID: 0000-0002-5608-179X

© 2019 Adam Sagan

This is an open access article distributed under the Creative Commons Attribution-NonCommercial-NoDerivs license (<http://creativecommons.org/licenses/by-nc-nd/3.0/>)

DOI: 10.15611/ead.2019.4.05

JEL Classification: C38, C83, M30

Abstract: In the process of sample selection, an important issue is the relationship between sample size and the type and complexity of the statistical model, which is the basis for testing research hypotheses. The paper presents methodological aspects of sample size determination in multilevel structural equation modelling (SEM) in the analysis of satisfaction with the banking products in Poland. The multilevel SEM results from the necessity to take into account both the sample size at the level of individual respondents, as well as at the higher level of analysis and the intraclass correlation coefficient. A comparison of factor loading bias based on the Monte Carlo simulation is made for different cluster sizes and the number of clusters.

Keywords: sampling, multilevel SEM, Monte Carlo simulations.

1. Introduction

The construction of multilevel models is especially related to cluster and area sampling. However, their application is dependent on the nature of the research question and type of variables used in the modelling process.

Two general approaches are used for SEM modelling based on cluster sampling. First, sample may be selected using cluster sampling method, but the analysis of data includes variables from an individual level only, and it is conducted at that level. In such a situation, the correction of standard errors for parameter estimates is only applied and thus a low intraclass correlations (ICC) are desirable to ensure adequate effective sample size.¹ The second approach involves a multilevel framework for clustered data analysis and the SEM model is developed on both “within” and

¹ In the Mplus program, for single-level models with standard errors corrections for complex samples, the required option of model is: TYPE=COMPLEX.

“between” levels. Therefore, high ICC’s are desirable. In general, factors influencing the sample size in scientific studies using SEM modelling are [Muthen, Muthen 2002]:

- Model complexity (number of degrees of freedom) – the more complex the model, the bigger the sample needed.
- Distribution of manifest variables – the stronger the deviations from the normal distribution, the bigger the sample size due to weight matrix size for WLS/DWLS estimation methods.
- The structure and type of missing data – the stronger the deviation from MAR, the bigger the sample size needed.
- Reliability of the measurement scales – the higher the measurement error, the larger the sample size.
- Effect size – the smaller the diagnosed effect size, the larger the sample size required.
- The power of the test – the stronger the power of the test needed, the larger the sample size.

Usually the sample size has a relatively smaller effect on model parameters bias, but it has an higher impact on standard errors of the parameters.

Random sampling is related to the possibility of assessment of the quality of the given complex sample. The quality of the sample is measured on the basis of two factors: 1) accuracy and 2) precision of estimation of the unknown population characteristics. The accuracy of an estimate is related to the assessment of the extent to which the value calculated from the sample reflects the true value in the general population. The precision of estimation concerns the reproducibility of the results obtained from the sample and is measured by the variance of estimates in repeated random samples – the smaller the variance of estimates, the more precise the measurement of an unknown characteristic in the population. Usually the criterion of comparison of accuracy and precision of the estimation in a given sampling scheme is simple random sampling. It is rarely used in practice, but it is a reference for the evaluation of complex sampling (strata or cluster).

The precision of estimation of an unknown population characteristic depends on the size of the random sample and the way it is selected. A measure of the precision of estimating an unknown parameter in the population depending on the type of random sample is the so-called design effect (deff), which is the ratio of variance of an estimator in a cluster (or stratified) sample to variance of an estimator in a simple random sample (srs) of the same size [Lehtonen, Pahkinen 2004, p. 15; Kish 2004, p. 33]:

$$deff = \frac{V(\bar{x})}{V_{srs}(\bar{x})}. \quad (1)$$

In general, stratified samples give more precise estimates in relation to simple random samples ($deff < 1$), and cluster/multistage samples give less precise estimates in relation to simple random samples ($deff > 1$).

In cluster samples the effect of a sampling scheme is a function of the group size (n_c) and the intraclass correlation coefficient (ICC) [Lehtonen, Pahkinen 2004, p. 45]:

$$deff = 1 + (n_c - 1)ICC. \quad (2)$$

In some analyses it is more convenient to use the square of *deff*, which determines the increase (inflation) of standard error for a parameter caused by correlated measurement errors of variables:

$$defft = \sqrt{1 + (n_c - 1)ICC}. \quad (3)$$

From this point of view, the sample size for multilevel analysis must be higher than for single-level analysis, because the effective sample size is lower in the cluster sample compared to the corresponding simple random sample, given the same level of precision for parameter estimation (these samples therefore require a higher sample size than simple random samples to achieve the same level of precision).

The effective sample size is the size of the simple random sample that gives the same precision of estimation as the complex sample used in the study. It is thus a measure of the degree of loss of estimates' precision in cluster samples. Effective cluster sample size is a function of sample size (n), average group size (n_c), and intraclass correlation coefficient (ICC):

$$n_{ef} = \frac{n}{1 + (n_c - 1)ICC}. \quad (4)$$

The effective sample size can be expressed as the ratio of the sample size to *deff* [Valiant, Dever, Kreuter 2013, p. 100]:

$$n_{ef} = \frac{n}{deff}. \quad (5)$$

Hence the *deff* is also the ratio of the actual sample size to the effective sample size. The intraclass correlation coefficient has a significant influence on the size of the cluster sample. In an extreme case, if $ICC = 0$, and each group has the same size, it is enough to draw one group (cluster of respondents) to estimate an unknown parameter in the population (all groups are identical).

On the other hand, if $ICC = 1$, then there is a maximum similarity among units in each group. In the process of sample selection, the relationship between sample size and the type and complexity of model is an important issue. The choice of the sampling method and sample size is largely dependent on the type of data analysis method, and complexity of the model. The problem of required sample size is especially important for structural equation models (SEM) and specifically, for the more general version of SEM that are multilevel structural equation models with latent variables.

2. Sample size in structural equation models

The relation of theoretical and statistical assumptions of the model and sample size is particularly important for structural equation modelling. This is related to the nature of the global test of exact fit and the domination of accept-support approach to the hypothesis testing in SEM. In the process of model testing, the investigator seeks not to reject the null hypothesis, which states that the standardized residuals of empirical covariance matrix S and the theoretical implied covariance matrix reproduced by the model parameters ($\Sigma(\theta)$) are equal 0, so $S = \Sigma(\theta)$.

The role of random sample size in the process of testing a structural model is particularly important because the test statistic, which is the basis for assessing the goodness-of-fit, is the product of the minimum value of the discrepancy function (F) and the random sample size (N). If the sample is "large enough" and the model is correctly specified, then the T statistic has a χ^2 distribution and is given by the formula: $\chi^2 = F(N-1)$.

Given the number of degrees of freedom, the higher the sample size, the higher χ^2 statistic and therefore, the higher the probability of rejecting H_0 . In the case of "too large" samples, the correct model may be rejected due to small differences between S and $\Sigma(\theta)$ and in the case of "too small" samples, even large discrepancies between S and $\Sigma(\theta)$ will not be recognized and the wrong model will not be rejected [Davey, Savla 2010].

Therefore the power of the χ^2 test and the specification of the SEM model are crucial factors in determining the sample size in testing the research hypotheses. There are several "rules of thumb" in sample size specification in SEM models, including the minimum sample size rule (e.g. $N > 200$), the minimum number of cases per parameter (10:1 rule), and the minimum number of cases per indicator (manifest variable) in the model (10:1 rule) [Bentler, Chou 1987]. All of the rules are usually selected in day-to-day research with SEM without any statistical or substantial justification.

The model-based approach to determining sample size is related to three basic approaches: a) the Satorra and Saris method [1985] related to the evaluation of the power of test on the basis of λ non-centrality parameter for a given structural model ($\lambda = \chi^2 - df$) that determines the model specification error, b) the MacCallum, Browne and Sugawara method [1996] based on the evaluation of the power of test at which the RMSEA value allowing for the model not to be rejected is achieved and c) the Monte Carlo simulation method [Muthén, Muthén 2002].

The power-based methods are commonly used in SEM modelling. This is a function of a) given significance level (α), b) type II error level (β), c) the expected strength of the effect size, d) the standard error for the parameter (σ) and e) the sample size (n). In general the power of test is assumed to be on the level of $\beta/\alpha = 4$, hence for $\alpha = 0.05$, the desired power is $1 - \beta = 0.8$ [Cohen 1992]. The necessary

random sample size for SEM can be determined on the basis of the power of the test and significance level. In the assessment of a simple random sample size for a SEM model, the evaluation of the model fit is made on the basis of the value of χ^2 statistics and the root mean square error of approximation (RMSEA). Due to the specificity of exact test of the H0 hypothesis, the "close fit" is tested in which the null hypothesis assumes that the RMSEA = 0.05 (and not 0.00) and the alternative hypothesis (H1) that RMSEA is = 0.08 (this is the upper limit of the "badness-of-fit" of the SEM model). Statistical programs, like GPower and Statistica, allow to assess the power of test at a given sample size, significance level and the hypotheses H0 and H1 or to determine the random sample size at a given power of test.

Figure 1 shows the relationship between sample size and effect size, given the number of manifest and latent variables (model complexity), the power of test and significance level. In simulations depicted in Figure 1, the sample sizes for SEM models (N) depend on effect sizes (ES) of 0.05 (very small), 0,1 (small), 0.3) (mode-

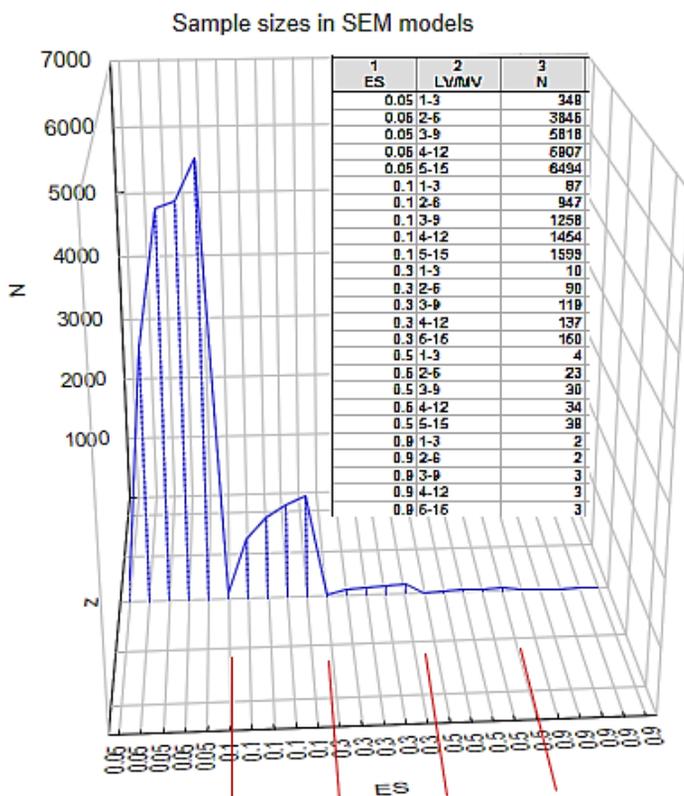


Fig. 1. Sample sizes in SEM

Source: own elaboration based on <https://www.danielsoper.com/statcalc>.

rate), 0.5 (strong), 0.9 (very strong) [Cohen 1988]. In the given effect size range, five models with different complexity are introduced (with the “three-indicators rule” for model identification): a) one latent and three manifest, b) two latent and six manifest, c) three latent and nine manifest, d) four latent and twelve manifest and e) five latent and fifteen manifest variables. The power of test = 0.8 and significance level = 0.05.

The simulations show that models with 0.05 ES need a sample size of between 348 and 6494. The sample size depends on the model complexity. Models with small ES (0.1) require N between 87 and 1599. However, for the models with moderate and strong ES (> 0.3) sufficient sample sizes are between 10 and 150 (for empirical identification of the model and avoiding Heywood cases in numeric estimation process, the minimum sample size is 200).

To sum up, for practical purposes, in scientific research using the SEM model, the minimum sample size for diagnosing the small effect size is around 1400, and where ES is moderate or strong, the required sample size is 200 observations.

3. Sample size in multilevel SEM models

The multilevel sample size should take into account the size of the primary (PSU) and secondary sampling units (SSU). The most stable solutions are obtained if the number of level II units is not too small and numbers about 50-100 groups [Eliason 1993], and the sample size at level I (individual) should be between 30 and 50 respondents in each group [Maas, Hox 2005]. This number is not strictly required, many studies (especially in family research) are conducted on groups of 1-3 persons.

The Bayesian estimation is often used for small sample sizes [Stegmueller, 2011; Hox, Van de Schoot 2013].

Usually the general rule of a multilevel sample size is 50 groups and 30 units in a group, which gives the sample size in multilevel analysis of about 1500 respondents.² However, the samples in multilevel SEM are more numerous and number from 12,000 to 17,000 respondents (in multilevel educational research). In the evaluation of the size of cluster samples, the power of statistical tests and sample weights, the appropriate computer programs are used such as sample size calculators (MaCorr, Raosoft, Statistica, macros etc.), power test calculation programs (GPower, SPSS, STATA, Statistica), programs for evaluation and weighing of random samples (SPSS, WeSVAR, SurveyGizmo, Zoho, QuestionPro, Typeform, Survey Anyplace, Wufoo, SurveyMoz, RationalSurvey, SoGoSurvey), and packages for multilevel sampling and data analysis (Mplus, LISREL, GLLAMM (Stata), MLWIN, NL MIXED (SAS)).

² In the Mplus program, the required option for the multilevel model is: TYPE=MULTILEVEL.

4. The Monte Carlo approach in sample size determination

The application of the Monte Carlo method results from the limitations of the power of the test to determine the sample size. The determination of the sample size based on power, significance level, H0 and H1 hypotheses, assumes the correct model specification and the lack of model parameters and standard errors biases. The parameter bias refers to the factors that cause the estimated parameter values in the sample to be different from the "true" values of parameters in the population.

The Monte Carlo approach enables the evaluation of model parameter bias and the power of test with the use of the generated "population" data on the basis of 1) prior parameter values determined in previous research, 2) meta-analysis or 3) estimated sample parameters. Once the population data is generated, the samples of a specific size are randomly selected.

On each of these sub-samples, the model parameters and their standard errors are estimated. Having the empirical distribution of the model parameters and the standard errors parameters and the standard errors bias, the coverage is also determined. The parameters' bias is related to the deviation of the "population" value estimate from its mean value from the replicated estimates:

$$\hat{\phi}_R^b = \frac{1}{R} \sum_{r=1}^R \frac{\hat{\beta}_x^r - \beta_x}{\beta_x} \times 100, \quad (6)$$

where: $\hat{\phi}_R^b$ – parameter bias in replications ,
 R – number of replications,
 $\hat{\beta}_x^r$ – mean parameter value in replications,
 β_x – population value of parameter.

The coverage specifies the quality of parameter estimates and represents the fraction of parameter estimates in replications not exceeding a 95% confidence interval.

$$\hat{\phi}_R^c = \frac{1}{R} \sum_{r=1}^R I \left[\hat{\beta}_x^r - 1,96se(\hat{\beta}_x^r) \leq \beta_x \leq \hat{\beta}_x^r + 1,96se(\hat{\beta}_x^r) \right], \quad (7)$$

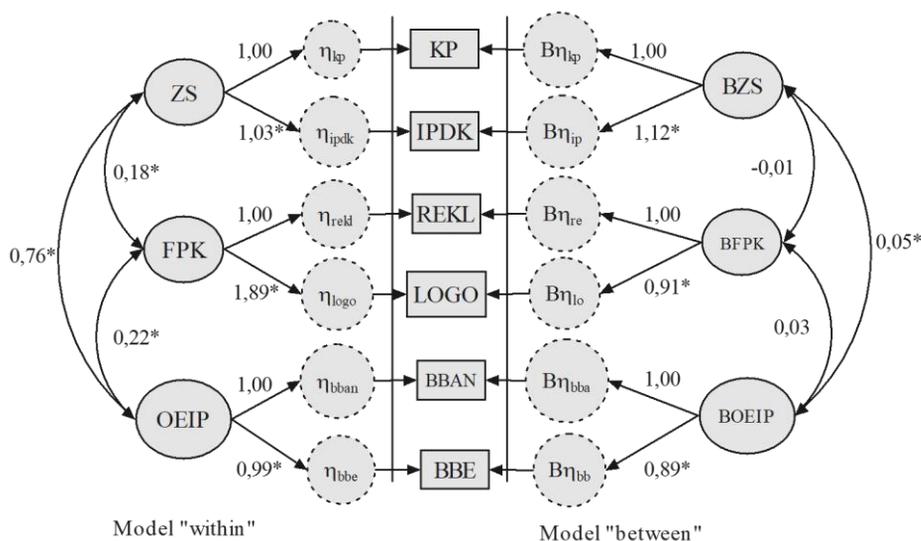
where: I – a binary indicator indicating the truthfulness (1) or falsity (0) of the inequality,
 $se(\hat{\beta}_x^r)$ – standard error of "population" value of parameter.

The use of the Monte Carlo method to determine the sample size required, takes into account the maximum acceptable bias of model parameters. L. Muthén and B. Muthén [2002] propose a set of criteria to assess the required sample size in the Monte Carlo simulations: a) the bias of parameters and standard errors should

not exceed 10%, b) the bias of standard error for the parameter for which the power of test is calculated should not exceed 5%, c) the coverage should be between 0.91 and 0.98. If these assumptions are met, the final sample size should be at the level at which the power of test is equal to 0.80.

5. Sample size in the multilevel SEM model in banking product research

The research of the image of banking products was conducted on the basis of a two-stage cluster sampling, where PSU are bank outlets located in specific geographical areas and SSU are the clients of the selected outlet. The total sample consisted of 1335 respondents nested in 258 bank outlets.³ The two-level SEM model of perceived bank image is presented in Figure 1.



Manifest variables: KP – employer competence, IPDK – individualized approach to the customer, REKL – advertising, LOGO – visual corporate identity, BBAN – ATM security, BBE – electronic banking security.

Latent variables: ZS/BZS – operant resources, FPK/BFPK – formal communication, OEIP/BOEIP – electronic banking services.

Fig. 1. Two-level CFA model of banking products

Source: own elaboration on the basis of the Mplus results.

³ The author expresses the gratitude for the raw data file to Dr. Izabela Szlis. The data are gathered for the research grant "Rzeczywisty a pożądaný wizerunek banku na rynku usług detalicznych" (Umowa nr 0435/B/H03/2009/37) [Szlis 2012]. The data is used with permission of the author.

All of the manifest variables were measured on 5-point Likert scales. The latent variable “operant resources” was measured by the KP and IDPK indicators, while the latent variable “formal communication” was measured by the REKL and LOGO indicators; the latent variable “electronic banking services“ was measured by the BBAN and BBE indicators.

On the left-hand side, the “within” model, and on the right-hand side the “between” model is presented. The intraclass correlation coefficients (ICC) for the indicators are as follows: KP = 0.03, IDPK = 0.04, REKL = 0.09, BBANK = 0.06, BBE = 0.02. ICC for the latent variables are: ZS = 0.05, FPK = 0.10, OEIP = 0.07. The average cluster size = 5.4.

The measurement model has an appropriate fit. The global χ^2 fit statistic, with the Satorra-Bentler correction with a scaling factor of 0.82, is 51.73 with 15 degrees of freedom (p -level = 0.00). The root mean square error of approximation (RMSEA) is 0.04 and the CFI and TLI incremental fit indices are respectively 0.977 and 0.953. The standardized mean square residuals (SMSR) for the “within” model is 0.03 and for the “between” model is 0.04.

The measurement model, both at the “within” and “between” levels, reflects the structure of the bank’s image. The dimensions of the bank’s image on the “within” level (ZS, FPK, OEIP), are significantly correlated, whereas the correlation on the “between” level are insignificant (BZS, BFPK, BOEIP).

The Monte Carlo analysis of the model presented in Figure 1 was performed on the basis of the parameters of the model (shown in Figure 1). The aim of the analysis is to assess the level of model parameters bias and the power of the χ^2 test for the existing sample size and to simulate the hypothetical sample size for the specific level of power.

In the first stage of the analysis, the model was estimated on the basis of empirical data from the sample (no theoretical assumptions related to the "real" values of the parameters in the general population and the results of previous studies). The values of the parameters were used to build a “population” model for the Monte Carlo simulations. The number of replications was 400.

The χ^2 values are given below:

Chi-Square Test of Model Fit

Degrees of freedom	15
Mean	12.798
Std Dev	4.604
Number of successful computations	400

Proportions		Percentiles	
Expected	Observed	Expected	Observed
0.990	0.975	5.229	4.117
0.980	0.968	5.985	4.740
0.950	0.915	7.261	6.733
0.900	0.843	8.547	7.535
0.800	0.695	10.307	9.124
0.700	0.532	11.721	10.251
0.500	0.295	14.339	12.072
0.300	0.152	17.322	14.234
0.200	0.090	19.311	16.105
0.100	0.045	22.307	18.759
0.050	0.020	24.996	21.627
0.020	0.003	28.259	24.981
0.010	0.000	30.578	26.865

Source: own elaboration on the basis of the Mplus results.

The means and standard deviations of the chi-square test statistic over the replications in the Monte Carlo analysis are given above. In the column labelled “Proportions expected” there are the probabilities of observing a chi-square value greater than the corresponding value in percentiles expected of the χ^2 value determined from the χ^2 distribution (column 3). In this output, the value of 0.05 in column 1 gives the probability that the chi-square value exceeds the percentile value (the critical value of the chi-square distribution) of 24.996 in column 3. The columns “Proportions observed” and “Percentiles observed” give the corresponding values observed in the Monte Carlo replications. Column 2 gives the proportion of replications for which the critical value is exceeded (in this example it is 0.02), close to the expected value of 0.05 which indicates that the chi-square distribution is well approximated in this case. The value of 21.627 in column 4 is the chi-square value at this percentile from the Monte Carlo analysis that has 5% of the values in the replications above it. The fact that it deviates only slightly from the theoretical value of 24.994 is an indication that the chi-square distribution is well approximated in this case.

The analysis of bias and coverage for the parameters involves the “within” and “between” model parameters. The parameter values (factor loadings) in the “within” model are given below.

MODEL RESULTS

ESTIMATES	S. E.	M. S. E.	95% % Sig	Population Average	Std. Dev.	Average Cover	Coeff
Within Level							
ZS BY							
KP	0.918	0.9207	0.0390	0.0506	0.0015	0.985	1.000
IPDK		0.943	0.9424	0.0430	0.0533	0.0018	0.985 1.000
FPK BY							
REKL		0.949	0.9452	0.1242	0.1505	0.0154	0.983 0.998
LOGO		1.792	1.8354	0.2375	0.2928	0.0582	0.967 0.998
OEIP BY							
BBANK		1.069	1.0695	0.0491	0.0586	0.0024	0.980 1.000
BBE		1.067	1.0652	0.0519	0.0663	0.0027	0.988 1.000
FPK WITH							
ZS	0.216	0.2104	0.0410	0.0531	0.0017	0.988	0.993
OEIP WITH							
ZS	0.777	0.7746	0.0301	0.0394	0.0009	0.990	1.000
FPK		0.217	0.2133	0.0428	0.0532	0.0018	0.985 0.990
Variances							
ZS	1.000	1.0000	0.0000	0.0000	0.0000	1.000	0.000
FPK		1.000	1.0000	0.0000	0.0000	0.0000	1.000 0.000
OEIP		1.000	1.0000	0.0000	0.0000	0.0000	1.000 0.000
Residual Variances							
KP	0.816	0.8039	0.0566	0.0696	0.0033	0.980	1.000
IPDK		0.960	0.9555	0.0617	0.0790	0.0038	0.985 1.000
REKL		2.361	2.3579	0.2294	0.2823	0.0525	0.983 1.000
LOGO		0.091	-0.1181	0.9301	1.1262	0.9067	0.942 0.070
BBANK		0.889	0.8845	0.0785	0.0961	0.0062	0.990 1.000
BBE		1.730	1.7242	0.0901	0.1161	0.0081	0.985 1.000

Source: own elaboration on the basis of the Mplus results.

In the first column (Population), the "population" model-based values of the estimated parameters are presented. The second column (Average) shows the average values of the parameters in the Monte Carlo simulation. These columns are the basis for determining the level of parameter bias. The following two columns show the standard deviations of the parameters in the replications, and the average of the standard error in the replications. Based on these values the standard errors bias can be calculated. The M.S.E. column gives the mean square error for each parameter, and it is calculated as the sum of the variance of the parameter in replications and the parameter bias.

The penultimate column (Cover) shows the coverage that is the proportion of replications for which the 95% confidence interval contains the "true" population value of a parameter. All parameter values, with the exception of the residual variance for the Logo variable, are above 0.95, which indicates the good quality of the estimates. The last column gives the replication ratios for which the null hypothesis for the parameter was rejected at the significance level of $\alpha = 0.05$, which indicates the power of test. All the parameter estimates indicate the appropriate power of the t -test. Only the test for the residual variance of the Logo variable is characterized by a very low power (very high probability of type II error). The values of the parameter and standard error bias are shown in Table 1.

Table 1. Parameters and standard errors bias in the "within" model

Path	Population	Average	Parameter bias	Population s.e.	Average s.e.	Standard error bias
ZS-KP	0.918	0.9207	-0.0029	0.0390	0.0506	-0.2974
ZS-IPDK	0.943	0.9424	0.0006	0.0430	0.0533	-0.2395
FPK-REKL	0.949	0.9452	0.0040	0.1242	0.1505	-0.2117
FPK-LOGO	1.792	1.8354	-0.0242	0.2375	0.2928	-0.2328
OEIP-BBANK	1.069	1.0695	-0.0004	0.0491	0.0586	-0.1934
OEIP-BBE	1.067	1.0652	0.0017	0.0519	0.0663	-0.2774
ZS-FPK	0.216	0.2104	0.0259	0.0410	0.0531	-0.2951
ZS-OEIP	0.777	0.7746	0.0031	0.0301	0.0394	-0.3090
FPK-OEIP	0.217	0.2133	0.0170	0.0428	0.0531	-0.2406

Source: own elaboration on the basis of the Mplus results.

The parameter bias does not exceed 0.1, and one can conclude that the sample size of 1335 is sufficient to obtain unbiased parameters of the "within" model. On the other hand, the standard errors are downwardly biased and exceed the acceptable level of 0.05. The parameter values in the "between" model are given below.

ESTIMATES S. E. M. S. E. 95% % Sig
 Population Average Std. Dev. Average Cover Coeff

Between Level

BZS BY

KP 0.244 0.2408 0.0533 0.0833 0.0028 0.995 0.880
 IPDK 0.275 0.2484 0.0675 0.0979 0.0053 0.985 0.780

BFPK BY

REKL 0.548 0.5354 0.0816 0.1175 0.0068 0.995 0.998
 LOGO 0.497 0.4928 0.0755 0.1044 0.0057 0.998 0.998

BOEIP BY

BBANK 0.383 0.3703 0.0765 0.1082 0.0060 0.998 0.935
 BBE 0.344 0.3411 0.0725 0.1056 0.0053 0.993 0.933

BFPK WITH

BZS -0.111 -0.0993 0.2895 0.3393 0.0838 0.967 0.032

BOEIP WITH

BZS 0.556 0.5647 0.2167 0.2832 0.0469 0.970 0.592
 BFPK 0.161 0.1745 0.2358 0.2771 0.0557 0.967 0.100

Intercepts

KP 5.991 5.9915 0.0406 0.0496 0.0016 0.985 1.000
 IPDK 5.831 5.8321 0.0462 0.0530 0.0021 0.975 1.000
 REKL 3.782 3.7803 0.0703 0.0798 0.0049 0.970 1.000
 LOGO 3.934 3.9331 0.0641 0.0762 0.0041 0.985 1.000
 BBANK 5.968 5.9704 0.0509 0.0609 0.0026 0.985 1.000
 BBE 5.718 5.7196 0.0555 0.0660 0.0031 0.985 1.000

Variances

BZS 1.000 1.0000 0.0000 0.0000 0.0000 1.000 0.000
 BFPK 1.000 1.0000 0.0000 0.0000 0.0000 1.000 0.000
 BOEIP 1.000 1.0000 0.0000 0.0000 0.0000 1.000 0.000

Residual Variances

KP 0.000 0.0001 0.0000 0.0000 0.0000 1.000 0.000
 IPDK 0.006 0.0126 0.0107 0.0317 0.0002 0.998 0.005
 REKL 0.030 0.0345 0.0281 0.0650 0.0008 1.000 0.015
 LOGO 0.000 0.0001 0.0000 0.0000 0.0000 1.000 0.000
 BBANK 0.023 0.0250 0.0183 0.0497 0.0003 1.000 0.005
 BBE 0.000 0.0001 0.0000 0.0000 0.0000 1.000 0.000

Source: own elaboration on the basis of the Mplus results.

In the “between” model (on the level of bank outlets) a too low power for a) the measurement model of operant resources (ZS), b) covariances between latent variables and c) residual variances are observed (very high probability of type II error). Therefore this suggests a too small sample size on the “between” level of the analysis. The values of the parameter and standard error bias are shown in Table 2.

Table 2. Parameters and standard errors bias in the “between” model

Path	Population	Average	Parameter bias	Population s.e.	Average s.e.	Standard error bias
ZS-KP	0.244	0.2408	0.0131	0.0533	0.0833	-0.5628
ZS-IPDK	0.275	0.2484	0.0967	0.0675	0.0975	-0.4444
FPK-REKL	0.548	0.5354	0.0229	0.0816	0.1175	-0.4399
FPK-LOGO	0.970	0.4928	0.0084	0.0755	0.1044	-0.3828
OEIP-BBANK	0.383	0.3703	0.0331	0.0765	0.1082	-0.4145
OEIP-BBE	0.344	0.3411	0.0084	0.0725	0.1056	-0.4565
BZS-BFPK	-0.111	-0.0993	0.1054	0.2895	0.3393	-0.1720
BZS-BOEIP	0.556	0.5647	-0.0156	0.2167	0.2832	-0.3069
BFPK-BOEIP	0.161	0.1745	-0.0838	0.2358	0.2771	-0.1751

Source: own elaboration on the basis of the Mplus results.

The parameter bias in the “between” model does not exceed 0.1. This means that the sample size at the II level allows for the estimation of the unbiased model parameters. In the case of standard errors, the systematic negative bias of standard errors is also observed, which exceeds the acceptable level of 0.05. To sum up, based on a sample of 1335 respondents and 258 bank outlets, the parameter estimates correctly reflect their population values. The estimates of standard errors are biased which increases the probability of type II error.

The final analysis was based on simulations of parameter (factor loadings) bias in the “within” and “between” models. In the simulations two factors are taken into account: 1) sample size and 2) relations between the number of cluster and the cluster size. The sample sizes range from between 500 and 3000 respondents. Two sets of cluster sizes are recommended. The first assumes $C/N_c = 60$, which means that the ratio between the number of clusters (C) to the number of respondents (N_c) within clusters, is 60 for the given sample size (e.g. 300 clusters and five respondents within each cluster = 1500 in total). The second option is $C/N_c = 0.6$, which means that the relation between the number of clusters and the number of respondents within the cluster is 0.6 (e.g. 30 clusters and 50 respondents in each cluster = 1500 in total). The first solution is used in multilevel modelling of small groups (families, small task-groups etc.). The second solution is more frequent in multilevel models in organizational B2B marketing, brand societies, public institution marketing (schools, hospitals etc.).

Figure 2 depicts the results of the Monte Carlo simulations of factor loadings bias.

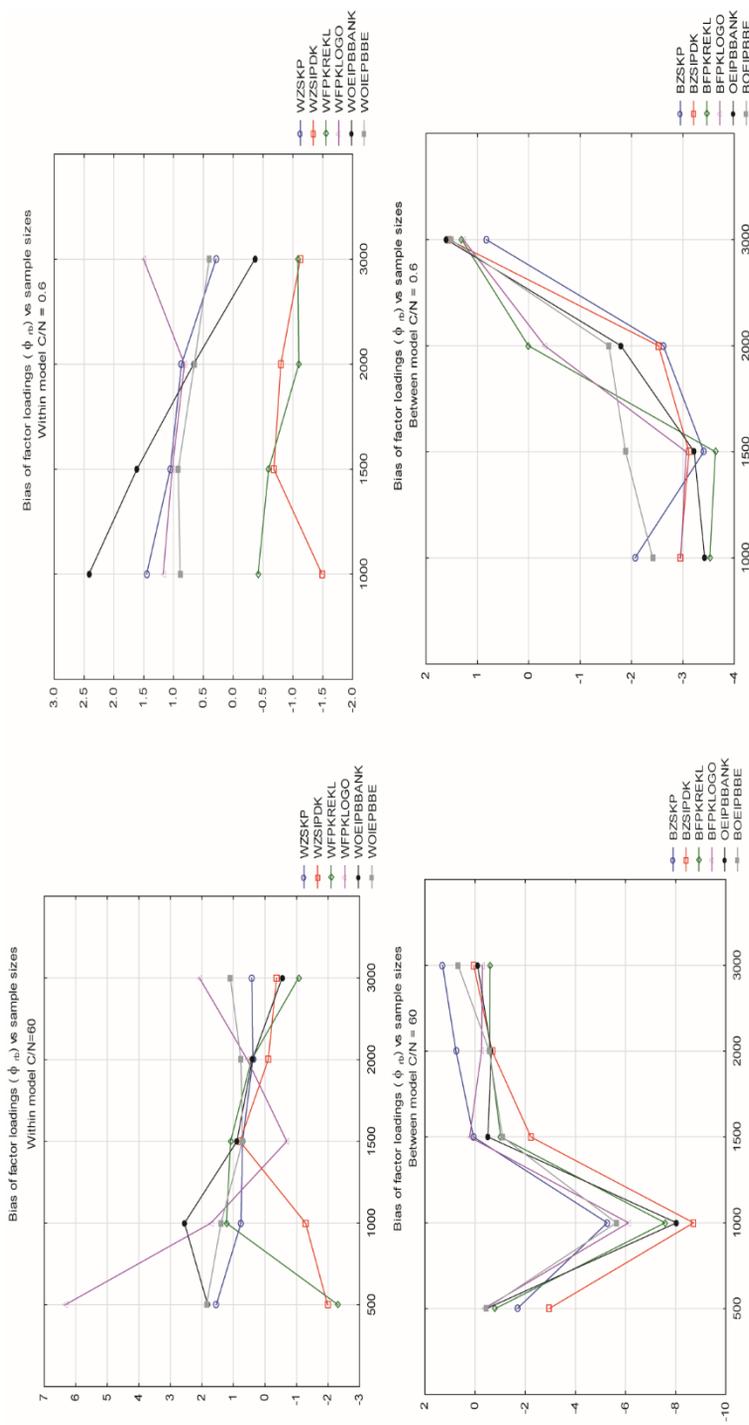


Fig. 2. Bias of factor loadings in the Monte Carlo simulation
 Source: own elaboration on the basis of the Mplus results.

The results of the simulations show that the larger the sample size, the smaller the bias of the parameters for the “between” models, especially when $C/Nc = 60$, which means the large number of clusters and the small number of observations within the clusters. In the case of the “within” models, the discrepancy of the parameter bias is much larger, nevertheless the $C/Nc = 60$ performs much better than the $C/Nc = 0.6$ rule (for the total sample of 500, the multilevel models cannot be even estimated). Taking into account both the “within” and “between” models, the optimum sample size for this multilevel model that minimizes the parameter bias for the $C/Nc = 60$ rule is 2000 respondents, and for the $C/Nc = 0.60$ rule it is around 2500. The existing sample size of 1335 respondents generates a downward bias of factor loadings and for measurement models of image of banking products.

6. Conclusion

The Monte Carlo simulation is a modern technique for the determination of sample size in structural equation modelling. The specificity of the accept-support approach in SEM model-building deserves special attention paid to the question of the sample size requirements. The commonly used rule of thumb (e.g. the 10:1 rule) that seems to persist in the used research literature, is replaced by approaches that involve model specification and complexity in determining the sample sizes.

In the development of contemporary SEM models, the single-level analysis is regarded only as a special case of more general multilevel framework of model building. The Monte Carlo simulations show that the standard errors bias is much higher than the parameters bias. Coverage values are in the range of 0.942 to 1.00 for the “within” model, and 0.967 to 1.00 for the “between” model. The power of the t -tests is between 0.993 and 1.00 for the “within” model, and between 0.032 and 1.00 for the “between” model. The optimal number of the multilevel sample is about 2000 respondents at $N/C = 60$ (that is maximizing the number of classes and minimizing the number of units in a given class). This research confirms the results of simulation studies that in multilevel multiple regression modelling, the number of groups is more important than the large number of individuals per group [Maas, Hox 2005; Snijders 2005]. However, this research is based on the real data and the more complicated multilevel CFA model. On the other hand, the general 0.6 rule (i.e. 30x50), seems to be not valid and depends on the type of the SEM model and its complexity.

Bibliography

- Bentler P.M., Chou C., 1987, *Practical issues in structural modeling*, Sociological Methods and Research, 16, pp. 78-117.
- Cohen J., 1988, *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates.
- Cohen J., 1992, *A power primer*, Psychological Bulletin, 112, pp. 155-159.

- Davey A., Savla J., 2010, *Power Analysis with Missing Data. A Structural Equation Modeling Approach*, Routledge.
- Eliason S.R., 1993, *Maximum Likelihood Estimation*, Newbury Park, CA, Sage
- Hox J., Van de Schoot R., 2013, *Robust Methods for Multilevel Analysis*, [in:] M.A. Scott, J.S., Simonoff, B.D. Marx, *The SAGE Handbook of Multilevel Modeling*, Sage.
- Kish L., 2004, *Statistical Design for Research*, Wiley
- Lehtonen R., Pahkinen E., 2004, *Practical Methods for Design and Analysis of Complex Surveys*, Wiley.
- Maas C.J.M., Hox J.J., 2005, *Sufficient sample sizes for multilevel modeling*, *Methodology* 1(3), pp. 86-92.
- MacCallum R.C., Browne M.W., Sugawara H.M., 1996, *Power analysis and determination of sample size for covariance structure modeling*, *Psychological Methods*, 1, pp. 130-149.
- Muthén L.K., Muthén B.O., 2002, *How to use a Monte Carlo study to decide on sample size and determine power*, *Structural Equation Modeling*, 4, pp. 599-620.
- Satorra A., Saris W.E., 1985, *Power of the likelihood ratio test in covariance structure analysis*, *Psychometrika*, 50, pp. 83-90.
- Snijders T.A.B., 2005, *Power and Sample Size in Multilevel Linear Models*, [in:] B.S. Everitt, D.C. Howell (eds.), *Encyclopedia of Statistics in Behavioral Science*. Volume 3, Chicester.
- Stegmuller D., 2013, *How many countries for multilevel modeling? A comparison of frequentist Bayesian and approaches*, *American Journal of Political Science*, 57(3), pp. 748-761
- Szlis I., 2012, *Aspiracje a rzeczywistość – wizerunek banku jako efekt sieci interakcji. Teoria – narzędzia pomiaru – praktyka*, CeDeWu, Warszawa.
- Valliant R., Dever J.A., Kreuter F., 2013, *Practical Tools for Designing and Weighting Survey Samples*, Springer.

LICZEBNOŚĆ PRÓBY W WIELOPOZIOMOWYM MODELOWANIU STRUKTURALNYM – PODEJŚCIE MONTE CARLO

Streszczenie: W procesie doboru i ocenie liczebności próby ważną kwestią jest ocena zależności między liczebnością próby, rodzajem i poziomem złożoności modelu empirycznego, który jest podstawą testowania stawianych hipotez. Założenia teoretyczne i statystyczne są szczególnie ważne w ocenie rozmiaru próby losowej w modelowaniu strukturalnym (SEM). Jest to związane z naturą globalnego testu dokładnego dopasowania modelu i potwierdzająco-akceptującego podejścia do testowania hipotez badawczych. Artykuł przedstawia metodologiczne problemy doboru próby w wielopoziomowym modelowaniu strukturalnym mającym zastosowanie w badaniach wizerunku produktów bankowych w Polsce. Został ukazany wpływ liczebności próby na I i II poziomie analizy w modelach wewnątrz- i zewnątrzgrupowym na poziom obciążenia parametrów modelu.

Słowa kluczowe: dobór próby, wielopoziomowe modelowania strukturalne, symulacja Monte Carlo.