Multigroup Analysis in Information Systems Research using PLS-PM: A Systematic Investigation of Approaches

Michael Klesel

University of Siegen, Germany and University of Twente, The Netherlands

Florian Schuberth

University of Twente, The Netherlands

Björn Niehaves

University of Siegen, Germany

Jörg Henseler

University of Twente, The Netherlands and Nova Information Management School, Portugal

Funding Information:

Jörg Henseler gratefully acknowledges financial support from FCT Fundação para a Ciência e a Tecnologia (Portugal), national funding through research grant Information Management Research Center – MagIC/NOVA IMS (UIDB/04152/2020). Moreover, he acknowledges a financial interest in ADANCO and its distributor, Composite Modeling.

Multigroup Analysis in Information Systems Research using PLS-PM: A Systematic Investigation of Approaches

Michael Klesel

University of Siegen, Germany and University of Twente, The Netherlands

Florian Schuberth

University of Twente, The Netherlands

Björn Niehaves

University of Siegen, Germany

Jörg Henseler

University of Twente, The Netherlands and Nova Information Management School, Portugal

Acknowledgments

Jörg Henseler acknowledges a financial interest in ADANCO and its distributor, Composite Modeling.

Abstract

Heterogeneity is a pertinent issue in Information Systems (IS) research because human behavior often differs across groups. In the context of partial least squares path modeling (PLS-PM), several approaches have been proposed to investigate potential group differences due to observed heterogeneity. Despite the availability of numerous approaches, literature that compares their efficacy is sparse. Consequently, IS research lacks guidance on which approach is best suited to detect group differences. We address this issue by presenting the results of an extensive Monte Carlo simulation study that juxtaposes the various approaches' behavior under numerous conditions such as group differences, no group differences, comparison of a single parameter, comparison of the complete structural model, and equally and unequally distributed sample sizes across groups. In doing so, we first provide an overview on existing approaches proposed for multigroup analysis (MGA) in the context of PLS-PM. Moreover, based on the simulation results, we derive important implications for applied research: Firstly, we show that the omnibus test of group differences (OTG) and approaches based on the comparison of confidence intervals are not recommendable for MGA. Secondly, we provide detailed information which approaches are suitable for comparing one specific path coefficient and which are recommended if the complete structural model is compared across groups. Finally, we show that approaches which are designed to compare a single parameter require an adjustment for multiple comparisons when used to compare more than two groups.

Keywords: Partial Least Squares Path Modeling; Multigroup Analysis; Monte Carlo Simulation; Distance-Based Permutation Test; OTG; Confidence Interval Comparison; Group Comparison.

Introduction

Over the last decades, variance-based estimators for structural equation models have gained increasing attention and popularity (Hwang & Takane, 2004; Tenenhaus, 2008). Among these estimators, partial least squares path modeling (PLS-PM) (Wold, 1975) is arguably the most wide-spread estimator and has been used for research in numerous fields, including marketing (Hair et al., 2012), business research (Hair et al., 2014), tourism (Müller et al., 2018), and information systems (IS) research (Benitez et al., 2020). PLS-PM was invented by Herman O.A. Wold (1975) and can emulate various of Kettengring's (1971) approaches to generalized canonical correlation analysis. Moreover, it is a consistent estimator for structural models in which the abstract concepts are represented by composites (Dijkstra, 2017).[1]

Many empirical studies in IS research that employ PLS-PM investigate samples collected from different populations (e.g., Ahuja & Thatcher, 2005; Dibbern et al., 2012), leading to heterogeneous datasets. In general, two types of heterogeneity can be distinguished: (i) unobserved heterogeneity, and (ii) observed heterogeneity. While the source of unobserved heterogeneity is unknown to the researcher (Wedel & Kamakura, 2000), observed heterogeneity can be traced back to observable characteristics, such as cultural background or gender (Hair et al., 2018). In both cases, ignoring heterogeneity can lead to severely biased results and therefore to questionable conclusions (Becker et al., 2013; Jedidi et al., 1997; Muthén, 1989).

Several approaches comprising statistical tests and comparisons of confidence intervals (CIs) have been proposed for multigroup analysis (MGA) in the context of PLS-PM to investigate whether observed heterogeneity is an issue that needs to be addressed. For example, a modified unpaired samples *t*-test can be used to compare a single parameter across two groups. Similarly, non-parametric tests were introduced to compare one parameter across two groups. Furthermore, the comparison of CIs was proposed to examine whether a parameter differs between two groups (Sarstedt et al. 2011). Since researchers often deal with more than two groups (e.g., Keil et al., 2000), the existing literature also suggests the omnibus test of group differences (OTG) to investigate whether a single parameter differs across multiple groups (Sarstedt et al., 2011). Because these approaches focus on a single parameter and not on the complete model, Klesel et al. (2019) recently introduced a distance-based permutation test, that compares all parameters simultaneously across groups. In doing so, they propose to consider the average squared Euclidean distance or the average geodesic distance of the model-implied indicator correlation matrix between groups. Hence, all model parameters are taken into account, and thus, the complete model is compared across groups.

Although previous literature has provided important advancements for applying MGA in the context of PLS-PM, there is a lack of a systematic comparison regarding existing approaches. To date, PLS-PM studies about MGA have primarily focused on proposing recommendations and demonstrating approaches' efficacy by means of an empirical examples (Hair et al., 2018; Matthews, 2017; Sarstedt et al., 2011). Moreover, they are limited to approaches that were available at the time of publication. Therefore, more current approaches including the distance-based permutation test (Klesel et al., 2019) have not been considered (Qureshi & Compeau, 2009). Hence, for IS researchers and applied researchers in general, it remains unclear which approach to favor under which condition.

To address this issue, we conduct an extensive Monte Carlo simulation to systematically investigate the performance of existing approaches. Since measurement invariance needs to be established before conducting MGA, the focus of our simulation study is on the structural model where, in principle, all parameters can vary across groups. To limit the scope of our simulation, we focus on the two most important scenarios: (i) only a specific parameter is compared across groups, and (ii) the complete structural model is compared across groups. The former is commonly applied in IS studies and is particularly relevant for IS researchers who have prior expectations about which effects in the model differ. For example, Sia et al. (2009) compare the effect of portal affiliation and peer endorsement on trusting beliefs across different cultures, i.e., individualist and collectivist. The comparison of the complete structural model is advantageous if researchers either have no prior expectations about which parts of the model differ or if they are interested in whether a complete postulated theory/model functions differently across groups. The latter might be a relevant research question itself. For example researcher might investigate whether the Unified Theory of Acceptance and Use of Technology (UTAUT) (Venkatesh et al., 2003) functions differently for different cultural backgrounds. Against this background, we examine the approaches' behavior when a single parameter and the complete structural model, respectively, is compared across groups.

This study contributes to extant literature in four ways. Firstly, we provide an overview of existing approaches to detect group differences in the context of PLS-PM. Secondly, we systematically investigate the performance of these approaches by means of an extensive Monte Carlo simulation. The results can be used to guide the selection of a specific test in empirical research. Thirdly, we show that the use of the OTG and the comparison of CIs are not recommendable as the former is misdesigned and the latter presents a misuse of CIs. Consequently, they either detect too often or too rarely group differences that do not exist. Fourthly, we show that approaches that are designed to compare a single parameter across groups require an adjustment for multiple comparisons if the complete structural model is compared. Otherwise, the family-wise error rate is inflated.

The remainder of this study is organized as follows: In Section 2, we give a brief overview of existing approaches that can be used to compare groups in the context of PLS-PM. Section 3 describes the design of our Monte Carlo simulation conducted to assess the approaches' performance. In Section 4, we present the results of our simulation. The paper closes with a discussion of the results, a recommendation which approach should be used in which situation, and an outlook on future research in Section 5.

Multigroup Analysis using PLS-PM

Generally, two types of heterogeneity are distinguished in the context of PLS-PM: (i) unobserved heterogeneity, and (ii) observed heterogeneity. As implied by its name, unobserved heterogeneity cannot be directly observed, and its source is often unknown to the researcher (Wedel & Kamakura, 2000). To uncover unobserved heterogeneity, several approaches have been proposed, such as the response based unit segmentation partial least squares (REBUS-PLS)(Esposito Vinzi et al., 2008), finite-mixture partial least squares (FIMIX-PLS) (Hahn et al., 2002), prediction-oriented segmentation (PLS-POS) (Becker et al., 2013), partial least squares genetic algorithm segmentation (PLS-GAS) (Ringle et al., 2010), and the Pathmox approach for PLS-PM (Lamberti et al., 2017). For an overview on approaches to reveal unobserved heterogeneity, we refer to Sarstedt et al. (2017). In contrast to unobserved heterogeneity, observed heterogeneity implies that heterogeneity stems from observable characteristics, such as cultural background (Srite & Karahanna, 2006), gender (Ahuja & Thatcher, 2005), or age (Lee & Kim, 2014). Usually, there are two ways to deal with observed heterogeneity. On the one hand, heterogeneity can be directly modeled, e.g., by including interaction terms (Henseler & Fassott, 2010). On the other hand, the observable characteristics can be used to partition the original dataset into groups (Henseler et al., 2009). Subsequently, the model is estimated for each group separately. Finally, to examine whether group differences due to observed characteristics exist in the population, MGA can be used.

To conduct a MGA in the context of PLS-PM various approaches have been proposed. An overview is given in Table 1 and a brief description is provided in the following. For a detailed discussion on the technical details of each approach, we refer to existing literature (Klesel et al., 2019; Matthews, 2017; Sarstedt et al., 2011) and the references mentioned in Table 1.

Insert Table 1 About Here

To compare a single model parameter across two groups, parametric and non-parametric tests have been proposed. The former comprises the *t*-test based on bootstrap standard errors for two independent samples assuming equal variances across groups (**P**arametric **T**est **E**qual variances) (Keil et al., 2000). Likewise, a modification of this test can be applied allowing for unequal variances across groups (**P**arametric **T**est **U**nequal variances)(Nitzl, 2010; Sarstedt et al., 2011). In both cases, the *p*-value to draw conclusions about the parameter difference is based on the Student's *t*-distribution. Examples in IS literature can be found in Hsieh et al. (2008), Ahuja and Thatcher (2005), and Keil et al. (2000).

Besides parametric tests, two non-parametric tests have been suggested to compare a parameter across two groups. Firstly, a one-sided bootstrap-based test that compares the absolute size order of the parameter estimates from the bootstrap runs across the two groups to draw conclusion about the parameter difference in the population (**N**on-parametric **B**ootstrap-based **T**est) (Henseler, 2007; Henseler et al., 2009). Secondly, a permutation test was proposed for comparing a single parameter across two groups (**N**on-parametric **P**ermutation-based **T**est) (Chin, 2003; Dibbern & Chin, 2005; Keil et al., 2000). As the name suggests, it builds on permutation to obtain the reference distribution of the parameter difference from which the critical values are drawn that are used to make the decision whether a parameter difference is statistically significant. An IS example for the NBT are given in Hew et al. (2017), and for the NPT, IS examples can be found in Wolf et al. (2012); Dibbern et al. (2012); Srite and Karahanna (2006).

Moreover, although not a statistical test, the comparison of CIs has been proposed to investigate for a parameter difference between two groups (Sarstedt et al, 2011). It comes with two flavors: Firstly, it is proposed to investigate whether the CIs constructed around a specific parameter estimate of the two groups overlap (**C**onfidence **I**nterval **O**verlap); if not, it is concluded that the parameter differs between the two groups. Secondly, it is suggested to examine whether the CI constructed around a specific parameter estimate of one group covers the corresponding parameter estimate of the other group, and vice versa (**C**onfidence **I**ntervals Covers **P**arameter); if not, it is concluded that, the parameter differs between the two groups. To the best of our knowledge, so far no IS study has applied this approach.

To compare a single parameter across more than two groups, the **O**mnibus **T**est of **G**roup Differences (OTG) was introduced, which, in principle, mimics the *F*-test known from analysis of variance (ANOVA) (ANOVA, Sarstedt et al., 2011). It calculates the mean of the bootstrap parameter estimates for each group and uses these means to compute the *F*-test statistic known from ANOVA. Subsequently, permutation is used to obtain reference distribution of the *F*-test statistic. An IS example is given in Papagiannidis et al. (2017).

Only recently, a new permutation test was introduced (**N**on-parametric **D**istance-based **T**est) (Klesel et al., 2019), which opens new research avenues, namely, that researchers can investigate whether their postulated theories/models function differently across groups. In doing so, the average distance between the model-implied indicator correlation matrices is investigated to draw conclusions about whether the complete model is different across two or more groups. Specifically, the test examines either the average squared Euclidean distance or the average geodesic distance of the model-implied indicator correlation matrix between groups. Similar to the NPT, it builds on permutation to obtain the reference distribution of the distances. If there are no group difference of the model-implied indicator correlation matrix in the population, both distances should be closely distributed around zero within the limits of sampling variation. In contrast, if the model-implied correlation matrix differs across groups, the differences should substantially exceed zero and the test should turn to be significant. As the test has been proposed only recently, to the best of our knowledge, it has not been applied in IS research yet.

Although the performance of some of these approaches has already been assessed by simulation studies, e.g., the efficacy of the PTE (Qureshi & Compeau, 2009), the NPT (Chin & Dibbern, 2010), and the NDT (Klesel et al., 2019), this has mainly been done in isolation, i.e., only a very limited selection of approaches was considered in the simulation. Moreover, numerous approaches remain unassessed. Hence, there is a lack of research that juxtapose the approaches' performance systematically. Therefore, it remains unclear which approach is most suited when it comes to MGA in the context of PLS-PM.

Monte Carlo Simulation

To address this shortcoming, we conducted an extensive Monte Carlo simulation to assess the performance of the approaches shown in Table 1**Error! Reference source not found.** for comparing the complete model and only a single parameter, respectively, across groups. In comparing the complete model, only the structural model is compared, as measurement invariance must be established before conducting MGA (Henseler et al., 2016; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000), see Henseler et al. (2016) on how to assess measurement invariance in the context of PLS-PM. Hence, all path coefficients and the correlations among the exogenous constructs are compared jointly across groups. Thus, we test the following null hypothesis: H_0: $\mathbf{\Sigma}_1$ = $\mathbf{\Sigma}_2$=…= $\mathbf{\Sigma}_G$, where $\mathbf{\Sigma}_i$ is the model-implied construct correlation matrix of group *i* and G is the total number of groups. This reflects a situation in applied research in which a researcher has established measurement invariance and wants to investigate whether a postulated theory functions differently across groups. Against this background, in case of the NDT, the model-implied construct correlation matrix instead of the model-implied indicator correlation matrix, as originally proposed, is compared across groups. For the approaches that are designed to compare only a single parameter across groups, the null hypothesis that the complete structural model is equal across groups is rejected if at least one difference of the path coefficient or the correlation among exogenous constructs is significantly different from zero. Since the NBT is a one-sided test, it requires a-priori expectations about the direction of the population parameter difference. As it is not clear how to compare several parameters across groups without such expectations, it was not taken into account for the case where the complete structural model is compared. Considering the comparison of CIs, the literature provides various types of CIs such as the normal CI and the percentile bootstrap CI (Davison & Hinkley, 1997; Efron & Tibshirani, 1993). In our simulation, we followed the advice of Sarstedt et al. (2011) and employed the bias-corrected and accelerated (BCa) bootstrap CI (Efron, 1987). Finally, the NDT was also employed for the scenarios where only a single path coefficient is compared across groups to examine its performance when only a single population path coefficient varies across groups. It is noted that the NDT has not been designed to compare a single path coefficient. Hence, as in the case of the complete model comparison, the model-implied construct correlation matrix is compared across groups in these scenarios. This represents a situation where a researcher expects the complete model to differ, but in fact only a single path coefficient is different across groups.

To examine the tests' performance, we investigated their type I error rates and power. The type I error refers to the rejection of the null hypothesis of no group differences when indeed no group differences exist in the population. In specific, we investigated whether the considered approaches yield rejection rate close to the predefined significance levels of 1%, 5%, and 10%, respectively. In contrast, the power of a test refers to a test's ability to reject the null hypothesis of no group differences when it is indeed false. Hence, large rejection rates are desired in the case of group differences in the population. According to Cohen (1988), the power of a test should be at least 80%. Although CIs are no statistical tests, they are often employed for statistical inference (Wood, 2005). This is also true for MGA in the context of PLS-PM, in which the comparison of CIs was proposed (Sarstedt et al., 2011). Hence, they were included in the following and the same criteria as for statistical test, namely type I error rate and power, were applied to judge their performance.

To examine different types and magnitudes of group differences, we considered various population models. Since measurement invariance must be established before conducting MGA, we only varied the magnitude of the path coefficients in the structural model. In designing the structural model, we used a model comprising two exogenous and five endogenous constructs. This reflects a model complexity that IS researchers employing PLS-PM typically encounter in their studies (Ringle et al., 2012). Figure 1 displays the structural model containing seven constructs. All constructs were specified as composites and, following Grace and Bollen (2008), are displayed as hexagons in the structural model to distinguish composites from latent variables.

Insert Figure 1 About Here

As illustrated in Figure 1, the correlation between the two exogenous composites η_1 and η_2 was set to zero. Moreover, the structural error terms (ζ) were assumed to be mutually independent and independent of the corresponding explanatory variables.

To investigate the approaches' performance in case of comparing the complete structural model across groups, we employed the path coefficients as displayed in Table 2. Group 1 represents the reference group. Compared to Group

1, for Groups 2, 3, 4, and 5, all path coefficients, except of β_{43} and β_{53}, were either increased or decreased by .05, .1, .15, and .2, respectively.

Insert Table 2 About Here

Similarly, in case of comparing only a single parameter across groups, we used the population parameters from Group 1, shown in Table 2, and only varied β_{31} from .2 to .6 by .1. Consequently, in case of the comparison of a single parameter, β_{31} was chosen as follows: $\beta_{31} = .2$ for Group 1, $\beta_{31} = .3$ for Group 2, $\beta_{31} = .4$ for Group 3, $\beta_{31} = .5 \text{ for Group } 4$, and $\beta_{31} = .6$ for Group 5.

Since the path coefficients were varied systematically, the explained variance of the endogenous constructs (R^2) was not fixed. As a consequence, the R^2 values varied across the groups. As shown in Table 3, the R^2 values range from 0.13 to 0.71, which are common observed R^2 values in the IS literature (e.g., Venkatesh et al., 2003).

Insert Table 3 About Here

The population weights to form the corresponding composites were chosen as follows: $\boldsymbol{w}_1' = (0.6 \quad 0.4 \quad 0.2)$, $\boldsymbol{w}_2' = (03. \quad 0.5 \quad 0.6)$, $\boldsymbol{w}_3' = (0.4 \quad 0.5 \quad 0.5)$, and $\boldsymbol{w}_4' = (0.4 \quad 0.5 \quad 0.5)$, $\boldsymbol{w}_5' = (0.6 \quad 0.4 \quad 0.2)$, $\boldsymbol{w}_6' = (0.3 \quad 0.5 \quad 0.6)$, and $\boldsymbol{w}_7' = (0.4 \quad 0.5 \quad 0.5)$.

Moreover, we varied the following aspects including those that already proved to be relevant in the context of MGA (e.g., Qureshi & Compeau, 2009): number of groups in the comparison; adjustment for multiple comparisons; total sample size; sample size distribution between groups; and the data distribution. An overview of the various design factors and their variations is provided in Table 4.

Insert Table 4 About Here

Number of groups in the comparison

In many situations, researchers deal with more than two groups. Therefore, we varied the number of compared groups from 2, 3 to 5 to examine whether the approaches are capable to detect differences across multiple groups.

To assess whether the approaches keep the predefined significance level, we compared Group 1 to itself, i.e., we draw different samples from the same population and therefore no group differences are present. In contrast, to examine the power of the considered approaches, we compared groups that function differently, i.e., group differences are present. Table 5 presents the made comparisons and displays the values of the average geodesic distance (d_G). It is noted that in the scenarios where a single parameter was compared across groups, only β_{31} was compared, except for the NDT which by design compares the model-implied construct correlation matrix across groups. Hence, we expect a lower statistical power of the NDT in these scenarios compared to tests that were originally designed to compare a single parameter. Overall, it is expected that the approaches produce rejection rates close to the significance level if no group differences are present and rejection rates above the significance level if group differences exist in the population. Moreover, the rejections are expected to increase in case of larger differences.

Insert Table 5 About Here

Multiple comparisons adjustment

We considered various approaches proposed for MGA in the context of PLS-PM to test whether the complete structural model or a single path coefficient differs across two or more groups. Since most of the approaches were originally designed to compare only one parameter across two groups, they face the well-known problem of multiple comparisons if applied to compare one parameter across more than two groups. Similarly, if they are used to compare several parameters jointly. This is also recognized in the PLS-PM literature on MGA, see for example Hair et al. (2018). For example, if the PTU was applied to compare the complete structural model across five groups, 120 single tests were conducted (10 group-comparisons times 12 compared parameters). As a consequence, the family-wise error rate will be inflated. Table 6 displays the number of conducted tests that are necessary to come up with the desired conclusion about the null hypothesis.

Insert Table 6 About Here

To address this issue, the literature provides various corrections for multiple comparisons that can be applied to adjust the *p*-values of the numerous conducted tests. In our study we employed the Bonferroni correction and the corrections proposed by Holm (1979), Hochberg (1988), Hommel (1988), and Benjamini and Hochberg (1995). In the following, adjustments for multiple comparisons were applied to the PTE, the PTU, the NPT, and the OTG. For the NBT no adjustment was applied, as it is a one-sided test and thus requires expectations about the sign of the parameter difference. Without such expectations, a proper adjustment of the *p*-values is not possible. Similarly, for the CIO, and the CIP no adjustments were applied. Although in general adjustments for CIs are possible (Benjamini & Yekutieli, 2005), the CIO and CIP present a misuse of CIs (see e.g., Altman, 2000), and thus, are expected to not perform well in general. In this case, the investigation of the performance of the various adjustments would only contribute marginally. Moreover, and as displayed in Table 5, the NDT is designed to compare the model-implied construct correlation matrix across several groups, hence, it does not require a correction for multiple comparisons. Finally, a proper p-value adjustment requires p-values different from 0. Otherwise, the adjusted p-values are also equal to 0 leading always to a rejection of the null hypothesis regardless of whether an adjustment was applied. Consequently, for the NPT, the number of permutation runs plays an important role in the context of p-value adjustment. The same is true for the number of bootstrap runs in the NBT if an adjustment for multiple comparisons is applied. A sufficient number of runs is required to obtain a sufficiently accurately estimated distribution of the test statistic under the null hypothesis. As proposed by Chin et al. (2003), we employed 1,000 permutation runs. As shown in Section 4.2, 1,000 permutation runs were not sufficient in the case of a large number of comparisons and led to an abnormal test's behavior, i.e., too high rejection rates.

As known from the literature on the multiple comparisons problem, we expect that the family-wise error rate is higher than the predefined significance level in case of no adjustment, while it is expected to be close to the significance level if a correction for multiple comparisons is applied.

Total sample size

The PLS-PM literature on MGA highlights the role of a sufficient sample size (Qureshi & Compeau, 2009). Therefore, we varied the total sample size, i.e., the sum of sample sizes across all groups, from 300, 600, 1500, to 3000. As expected from statistical significance tests, with increasing sample size, the statistical power should increase in case of group differences, and in case of no group differences, the predefined significance level should be maintained with more precision.

Sample size distribution between groups

The existing literature on PLS-PM showed that tests for group comparisons suffer from a loss of statistical power in the case of unequal group sizes (Chin & Dibbern, 2010). Therefore, we also took differences in the sample size distribution between groups into account. In doing so, we considered equally distributed, moderately unequally distributed, and severely unequally distributed sample sizes across groups. An overview of the different sample size distributions for the two and three group-comparison is shown in Table 7 and for the five group-comparison, it is given in Table 8. It is cautioned that for some comparisons the sample size per group is very small. This is particularly relevant in terms of multiple groups which result in a high number of comparisons. Since studies employing PLS-PM often face small sample sizes (see e.g., Ringle et al., 2012), we deliberately investigated the tests' performance in these situations although the tests are expected to show a very low statistical power. Likewise, we expect that compared to equally distributed sample sizes, the statistical power decreases in the case of unequal sample sizes across groups.

Insert Table 7 About Here

Insert Table 8 About Here

Data distribution

In empirical studies, researchers rarely deal with samples that stem from a normal distribution. Hence, we additionally investigated the approaches' performance in the case of non-normally distributed data. To generate these datasets, we rescaled each standard normally distributed indicator by a scaling factor, as proposed by Dijkstra and Henseler (2015a), leading to an excess kurtosis of approximately 1.71. In the case of non-normally distributed datasets, we expect a lower statistical power of the approaches compared to a situation in which normality is given.

Data generation and analysis

The complete simulation was conducted in the statistical programming environment R (R Core Team, 2020). The multivariate standard-normally distributed datasets were drawn using the *mvrnorm* function of the MASS package (Venables & Ripley, 2002). We used the *csem* function provided by the cSEM package to conduct the PLS-PM estimations (Rademaker & Schuberth, 2020). In doing so, all composite models were estimated by Mode B and the path weighting scheme was used for inner weighting. Mode B was applied because it provides consistent estimates for composite models (Dijkstra, 2017), regardless of whether the composite was formed by correlation or regression weights in the population. Composites formed by correlation weights are a special case of composites formed by regression weights (Cho & Choi, 2020). Since the cSEM package is still under development, we cross-checked the initial estimations with ADANCO (Henseler, 2017a) and SmartPLS (Ringle et al., 2015). Since all software packages provided the identical results for the estimation, we carried out the complete simulations with the cSEM package. The approaches for MGA were conducted by using the *testMGD* function from the same package. For each condition, we executed 500 simulation runs. The number of bootstrap runs was set to 1000. Similarly, as proposed by Chin (2003), we set the number of permutation runs to 1000. Finally, inadmissible estimations, i.e., estimations that have not converged or produced no positive semi-definite model-implied indicator/construct correlation matrix, were replaced by admissible one. As a consequence, each simulation run is based on 1,000 valid bootstrap and permutation runs, respectively.

Results

In this section we highlight the most important findings of our simulation study. The complete results can be found the Supplementary Material. Since applied research oftentimes deals with non-normally distributed data and unequal group-sizes, the rejection rates presented in the following are based on non-normally distributed datasets, moderately unequally distributed sample sizes across groups, and a significance level of 5% if not explicitly indicated otherwise. Moreover, as shown in Section 4.3, the type of adjustment hardly influences the results. Therefore, only the results based on the adjustment proposed by Holm (1979) are reported in case that a correction for multiple comparisons is applied. The Holm correction is well established and more powerful compared to the well-known Bonferroni correction (Aickin & Gensler, 1996).

The following Subsections 4.1 and 4.2 show the results in case of comparing only single path coefficient and the complete structural model, respectively, across groups. Finally, in Subsections 4.3 and 4.4 we provide specific results related to the performance of the various adjustments for multiple comparisons and the impact of the data and sample size distribution.

Comparison of one path coefficient

Figure 2 shows the rejection rates for all approaches in case of no group differences and different numbers of group comparisons. The dashed line represents the predefined significance levels of 5%. Additionally, the grey area highlights the 95% normal CIs constructed around the reject rates.[2]

Considering the two group-comparison, the results show that most of the tests, namely the PTE, PTU, NBT, NPT, and the NDT[3] produce rejection rates close to the predefined significance levels. In contrast, the CIO indicates too rarely group differences, i.e., it is too conservative. Similarly, the CIP and the OTG indicate group differences too often although they do not exist. This is especially striking for the OTG, which produces rejection rates close to 100%.

With regard to three and five group-comparisons, the PTE and the PTU show rejection rates that are bit too low compared to the assumed significance levels. However, for an increasing total sample size, the rejection rates seem to converge towards the significance level. The NDT produces rejection rates close to the predefined significance level. Similarly, the rejection rates of the NPT are close to predefined significance levels, while the rejection rates produced by the NBT are too high, which becomes worse for the five group-comparison. This highlights the importance of the multiple comparisons adjustment, which was not applied to the NBT. Similar to the two group-comparison, the CIP and the OTG show rejection rates that are much higher than the assumed significance level. Considering the CIO, for the three group-comparison, it produces rejections rates that are below the desired 5% level. In contrast, for the five group-comparison, it produces rejection rates close to predefined significance level. This can be explained by the fact that the CIO originally indicates too rarely group differences (see the two group-comparison), which is counterbalanced by the number of comparisons for which it was not corrected in the three and five group-comparison. As more pairs of CIs need to be compared in the five group-comparison, it is more likely that CIs overlap by chance. Overall, the total sample size has only a minor impact on the approaches' performance with regard to maintaining the predefined significance level.

Insert Figure 2 About Here

Figure 3 depicts the rejection rates for the different number of group-comparisons in case that the considered parameter β_{31} differs across groups. As expected, the rejection rates of all approaches increase with an increasing total sample size. Similarly, the rejection rates increase with an increasing parameter difference across groups. However, small and small-medium parameter differences are oftentimes undetected, i.e., rejection rates below 80% are observed. This is not true for the CIP and the OTG, which overall produce the highest rejection rates. Especially, the OTG produces rejection rates of almost 100%, regardless of the parameter difference. Although the rejection rates are fairly similar between the PTE, the PTU, the NBT, and the NPT in case of the two group-comparison, the latter is slightly more powerful. In contrast, for the cases that more than 2 groups are compared, the NBT produces the highest rejection rates among the four tests as no adjustment for multiple comparisons is applied to this test in these cases. Finally, the NDT, which compares the complete structural model in form of the model-implied construct correlation matrix across groups, shows the lowest rejection rates compared to the other approaches. This can be explained by the fact that the NDT compares effectively the complete structural model.

Insert Figure 3 About Here

Comparison of the complete structural model

In the following, we present the results with regard to the comparison of the complete structural model. Figure 4 presents the rejection rates for the case that the complete structural model does not differ across groups. As can be seen, the PTE, the PTU, the NPT, and the NDT keep the 5% significance level when two groups are compared.

In case that more than two groups are compared, the rejection rates of the PTE and the PTU are a bit too low, despite the adjustment for multiple comparisons. However, for an increasing total sample size, the rejection rates converge towards the desired 5% level. Considering the CIO, the CIP, and the OTG, they produce rejection rates remarkably above the predefined significance level of 5%. For the former two, this can be explained by the fact that no adjustment for multiple comparisons is applied for the approaches that rely on CIs. Although in case of CIO and the two-group comparison the rejection rates seems to meet the desired 5%, it should not be generally concluded that it keeps the predefined significance level. This can be explained by the fact that the number of comparisons counterbalances the too low rejection rates in case of comparing a single parameter across two groups leading to rejection rates that seem to keep the 5% significance level.

Insert Figure 4 About Here

Considering the NPT, it produces rejection rates close the assumed significance level in the case of three compared groups. However, if five groups are compared the rejection rates are a bit too high. This can be explained by the ratio of permutation runs to the number of involved comparisons. If the number of permutation runs is not sufficiently large compared to the number of comparison, in some instances *p*-values of exactly zero are produced by the NPT.[4] As a consequence, these *p*-values remain zero after the *p*-value adjustment for multiple comparisons leading to too high rejection rates. Figure 5 illustrates this behavior of the NPT in case of the five group-comparison and no group differences. In doing so, the rejection rates obtained for 1,000 permutation runs (see Figure 4) are contrasted to those obtained for 5,000 permutation runs. As can be seen, 1,000 permutation runs are not sufficient to maintain the 5% significance level, however, if 5,000 permutation runs are employed, the rejection rates are close to the desired 5%. This emphasize the importance of choosing a sufficiently large number of permutation runs. It is noted, that this problem does not occur for the PTE and the PTU, as their p-values are derived from a *t*-distribution. As a consequence, the *p*-value of single test is never exactly equal to zero.

Insert Figure 5 About Here

Figure 6 displays the approaches' rejection rates in case that the structural model differs across groups. Similar to the comparison of one path coefficient, for the comparison of the complete structural model, the results show that for an increasing total sample size and group differences, the rejection rates increase. For small total sample sizes (particularly 300 observations in total), none of the approaches, except the CIP and the OTG, are capable to reliably detect small and small-medium group differences. This is not surprising, as shown in the case of no group differences, the CIP and the OTG lead to the highest type I error rates. Among the approaches that reliably kept the significance level in case of no group-difference, the NDT show most of the times the highest rejection rates.

Insert Figure 6 About Here

Addressing multiple comparison issue

We highlighted the fact that researchers conducting a MGA in the context of PLS-PM might face multiple comparison issues. In this case, a proper way of adjusting p-values for multiple comparisons is required. To shed more light into the various adjustments' performance, we selected a specific condition which is representative for the multiple comparison issue to show the similarities and differences of the various corrections. It is noted that the adjustments' behavior is similar in the remaining conditions where an adjustment was applied.

Figure 7 depicts the rejection rates for the PTE, the PTU, and the NPT using five different p-value adjustment and the original rejection rates where no adjustment is applied. In specific, the rejection rates for the comparison of the complete structural model across three groups are contrasted for the cases of no group-differences and medium-large group differences. The results clearly show that without an adjustment, the family-wise error rate is inflated. However, this can be overcome by applying an adjustment for multiple comparisons. Moreover, the results suggest that in our case, all adjustments fulfill their purpose and there is no superior adjustment, i.e., all adjustments lead to almost identical rejection rates regardless of whether group differences are present.

Insert Figure 7 About Here

Data and sample size distribution

As observed in the literature and argued earlier, data and sample size distribution can have negative effects on the approaches employed in MGA (Chin & Dibbern, 2010). To highlight these effects, Figure 8 contrasts the performance of various tests applied to compare the complete structural model across three groups in case of small-medium group differences. It can be observed that all approaches produce the highest rejection rates if the data is normally distributed and the sample sizes are equal between groups. In contrast, the rejection are the lowest if the data is non-normally distributed and the sample size is severely unequally distributed between groups. The difference in the approaches' performance can be substantial. For example, in case of a total sample size of 1,500 observations, the difference in the rejection rates can be almost 40%-points implying that tests do not reliably detect group differences, i.e., the rejection rates are below the desired threshold of 80%. However, it is noted that for an increasing total sample size, the difference in the rejection rates decreases. Similar effects can be observed in the remaining conditions.

Insert Figure 8 About Here

Discussion and Outlook

Discussion of the results

It is well known that ignoring observed heterogeneity can lead to severely biased results, and therefore to erroneous conclusions (Jedidi et al., 1997; Muthén, 1989). Consequently, investigating group differences has become an essential endeavor when dealing with datasets that stem from potentially different populations. Several approaches have been suggested for MGA in the context of PLS-PM including statistical tests and the comparison of CIs. This study contributes to the existing body of knowledge presenting the results of a systematic comparison of existing approaches within an experimental setting. Therefore, we substantiate existing recommendations and raise issues that have not been uncovered yet. We also contribute to studies which have investigated a very limited number of approaches. For instance, Klesel et al. (2019) exclusively focused on the NDT in their simulation study, Chin and Dibbern (2010) studied only the performance of the NPT, and Qureshi and Compeau (2009) compared only two approaches including the PTE. To gain more detailed insights about the approaches' performance and the current recommendations, the study at hand is the first that compares the performance of a broad range of approaches in a controlled environment.

Based on the review of existing approaches, applied researchers employing PLS-PM can choose from a variety of approaches to compare a single parameter or the complete structural model across groups. While the usefulness of comparing single parameters across groups is widely acknowledged, the comparison of the complete structural model is particularly relevant if researchers either have no prior expectations about which parameters differ across groups or if they want to examine whether a proposed theory/model functions differently across groups, e.g., females and males. To the best of our knowledge, the comparison of the complete structural model has not been done so far in empirical research. This might be due to the fact that before the development of the NDT, no approach has been available that is designed for that specific purpose. As a consequence, researchers might be not aware of that opportunity.

While this diversity provides manifold opportunities, it is important to choose the approach in accordance to the research question, i.e., comparing the complete structural model or only a subset of single parameters, to avoid pitfalls such as an inflated family-wise type I error rate. In general, the simulation results are in line with our expectations, i.e., the power increases for an increasing sample size and an increasing group difference. In contrast, the power decreases in case of non-normally distributed samples compared to datasets that stem from a normally distributed population. Similarly, the power decreases if the sample size distribution varies between groups. As consequence, researchers should strive for a sufficiently large total sample size if they conduct a MGA in the context of PLS-PM, particularly if their samples are non-normally distributed and/or the number of observations substantially varies between groups. Otherwise, researchers will likely face approaches that are statistically underpowered for detecting postulated group differences. This is in line with the results of previous simulation studies (e.g., Qureshi & Compeau, 2009) and particularly important, if researchers only assume a small difference between groups, regardless of whether a single parameter or the complete structural model is compared. As our simulation study showed, small group differences mainly remain undetected, even for total sample sizes of 3,000 observations. Our findings are particularly important in the context of PLS-PM because a small sample size is often used to justify its application (see e.g., Nitzl, 2016; Ringle et al., 2012). To address this issue, it is recommended to determine the necessary sample size in advance to achieve a sufficient statistical power of the applied statistical tests. This can be done in several ways, for example by applying heuristic rules such as Cohen's power tables (Cohen, 1992) or conducting a Monte Carlo simulation.

Based on our simulation findings, if researchers are interested in comparing a single parameter between two groups, we recommend to use the NPT. Although our simulation results showed that the NPT, the PTU, the PTE and the NBT perform similar in this situation, the NPT showed a slightly higher power while maintaining the predefined significance level compared to the other approaches that are designed to compare a single parameter. This recommendation is in line with current recommendations from the PLS-PM literature (e.g., Hair et al., 2018; Matthews, 2017). Moreover, in contrast to the PTE and the PTU, it does not rely on distributional assumptions. If the NPT is used to compare one or more parameters across more than two groups, an adjustment should be employed to avoid an inflation of the family-wise error rate. This is in line with literature related to the multiple comparisons issue (e.g., Aickin & Gensler, 1996) and has been echoed in the PLS-PM literature (Hair et al., 2018). To employ the NPT in this situation, it is emphasized that researchers should use a sufficiently large number of permutation runs to ensure that the adjustment for multiple comparisons works properly. Consequently, researchers employing the NPT for multiple comparisons should examine before the adjustment whether some of the p-values are exactly equal to zero. If this is the case, it is recommended to increase the number of permutation runs. With regard to the p-value adjustments, our simulation study showed that all of them produced very similar results.

In contrast, if researchers want to compare the complete structural model, they should rely on the NDT, which was specifically designed for that purpose, and showed the highest power in most of the situation. This is particularly apparent, in case of small group differences and a large total sample size and in case of smaller total sample sizes if the group differences are moderate and large, i.e., small-medium to large. Subsequently, they can use the NPT in combination with a correction for multiple comparison to examine which parameters differ. This procedure resembles the approach commonly used in the context of ANOVA, where first a *F*-test is conducted and subsequently, multiple comparison procedures are employed to investigate which means actually differ significantly.

Considering the OTG, the CIO, and the CIP, their use is not recommended, regardless of whether only a single parameter or the complete structural model is compared. As shown by our simulation, the OTG proposed by Sarstedt et al. (2011) almost always rejects the null hypothesis of no group differences. This can be explained by a misconception in the design of the test. The OTG investigates whether the average bootstrap estimates of a parameter differ significantly across groups, but not the estimated parameters. Typically, the mean of the bootstrap estimates of a specific parameter is close to the corresponding parameter estimate. Therefore, for an increasing number of bootstrap runs the value of the ANOVA *F*-test statistic, which compares the bootstrap means across groups, increases, as the within group variation, i.e., variance of the bootstrap means, decreases, even in case of no group-difference in the population. In contrast, the *F*-test statistics based on the permutation samples that are used to approximate the reference distribution under the null hypothesis of no group differences, are relatively small and therefore closely distributed right of the zero. As a consequence, for a large enough number of bootstrap runs, the OTG almost always rejects the null hypothesis of no parameter difference across groups. Similarly, the CIO and the CIP generally do not keep the predefined significance level. This is due to the fact that CIs are misused as they are compared across groups to draw inferences (see e.g., Altman, 2000). To overcome this problem, researchers can construct a CI around the parameter difference to examine whether it covers the zero, instead of investigating whether two CIs overlap.

Recommendations

Based on the results of this simulation, we propose several guidelines for MGA in the context of PLS-PM. First, the scope of the analysis should be clearly articulated. Specifically, it should be clarified whether a single path, multiple paths or the complete model is compared. Second, a test procedure should be selected that matches the scope of the analysis. For instance, if a complete model is compared, it is beneficial to employ the NDT. However, if only a single parameter is compared across two groups the NPT is recommended. Third, our study highlights the importance of a sufficient sample size to achieve a satisfactory test's power. Particularly if researchers expect only small group differences, more than 1,000 observations per group are required. Therefore, researchers are advised to determine the sample size before the data collection that is necessary to achieve a satisfactory statistical power for the employed testing procedure. Fourth, if tests whose p-values are directly based on the bootstrap or the permutation runs are combined with a p-value adjustment for multiple comparison, special attention should be given to the number of bootstrap and permutation runs, respectively. As shown by our results, an insufficiently small number of runs can lead to p-values that are exactly equal to 0 and thus render any adjustment useless. Therefore, it is recommended to examine the size of the p-values before the adjustment and rerun the test with an increased number of runs if p-values are observed that are exactly equal to zero.

Limitations and future research

A Monte Carlo simulation is always limited to its design which invites future research. In specific, applied researchers using PLS-PM also face structural models that contain both latent variables and composites (Benitez et al., 2020). Therefore, future research could investigate the approaches' performance for structural models containing a mixture of composites and latent variables. To estimate models containing latent variables, PLSc should be employed as original PLS-PM is known to produce inconsistent estimates (Dijkstra, 1981). Moreover, future studies investigating the approaches' performance, might consider the effect of model complexity, i.e., number of path coefficients and constructs, on the approaches' performance. This is particularly relevant for the case that the complete structural model is compared across groups. In our simulation we either varied one path coefficient or all path coefficients across population groups. For future research, situations should be examined in which a researcher compares the complete structural model across groups but only a subset of path coefficients differs across groups in the population. In doing so, it is of particular interest to compare the performance between the NDT and tests that have been designed to compare only a single coefficient to figure out when the use of the NDT becomes advantageous. Moreover, in this study, we varied the population path coefficients across groups without fixing the R^2 values of the endogenous constructs. Future research should investigate the effect of not fixing the R^2 values on the tests' performance.[5] Considering the NDT, future research might examine and contrast its performance for other discrepancy measures than the (average) geodesic distance. Furthermore, our study solely focused on PLS-PM. Hence, further simulation studies are necessary to examine how the approaches perform in combination with other estimators for composite models such as generalized structured component analysis (Hwang et al., 2017; Hwang et al., 2020; Hwang & Takane, 2004). Finally, it might be worthwhile for future research to design a two-sided version of the NBT which facilitates its use for comparing the complete structural model.

Notes

[1] The literature on structural equation modeling usually distinguishes two ways of representing abstract concepts, namely, (i) by a latent variable, and (ii) by a composite (Benitez et al., 2020; Henseler, 2017b, 2021; Henseler & Schuberth, 2020; McDonald, 1996; Rhemtulla et al., 2020; Rigdon, 2012; Rigdon et al., 2017; Schuberth et al., 2018a). Originally, PLS-PM estimates consistently structural models that contain only composites. However, in its current form, known as consistent partial least squares (PLSc), it is a consistent estimator for structural equation models containing both latent variables and composites (Dijkstra & Henseler, 2015b; Rademaker et al., 2019). While parameters that are related to latent variables are corrected for attenuation by PLSc, the parameters associated with composites remain untouched, see e.g., Schuberth et al. (2018b).

[2] The 95% normal confidence interval is calculated as $\hat{p} \pm \Phi^{-1}(0.975)\sqrt{\hat{p}(1-\hat{p})/500}$, where $\hat{p}$ represents the rejection rate and $\Phi^{-1}(\,)$ is the quantile function of the standard normal distribution.

[3] It is noted that for the NDT the model-implied construct correlation matrix and not a single path coefficient was compared across groups.

[4] For the NPT, the *p*-value is calculated as the sum of shares of parameter differences from the permutation runs that exceed the positive original parameter difference and fall below the negative original parameter difference. If a large number of comparisons is performed, it is likely that in some instances, the original absolute parameter difference is larger than all absolute parameter differences from the permutation runs. As a consequence, a *p*-value of exactly zero is produced, which cannot be properly adjusted for multiple comparisons.

[5] We thank an anonymous reviewer for this suggestion.

References

Ahuja, M., & Thatcher, J. B. (2005). Moving beyond intentions and toward the theory of trying: Effects of work environment and gender on post-adoption information technology use. *MIS Quarterly*, *29*(3), 427–459. https://doi.org/10.2307/25148691

Aickin, M., & Gensler, H. (1996). Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods. *American Journal of Public Health*, *86*(5), 726–728. https://doi.org/10.2105/ajph.86.5.726

Altman, D. G. (2000). Confidence intervals in practice. In D. G. Altman, D. Machin, T. Bryant, & M. Gardner (Eds.), *Statistics with confidence: Confidence intervals and statistical guidelines* (2nd ed., pp. 6–14). BMJ Books.

Becker, J.-M., Rai, A., Ringle, C. M., & Völckner, F. (2013). Discovering unobserved heterogeneity in structural equation models to avert validity threats. *MIS Quarterly*, *37*(3), 665–694. https://doi.org/10.25300/misq/2013/37.3.01

Benitez, J., Henseler, J., Castillo, A., & Schuberth, F. (2020). How to perform and report an impactful analysis using partial least squares: Guidelines for confirmatory and explanatory IS research. *Information & Management*, *57*(2), 103168. https://doi.org/10.1016/j.im.2019.05.003

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300.

Benjamini, Y., & Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, *100*(469), 71–81. https://doi.org/10.1198/016214504000001907

Chin, W. W. (2003). A permutation procedure for multi-group comparison of PLS models. In M. Vilares, M. Tenenhaus, P. Coelho, V. Esposito Vinzi, & A. Morineau (Eds.), *Proceedings of the International Symposium PLS'03. PLS and related methods* (pp. 33–43).

Chin, W. W., & Dibbern, J. (2010). An introduction to a permutation based procedure for multi-group PLS analysis: Results of tests of differences on simulated data and a cross cultural analysis of the sourcing of information system services between Germany and the USA. In V. Esposito Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of partial least squares* (pp. 171–193). Springer.

Cho, G., & Choi, J. Y. (2020). An empirical comparison of generalized structured component analysis and partial least squares path modeling under variance-based structural equation models. *Behaviormetrika*, *47*(1), 243–272. https://doi.org/10.1007/s41237-019-00098-0

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203771587

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037//0033-2909.112.1.155

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press.

Dibbern, J., & Chin, W. W. (2005). Multi-group comparison: Testing a PLS model on the sourcing of application software services across Germany and the USA using a permutation based algorithm. In F. Bliemel, A. Eggert, G. Fassott, & J. Henseler (Eds.), *Handbuch PLS-Pfadmodellierung: Methode, Anwendung, Praxisbeispiele* (pp. 135–160). Schäffer-Poeschel Verlag.

Dibbern, J., Chin, W. W., & Heinzl, A. (2012). Systemic determinants of the information systems outsourcing decision: A comparative study of German and United States firms. *Journal of the Association for Information Systems*, *13*(6), 466–497. https://doi.org/10.7892/BORIS.43287

Dijkstra, T. K. (1981). *Latent variables in linear stochastic models: Reflections on ‘‘maximum likelihood’’ and ‘‘partial least squares’’ methods (Ph.D. thesis)*. Groningen University, Groningen, a second edition was published in 1985 by Sociometric Research Foundation.

Dijkstra, T. K. (2017). A perfect match between a model and a mode. In H. Latan & R. Noonan (Eds.), *Partial least squares path modeling: Basic concepts, methodological issues and applications* (pp. 55–80). Springer International Publishing.

Dijkstra, T. K., & Henseler, J. (2015a). Consistent and asymptotically normal PLS estimators for linear structural equations. *Computational Statistics & Data Analysis*, *81*, 10–23. https://doi.org/10.1016/j.csda.2014.07.008

Dijkstra, T. K., & Henseler, J. (2015b). Consistent partial least squares path modeling. *MIS Quarterly*, *39*(2), 297–316. https://doi.org/10.25300/MISQ/2015/39.2.02

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185. https://doi.org/10.1080/01621459.1987.10478410

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the bootstrap*. Chapman & Hall.

Esposito Vinzi, V., Trinchera, L., Squillacciotti, S., & Tenenhaus, M. (2008). REBUS-PLS: a response-based procedure for detecting unit segments in PLS path modelling. *Applied Stochastic Models in Business and Industry*, *24*(5), 439–458. https://doi.org/10.1002/asmb.728

Grace, J. B., & Bollen, K. A. (2008). Representing general theoretical concepts in structural equation models: The role of composite variables. *Environmental and Ecological Statistics*, *15*(2), 191–213. https://doi.org/10.1007/s10651-007-0047-7

Hahn, C., Johnson, M. D., Herrmann, A., & Huber, F. (2002). Capturing customer heterogeneity using a finite mixture PLS approach. *Schmalenbach Business Review*, *54*(3), 243–269. https://doi.org/10.1007/BF03396655

Hair, J. F., Sarstedt, M., Hopkins, L., & Kuppelwieser, V. G. (2014). Partial least squares structural equation modeling (PLS-SEM). *European Business Review*, *26*(2), 106–121. https://doi.org/10.1108/EBR-10-2013-0128

Hair, J. F., Sarstedt, M., Ringle, C. M., & Gudergan, S. P. (2018). *Advanced issues in partial least squares structural equation modeling*. Sage.

Hair, J. F., Sarstedt, M., Ringle, C. M., & Mena, J. A. (2012). An assessment of the use of partial least squares structural equation modeling in marketing research. *Journal of the Academy of Marketing Science*, *40*(3), 414–433. https://doi.org/10.1007/s11747-011-0261-6

Henseler, J. (2007). A new and simple approach to multi-group analysis in partial least squares path modeling. In H. Martens, T. Naes, & M. Martens (Eds.), *PLS’07 international symposium on PLS and related methods - causalities explored by indirect observation.*

Henseler, J. (2012). PLS-MGA: A non-parametric approach to partial least squares-based multi-group analysis. In W. A. Gaul, A. Geyer-Schulz, L. Schmidt-Thieme, & J. Kunze (Eds.), *Challenges at the interface of data analysis, computer science, and optimization* (pp. 495–501). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-24466-7_50

Henseler, J. (2017a). ADANCO 2.0.1. *Kleve, Germany: Composite Modeling*.

Henseler, J. (2017b). Bridging design and behavioral research with variance-based structural equation modeling. *Journal of Advertising*, *46*(1), 178–192. https://doi.org/10.1080/00913367.2017.1281780

Henseler, J. (2021). *Composite-based structural equation modeling: Analyzing latent and emergent variables*. The Guilford Press.

Henseler, J., & Fassott, G. (2010). Testing moderating effects in PLS path models: An illustration of available procedures. In V. Esposito Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of partial least squares* (Vol. 51, pp. 713–735). Springer. https://doi.org/10.1007/978-3-540-32827-8_31

Henseler, J., Ringle, C. M., & Sarstedt, M. (2016). Testing measurement invariance of composites using partial least squares. *International Marketing Review*, *33*(3), 405–431. https://doi.org/10.1108/imr-09-2014-0304

Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. In R. R. Sinkovics & P. N. Ghauri (Eds.), *Advances in International Marketing* (Vol. 20, pp. 277–319). Emerald Group Publishing Limited. https://doi.org/10.1108/S1474-7979(2009)0000020014

Henseler, J., & Schuberth, F. (2020). Using confirmatory composite analysis to assess emergent variables in business research. *Journal of Business Research*, *120*, 147–156. https://doi.org/10.1016/j.jbusres.2020.07.026

Hew, J.-J., Badaruddin, M. N. B. A., & Moorthy, M. K. (2017). Crafting a smartphone repurchase decision making process: Do brand attachment and gender matter? *Telematics and Informatics*, *34*(4), 34–56. https://doi.org/10.1016/j.tele.2016.12.009

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, *75*(4), 800–802. https://doi.org/10.2307/2336325

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, *75*(2), 383–386. https://doi.org/10.2307/2336190

Hsieh, Rai, A., & Keil (2008). Understanding digital inequality: Comparing continued use behavioral models of the socio-economically advantaged and disadvantaged. *MIS Quarterly*, *32*(1), 97. https://doi.org/10.2307/25148830

Hwang, H., Cho, G., Jung, K., Falk, C., Flake, J., Jin, M., & Lee, S.-H. (2020). An approach to structural equation modeling with both factors and components: Integrated generalized structured component analysis. *Psychological Methods*(forthcoming).

Hwang, H., & Takane, Y. (2004). Generalized structured component analysis. *Psychometrika*, *69*(1), 81–99. https://doi.org/10.1007/BF02295841

Hwang, H., Takane, Y., & Jung, K. (2017). Generalized structured component analysis with uniqueness terms for accommodating measurement error. *Frontiers in Psychology*, *8*(2137). https://doi.org/10.3389/fpsyg.2017.02137

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, *16*(1), 39–59. https://doi.org/10.1287/mksc.16.1.39

Keil, M., Tan, B. C. Y., Wei, K.-K., Saarinen, T., Tuunainen, V., & Wassenaar, A. (2000). A cross-cultural study on escalation of commitment behavior in software projects. *MIS Quarterly*, *24*(2), 299–325. https://doi.org/10.2307/3250940

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika*, *58*(3), 433–451. https://doi.org/10.1093/biomet/58.3.433

Klesel, M., Schuberth, F., Henseler, J., & Niehaves, B. (2019). A test for multigroup comparison in partial least squares path modeling. *Internet Research*, *29*(3), 464–477. https://doi.org/10.1108/IntR-11-2017-0418

Lamberti, G., Banet Aluja, T., & Sanchez, G. (2017). The Pathmox approach for PLS path modeling: Discovering which constructs differentiate segments. *Applied Stochastic Models in Business and Industry*, *33*(6), 674–689. https://doi.org/10.1002/asmb.2270

Lee, J. H., & Kim, J. (2014). Socio-demographic gaps in mobile use, causes, and consequences: A multi-group analysis of the mobile divide model. *Information, Communication & Society*, *17*(8), 917–936. https://doi.org/10.1080/1369118x.2013.860182

Matthews, L. (2017). Applying multigroup analysis in PLS-SEM: A step-by-step process. In H. Latan & R. Noonan (Eds.), *Partial least squares path modeling: Basic concepts, methodological issues and applications* (pp. 219–243). Springer International Publishing. https://doi.org/10.1007/978-3-319-64069-3_10

McDonald, R. P. (1996). Path analysis with composite variables. *Multivariate Behavioral Research*, *31*(2), 239–270. https://doi.org/10.1207/s15327906mbr3102_5

Müller, T., Schuberth, F., & Henseler, J. (2018). PLS path modeling – a confirmatory approach to study tourism technology and tourist behavior. *Journal of Hospitality and Tourism Technology*, *9*(3), 249–266. https://doi.org/10.1108/JHTT-09-2017-0106

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585. https://doi.org/10.1007/bf02296397

Nitzl, C. (2010). Eine anwenderorientierte Einführung in die Partial Least Square (PLS)-Methode. *Universität Hamburg, Institut Für Industrielles Management, Hamburg*.

Nitzl, C. (2016). Partial least squares structural equation modelling (PLS-SEM) in management accounting research: Critical analysis, advances, and future directions. *Journal of Accounting Literature*, *37*, 19–35. https://doi.org/10.2139/ssrn.2469802

Papagiannidis, S., Pantano, E., See-To, E. W., Dennis, C., & Bourlakis, M. (2017). To immerse or not? Experimenting with two virtual retail environments. *Information Technology & People*, *30*(1), 163–188. https://doi.org/10.1108/ITP-03-2015-0069

Qureshi, I., & Compeau, D. (2009). Assessing between-group differences in information systems research: A comparison of covariance- and component-based SEM. *MIS Quarterly*, *33*(1), 197–214. https://doi.org/10.2307/20650285

R Core Team. (2020). *R: A language and environment for statistical computing*. https://www.R-project.org/

Rademaker, M. E., & Schuberth, F. (2020). *cSEM: Composite-based Structural Equation Modeling* (Version: 0.2.0.9000). https://m-e-rademaker.github.io/cSEM/

Rademaker, M. E., Schuberth, F., & Dijkstra, T. K. (2019). Measurement error correlation within blocks of indicators in consistent partial least squares. *Internet Research*, *29*(3), 448–463. https://doi.org/10.1108/IntR-12-2017-0525

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30–45. https://doi.org/10.1037/met0000220

Rigdon, E. E. (2012). Rethinking partial least squares path modeling: In praise of simple methods. *Long Range Planning*, *45*(5-6), 341–358. https://doi.org/10.1016/j.lrp.2012.09.010

Rigdon, E. E., Sarstedt, M., & Ringle, C. M. (2017). On comparing results from CB-SEM and PLS-SEM: Five perspectives and five recommendations. *Marketing ZFP*, *39*(3), 4–16. https://doi.org/10.15358/0344-1369-2017-3-4

Ringle, C. M., Sarstedt, M., & Schlittgen, R. (2010). Finite mixture and genetic algorithm segmentation in partial least squares path modeling: Identification of multiple segments in complex path models. In A. Fink, B. Lausen, W. Seidel, & A. Ultsch (Eds.), *Advances in data analysis, data handling and business intelligence* (Vol. 14, pp. 167–176). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-01044-6_15

Ringle, C. M., Sarstedt, M., & Straub, D. W. (2012). Editor's comments: A critical look at the use of PLS-SEM in MIS Quarterly. *MIS Quarterly*, *36*(1), iii-xiv. https://doi.org/10.2307/41410402

Ringle, C. M., Wende, S., & Becker, J.-M. (2015). SmartPLS 3. www.smartpls.com

Sarstedt, M., Henseler, J., & Ringle, C. M. (2011). Multigroup analysis in partial least squares (PLS) path modeling: Alternative methods and empirical results. *Advances in International Marketing*, *22*, 195–218. https://doi.org/10.1108/s1474-7979(2011)0000022012

Sarstedt, M., Ringle, C. M., & Hair, J. F. (2017). Treating unobserved heterogeneity in PLS-SEM: A multi-method approach. In H. Latan & R. Noonan (Eds.), *Partial least squares path modeling: Basic concepts, methodological issues and applications* (pp. 197–217). Springer International Publishing.

Schuberth, F., Henseler, J., & Dijkstra, T. K. (2018a). Confirmatory composite analysis. *Frontiers in Psychology*, *9*(2541). https://doi.org/10.3389/fpsyg.2018.02541

Schuberth, F., Henseler, J., & Dijkstra, T. K. (2018b). Partial least squares path modeling using ordinal categorical indicators. *Quality & Quantity*, *52*(1), 9–35. https://doi.org/10.1007/s11135-016-0401-7

Sia, C. L., Lim, K. H., Leung, K., Lee, M. K., Huang, W. W., & Benbasat, I. (2009). Web strategies to promote internet shopping: Is cultural-customization needed? *MIS Quarterly*, *33*(3), 491–512. https://doi.org/10.2307/20650306

Srite, M., & Karahanna, E. (2006). The role of espoused national cultural values in technology acceptance. *MIS Quarterly*, *30*(3), 679–704. https://doi.org/10.2307/25148745

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–107. https://doi.org/10.1086/209528

Tenenhaus, M. (2008). Component-based structural equation modelling. *Total Quality Management & Business Excellence*, *19*(7-8), 871–886. https://doi.org/10.1080/14783360802159543

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. *Statistics and computing*. Springer.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540

Wedel, M., & Kamakura, W. A. (2000). *Market segmentation conceptual and methodological foundations* (Vol. 8). Springer US. https://doi.org/10.1007/978-1-4615-4651-1

Wold, H. (1975). Path models with latent variables: the NIPALS approach. In H. M. Blalock, A. Aganbegian, F. M. Borodkin, R. Boudon, & V. Capecchi (Eds.), *Quantitative sociology* (pp. 307–357). Academic Press. https://doi.org/10.1016/B978-0-12-103950-9.50017-4

Wolf, M., Beck, R., & Pahlke, I. (2012). Mindfully resisting the bandwagon: Reconceptualising IT innovation assimilation in highly turbulent environments. *Journal of Information Technology*, *27*(3), 213–235. https://doi.org/10.1057/jit.2012.13

Wood, M. (2005). Bootstrapped confidence intervals as an approach to statistical inference. *Organizational Research Methods*, *8*(4), 454–470. https://doi.org/10.1177/1094428105280059

About the Authors

Michael Klesel Michael Klesel is IT Consultant and is visiting Scholar at the University of Twente, The Netherlands.

His research interests include the individualization of information systems and structural equation modeling. He has published in various journals including *Internet Research* or *Communications of the Association of Information Systems (CAIS)* and in leading conferences including the *International Conference on Information Systems* (ICIS), the *European Conference on Information Systems* (ECIS) and the *American Conference on Information Systems* (AMCIS).

Florian Schuberth obtained his PhD in Econometrics in the Faculty of Business Management and Economics at the University of Würzburg, Germany. Currently, he is Assistant Professor in the Faculty of Engineering Technology at the University of Twente, the Netherlands. His main research interests are focused on SEM, in particular on composite-based estimators and their enhancement. His work has been published in various journals such as *Behaviormetrika*, *Information and Management*, *Internet Research*, *Journal of Business Research*, and *Quality & Quantity*.

Björn Niehaves is Full Professor and holds the Chair of Information Systems at the University of Siegen, Germany. He received a PhD Degree in Information Systems and a PhD Degree in Political Science from the University of Münster, Germany. Björn holds or held visiting positions at Harvard University (USA), the London School of Economics and Political Science (UK), Waseda University (Japan), Royal Institute of Technology (Sweden), Copenhagen Business School (Denmark), and Aalto University (Finland). He has published more than 200 research articles.

Jörg Henseler holds the Chair of Product-Market Relations in the Faculty of Engineering Technology at the University of Twente, the Netherlands, and he is Visiting Professor at NOVA Information Management School, Universidade Nova de Lisboa, Portugal. His broad-ranging research interests encompass empirical methods of Marketing and Design research as well as the management of design, products, services, and brands. He is co-inventor of consistent partial least squares (PLSc), the heterotrait-monotrait ratio of correlations (HTMT), and confirmatory composite analysis (CCA). He is a highly cited researcher according to Web of Science; his work has been published in *Computational Statistics and Data Analysis*, *European Journal of Information Systems*, *International Journal of Research in Marketing, Journal of the Academy of Marketing Science, Journal of Supply Chain Management, MIS Quarterly, Organizational Research Methods*, and *Structural Equation Modeling-A Multidisciplinary Journal*, among others. He chairs the Scientific Advisory Board of ADANCO, software for composite-based SEM (http://www.composite-modeling.com).