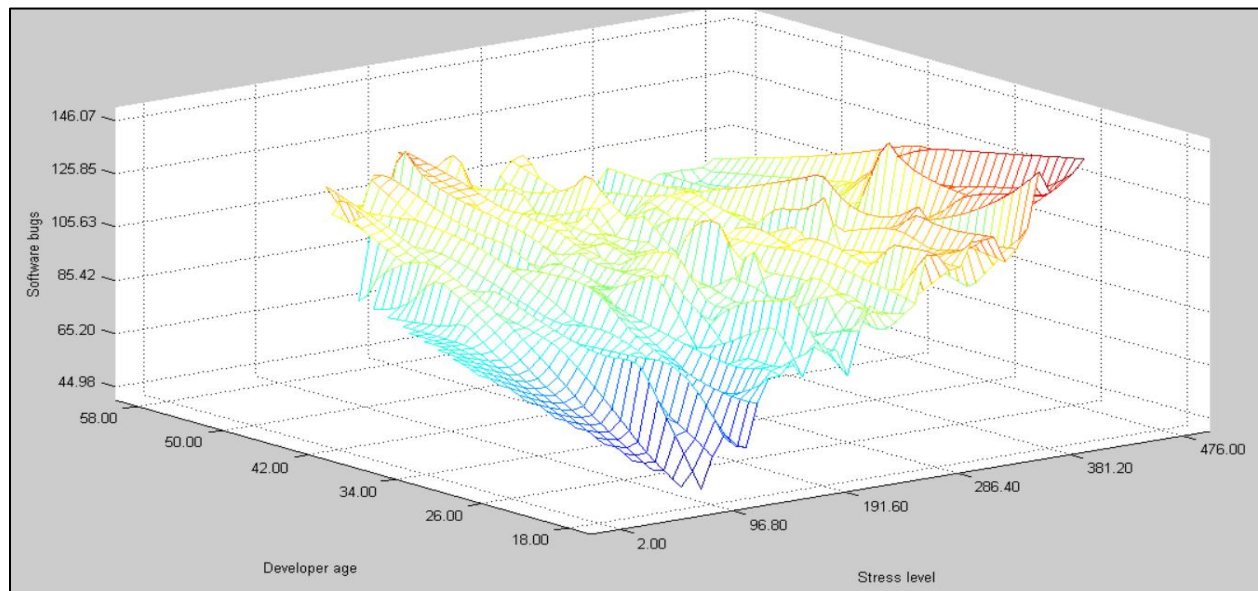


# Model-Driven Data Analytics: Applications with WarpPLS



**Ned Kock**

# Model-Driven Data Analytics: Applications with WarpPLS

December 2024

Ned Kock



*ScriptWarp Systems* <sup>TM</sup>

Laredo, Texas

USA

Model-Driven Data Analytics: Applications with WarpPLS

**Model-Driven Data Analytics: Applications with WarpPLS, Editions 1 – 2, November 2022  
– October 2023, Edition 3, December 2024, Copyright © by Ned Kock**

All rights reserved worldwide.

**For more information:**

ScriptWarp Systems  
P.O. Box 452428  
Laredo, Texas, 78045  
USA  
[www.scriptwarp.com](http://www.scriptwarp.com)

## Table of contents

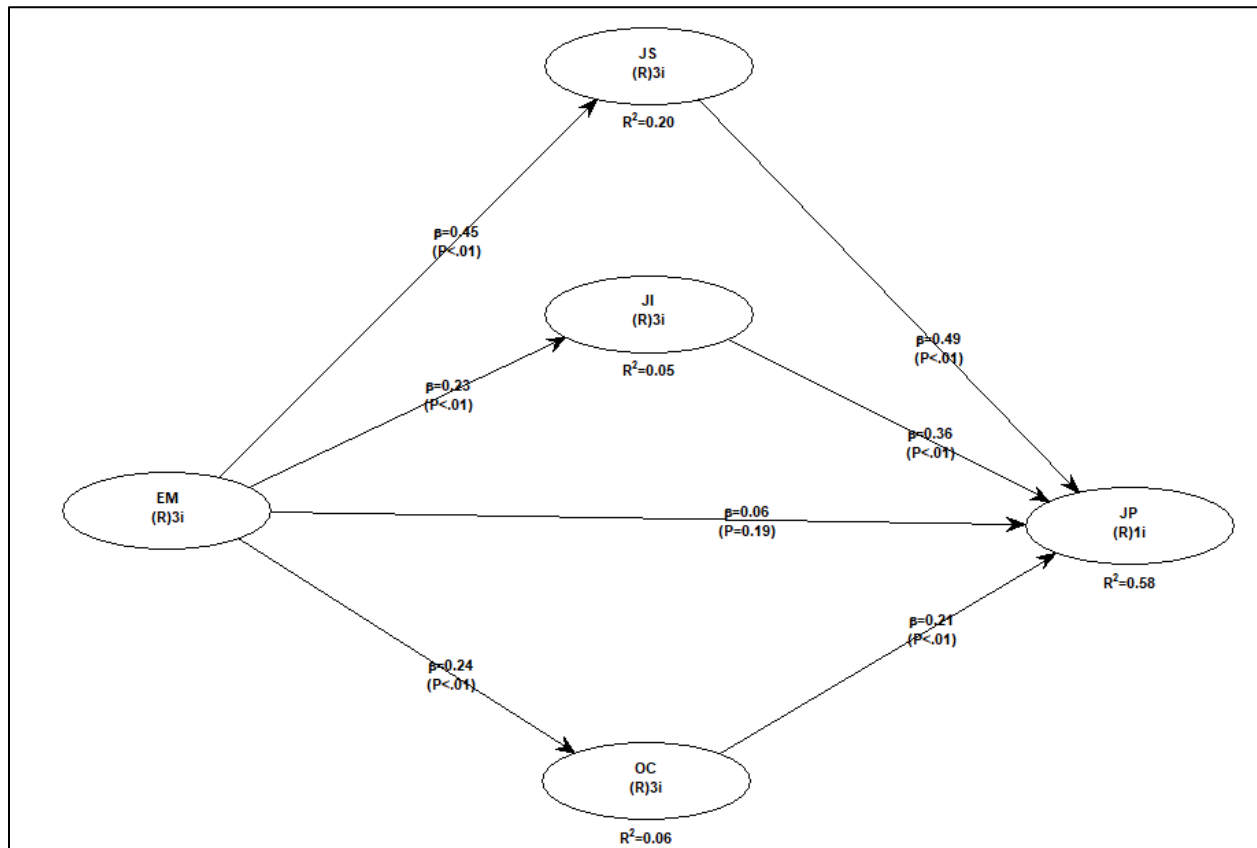
<b>PART 1: INTRODUCTION .....</b>	<b>5</b>
1.1. MODEL-DRIVEN DATA ANALYTICS .....	6
1.2. CREATING A MODEL WITH WARPPLS .....	8
1.3. CHOOSING ANALYSIS SETTINGS IN WARPPLS.....	11
1.4. INSPECTING GRAPHS IN WARPPLS.....	13
1.5. ASSESSING COLLINEARITY, VALIDITY, AND RELIABILITY .....	14
1.6. OTHER INTRODUCTORY REMARKS .....	16
<b>PART 2: APPLICATIONS.....</b>	<b>17</b>
2.1. INCREASING SAT SCORES IN A U.S. STATE .....	18
2.2. IMPROVING SATISFACTION WITH CAR PART DELIVERY .....	25
2.3. IMPROVING JOB PERFORMANCE THROUGH EMPATHETIC MANAGEMENT.....	34
2.4. IMPROVING SOFTWARE DEVELOPMENT BY EMPLOYING OLDER CODERS.....	50
2.5. DECIDING ON A MALL LOCATION TO ESTABLISH A HOT DOG KIOSK.....	58
2.6. ORGANIZING GROCERY STORE ITEMS TO INCREASE SALES .....	66
<b>PART 3: CONCLUDING REMARKS .....</b>	<b>74</b>
3.1. APPLIED MODEL-DRIVEN DATA ANALYTICS .....	75
3.2. USE OF SIMULATED DATA .....	76
<b>GLOSSARY .....</b>	<b>77</b>
<b>ACKNOWLEDGEMENTS .....</b>	<b>83</b>
<b>BIBLIOGRAPHY.....</b>	<b>84</b>

# **Part 1: Introduction**

## 1.1. Model-driven data analytics

The technique of model-driven data analytics (MDDA) involves the creation of a path model expressing an applied theory, and testing the model using path analysis with latent variables. The latter, path analysis with latent variables, is generally known as structural equation modeling (SEM). Figure 1.1 displays an example of a model with results.

Figure 1.1. Example of model with results



The path model expressing an applied theory is typically made up of several latent variables that are causally linked. The latent variables may be measured through one or multiple indicators, which are usually available as columns of numeric data on a table-like dataset. Multiple indicators often help reduce the impact of measurement error on the various model parameters that are estimated.

The applied theory, expressed through the model, usually comes from organizational stakeholders. This applied theory typically does not come from the data analysts, because the main expertise of the analysts is usually on data analysis techniques, not on the applied domain for which data is being analyzed. Normally the applied domain consists of one or more organizations, and the key stakeholders are the employees and managers of those organizations. It is via discussions and interviews with those stakeholders that the analysts obtain the necessary applied knowledge to build a model.

MDDA emerged from the work of a special category of users of the software WarpPLS – data analysis consultants, who regularly work with organizations to provide data-driven

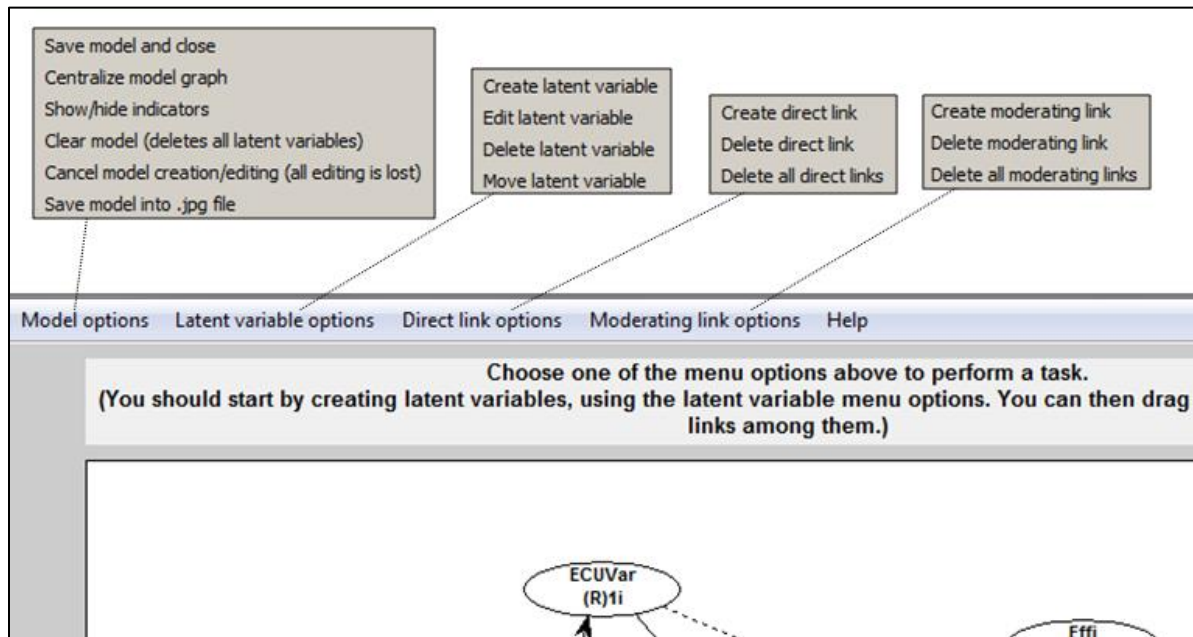
## Model-Driven Data Analytics: Applications with WarpPLS

recommendations. While MDDA can be implemented through a variety of software tools, it has found wide adoption among WarpPLS users, because of the many powerful features of this software that can be used in this context. Moreover, in WarpPLS all analyses are model-driven, which makes this software much more user-friendly than other software tools that rely on extensive scripting to conduct analyses.

## 1.2. Creating a model with WarpPLS

The window used to create a model is shown in Figure 1.2. A model can be edited if it has been created and saved before as part of a WarpPLS project (these projects include the data used in the analysis, as well as the analysis results). While editing or creating a model you can choose from a number of menu options related to overall model functions, latent variable functions, direct link functions, and moderating link functions. As with other windows in WarpPLS, there is a help menu option that provides access to its manual, displayed as a PDF file. The help menu option also provides links to Web resources.

Figure 1.2. Options to create a model



A guiding text box is shown at the top of the model editing and creation window. The content of this guiding text box changes depending on the menu option you choose, guiding you through the sub-steps related to each option. For example, if you choose the option “Create latent variable”, the guiding text box will change color, and tell you to select a location for the latent variable on the model graph.

**Direct links are displayed as full arrows** in the model graph, and **moderating links as dashed arrows**. Each latent variable is displayed in the model graph within an oval symbol, where its name is shown above a combination of alphanumerical characters with this general format: “(F)16i”. The “F” refers to the measurement model; where “F” means formative, and “R” reflective. The “16i” reflects the number of indicators of the latent variable, which in this case is 16.

A **reflective** latent variable is one in which all of the indicators are expected to be highly correlated with the latent variable, and also highly correlated with one another. For example, the answers to certain question-statements by a group of people, measured on a 1 to 7 scale (1=strongly disagree; 7 strongly agree) and answered after a meal, are expected to be highly correlated with the latent variable “satisfaction with a meal”. The question-statements are: “I am



satisfied with this meal”, and “After this meal, I feel full”. Therefore, the latent variable “satisfaction with a meal”, can be said to be reflectively measured through these indicators.

A **formative** latent variable is one in which the indicators are expected to measure certain attributes of the latent variable, but the indicators are not expected to be highly correlated with the latent variable, because they (i.e., the indicators) are not expected to be correlated with one another. For example, let us assume that the latent variable “Satisf” (“satisfaction with a meal”) is measured using the two following question-statements: “I am satisfied with the main course” and “I am satisfied with the dessert”. Both main course and dessert make up the meal (i.e., they are part of the same meal) but their satisfaction indicators are not expected to be highly correlated with each other. Some people may like the main course, and not like the dessert, or vice-versa.

**Save model and close.** This option saves the model within the project, and closes the model editing and creation window. This option does not, however, save the project file. That is, the project file has to be saved for a model to be saved as part of it. This allows you to open a project file, change its model, run a SEM analysis, and discard all that you have done, if you wish to do so, reverting back to the previous project file.

**Centralize model graph.** This option centralizes the model graph, and is useful when you are building complex models and, in the process of doing so, end up making the model visually unbalanced. For example, you may move variables around so that they are all accidentally concentrated on the left part of the screen. This option corrects that by automatically redrawing all symbols in the model graph so that the center of the model graph coincides with the center of the model screen.

**Show/hide indicators.** This option shows or hides the list of indicators for each latent variable. The indicators are shown on a vertical list next to each latent variable, and without the little boxes that are usually shown in other SEM software. This display option is used to give the model graph a cleaner look. It also has the advantage that it saves space in the model graph for latent variables. Normally you will want to keep the indicators hidden, except when you are checking whether the right indicators were selected for the right latent variables. That is, normally you will show the indicators to perform a check, and then hide them during most of the model building process.

**Clear model (deletes all latent variables).** This option deletes all latent variables, essentially “clearing” the model. Given that choosing this option by mistake can potentially cause some serious loss of work (not to mention some major user aggravation), the software shows a dialog box asking you to confirm that you want to clear the model before it goes ahead and deletes all latent variables. Even if you choose this option by mistake, and confirm your choice also by mistake (a double mistake), you can still undo it by choosing the option “Cancel model creation/editing (all editing is lost)” immediately after clearing the model.

**Cancel model creation/editing (all editing is lost).** This option cancels the model creation or editing, essentially undoing all of the model changes you have made.

**Save model into image file.** This option allows you to save the model graph into an image file (e.g., a .jpg or .png file). You will be asked to select the file name and the folder where the file will be saved. After saved, this file can then be viewed and edited with standard picture viewers, as well as included as a picture into reports in other files (e.g., a Word file). Users can also generate model graph files by copying the model screen into a picture-editing application (e.g., Paint), cropping it to leave out unnecessary or unneeded areas, saving it into a picture file (e.g., .jpg or .png), and then importing that file into reports.

**Create latent variable.** This option allows you to create a latent variable, and is discussed in more detail below. Once a latent variable is created it can be dragged and dropped anywhere within the window that contains the model.

**Edit latent variable.** This option allows you to edit a latent variable that has already been created, and thus that is visible on the model graph.

**Delete latent variable.** This option allows you to delete an existing latent variable. All links associated with the latent variable are also deleted.

**Move latent variable.** This option is rarely used since, once a latent variable is created, it can be easily dragged and dropped with the pointing device (e.g., mouse) anywhere within the window that contains the model. This option is a carryover from a previous version, maintained for consistency and for those users who still want to use it. It allows a user to move a latent variable across the model by first clicking on the variable and then on the destination position.

**Create direct link.** This option allows you to create a direct link between one latent variable and another. The arrow representing the link points from the predictor latent variable to the criterion latent variable. Direct links are usually associated with direct cause-effect hypotheses; testing a direct link's strength (through the calculation of a path coefficient) and statistical significance (through the calculation of a P value) is equivalent to testing a direct cause-effect hypothesis.

**Delete direct link.** This option allows you to delete an existing direct link. You will click on the direct link that you want to delete, after which the link will be deleted.

**Delete all direct links.** This option deletes all direct links. Given that choosing this option by mistake is a possibility, the software shows a dialog box asking you to confirm that you want to execute it before it proceeds. Even if you choose this option by mistake, and confirm your choice also by mistake, you can still undo it by choosing the option "Cancel model creation/editing (all editing is lost)".

**Create moderating link.** This option allows you to create a link between a latent variable and a direct link. With some exceptions, both formative and reflective latent variables can be part of moderating links. Moderating links are typically associated with moderating cause-effect hypotheses, or interaction effect hypotheses. Testing a moderating link's strength (through the calculation of a path coefficient) and statistical significance (through the calculation of a P value) is equivalent to testing a moderating cause-effect or interaction effect hypothesis. **Moderating links should be used with moderation (no pun intended)**, because they may introduce multicollinearity into the model, and also because they tend to add nonlinearity to the model. By introducing multicollinearity into the model, they may make some model parameter estimates unstable and biased.

**Delete moderating link.** This option allows you to delete an existing moderating link. You will click on the moderating link that you want to delete, after which the link will be deleted.

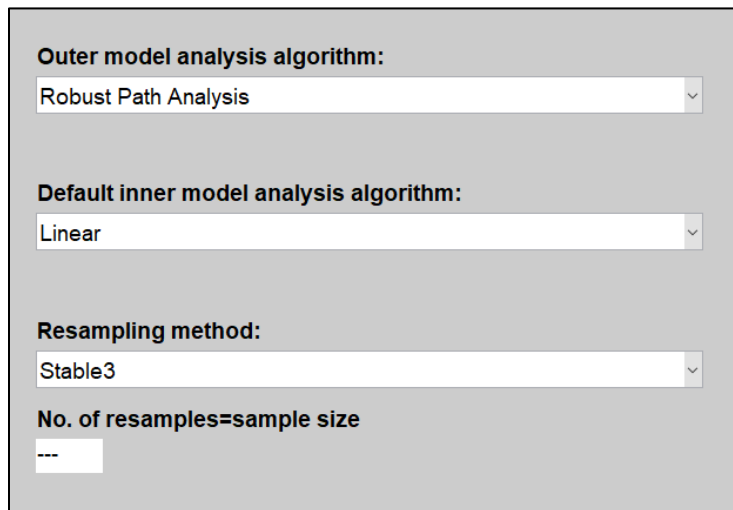
**Delete all moderating links.** This option deletes all moderating links. Given that choosing this option by mistake is a possibility, the software shows a dialog box asking you to confirm that you want to execute it before it proceeds. Even if you choose this option by mistake, and confirm your choice also by mistake, you can still undo it by choosing the option "Cancel model creation/editing (all editing is lost)".

After you create a model and choose the option "Save model and close" **a wait bar will be displayed on the screen telling you that the SEM model structure is being created.** This is an important sub-step where a number of checks are made.

### 1.3. Choosing analysis settings in WarpPLS

The options shown in Figure 1.3 are common in MDDA analyses that employ latent variables with single indicators, which is a common feature of these analyses. They are set through the “**View or change general settings**” menu option, which allows users to set the outer model analysis algorithm, default inner model analysis algorithm, resampling method, and number of resamples. Through these sub-options, users can set outer and default inner model algorithms separately. Users are also allowed to set inner model algorithms for individual paths, but through a different settings option. If users choose not to set inner model algorithms for individual paths, their choice of default inner model algorithm is automatically used for all paths.

Figure 1.3. Choosing analysis settings in WarpPLS



The screenshot shows a settings dialog box with the following options:

- Outer model analysis algorithm:** Robust Path Analysis
- Default inner model analysis algorithm:** Linear
- Resampling method:** Stable3
- No. of resamples=sample size:** ---

In a SEM analysis implementing MDDA, the **inner model** is the part of the model that describes the relationships among the latent variables that make up the model. In this sense, the path coefficients are inner model parameter estimates. These path coefficients are obtained by regressing a latent variable on the latent variables that are assumed to predict it; i.e., that point at it. The **outer model** is the part of the model that describes the relationships among the latent variables that make up the model and their indicators. In this sense, the weights and loadings are outer model parameter estimates. The weights are obtained by regressing a latent variable on its indicators; the loadings by regressing the indicators on their latent variable.

The **Robust Path Analysis** algorithm is a simplified algorithm in which latent variable scores are calculated by averaging the scores of the indicators associated with the latent variables. This algorithm is called “robust” path analysis, because a standard path analysis (where all latent variables are measured through single indicators) can be conducted through it, and the P values can be calculated through the nonparametric resampling or stable methods implemented through the software. If all latent variables are measured with single indicators, the Robust Path Analysis algorithm will yield latent variable scores and various parameters that are identical to those generated through the other algorithms, but with greater computational efficiency.

The **Linear** algorithm does not perform any warping of relationships; that is, it does not model the relationships as nonlinear. Other default inner model analysis algorithms can be employed to do that. For example, the Warp2 algorithm tries to identify U-curve relationships among linked

latent variables, and, if those relationships exist, the algorithm transforms (or “warps”) the scores of the predictor latent variables so as to better reflect the U-curve relationships in the estimated path coefficients in the model.

The **Stable3** method is the default resampling method of the software. Several Monte Carlo experiments show that the **Stable2** and **Stable3** methods yield estimates of the actual standard errors that are consistent with those obtained via bootstrapping, in many cases yielding more precise estimates of the actual standard errors. These standard errors are then used to produce P values, which tell the analyst whether the corresponding coefficients (e.g., path coefficients) refer to effects that appear to be “real” (i.e., not due to chance). Typically P values equal to or lower than 0.05 will refer to effects are “real” in this sense. **The more accurate of the two methods seems to be the Stable3 method**, which is why it is the default method.

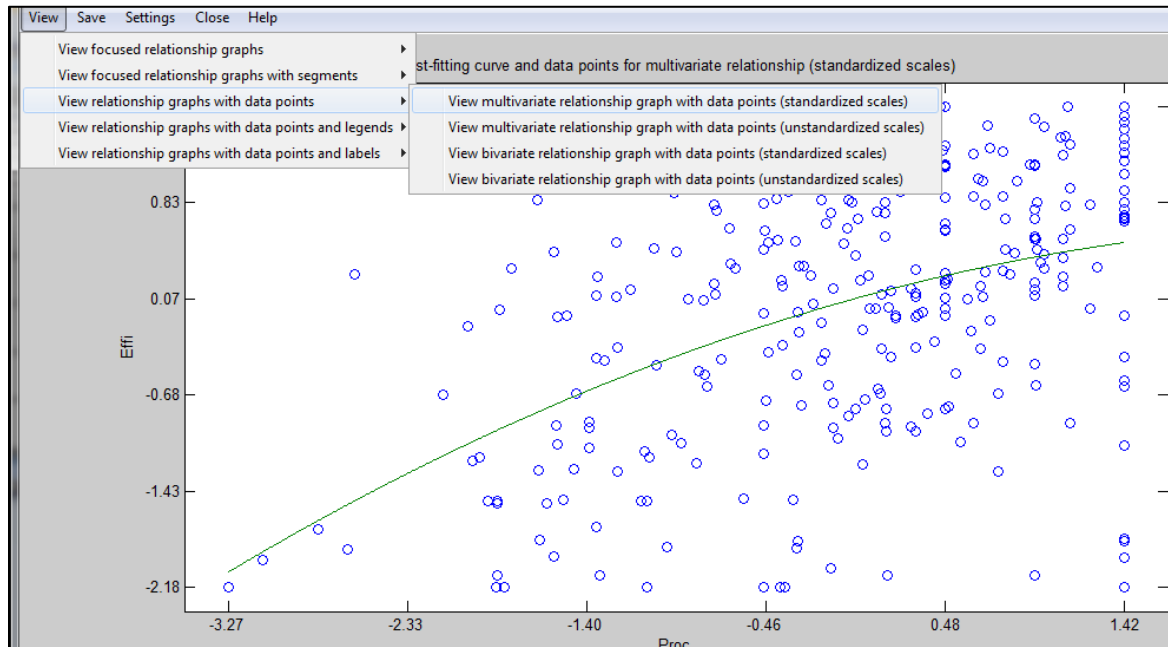
## 1.4. Inspecting graphs in WarpPLS

Choosing the menu option “**View/plot linear and nonlinear relationships among latent variables**” causes the software to show a table with the types of relationships, warped or linear, between latent variables that are linked in the model (see Figure 1.4.1). The term “warped” is used for relationships that are clearly nonlinear, and the term “linear” for linear or quasi-linear relationships. Quasi-linear relationships are slightly nonlinear relationships, which look linear upon visual inspection on plots of the regression curves that best approximate the relationships.

Figure 1.4.1. Linear and nonlinear (“warped”) relationships among latent variables window

Click on a "Linear" or "Warped" relationship cell to view plot					
	ECUVar	Proc	Effi	Effe	Effi*Proc
ECUVar					
Proc	Warped				
Effi	Warped	Warped			
Effe		Warped			Warped
Effi*Proc					

Figure 1.4.2. Graph options for direct effects including one with points and best-fitting curve



Several graphs (a.k.a. plots) for direct effects can be viewed by clicking on a cell containing a relationship type description. These cells are the same as those that contain path coefficients, in the path coefficients table that is available from the software under the option “View path coefficients and P values”; the path coefficients are also shown on the model graph with results, but with reduced two-decimals precision. Among the options available are graphs showing the points as well as the curves that best approximate the relationships (see Figure 1.4.2). Also graphs with both standardized and unstandardized scales are available.

## 1.5. Assessing collinearity, validity, and reliability

**Assessing collinearity.** Variance inflation factors (VIFs) are measures of the degree of collinearity (or multicollinearity) among variables, including both indicators and latent variables. With latent variables, collinearity can take two main forms: vertical and lateral collinearity. Vertical, or classic, collinearity is predictor-predictor latent variable collinearity in individual latent variable blocks. (A latent variable block is a set of latent variables including those pointing at a criterion latent variable, plus that criterion latent variable.)

Lateral collinearity is a term that refers to predictor-criterion latent variable collinearity; a type of collinearity that can lead to particularly misleading results. Full collinearity VIFs allow for the simultaneous assessment of both vertical and lateral collinearity in a SEM model. They can also be used for common method bias and discriminant validity assessment.

Full collinearity VIFs of 3.3 or lower suggest the existence of no multicollinearity in the model. A more relaxed threshold would be 5. This means that all of the latent variables in the model measure different “things” (e.g., mental concepts, or ideas), which is an important precondition for a valid analysis. Full collinearity VIFs of 10 or higher suggest the existence of multicollinearity in the model. Multicollinearity at this level, with full collinearity VIFs of 10 or higher, tends to distort coefficients of association, such as path coefficients.

**Assessing validity and reliability.** A **reflective** latent variable is one in which all of the indicators are expected to be highly correlated with the latent variable, and also highly correlated with one another. For example, the answers to certain question-statements by a group of people, measured on a 1 to 7 scale (1=strongly disagree; 7 strongly agree) and answered after a meal, are expected to be highly correlated with the latent variable “satisfaction with a meal”. The question-statements are: “I am satisfied with this meal”, and “After this meal, I feel full”. Therefore, the latent variable “satisfaction with a meal”, can be said to be reflectively measured through these indicators.

A **formative** latent variable is one in which the indicators are expected to measure certain attributes of the latent variable, but the indicators are not expected to be highly correlated with the latent variable, because they (i.e., the indicators) are not expected to be correlated with one another. For example, let us assume that the latent variable “Satisf” (“satisfaction with a meal”) is measured using the two following question-statements: “I am satisfied with the main course” and “I am satisfied with the dessert”. Both main course and dessert make up the meal (i.e., they are part of the same meal) but their satisfaction indicators are not expected to be highly correlated with each other. Some people may like the main course, and not like the dessert, or vice-versa.

The assessment of **convergent validity, discriminant validity, and reliability**, as discussed in the next few paragraphs, **applies to reflective measurement** of latent variables (which is much more common than formative measurement). Reflective latent variable indicators that do not satisfy the criteria discussed below may be considered for removal, or, in some cases, re-allocation to other latent variables. Examples of the use of these criteria are discussed later, in the context of one or more applications.

**Convergent validity** is a measure of the quality of a measurement instrument; the instrument itself is typically a set of question-statements (i.e., a questionnaire). A measurement instrument has good convergent validity if the question-statements (or other measures) associated with each latent variable are understood by the respondents in the same way as they were intended by the designers of the question-statements. Two criteria are recommended as the basis for concluding

that a measurement model has acceptable convergent validity: that the **P values associated with the loadings be equal to or lower than 0.05**; and that the **loadings be equal to or greater than 0.5**.

**Discriminant validity** is also a measure of the quality of a measurement instrument. A measurement instrument has good discriminant validity if the question-statements (or other measures) associated with each latent variable are not confused by the respondents, in terms of their meaning, with the question-statements associated with other latent variables. The following criterion is recommended for discriminant validity assessment: **for each latent variable, the square root of the average variance extracted (AVE) should be higher than any of the correlations involving that latent variable**.

**Reliability** is yet another measure of the quality of a measurement instrument; the instrument itself is typically a set of question-statements. A measurement instrument has good reliability if the question-statements (or other measures) associated with each latent variable are understood in the same way by different respondents. The following criterion is suggested in the assessment of the reliability of a measurement instrument: **either the composite reliability or the Cronbach's alpha coefficient should be equal to or greater than 0.6**.

For **formative measurement, different criteria are used**, notably the following two. It is recommended that **weights with P values that are equal to or lower than 0.05** be considered valid items in a formative latent variable measurement item subset. Formative latent variable indicators whose weights do not satisfy this criterion may be considered for removal.

In addition to P values, **VIFs are provided** for the indicators of all latent variables, including moderating latent variables. These can be used for indicator redundancy assessment. In reflective latent variables indicators are expected to be redundant. This is not the case with formative latent variables. In formative latent variables indicators are expected to measure *different* facets of the same construct, which means that they should *not* be redundant. Therefore, here the **VIF threshold of 3.3 is recommended**. Formative latent variable indicators whose weights do not satisfy this criterion may be considered for removal.

## 1.6. Other introductory remarks

The applications presented in this document show how MDDA can be employed in a variety of different contexts, where typically data is collected from organizations with the goal of answering questions that ultimately affect the ability of the organizations to grow their sales and profits.

As you will see, **the applications follow a similar set of steps**, such as: create model, choose general settings, assess collinearity, assess validity and reliability, inspect path coefficients, inspect graphs, and provide advice.

There is **repetition across applications**, of both steps and the text that describes them. Our experience is that this repetition helps with the internalization of complex concepts and techniques, while at the same time making each application section fairly self-contained. That is, one can jump from one application to another without having to review all of them in sequence.

Revised text and other materials from previously published documents by the author have been used in the development of this book. Some of the data discussed here have been compiled based on publicly available sources, some have been created via Monte Carlo simulations based on empirical studies, and some have been produced as a mix of both approaches.

For ethical reasons, and to **protect individual privacy**, all of the individual-level data have been created via Monte Carlo simulations, based on empirical studies – to mimic what happened with real data.

A **glossary** is available at the end of this document. This glossary includes terms that go beyond those used in the applications discussed in this document. We are including this extended set of terms here because some readers may want to go beyond the features discussed in the applications, and explore other more advanced features that refer to some of the terms in this extended set.



## **Part 2: Applications**

## 2.1. Increasing SAT scores in a U.S. state

Exhibit 2.1 displays the scenario, question, and variables related to the sample dataset used to illustrate how MDDA can be used to address the need for increasing SAT scores in a U.S. state. The acronym SAT originally referred to the term “Scholastic Aptitude Test”; a standardized test widely used for college admissions in the U.S.

### Exhibit 2.1. Scenario, question, and variables

#### Scenario

- The Department of Education of a state in the USA believes that SAT scores are strongly influenced by two predictor variables.
- Data is collected from a number of school districts in the state, for a given year, and a data analysis is commissioned.

#### Question

- What is the order of importance of the predictors with respect to SAT scores?

#### Variables

- TchExp: Average years of teaching experience by the school district teachers.
- Susp: Number of suspensions in the district due to student behavioral problems.
- SAT: Average SAT score in school district.

The problem to be solved is a common one in MDDA applications, namely to assess the order of importance of likely predictors of a main criterion variable. In this case, the predictors are TchExp (average years of teaching experience by the school district teachers) and Susp (number of suspensions in the district due to student behavioral problems). The main criterion variable is SAT (average SAT score in school district).

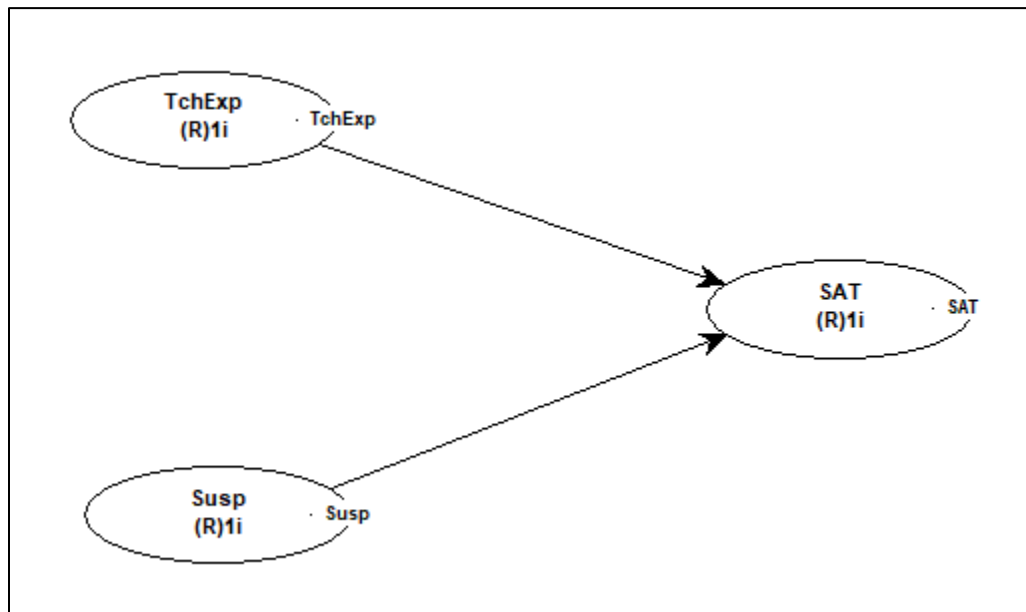
The main client of this analysis was the Department of Education of a state in the U.S. This organization wanted to know what they could do to increase SAT scores in the school districts under its jurisdiction. The expectation that the two predictors (TchExp and Susp) influenced SAT scores came from interviews with various stakeholders in the Department of Education and the school districts.

The importance of this analysis came from the costs associated with changing the predictors with the goal of increasing average SAT scores in the school districts. Changing the average years of teaching experience by the school districts’ teachers, presumably by increasing them, would probably require financial incentives (e.g., higher pay, better benefits). Changing the number of suspensions in the districts due to student behavioral problems, presumably by decreasing them, would probably require additional labor resources (e.g., additional security workers, more counseling personnel).

### 2.1.1. Create the model

Figure 2.1.1 shows the model that was built to serve as the basis for our analysis. It contains two predictor latent variables, named TchExp and Susp, pointing at one criterion latent variable, named SAT. The latent variables have only one indicator each, which essentially means that they are assumed to be measured through their single indicators without error.

Figure 2.1.1. Create the model

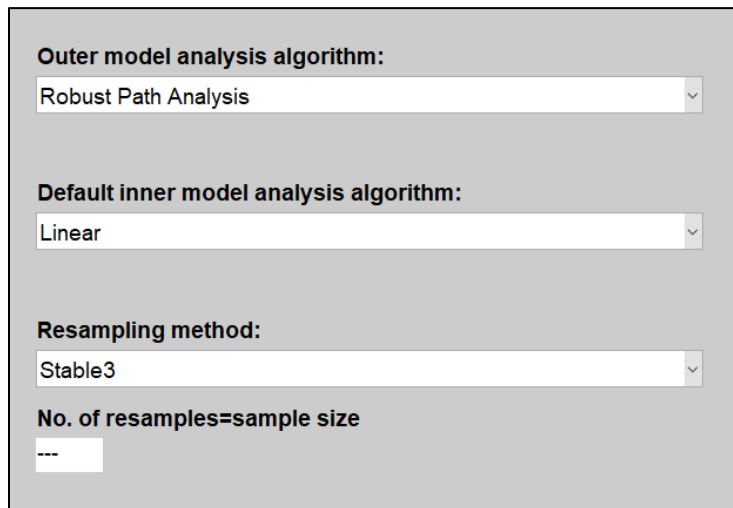


As you can see, the latent variables have the same names as their indicators. This is not a requirement. The names could have been different. In fact, they will typically be different for latent variables that are measured through multiple indicators. They will also be different if the indicators' names are longer than 8 characters, which is the maximum allowed for latent variables names. This limitation is to give model graphs a cleaner look.

## 2.1.2. Choose general settings

The options shown in Figure 2.1.2 were the ones chosen for this analysis. They are common in analyses that employ latent variables that are all measured through single indicators, as is the case in our model. The options can be selected through the “**View or change general settings**” menu option.

Figure 2.1.2. Choose general settings



The screenshot shows a settings dialog box with the following options:

- Outer model analysis algorithm:** Robust Path Analysis
- Default inner model analysis algorithm:** Linear
- Resampling method:** Stable3
- No. of resamples=sample size:** ---

The **Robust Path Analysis** outer model analysis algorithm is a simplified algorithm with very good computational efficiency. The **Linear** default inner model analysis algorithm does not perform any warping of relationships; that is, it does not model the relationships as nonlinear. This helps with the interpretation of the results. The **Stable3** method is the default resampling method of the software, because of its high accuracy and robustness to deviations of normality. An advantage of this method is that it does not assume that the data is normally distributed, which is often the case with empirical data. That is, empirical data is typically *not* normally distributed, even though many data analysis techniques assume that it is.

### 2.1.3. Assess collinearity

As soon as the SEM analysis is completed, the software shows the results in graphical format on a window, which also contains a number of menu options that allow you to view and save more detailed results. One of these menu options is the “**View latent variable coefficients**”, which shows the results in the table in Figure 2.1.3. The last row of the table in the figure shows the **full collinearity VIFs** for all latent variables in the model.

Figure 2.1.3. Assess collinearity

	TchExp	Susp	SAT
R-squared			0.059
Adj. R-squared			0.029
Composite reliab.	1.000	1.000	1.000
Cronbach's alpha	1.000	1.000	1.000
Avg. var. extrac.	1.000	1.000	1.000
Full collin. VIF	1.099	1.049	1.062

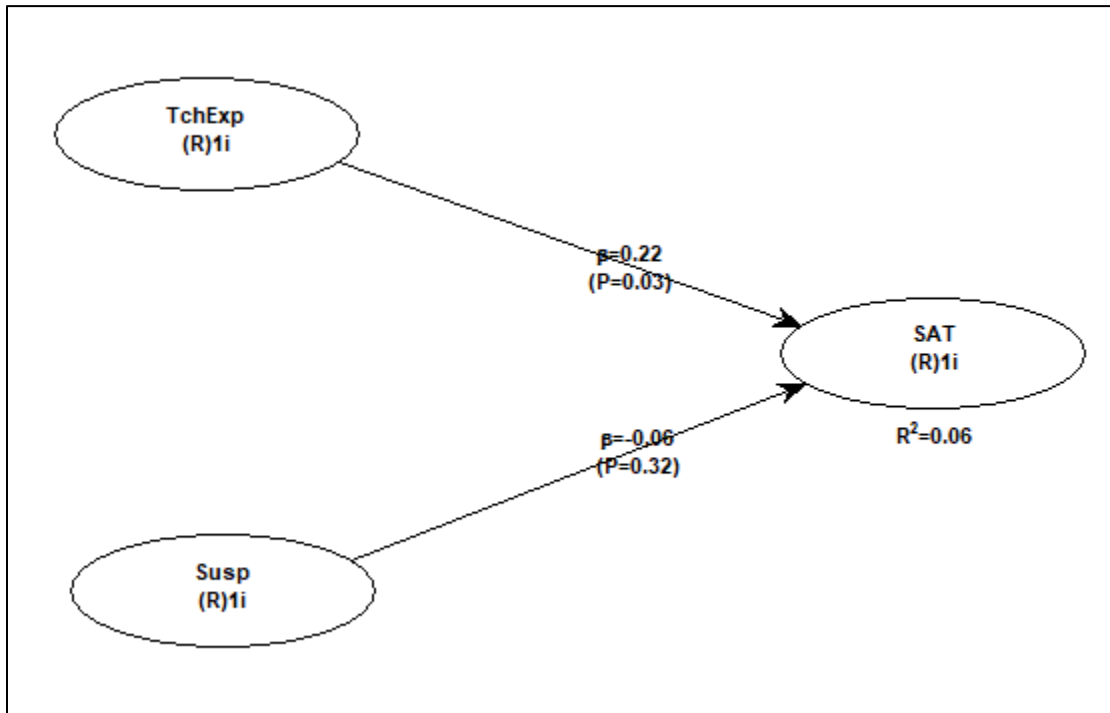
**Full collinearity VIFs of 3.3 or lower suggest** the existence of **no multicollinearity** in the model. A more **relaxed threshold would be 5**. This means that all of the latent variables in the model measure different things, which is an important precondition for a valid analysis. **Full collinearity VIFs of 10 or higher suggest** the existence of **multicollinearity** in the model. Multicollinearity at this level, with full collinearity VIFs of 10 or higher, tends to distort coefficients of association, such as path coefficients.

As we can see, the highest full collinearity VIF in the model is 1.099, well below the conservative threshold of 3.3, which allows us to conclude that all of the latent variables in the model measure different things. That is, the latent variables in the model measure constructs that appear to be conceptually different from one another.

### 2.1.4. Inspect path coefficients

As soon as the SEM analysis is completed, the software shows some of the main results in graphical format on a window. This graphical representation is shown in Figure 2.1.4. The graph with the results shows path coefficients, respective P values, and R-squared coefficients. Users can also show or hide indicators weights, loadings and names.

Figure 2.1.4. Inspect path coefficients



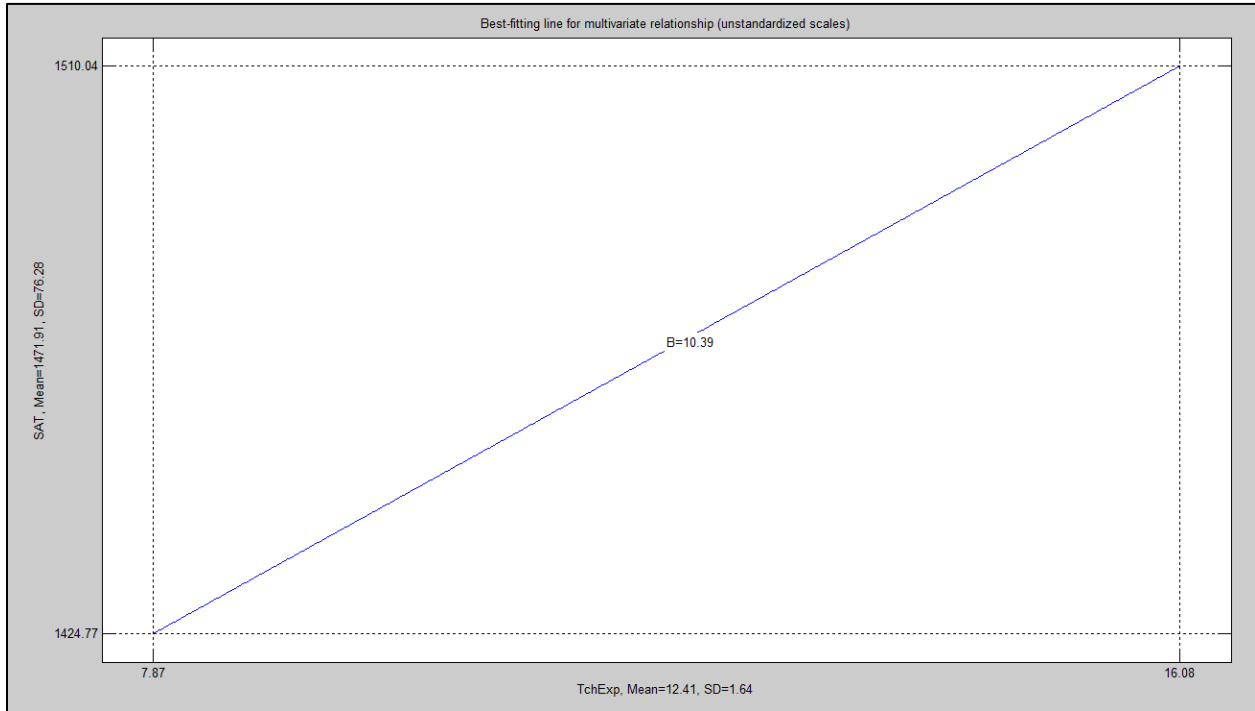
The path coefficients are noted as beta coefficients. “Beta coefficient” is another term often used to refer to path coefficients in SEM analyses; this term is commonly used in multiple regression analyses as well. The P values are displayed below the path coefficients, within parentheses. The R-squared coefficients are shown below each endogenous latent variable (i.e., a latent variable that is hypothesized to be affected by one or more other latent variables), and reflect the percentage of the variance in the latent variable that is explained by the latent variables that are hypothesized to affect it.

Path coefficients associated with P values equal to or lower than 0.05 are deemed to refer to real effects, as opposed to effects that are to be interpreted as “zero”. The path coefficient for the link TchExp > SAT is positive and associated with a P value of 0.03, therefore it refers to a real effect that is positive. This path coefficient is 0.22, meaning that each standard deviation increase in TchExp is associated with a 0.22 standard deviation variation in SAT. The path coefficient for the link Susp > SAT is associated with a P value of 0.32, therefore it refers to an effect that is to be interpreted as “zero” – or no effect.

### 2.1.5. Inspect graphs

As soon as the SEM analysis is completed, the software shows the results in graphical format on a window, where the menu option menu option “**View/plot linear and nonlinear relationships among latent variables**” becomes available. One of the graphs that is particularly useful is available from the option “**View focused multivariate relationship graph with segments (unstandardized scales)**”. This graph representation is shown in Figure 2.1.5.

Figure 2.1.5. Inspect graphs



Unlike a path coefficient, which is standardized, the graph shows the corresponding unstandardized regression coefficient. This type of coefficient tends to be more telling to stakeholders. In this case, the unstandardized regression coefficient for the link TchExp > SAT is 10.39. The meaning of this is that, for each increase of 1 year in the variable TchExp (average years of teaching experience by the school district teachers), there is a corresponding increase in 10.39 points in the variable SAT (average SAT score in school district).

### **2.1.6. Provide advice**

One of the goals of the analysis was to answer the question: What is the order of importance of the predictors with respect to SAT scores? The inspection of the path coefficients suggests that TchExp (average years of teaching experience by the school district teachers) is the most important predictor of the two. Moreover, the inspection of the path coefficients suggests that Susp (number of suspensions in the district due to student behavioral problems) has no effect on SAT scores.

Given this, the Department of Education of the state in the U.S. from which the data was obtained should arguably be advised to try to increase the average years of teaching experience by the school district teachers, so that SAT scores would increase. This could be done by providing financial incentives, such as higher pay and better benefits, to retain teachers for as long as possible.

Also, the Department of Education should arguably be advised to ignore the number of suspensions due to student behavioral problems, at least as far as improving SAT scores is concerned. The finding that number of suspensions is unrelated to SAT scores could have an explanation, but the data does not allow us to reach a conclusion about this with confidence. We could speculate that very smart students are more rebellious, which is an association that could offset the possible problem caused by those students being suspended more frequently than other students.



## 2.2. Improving satisfaction with car part delivery

Exhibit 2.2 displays the scenario, question, and variables related to the sample dataset used to illustrate how MDDA can be used to address the need for improving the delivery of car parts by a manufacturer to its main customer – an automaker. The problem to be solved is a common one in MDDA applications, namely to assess the order of importance of likely predictors of a main criterion variable.

### Exhibit 2.2. Scenario, question, and variables

#### Scenario

- A car parts manufacturer has been receiving multiple complaints from its main customer, an automaker, regarding part delivery.
- The car parts manufacturer decides to collect data about satisfaction with car part delivery, based on 4 possible predictor variables.

#### Question

- What is the order of importance of the predictors with respect to satisfaction with car part delivery?

#### Variables

- Packg: The quality of the packaging of the car part(s), as perceived by the automaker on a 1-7 scale.
- Crtsy: The courtesy of the deliverer, as perceived by the automaker on a 1-7 scale.
- Cost: The freight cost, measured in dollars.
- Late: How late the delivery is, measured in days.
- Satsf: The satisfaction with the delivery, as perceived by the automaker on a 1-7 scale.

In this case, the predictors are Packg (the quality of the packaging of the car part(s), as perceived by the automaker on a 1-7 scale), Crtsy (the courtesy of the deliverer, as perceived by the automaker on a 1-7 scale), Cost (the freight cost, measured in dollars), and Late (how late the delivery is, measured in days). The main criterion variable is Satsf (the satisfaction with the delivery, as perceived by the automaker on a 1-7 scale).

The main client of this analysis was the car parts manufacturer. This organization wanted to know what they could do to increase satisfaction with car part delivery by its main customer – an automaker. Automakers like Ford typically outsource the manufacturing of car parts (e.g., exhaust pipes, breaks), which they use for assembly – i.e., most car parts are not manufactured directly by the automakers.

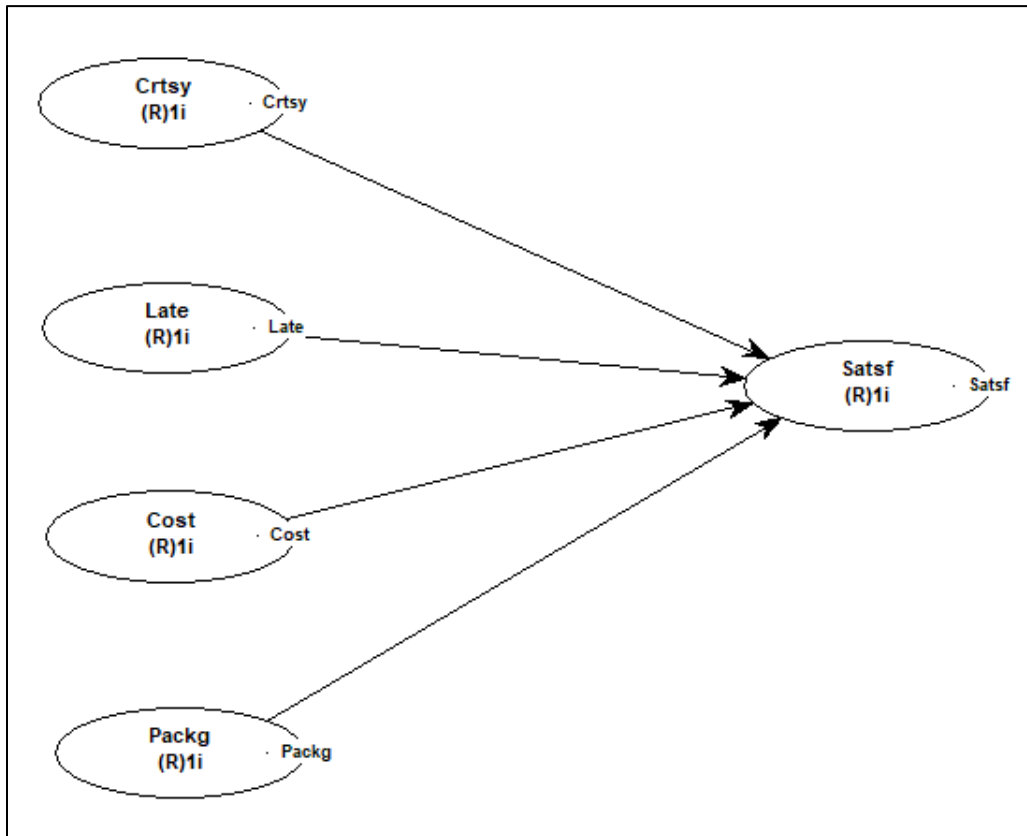
The expectation that the four predictors (Packg, Crtsy, Cost and Late) influence the satisfaction with the deliveries came from interviews with various stakeholders in the car parts manufacturer and the automaker.

The importance of this analysis came from the different costs associated with changing the predictors with the goal of increasing satisfaction with the deliveries. For example, changing the courtesy of the deliverer, through a training program, could be significantly less expensive than changing the freight cost. The latter might require the car parts manufacturer to cover a portion of that cost, possibly reducing its profit margin in a material way.

## 2.2.1. Create the model

Figure 2.2.1 shows the model that was built to serve as the basis for our analysis. It contains four predictor latent variables – named Packg, Crtsy, Cost and Late – pointing at one criterion latent variable, named Satsf. The latent variables have only one indicator each, which essentially means that they are assumed to be measured through their single indicators without error.

Figure 2.2.1. Create the model



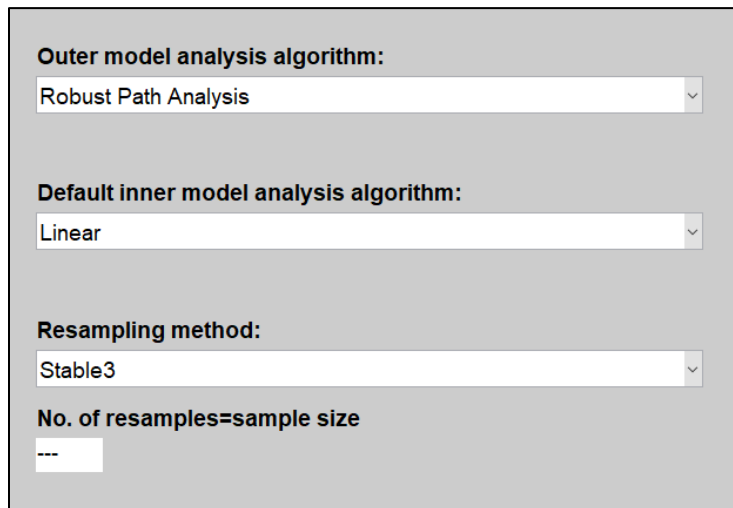
The assumption that a latent variable is measured without error may not be entirely correct, especially when what are measured are perceptions (e.g., satisfaction with something), and frequently may be made for convenience. Whenever possible, multiple indicators should be used with perception-based latent variables, because that enables SEM to minimize the effect of measurement error on the parameters being estimated (e.g., path coefficients). If that is not possible, data analysts have to do the best that they can with what they have available, recognizing the limitations of what they are doing.

As you can see, the latent variables have the same names as their indicators. This is not a requirement. The names could have been different. In fact, they will typically be different for latent variables that are measured through multiple indicators. They will also be different if the indicators' names are longer than 8 characters, which is the maximum allowed for latent variables names. This limitation is to give model graphs a cleaner look.

## 2.2.2. Choose general settings

The options shown in Figure 2.2.2 were the ones chosen for this analysis. They are common in analyses that employ latent variables that are all measured through single indicators, as is the case in our model. The options can be selected through the “**View or change general settings**” menu option.

Figure 2.2.2. Choose general settings



The screenshot shows a settings dialog box with the following options:

- Outer model analysis algorithm:** Robust Path Analysis
- Default inner model analysis algorithm:** Linear
- Resampling method:** Stable3
- No. of resamples=sample size:** ---

The **Robust Path Analysis** outer model analysis algorithm is a simplified algorithm with very good computational efficiency. The **Linear** default inner model analysis algorithm does not perform any warping of relationships; that is, it does not model the relationships as nonlinear. This helps with the interpretation of the results. The **Stable3** method is the default resampling method of the software, because of its high accuracy and robustness to deviations of normality. An advantage of this method is that it does not assume that the data is normally distributed, which is often the case with empirical data. That is, empirical data is typically *not* normally distributed, even though many data analysis techniques assume that it is.

### 2.2.3. Assess collinearity

As soon as the SEM analysis is completed, the software shows the results in graphical format on a window, which also contains a number of menu options that allow you to view and save more detailed results. One of these menu options is the “**View latent variable coefficients**”, which shows the results in the table in Figure 2.2.3. The last row of the table in the figure shows the **full collinearity VIFs** for all latent variables in the model.

Figure 2.2.3. Assess collinearity

	Packg	Crtsy	Cost	Late	Satsf
R-squared					0.539
Adj. R-squared					0.536
Composite reliab.	1.000	1.000	1.000	1.000	1.000
Cronbach's alpha	1.000	1.000	1.000	1.000	1.000
Avg. var. extrac.	1.000	1.000	1.000	1.000	1.000
Full collin. VIF	1.080	1.635	1.159	1.385	2.171

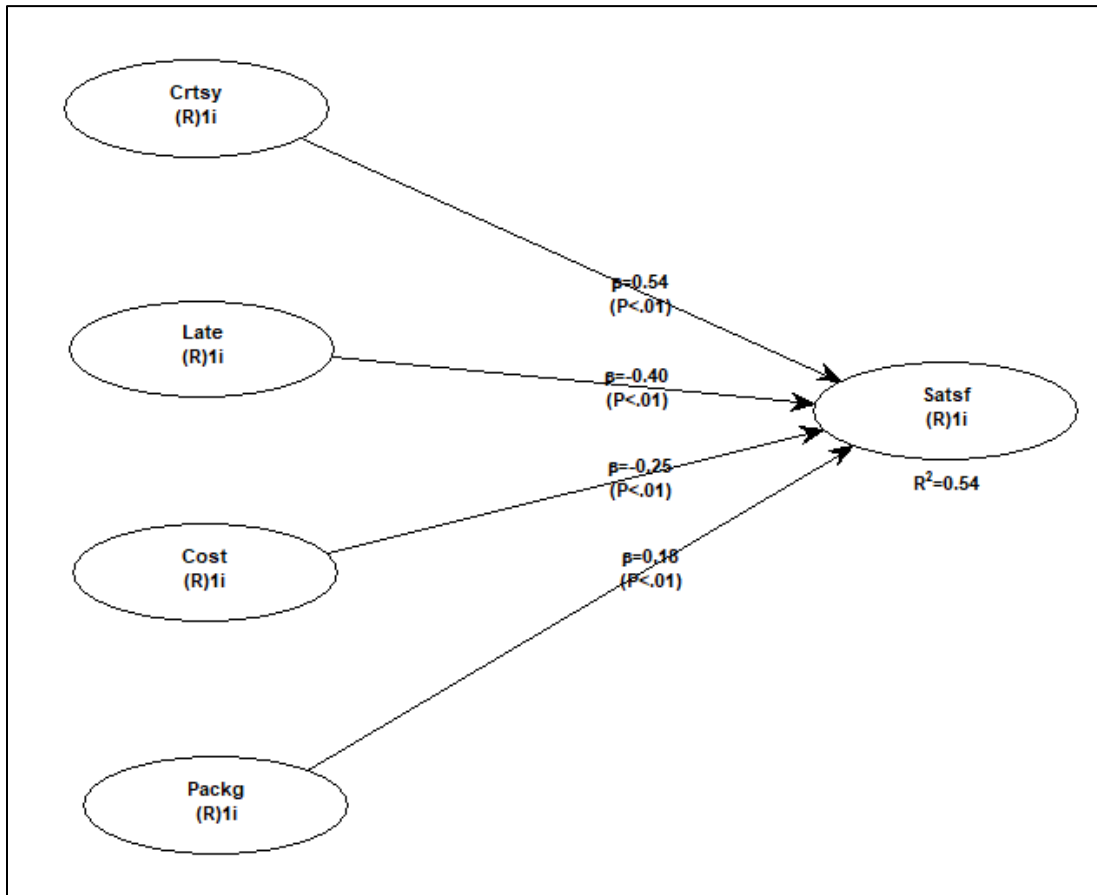
**Full collinearity VIFs of 3.3 or lower suggest** the existence of **no multicollinearity** in the model. A more **relaxed threshold would be 5**. This means that all of the latent variables in the model measure different things, which is an important precondition for a valid analysis. **Full collinearity VIFs of 10 or higher suggest** the existence of **multicollinearity** in the model. Multicollinearity at this level, with full collinearity VIFs of 10 or higher, tends to distort coefficients of association, such as path coefficients.

As we can see, the highest full collinearity VIF in the model is 2.171, well below the conservative threshold of 3.3, which allows us to conclude that all of the latent variables in the model measure different things. That is, the latent variables in the model measure constructs that appear to be conceptually different from one another.

### 2.2.4. Inspect path coefficients

As soon as the SEM analysis is completed, the software shows some of the main results in graphical format on a window. This graphical representation is shown in Figure 2.2.4. The graph with the results shows path coefficients, respective P values, and R-squared coefficients. Users can also show or hide indicators weights, loadings and names.

Figure 2.2.4. Inspect path coefficients



The path coefficients are noted as beta coefficients. “Beta coefficient” is another term often used to refer to path coefficients in SEM analyses; this term is commonly used in multiple regression analyses as well. The P values are displayed below the path coefficients, within parentheses. The R-squared coefficients are shown below each endogenous latent variable (i.e., a latent variable that is hypothesized to be affected by one or more other latent variables), and reflect the percentage of the variance in the latent variable that is explained by the latent variables that are hypothesized to affect it.

Path coefficients associated with P values equal to or lower than 0.05 are deemed to refer to real effects, as opposed to effects that are to be interpreted as “zero”. In this sense, all of the path coefficients in this model refer to effects that appear to be real. The strongest path coefficient is for the link Crtsy > Satsf; the higher is the courtesy of the deliverer, the higher is the satisfaction with the delivery. The strength of a path coefficient depends on its absolute value; i.e., the sign is disregarded. Thus, the second strongest path coefficient is for the link Late > Satsf; the later the

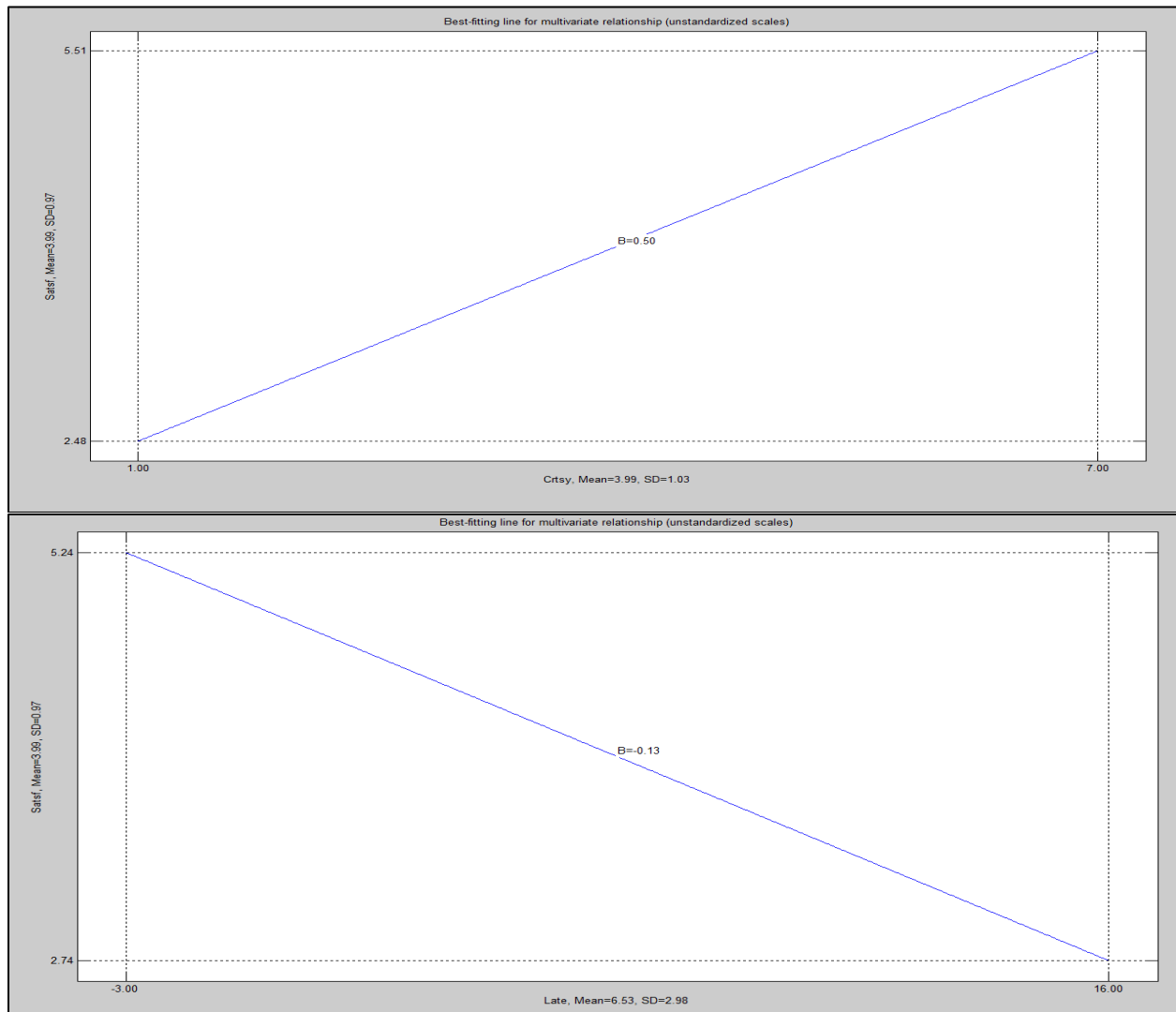
## Model-Driven Data Analytics: Applications with WarpPLS

delivery is completed, the lower is the satisfaction with the delivery. Cost (the freight cost) and Packg (the quality of the packaging) are far behind the other two predictors in terms of the strength of their associations with Satsf (the satisfaction with the delivery).

### 2.2.5. Inspect graphs

As soon as the SEM analysis is completed, the software shows the results in graphical format on a window, where the menu option menu option “**View/plot linear and nonlinear relationships among latent variables**” becomes available. One of the graphs that is particularly useful is available, under this option, from the option “**View focused multivariate relationship graph with segments (unstandardized scales)**”. This graph representation is shown in Figure 2.2.5; with two graphs, for the links Crtsy > Satsf and Late > Satsf, respectively at the top and bottom.

Figure 2.2.5. Inspect graphs



Unlike a path coefficient, which is standardized, the graph shows the corresponding unstandardized regression coefficient. This type of coefficient tends to be more telling to stakeholders. In this case, the unstandardized regression coefficient for the link Crtsy > Satsf is 0.50. The meaning of this is that, for each increase of 1 point in the variable Crtsy (the courtesy of the deliverer, as perceived by the automaker on a 1-7 scale), there is a corresponding increase

## Model-Driven Data Analytics: Applications with WarpPLS

of 0.50 points in the variable Satsf (the satisfaction with the delivery, as perceived by the automaker on a 1-7 scale).

The unstandardized regression coefficient for the link Late > Satsf is -0.13. The meaning of this is that, for each increase of 1 day in the variable Late (how late the delivery is, measured in days), there is a corresponding decrease of -0.13 points in the variable Satsf (the satisfaction with the delivery, as perceived by the automaker on a 1-7 scale).

Note that Late (how late the delivery is, measured in days) assumes negative values; these are cases in which the delivery occurred early, or before the expected date of delivery. As you can see, when unstandardized values are considered, the importance of the link Crtsy > Satsf over Late > Satsf becomes more apparent than when we inspected only the standardized path coefficients. The unstandardized values are farther apart, in absolute terms, than the corresponding standardized values.



### **2.2.6. Provide advice**

One of the goals of the analysis was to answer the question: What is the order of importance of the predictors with respect to satisfaction with car part delivery? The inspection of the path coefficients suggests that Crtsy (the courtesy of the deliverer, as perceived by the automaker on a 1-7 scale) is the most important predictor, followed by these predictors in order of importance: Late (how late the delivery is, measured in days); Cost (the freight cost, measured in dollars); and Packg (the quality of the packaging of the car part(s), as perceived by the automaker on a 1-7 scale).

Given this, the car parts manufacturer from which the data was obtained should arguably be advised to try to increase the courtesy of the deliverers of its car parts, so that satisfaction with car part deliveries would increase. This could be done through a training program focused on courtesy in the context of car part deliveries. This type of intervention could be significantly less expensive than initiatives aimed at changing the other predictors.

Also, the car parts manufacturer should arguably be advised to target only courtesy in the context of car part deliveries at first, and then conduct a follow-up analysis. The reason for this is that the importance of the other variables (Packg, Cost and Late) may increase as Crtsy reaches a higher level at which its variation decreases – i.e., all deliveries are equally high in terms of courtesy.

As the variation in Crtsy decreases, it may become a less important predictor of satisfaction with car part delivery when compared with the other predictors. Also, the relative importance of the other predictors may change (e.g., one may become significantly more important than it was before), calling for a different prioritization scheme.

Analogously, one could speculate that something similar to the above scenario might have been the case with Packg. That is, its relatively weak influence as a predictor in this analysis might simply have been due to the fact that the quality of the packaging of the car part(s) was good enough in this first analysis, and/or showed only a small amount of variation, compared with the other predictors.

## 2.3. Improving job performance through empathetic management

Exhibit 2.3 displays the scenario, question, and variables related to the sample dataset used to illustrate how MDDA can be used to address the need of a large insurance company for improving job performance, via techniques that emphasize empathetic management.

### Exhibit 2.3. Scenario, question, and variables

#### Scenario

- A large insurance company wants to improve the performance of its employees by employing "empathetic management" techniques.
- Empathetic management techniques emphasize the development of positive emotions in employees, as opposed to fear of termination.
- The insurance company decides to train the managers of a 250-employee unit on the use of empathetic management techniques.
- The insurance company also decides to collect data about 5 variables, 4 of which are measured through multiple indicators (i.e., they are "latent" variables).
- It is believed that empathetic management will improve job satisfaction, job innovativeness, and organizational commitment.
- It is also believed that, through the mediated effects above, empathetic management will improve job performance.

#### Question

- Is the effect of empathetic management on job performance mediated? If yes, is the mediation full or partial?

#### Variables

EM: Empathetic management, as perceived by employees on 1-7 scales via the question-statements below (indicators).

- EM1: My supervisor gives me praise for my good work.
- EM2: My supervisor shows me encouragement for my work efforts.
- EM3: My supervisor shows concern about my job satisfaction.

JS: Job satisfaction, as perceived by employees on 1-7 scales via the question-statements below (indicators).

- JS1: I always feel satisfied with my job.
- JS2: I like my job.
- JS3: I do not want to change my job.

JI: Job innovativeness, as perceived by employees on 1-7 scales via the question-statements below (indicators).

- JI1: I try new ideas and approaches to problems.
- JI2: I welcome uncertainty and unusual circumstances related to my tasks.
- JI3: I can be counted on to find a new use for existing methods or equipment.

OC: Organizational commitment, as perceived by employees on 1-7 scales via the question-statements below (indicators).

- OC1: I would be very happy to spend the rest of my career with this organization.
- OC2: I feel a strong sense of belonging to my organization.
- OC3: I feel 'emotionally attached' to this organization.

JP: Job performance, as perceived by employees' supervisors on a 1-7 scale via the question below (from annual evaluation).

- JP1: How would you rate the performance of this employee?

Here the MDDA analysis is aimed at finding out if the effect of empathetic management on job performance is mediated by intermediate effects on job satisfaction, job innovativeness, and

organizational commitment. Also, the organization wanted to find out if the mediation is full or partial. If the mediation is full, this would mean that no mediating variable has been omitted, and that the model is complete in that respect. If the mediation is partial, then one or more variables would have been omitted, and it might be a good idea to try to identify them in future analyses.

The variables used in this analysis are EM (empathetic management, as perceived by employees on 1-7 scales via 3 question-statements), JS (job satisfaction, as perceived by employees on 1-7 scales via 3 question-statements), JI (job innovativeness, as perceived by employees on 1-7 scales via 3 question-statements), and OC (organizational commitment, as perceived by employees on 1-7 scales via 3 question-statements), and JP (job performance, as perceived by employees' supervisors on a 1-7 scale via 1 question from their annual evaluations).

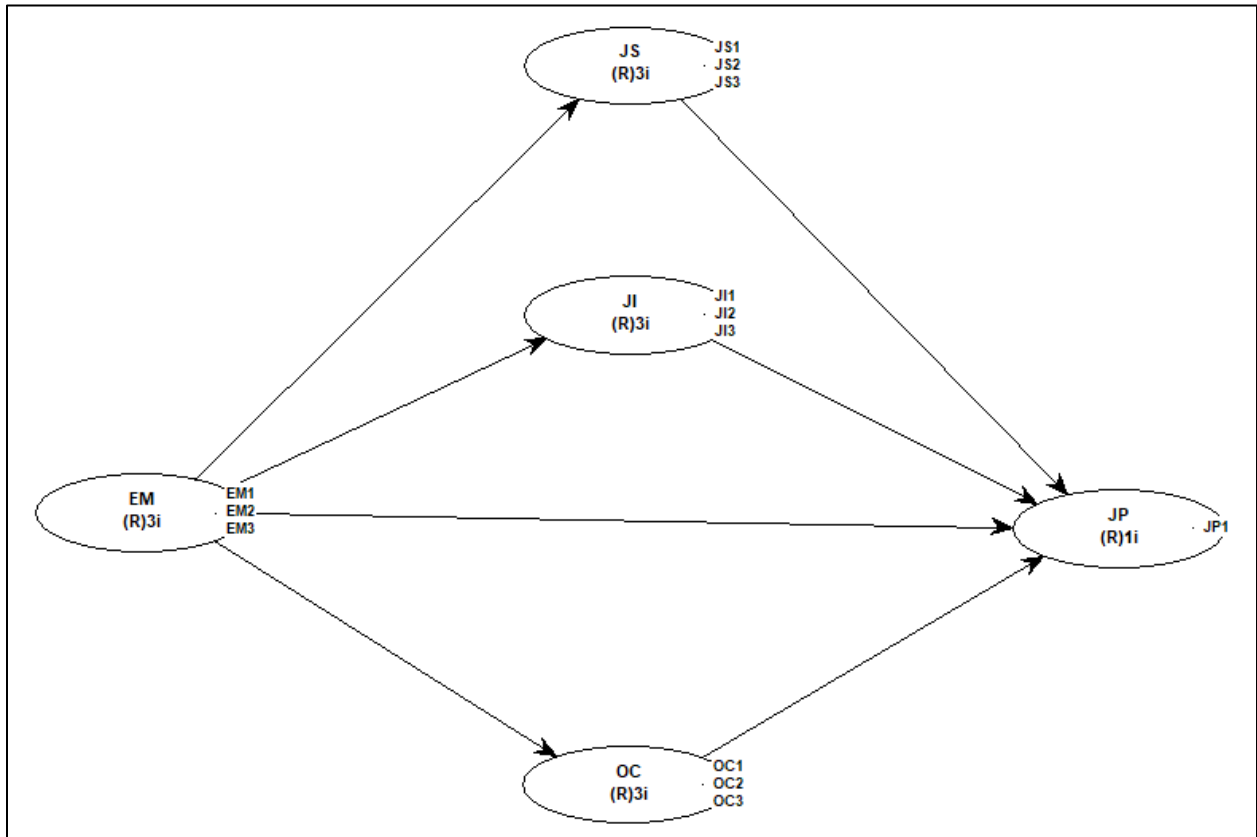
The importance of this analysis comes from the interest of the large insurance company in ways to improve the performance of its employees by employing management techniques that emphasize the development of positive emotions in employees, as opposed to fear of termination. These empathetic management techniques essentially entail supervisors giving employees praise for good work, encouragement for work efforts, as well as showing concern for the employee's job satisfaction.

If these techniques are successful, they can potentially make the insurance company a "best place to work". This would attract labor talent, and generate positive publicity for the company, among other advantages. All of these would likely make the company more competitive, which would ultimately lead to relative gains in sales and profits compared with other companies in the same industry.

### 2.3.1. Create the model

Figure 2.3.1 shows the model that was built to serve as the basis for our analysis. It contains five latent variables – named EM (empathetic management), JS (job satisfaction), JI (job innovativeness), OC (organizational commitment), and JP (job performance). The variables JS, JI and OC are assumed by the model to mediate the relationship between EM and JP. Four of the latent variables (EM, JS, JI and OC) have three indicators each, which essentially means that they are assumed to be measured with error. One of the latent variables (JP) has only one indicator, which means that it is assumed to be measured without error.

Figure 2.3.1. Create the model



The assumption that a latent variable is measured without error may not be entirely correct, and frequently may be made for convenience. Whenever possible, multiple indicators should be used, because that enables SEM to minimize the effect of measurement error on the parameters being estimated (e.g., path coefficients). If that is not possible, data analysts have to do the best that they can with what they have available, recognizing the limitations of what they are doing.

In our illustrative case, JP is assumed to be measured without error because there was only one supervisor evaluation score for each employee in our dataset. One could reasonably argue that those supervisor evaluation scores were imprecise measures of the employees' actual performance, violating the measurement without error assumption. This would tend to suppress the values of the path coefficients for the links pointing at JP. That is a limitation that would

## Model-Driven Data Analytics: Applications with WarpPLS

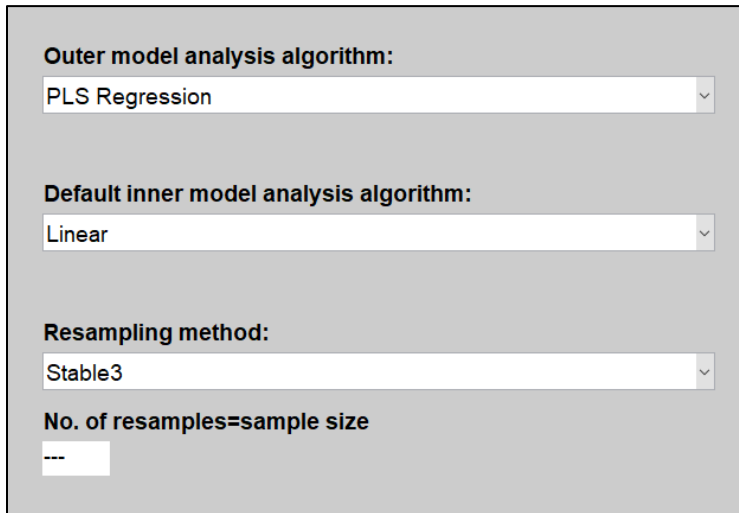
make the data analysis results more conservative; i.e., one may not find an effect, when in reality there is a real underlying effect.

As you can see, the latent variables have the same names as their indicators, minus the number identifies (e.g., “1” for JS1). This is not a requirement. The names could have been different. In fact, they will typically be different for latent variables that are measured through multiple indicators when those indicators’ names differ from one another. They will also be different if the indicators’ names are longer than 8 characters, which is the maximum allowed for latent variables names. This limitation is to give model graphs a cleaner look.

### 2.3.2. Choose general settings

The options shown in Figure 2.3.2 were the ones chosen for this analysis. They are common in exploratory analyses that employ one or more latent variables that are measured through multiple indicators, as is the case in our model. The options can be selected through the “**View or change general settings**” menu option.

Figure 2.3.2. Choose general settings



The screenshot shows a dialog box with a light gray background. It contains four settings, each with a label and a dropdown menu:

- Outer model analysis algorithm:** The dropdown menu is set to "PLS Regression".
- Default inner model analysis algorithm:** The dropdown menu is set to "Linear".
- Resampling method:** The dropdown menu is set to "Stable3".
- No. of resamples=sample size:** The dropdown menu is set to "---".

**PLS Regression** has been the default outer model algorithm since the software’s inception. This algorithm iterates by making the outer model weights directly proportional to the loadings, until the weights become stable. This algorithm does not let the inner model influence the outer model. The weights are obtained by regressing the latent variables on their indicators, and the loadings by regressing the indicators on the latent variables.

The **Linear** default inner model analysis algorithm does not perform any warping of relationships; that is, it does not model the relationships as nonlinear. This helps with the interpretation of the results. The **Stable3** method is the default resampling method of the software, because of its high accuracy and robustness to deviations of normality. An advantage of this method is that it does not assume that the data is normally distributed, which is often the case with empirical data. That is, empirical data is typically *not* normally distributed, even though many data analysis techniques assume that it is.

### 2.3.3. Assess collinearity

As soon as the SEM analysis is completed, the software shows the results in graphical format on a window, which also contains a number of menu options that allow you to view and save more detailed results. One of these menu options is the “**View latent variable coefficients**”, which shows the results in the table in Figure 2.3.3. The last row of the table in the figure shows the **full collinearity VIFs** for all latent variables in the model.

Figure 2.3.3. Assess collinearity

	EM	JS	JL	OC	JP
R-squared		0.200	0.051	0.055	0.579
Adj. R-squared		0.196	0.047	0.052	0.572
Composite reliab.	0.848	0.862	0.845	0.827	1.000
Cronbach's alpha	0.732	0.761	0.724	0.685	1.000
Avg. var. extrac.	0.651	0.676	0.646	0.614	1.000
Full collin. VIF	1.334	1.857	1.399	1.165	2.376

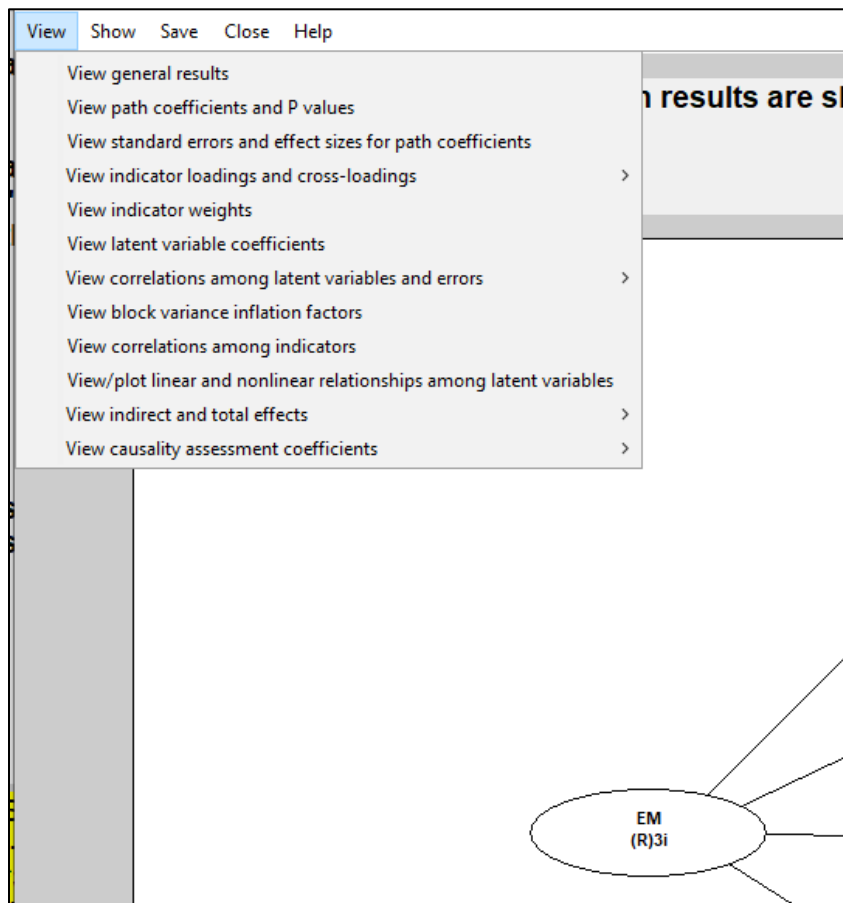
**Full collinearity VIFs of 3.3 or lower suggest** the existence of **no multicollinearity** in the model. A more **relaxed threshold would be 5**. This means that all of the latent variables in the model measure different things, which is an important precondition for a valid analysis. **Full collinearity VIFs of 10 or higher suggest** the existence of **multicollinearity** in the model. Multicollinearity at this level, with full collinearity VIFs of 10 or higher, tends to distort coefficients of association, such as path coefficients.

As we can see, the highest full collinearity VIF in the model is 2.376, well below the conservative threshold of 3.3, which allows us to conclude that all of the latent variables in the model measure different things. That is, the latent variables in the model measure constructs that appear to be conceptually different from one another.

### 2.3.4. Assess validity and reliability

As soon as the SEM analysis is completed, the software shows the results in graphical format on a window, which also contains a number of menu options that allow you to view and save more detailed results. If one wants to assess the quality of a measurement instrument where the indicators reflectively measure their corresponding latent variables, like a typical questionnaire, one will usually want to assess convergent validity, discriminant validity, and reliability. The menu options, shown in Figure 2.3.4, that are normally used for those assessments are: “**View indicator loadings and cross-loadings**”, “**View latent variable coefficients**”, and “**View correlations among latent variables and errors**”.

Figure 2.3.4. Assess collinearity



A **reflective** latent variable is one in which all of the indicators are expected to be highly correlated with the latent variable, and also highly correlated with one another. For example, the answers to certain question-statements by a group of people, measured on a 1 to 7 scale (1=strongly disagree; 7 strongly agree) and answered after a meal, are expected to be highly correlated with the latent variable “satisfaction with a meal”. The question-statements are: “I am satisfied with this meal”, and “After this meal, I feel full”. Therefore, the latent variable “satisfaction with a meal”, can be said to be reflectively measured through these indicators.

A **formative** latent variable is one in which the indicators are expected to measure certain attributes of the latent variable, but the indicators are not expected to be highly correlated with



the latent variable, because they (i.e., the indicators) are not expected to be correlated with one another. For example, let us assume that the latent variable “Satisf” (“satisfaction with a meal”) is measured using the two following question-statements: “I am satisfied with the main course” and “I am satisfied with the dessert”. Both main course and dessert make up the meal (i.e., they are part of the same meal) but their satisfaction indicators are not expected to be highly correlated with each other. Some people may like the main course, and not like the dessert, or vice-versa.

The assessment of **convergent validity, discriminant validity, and reliability**, as discussed here, applies to **reflective** measurement of latent variables.

**Convergent validity** is a measure of the quality of a measurement instrument; the instrument itself is typically a set of question-statements (i.e., a questionnaire). A measurement instrument has good convergent validity if the question-statements (or other measures) associated with each latent variable are understood by the respondents in the same way as they were intended by the designers of the question-statements.

**Discriminant validity** is also a measure of the quality of a measurement instrument. A measurement instrument has good discriminant validity if the question-statements (or other measures) associated with each latent variable are not confused by the respondents, in terms of their meaning, with the question-statements associated with other latent variables.

**Reliability** is yet another measure of the quality of a measurement instrument; the instrument itself is typically a set of question-statements. A measurement instrument has good reliability if the question-statements (or other measures) associated with each latent variable are understood in the same way by different respondents.

### 2.3.4.1. Convergent validity

The “**View indicator loadings and cross-loadings**” menu options allow users to view various variations of loadings and cross-loadings: combined loadings and cross-loadings, normalized combined loadings and cross-loadings, pattern loadings and cross-loadings, normalized pattern loadings and cross-loadings, structure loadings and cross-loadings, and normalized structure loadings and cross-loadings. The option with combined loadings and cross-loadings is typically the one used for convergent validity assessment (see Figure 2.3.4.1).

**Figure 2.3.4.1. Combined loadings and cross-loadings**

	EM	JS	J1	OC	JP	Type (as defined)	SE	P value
EM1	(0.824)	0.075	0.041	0.036	0.023	Reflective	0.055	<0.001
EM2	(0.824)	0.004	-0.086	-0.134	-0.024	Reflective	0.055	<0.001
EM3	(0.772)	-0.085	0.049	0.105	0.000	Reflective	0.055	<0.001
JS1	0.019	(0.835)	-0.097	-0.067	0.283	Reflective	0.055	<0.001
JS2	0.006	(0.822)	0.016	-0.036	0.048	Reflective	0.055	<0.001
JS3	-0.026	(0.810)	0.084	0.106	-0.341	Reflective	0.055	<0.001
J11	-0.002	-0.091	(0.845)	-0.052	0.189	Reflective	0.055	<0.001
J12	0.013	-0.077	(0.807)	0.019	0.126	Reflective	0.055	<0.001
J13	-0.011	0.184	(0.756)	0.038	-0.345	Reflective	0.056	<0.001
OC1	-0.036	-0.004	0.000	(0.816)	0.014	Reflective	0.055	<0.001
OC2	-0.053	0.019	-0.003	(0.796)	0.080	Reflective	0.055	<0.001
OC3	0.098	-0.017	0.003	(0.737)	-0.101	Reflective	0.056	<0.001
JP1	0.000	0.000	0.000	0.000	(1.000)	Reflective	0.053	<0.001

A measurement instrument has good convergent validity if the question-statements (or other measures) associated with each latent variable are understood by the respondents in the same way as they were intended by the designers of the question-statements. In this respect, two criteria are recommended as the basis for concluding that a measurement model has acceptable convergent validity: that the **P values associated with the loadings be equal to or lower than 0.05**; and that the **loadings be equal to or greater than 0.5**. The loadings associated with each latent variable are shown within parentheses.

As we can see, all of the loadings are equal to or greater than 0.5. In fact, they are well above this threshold; the lowest loading in the whole table is 0.737. The loading shown as 1.000 is for a latent variable that is measured through a single indicator; loadings of 1.000 are always shown in these cases, by convention, and do not mean anything (only loadings for latent variables with multiple indicators have meaning in this context). Also, we can see that the P values associated with the loadings are all equal to or lower than 0.05. In fact, they are all lower than 0.001. Therefore, we can say that our measurement instrument has good convergent validity.

### 2.3.4.2. Discriminant validity

The “**View correlations among latent variables and errors**” menu options allow users to view tables containing correlations among latent variables, the P values associated with those correlations, square roots of average variances extracted (AVEs), correlations among latent variable error terms (or residuals), and the VIFs associated with latent variable error terms. The table containing correlations among latent variables and square roots of AVEs is typically the one used for discriminant validity assessment (see Figure 2.3.4.2).

**Figure 2.3.4.2. Correlations among latent variables with square roots of AVEs**

	EM	JS	JI	OC	JP
EM	(0.807)	0.447	0.226	0.236	0.407
JS	0.447	(0.822)	0.245	0.128	0.631
JI	0.226	0.245	(0.803)	0.038	0.505
OC	0.236	0.128	0.038	(0.784)	0.300
JP	0.407	0.631	0.505	0.300	(1.000)

A measurement instrument has good discriminant validity if the question-statements (or other measures) associated with each latent variable are not confused by the respondents answering the questionnaire with the question-statements associated with other latent variables, particularly in terms of the meaning of the question-statements. The following criterion is recommended for discriminant validity assessment: **for each latent variable, the square root of the AVE should be higher than any of the correlations involving that latent variable.** That is, the values on the diagonal of the table containing correlations among latent variables, which are the square roots of the AVEs for each latent variable, should be higher than any of the values above or below them, in the same column.

As we can see, for all latent variables, the square roots of the AVEs are higher than any of the correlations involving those latent variables. The highest correlation among latent variables in the table is 0.631, between JS and JP, and the corresponding square roots of the AVEs for JS and JP are respectively 0.822 and 1.000. When a latent variable is measured through a single indicator, the square root of its AVE is always shown as 1.000. In any event, since the recommended criterion was met, we can say that our measurement instrument has good discriminant validity.

### 2.3.4.3. Reliability

The “**View latent variable coefficients**” menu option allow users to view several estimates that are provided for each latent variable. Among these are coefficients that can be used to assess a measurement instrument’s reliability. These are the composite reliability and Cronbach’s alpha coefficients, shown on the last two rows of the table in Figure 2.3.4.3.

**Figure 2.3.4.3. Reliability coefficients**

	EM	JS	Jl	OC	JP
R-squared		0.200	0.051	0.055	0.579
Adj. R-squared		0.196	0.047	0.052	0.572
Composite reliab.	0.848	0.862	0.845	0.827	1.000
Cronbach's alpha	0.732	0.761	0.724	0.685	1.000

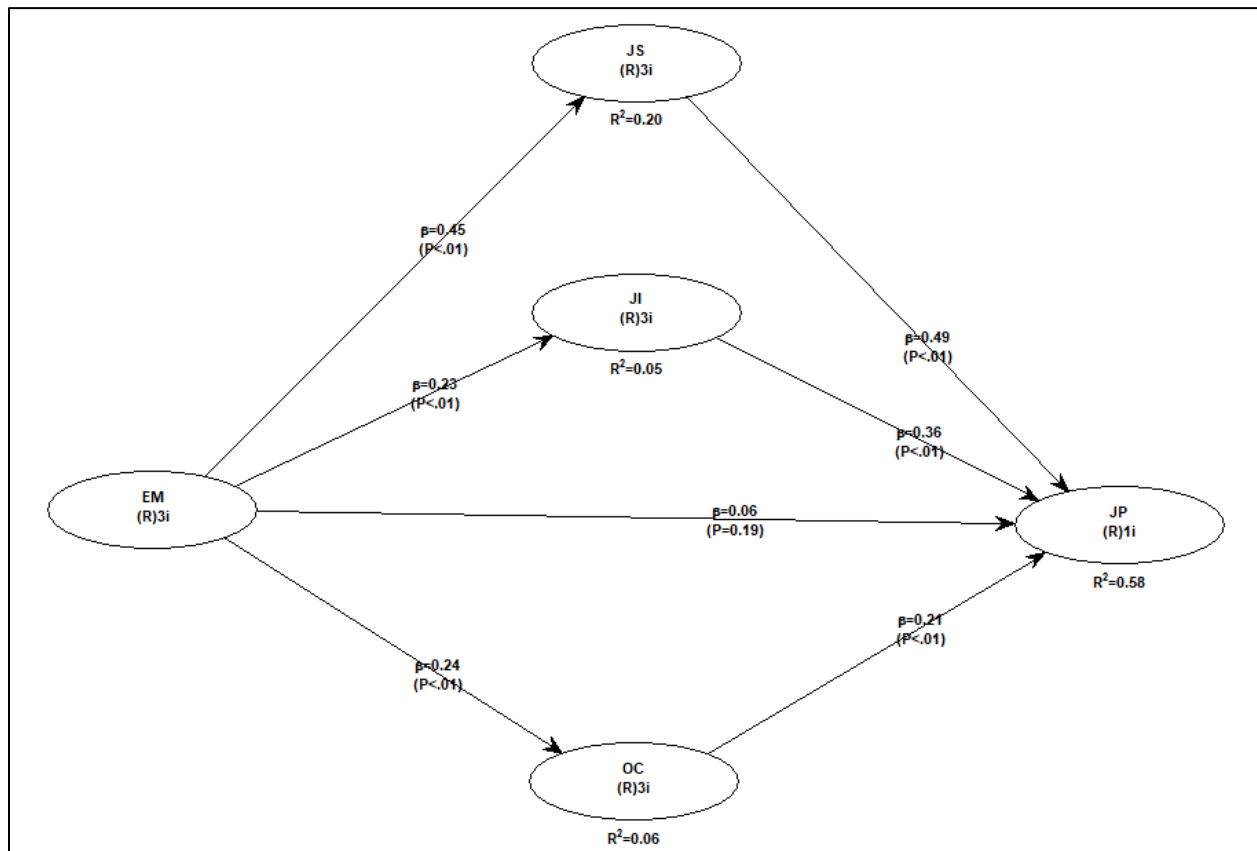
A measurement instrument has good reliability if the question-statements (or other measures) associated with each latent variable are understood in the same way by different respondents. The following criterion is suggested in the assessment of the reliability of a measurement instrument: **either the composite reliability or the Cronbach’s alpha coefficient should be equal to or greater than 0.6.**

As we can see, for all latent variables, either the composite reliability or the Cronbach’s alpha coefficient is equal to or greater than 0.6. In fact, in this case all of the composite reliability and Cronbach’s alpha coefficients are greater than 0.6. The lowest such coefficient in the table is 0.685, which is the Cronbach’s alpha coefficient for OC. Therefore, since the recommended criterion was met, we can say that our measurement instrument has good reliability.

### 2.3.5. Inspect path coefficients

As soon as the SEM analysis is completed, the software shows some of the main results in graphical format on a window. This graphical representation is shown in Figure 2.3.5. The graph with the results shows path coefficients, respective P values, and R-squared coefficients. Users can also show or hide indicators weights, loadings and names.

Figure 2.3.5. Inspect path coefficients



The path coefficients are noted as beta coefficients. “Beta coefficient” is another term often used to refer to path coefficients in SEM analyses; this term is commonly used in multiple regression analyses as well. The P values are displayed below the path coefficients, within parentheses. The R-squared coefficients are shown below each endogenous latent variable (i.e., a latent variable that is hypothesized to be affected by one or more other latent variables), and reflect the percentage of the variance in the latent variable that is explained by the latent variables that are hypothesized to affect it.

Path coefficients associated with P values equal to or lower than 0.05 are deemed to refer to real effects, as opposed to effects that are to be interpreted as “zero”. In this sense, all of the path coefficients in this model refer to effects that appear to be real, except for the link EM > JP. The strongest path coefficients are for the links EM > JS and JS > JP. That is, the higher is the use of empathetic management, the higher is job satisfaction. And, the higher is job satisfaction, the higher is job performance. The strength of a path coefficient depends on its absolute value; i.e.,

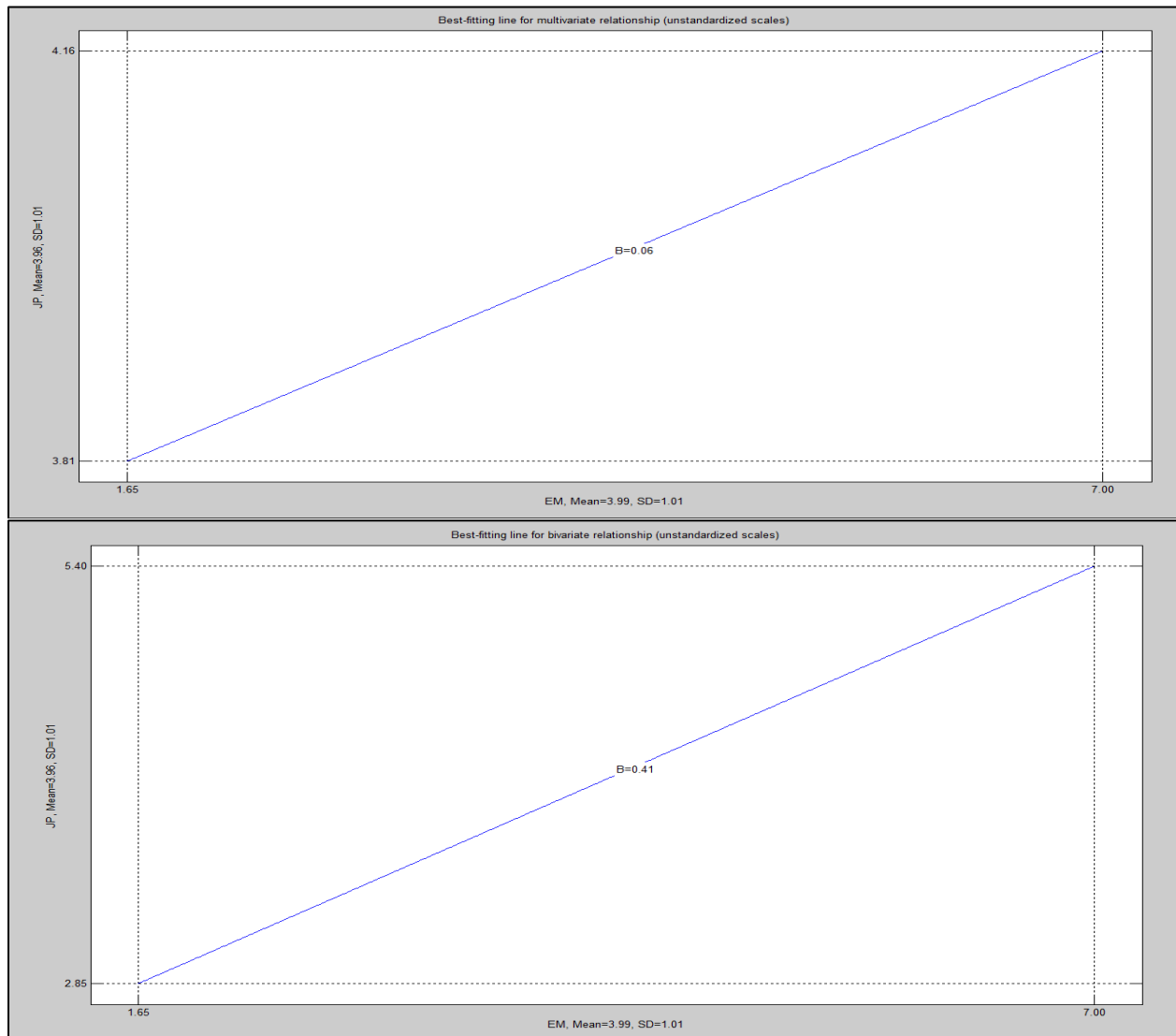
## Model-Driven Data Analytics: Applications with WarpPLS

the sign is disregarded. In our case this is not particularly relevant, because all path coefficients are positive.

### 2.3.6. Inspect graphs

As soon as the SEM analysis is completed, the software shows the results in graphical format on a window, where the menu option “**View/plot linear and nonlinear relationships among latent variables**” becomes available. Two of the graphs that are particularly useful in this analysis are available, under this option, from the options “**View focused multivariate relationship graph with segments (unstandardized scales)**” and “**View focused bivariate relationship graph with segments (unstandardized scales)**”. These are shown in Figure 2.3.6, respectively at the top and bottom, both for the link EM > JP.

Figure 2.3.6. Inspect graphs



**Multivariate and bivariate** relationship graphs usually differ only when two or more predictor latent variables point at one criterion latent variable in a latent variable block. The addition of predictors will normally reduce the path coefficients in a latent variable block.

Because of this, **typically a multivariate relationship graph will have a lower overall inclination (or steepness) than its corresponding bivariate relationship graph.**

The multivariate unstandardized regression coefficient for the link EM > JP is 0.06. The meaning of this is that, for each increase of 1 point in the variable EM (empathetic management, as perceived by employees on 1-7 scales via 3 question-statements), there is a corresponding increase of 0.06 points in the variable JP (job performance, as perceived by employees' supervisors on a 1-7 scale via 1 question from their annual evaluations).

The tiny increase above factors out the mediating effects via JS (job satisfaction), JI (job innovativeness), and OC (organizational commitment). The total effect, however, does not. This effect is given by the bivariate unstandardized regression coefficient for the link EM > JP, which is a much higher 0.41. For each increase of 1 point in the variable EM (empathetic management), there is a corresponding increase of 0.41 points in the variable JP (job performance).



### **2.3.7. Provide advice**

One of the goals of the analysis was to answer two questions: Is the effect of empathetic management on job performance mediated? If yes, is the mediation full or partial? The inspection of path coefficients suggests that the effect of empathetic management on job performance is indeed mediated by intermediate effects on JS (job satisfaction), JI (job innovativeness), and OC (organizational commitment). Given that the path coefficient for the link EM > JP was found to be statistically nonsignificant, when the mediating variables (JS, JI and OC) were controlled for, we can conclude that the mediation is full.

That is, we can conclude that the model is complete, and that no other “hidden” mediating variables were missed. This conclusion validates the decision by the insurance company to train the managers of a 250-employee unit on the use of empathetic management techniques, and supports the expansion of this training. The conclusion is an endorsement of the insurance company’s management; not only were they apparently right about the effect of empathetic management on job performance, they were also correct about the mechanisms by which this effect worked – via mediation by precisely three variables (JS, JI and OC).

The analysis also provides support for the interest of the large insurance company in ways to improve the performance of its employees by employing management techniques that emphasize the development of positive emotions in employees, as opposed to fear of termination. The analysis suggests that this class of techniques, if broadly employed in the company, could make the insurance company a “best place to work”, with other related advantages: attract labor talent, and generate positive publicity for the company. All of these would likely make the company more competitive, which would ultimately lead to relative gains in sales and profits compared with other companies in the same industry.

## 2.4. Improving software development by employing older coders

Exhibit 2.4 displays the scenario, questions, and variables related to the sample dataset used to illustrate how MDDA can be used to address the need of a large software developer to understand the impact of stress on software development performance, and the role of age in this area, with the end goal of reducing the number of bugs in software modules.

### Exhibit 2.3. Scenario, question, and variables

#### Scenario

- A large software developer wants to reduce the number of bugs in software modules.
- Some software modules must be developed faster than others (due to deadlines), which is a source of stress.
- The software developer decides to study the moderating effect that age has on the impact of stress on number of bugs.

#### Questions

- Does stress have an impact on the number of software bugs?
- If yes to the above, does age moderate the impact of stress on the number of bugs?

#### Variables

Strs: Stress of the software developer, measured as blood cortisol concentration in micrograms per decilitre (mcg/dL).

Age: Age of the software developer, measured in years.

Bugs: Number of bugs per 1,000 lines of code written by the software developer.

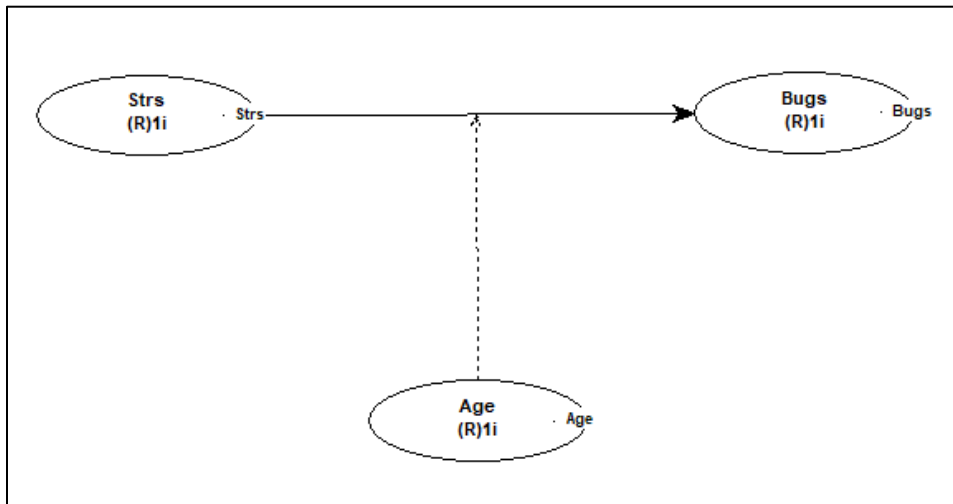
Here the MDDA analysis is aimed at finding out if stress has an impact on the number of software bugs, and, if yes, whether age moderates this relationship – i.e., moderates the impact of stress on the number of bugs. The variables used in this analysis are Strs (stress of the software developer, measured as blood cortisol concentration in micrograms per decilitre – mcg/dL), Age (age of the software developer, measured in years), and Bugs (number of bugs per 1,000 lines of code written by the software developer).

The importance of this analysis comes from the interest of the large software developer in ways to improve software development performance, particularly under stress (due to deadlines). Some anecdotal evidence in the company suggested that older software developers tend to perform well under stress, an effect that the company wants to assess and quantify. If this effect is indeed real, the company may start hiring more older software developers, which would go counter the industry trend of hiring mostly younger employees – a trend that generally violates age discrimination laws.

### 2.4.1. Create the model

Figure 2.4.1 shows the model that was built to serve as the basis for our analysis. It contains one predictor latent variable, namely Strs (stress of the software developer); one moderating latent variable, namely Age (age of the software developer); and one criterion latent variable, namely Bugs (number of bugs per 1,000 lines of code). The latent variables have only one indicator each, which essentially means that they are assumed to be measured through their single indicators without error.

Figure 2.4.1. Create the model



The assumption that a latent variable is measured without error may not be entirely correct, and frequently may be made for convenience. Whenever possible, multiple indicators should be used, because that enables SEM to minimize the effect of measurement error on the parameters being estimated (e.g., path coefficients). If that is not possible, data analysts have to do the best that they can with what they have available, recognizing the limitations of what they are doing. In this particular analysis, no perception-based variables were used, which mitigates the problem possibly caused by using only one indicator per latent variable.

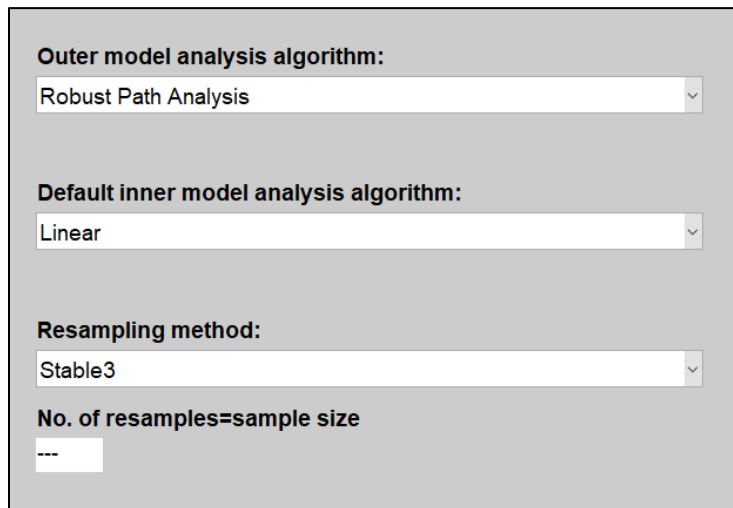
As you can see, the latent variables have the same names as their indicators. This is not a requirement. The names could have been different. In fact, they will typically be different for latent variables that are measured through multiple indicators. They will also be different if the indicators' names are longer than 8 characters, which is the maximum allowed for latent variables names. This limitation is to give model graphs a cleaner look.

Note that there is no direct Age > Bugs link in the model, which means that the analysis assumes that the variable Age influences Bugs only indirectly, as a moderator of the direct link Strs > Bugs. This is a valid assumption, which could be also tested via the creation of a direct Age > Bugs link, which would be in addition to the moderating link Age > (Strs > Bugs). However, the anecdotal data does not support the Age > Bugs link, which is why it was not explicitly included. If this directly link was included, it would compete with the moderating link for the explained variance in the variable Bugs, and this could artificially suppress that link to the point of making it appear to be nonexistent. This illustrates the importance of applied theories in analysis.

## 2.4.2. Choose general settings

The options shown in Figure 2.4.2 were the ones chosen for this analysis. They are common in analyses that employ latent variables that are all measured through single indicators, as is the case in our model. The options can be selected through the “**View or change general settings**” menu option.

Figure 2.4.2. Choose general settings



The screenshot shows a settings dialog box with the following options:

- Outer model analysis algorithm:** Robust Path Analysis
- Default inner model analysis algorithm:** Linear
- Resampling method:** Stable3
- No. of resamples=sample size:** ---

The **Robust Path Analysis** outer model analysis algorithm is a simplified algorithm with very good computational efficiency. The **Linear** default inner model analysis algorithm does not perform any warping of relationships; that is, it does not model the relationships as nonlinear. This helps with the interpretation of the results. The **Stable3** method is the default resampling method of the software, because of its high accuracy and robustness to deviations of normality. An advantage of this method is that it does not assume that the data is normally distributed, which is often the case with empirical data. That is, empirical data is typically *not* normally distributed, even though many data analysis techniques assume that it is.

### 2.4.3. Assess collinearity

As soon as the SEM analysis is completed, the software shows the results in graphical format on a window, which also contains a number of menu options that allow you to view and save more detailed results. One of these menu options is the “**View latent variable coefficients**”, which shows the results in the table in Figure 2.4.3. The last row of the table in the figure shows the **full collinearity VIFs** for all latent variables in the model.

Figure 2.4.3. Assess collinearity

	Strs	Bugs	Age	Age*Strs
R-squared		0.459		
Adj. R-squared		0.457		
Composite reliab.	1.000	1.000	1.000	1.000
Cronbach's alpha	1.000	1.000	1.000	1.000
Avg. var. extrac.	1.000	1.000	1.000	1.000
Full collin. VIF	1.324	1.855	1.005	1.593

**Full collinearity VIFs of 3.3 or lower suggest** the existence of **no multicollinearity** in the model. A more **relaxed threshold would be 5**. This means that all of the latent variables in the model measure different things, which is an important precondition for a valid analysis. **Full collinearity VIFs of 10 or higher suggest** the existence of **multicollinearity** in the model. Multicollinearity at this level, with full collinearity VIFs of 10 or higher, tends to distort coefficients of association, such as path coefficients.

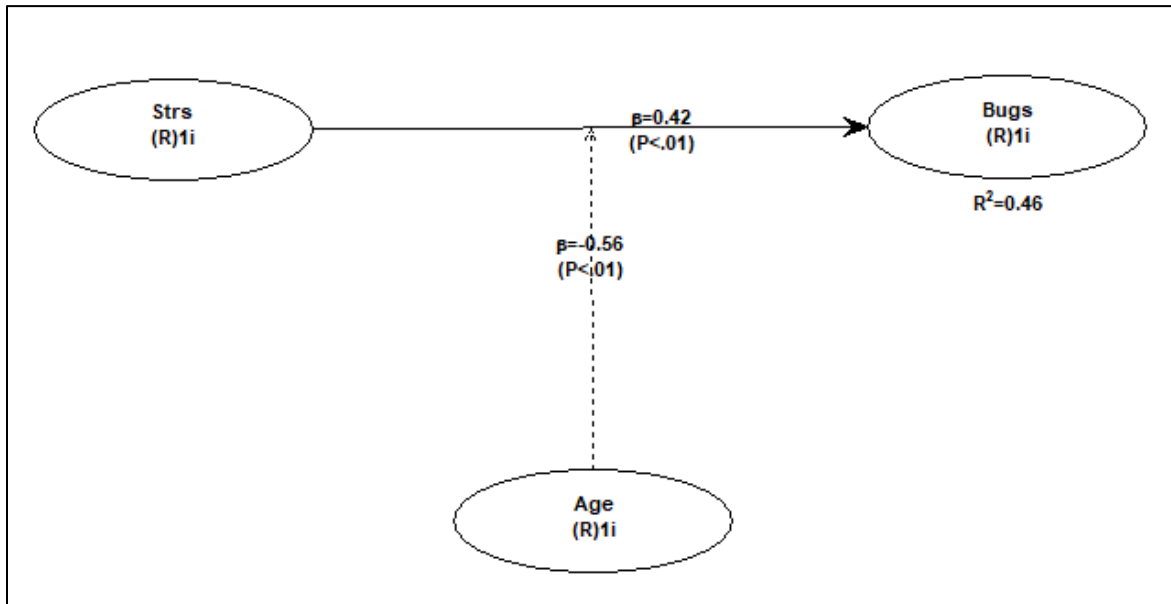
As we can see, the highest full collinearity VIF in the model is 1.855, well below the conservative threshold of 3.3, which allows us to conclude that all of the latent variables in the model measure different things. That is, the latent variables in the model measure constructs that appear to be conceptually different from one another.

Note that the full collinearity VIF for the moderating effect, noted as the product (interaction) variable Age\*Strs, is also included in this collinearity assessment. This is done because we want to ensure that the product variable (i.e., Age\*Strs) also measures something that is conceptually different from the other variables.

### 2.4.4. Inspect path coefficients

As soon as the SEM analysis is completed, the software shows some of the main results in graphical format on a window. This graphical representation is shown in Figure 2.4.4. The graph with the results shows path coefficients, respective P values, and R-squared coefficients. Users can also show or hide indicators weights, loadings and names.

Figure 2.4.4. Inspect path coefficients



The path coefficients are noted as beta coefficients. “Beta coefficient” is another term often used to refer to path coefficients in SEM analyses; this term is commonly used in multiple regression analyses as well. The P values are displayed below the path coefficients, within parentheses. The R-squared coefficients are shown below each endogenous latent variable (i.e., a latent variable that is hypothesized to be affected by one or more other latent variables), and reflect the percentage of the variance in the latent variable that is explained by the latent variables that are hypothesized to affect it.

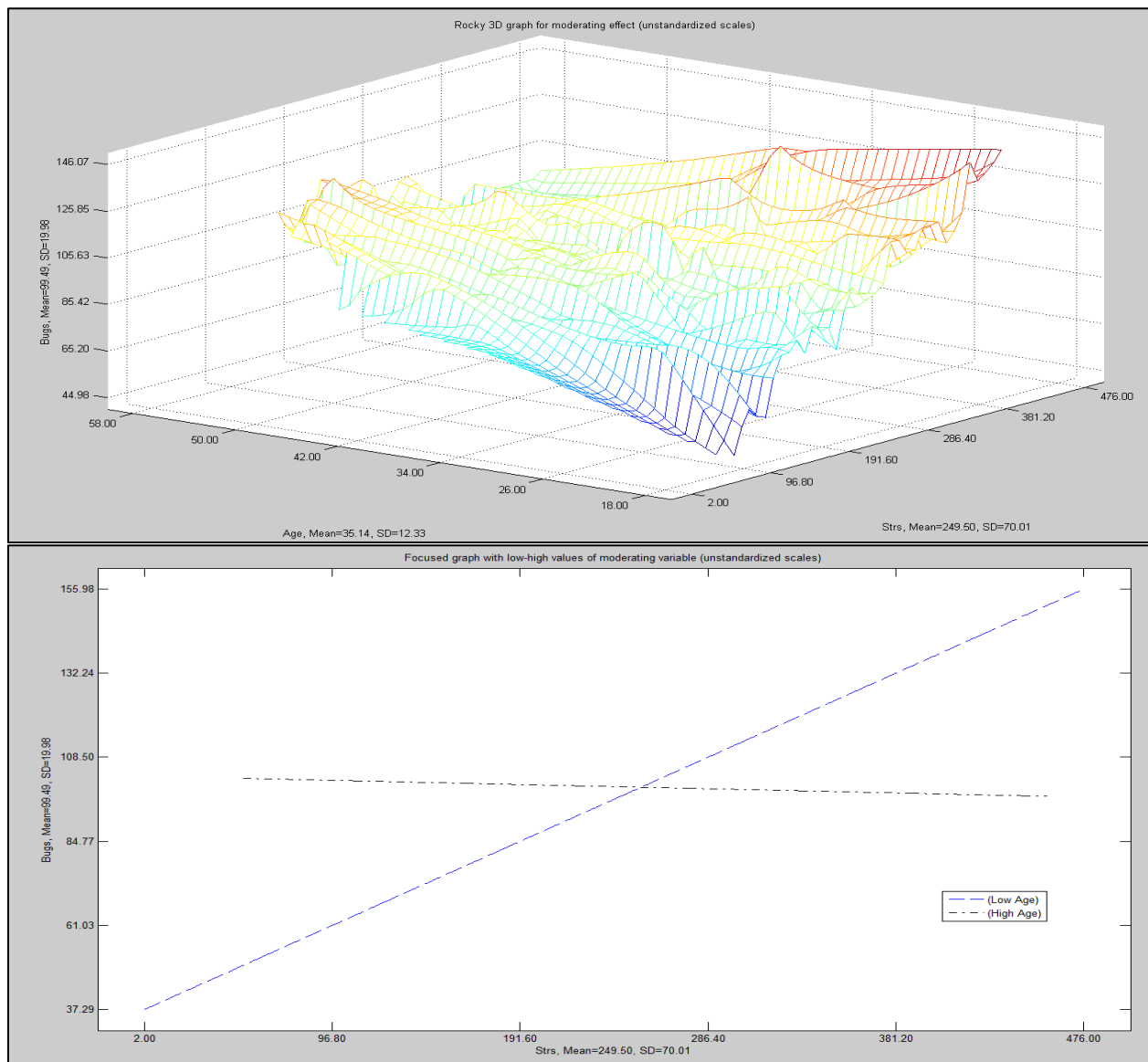
Path coefficients associated with P values equal to or lower than 0.05 are deemed to refer to real effects, as opposed to effects that are to be interpreted as “zero”. In this sense, all of the path coefficients in this model refer to effects that appear to be real. The strength of a path coefficient depends on its absolute value; i.e., the sign is disregarded.

The strongest path coefficient is for the moderating link Age > (Strs > Bugs), and this coefficient is negative; the higher is the age of the software developer, the weaker is the positive association between the developer’s stress and the number of bugs in the software. The other path coefficient is for the direct link Strs > Bugs, and this coefficient is positive; the higher the developer’s stress, the higher is number of bugs in the software.

### 2.4.5. Inspect graphs

As soon as the SEM analysis is completed, the software shows the results in graphical format on a window, where the menu option menu option “**View/plot linear and nonlinear relationships among latent variables**” becomes available. Two of the graphs that are particularly useful in analyses like this are available, under this option, from the options “**View rocky 3D graph for moderating effect (unstandardized scales)**” and “**View focused graph with low-high values of moderating variable (unstandardized scales)**”. These graph representations are respectively shown at the top and bottom of Figure 2.4.5; for the moderating link Age > (Strs > Bugs).

Figure 2.4.5. Inspect graphs



Unlike graphs for direct links, graphs for moderating links show how the strength of a direct link (Strs > Bugs) varies as the moderating link variable (Age) goes from low to high. strength of a direct link is reflected in its inclination, whether it is positive or negative. The top graph is a 3D graph that shows that, as age goes from 18 to 58, the inclination, of the (Strs > Bugs) relationship goes from steep (right part of the graph) to almost flat (left part of the graph). The bottom graph is a 2D graph that shows the same, but now with the sample split into “Low Age” and “High Age”.

So, clearly Age has a significant moderating effect on the (Strs > Bugs) relationship. But, at this point the reader may be wondering whether Age is directly associated with Bugs. We can check that by inspecting the table containing correlations among latent variables and square roots of AVEs, which is available under the “**View correlations among latent variables and errors**” menu option. In our case, the correlation between the variables Age and Bugs is a very small 0.041, which is not statistically significant. This allows us to conclude the Age has no influence on Bugs, even though it strongly influences the (Strs > Bugs) relationship.



### **2.4.6. Provide advice**

One of the goals of the analysis was to answer two questions: Does stress have an impact on the number of software bugs? If yes to the above, does age moderate the impact of stress on the number of bugs? The inspection of path coefficients suggests that stress does have an impact on the number of software bugs, and that age does indeed moderate the impact of stress on the number of bugs. Moreover, the inspection of correlations among latent variables suggests that age is not directly associated with the number of bugs, even though it is indirectly related as a moderator.

The analysis also provides support for the belief, earlier based on anecdotal evidence from the large software developer, that older software developers tend to perform well under stress, an effect that our analysis helped assess and quantify. This provides the basis for the recommendation that the company may start hiring more older software developers, which would go counter the industry trend of hiring mostly younger employees. Since this trend generally violates age discrimination laws, following the recommendation could make the company a “best place to work” that is also in compliance with the law, with other related advantages: it would potentially attract labor talent that was not being utilized due to a stereotype, and generate positive publicity for the company. All of these would likely make the company more competitive, which would ultimately lead to relative gains in sales and profits compared with other companies in the same industry.

## 2.5. Deciding on a mall location to establish a hot dog kiosk

Exhibit 2.5 displays the scenario, question, and variables related to the sample dataset used to illustrate how MDDA can be used to address the need of a new vegan hot dog producer to establish a kiosk in one of two malls, in the north or in the south part of a city, with the goal of maximizing sales and profits.

### Exhibit 2.5. Scenario, question, and variables

#### Scenario

- A new vegan hot dog producer wants to establish a kiosk in one of two malls, in the north or in the south part of a city.
- To test the market the producer provides free samples in both malls, and asks customers how much they would be willing to pay for the hot dog.

#### Questions

- In which mall, north or south, should the kiosk be established?
- At what price should the hot dog be sold?

#### Variables

- MallN1S0: Location of data collection: 1=North Mall, 0=South Mall.
- Paymt: Amount customer is willing to pay for vegan hot dog, measured in dollars.

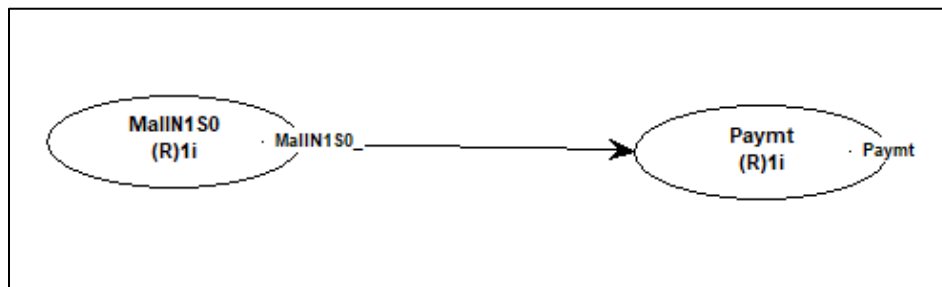
Here the MDDA analysis is aimed at finding out the best mall, north or south, where the vegan hot dog kiosk should be established; and at what price the hot dog should be sold. The variables used in this analysis are MallN1S0 (the location of data collection: 1=North Mall, 0=South Mall) and Paymt (the amount a customer is willing to pay for the vegan hot dog, measured in dollars).

The importance of this analysis comes from the need of the vegan hot dog producer to establish a kiosk in one of two malls, in the north or in the south part of a city, but not in both malls (at least not right away). The expectation of the vegan hot dog producer is that choosing the right mall will maximize the sales and profits that they can get in the city. This is particularly important because, as a new company, they have limited resources; and, therefore, cannot afford costly mistakes early on in their business cycle.

## 2.5.1. Create the model

Figure 2.5.1 shows the model that was built to serve as the basis for our analysis. It contains one predictor latent variable, MallN1S0 (location of data collection: 1=North Mall, 0=South Mall); and one criterion latent variable, namely Paymt (amount a customer is willing to pay for the vegan hot dog, measured in dollars). The latent variables have only one indicator each, which essentially means that they are assumed to be measured through their single indicators without error.

Figure 2.5.1. Create the model



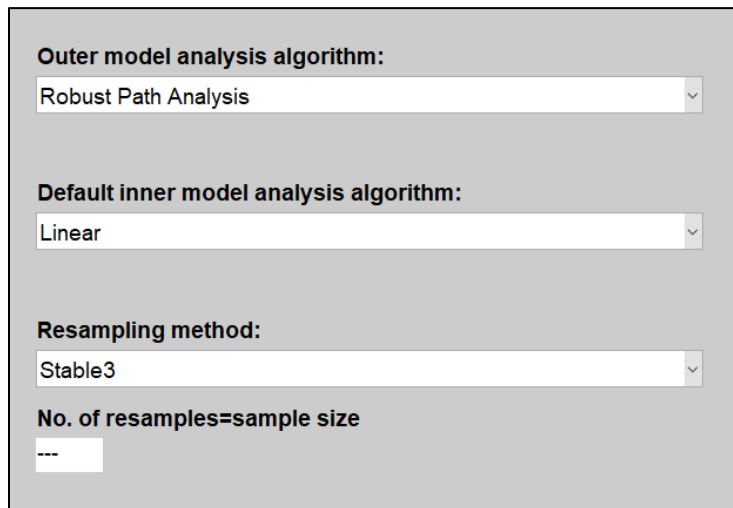
The assumption that a latent variable is measured without error may not be entirely correct, and frequently may be made for convenience. Whenever possible, multiple indicators should be used, because that enables SEM to minimize the effect of measurement error on the parameters being estimated (e.g., path coefficients). If that is not possible, data analysts have to do the best that they can with what they have available, recognizing the limitations of what they are doing. In this particular analysis, no perception-based variables were used, which mitigates the problem possibly caused by using only one indicator per latent variable.

As you can see, the latent variables have the same names as their indicators. This is not a requirement. The names could have been different. In fact, they will typically be different for latent variables that are measured through multiple indicators. They will also be different if the indicators' names are longer than 8 characters, which is the maximum allowed for latent variables names. This limitation is to give model graphs a cleaner look.

## 2.5.2. Choose general settings

The options shown in Figure 2.5.2 were the ones chosen for this analysis. They are common in analyses that employ latent variables that are all measured through single indicators, as is the case in our model. The options can be selected through the “**View or change general settings**” menu option.

Figure 2.5.2. Choose general settings



The screenshot shows a settings dialog box with the following options:

- Outer model analysis algorithm:** Robust Path Analysis
- Default inner model analysis algorithm:** Linear
- Resampling method:** Stable3
- No. of resamples=sample size:** ---

The **Robust Path Analysis** outer model analysis algorithm is a simplified algorithm with very good computational efficiency. The **Linear** default inner model analysis algorithm does not perform any warping of relationships; that is, it does not model the relationships as nonlinear. This helps with the interpretation of the results. The **Stable3** method is the default resampling method of the software, because of its high accuracy and robustness to deviations of normality. An advantage of this method is that it does not assume that the data is normally distributed, which is often the case with empirical data. That is, empirical data is typically *not* normally distributed, even though many data analysis techniques assume that it is.

### 2.5.3. Assess collinearity

As soon as the SEM analysis is completed, the software shows the results in graphical format on a window, which also contains a number of menu options that allow you to view and save more detailed results. One of these menu options is the “**View latent variable coefficients**”, which shows the results in the table in Figure 2.5.3. The last row of the table in the figure shows the **full collinearity VIFs** for all latent variables in the model.

Figure 2.5.3. Assess collinearity

	MallN1S0	Paymt
R-squared		0.353
Adj. R-squared		0.351
Composite reliab.	1.000	1.000
Cronbach's alpha	1.000	1.000
Avg. var. extrac.	1.000	1.000
Full collin. VIF	1.545	1.545

**Full collinearity VIFs of 3.3 or lower suggest** the existence of **no multicollinearity** in the model. A more **relaxed threshold would be 5**. This means that all of the latent variables in the model measure different things, which is an important precondition for a valid analysis. **Full collinearity VIFs of 10 or higher suggest** the existence of **multicollinearity** in the model. Multicollinearity at this level, with full collinearity VIFs of 10 or higher, tends to distort coefficients of association, such as path coefficients.

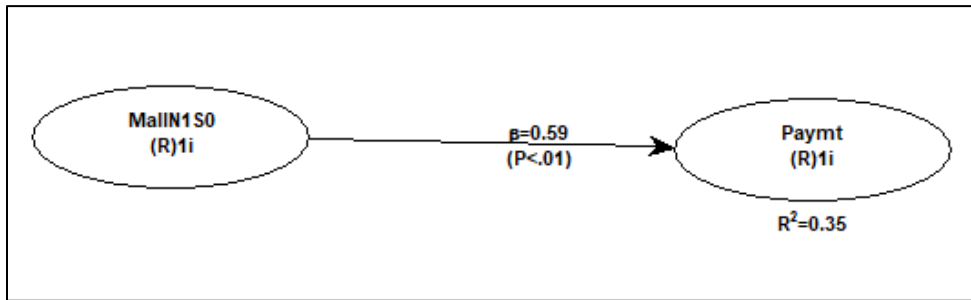
As we can see, the highest full collinearity VIF in the model is 1.545, well below the conservative threshold of 3.3, which allows us to conclude that all of the latent variables in the model measure different things. That is, the latent variables in the model measure constructs that appear to be conceptually different from one another.

Note that the full collinearity VIFs are identical in this case, because the model has only two variables. This is due to the way in which full collinearity VIFs are calculated, namely as a function of the variance explained in each variable by all of the other variables in the model. In a model with only two variables, that variance explained will be the same for both variables.

### 2.5.4. Inspect path coefficients

As soon as the SEM analysis is completed, the software shows some of the main results in graphical format on a window. This graphical representation is shown in Figure 2.5.4. The graph with the results shows path coefficients, respective P values, and R-squared coefficients. Users can also show or hide indicators weights, loadings and names.

Figure 2.5.4. Inspect path coefficients



The path coefficients are noted as beta coefficients. “Beta coefficient” is another term often used to refer to path coefficients in SEM analyses; this term is commonly used in multiple regression analyses as well. The P values are displayed below the path coefficients, within parentheses. The R-squared coefficients are shown below each endogenous latent variable (i.e., a latent variable that is hypothesized to be affected by one or more other latent variables), and reflect the percentage of the variance in the latent variable that is explained by the latent variables that are hypothesized to affect it.

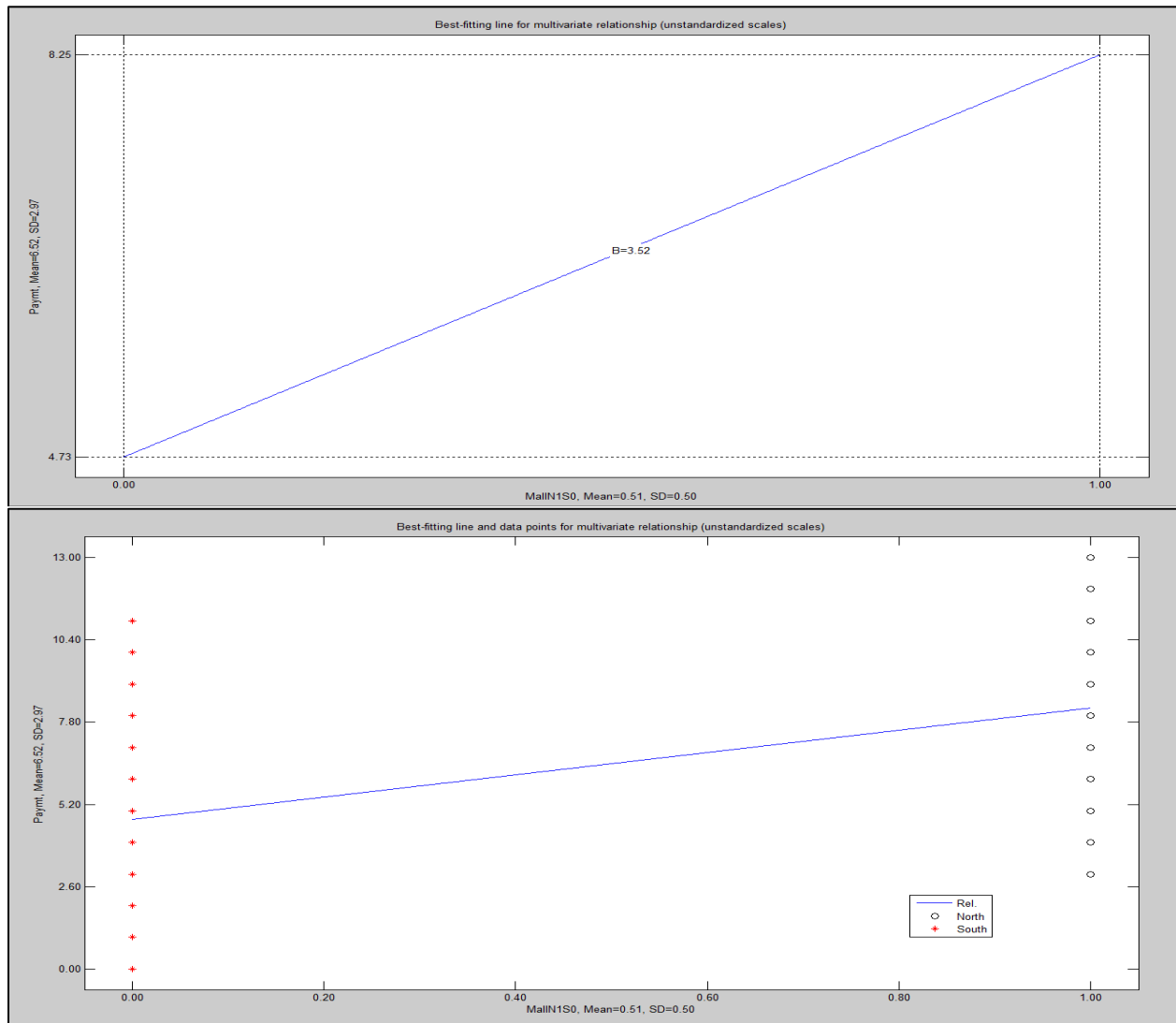
Path coefficients associated with P values equal to or lower than 0.05 are deemed to refer to real effects, as opposed to effects that are to be interpreted as “zero”. In this sense, all of the path coefficients in this model refer to effects that appear to be real. The strength of a path coefficient depends on its absolute value; i.e., the sign is disregarded.

The path coefficient for the only link in the model, namely the link  $MallN1S0 > Paymt$ , is fairly strong, statistically significant, and positive. As we know, the variable MallN1S0 was dummy coded as: 1=North Mall, 0=South Mall. This means that the positive coefficient is interpreted as the north mall being associated with higher values of Paymt than the south mall. Stated differently, customers are willing to pay significantly more for the vegan hot dog in the north mall than in the south mall.

### 2.5.5. Inspect graphs

As soon as the SEM analysis is completed, the software shows the results in graphical format on a window, where the menu option menu option “**View/plot linear and nonlinear relationships among latent variables**” becomes available. Two of the graphs that are particularly useful are available, under this option, from the options “**View focused multivariate relationship graph with segments (unstandardized scales)**” and “**View multivariate relationship graphs with data points and legends (unstandardized scales)**”. These graph representations are respectively shown at the top and bottom of Figure 2.5.5; for the link MallN1S0 > Paymt.

Figure 2.5.5. Inspect graphs



Unlike a path coefficient, which is standardized, the graph at the top shows the corresponding unstandardized regression coefficient. This type of coefficient tends to be more telling to stakeholders. In this case, the unstandardized regression coefficient for the link MallN1S0 >

Paymt is 3.52. The meaning of this is that, for each increase of 1 point in the variable MallN1S0 (location of data collection: 1=North Mall, 0=South Mall), there is a corresponding increase of 3.52 points in the variable Paymt (amount a customer is willing to pay for the vegan hot dog, measured in dollars). In other words, customers are willing to pay 3.52 dollars more, on average, for the vegan hot dog in the north mall than in the south mall.

The graph at the bottom illustrates the different distribution of answers for the North and South malls, and also uses unstandardized scales. Note that it provides the means and standard deviations for both variables, next to the corresponding axes. The mean for Paymt is listed as 6.52 dollars. This graph employs a data label variable, selected with the “**Settings**” option available from the graph menu options.

**Data labels** are text identifiers that are entered by you separately, through one of the “**Modify**” menu options. Like the original numeric dataset, the data labels are stored in a table. Each column of this table refers to one data label, and each row to the corresponding row of the original numeric dataset. Data labels can be shown on graphs, either next to each data point that they refer to, or as part of a graph’s legend (as done here).

The “**Modify**” menu options allow you to add new data labels to your dataset. Data labels can be read from the clipboard or from a file, but **only one column of labels can be read at a time**. Data label cells cannot be empty, contain spaces, or contain only numbers; **they must be combinations of letters, or of letters and numbers**. Valid examples are the following: “Age>17”, “Y2001”, “AFR”, and “HighSuccess”. These would normally be entered without the quotation marks, which are used here only for clarity. Some invalid examples: “123”, “Age > 17”, and “Y 2001”.



### **2.5.6. Provide advice**

One of the goals of the analysis was to answer two questions: In which mall, north or south, should the kiosk be established? At what price should the hot dog be sold? The inspection of path coefficients suggests that customers are willing to pay more for the vegan hot dog in the north mall than in the south mall, so the answer to the first question is that the kiosk should be in the north mall. The inspection of graphs suggests that the mean for the Paymt variable is 6.52 dollars. Arguably a good answer to the second question is 6.52 dollars, because this is the mean for both malls and thus a competitive price for the north mall.

The expectation of the vegan hot dog producer is that choosing the right mall will maximize the sales and profits that they can get in the city. If demand is high for the hot dogs at the 6.52 dollars price, it would make sense to incrementally increase that price until demand starts going down, even if input costs (for the hot dog production) are stable. In this scenario, profits margins will gradually increase, allowing the vegan hot dog producer to expand, with additional kiosks in the north mall – and even possibly the south mall.

## 2.6. Organizing grocery store items to increase sales

Exhibit 2.6 displays the scenario, question, and variables related to the sample dataset used to illustrate how MDDA can be used to address the need of a small grocery store to organize several of the items it sells by placing the items that are purchased together near one another, with the goal of maximizing sales and profits.

### Exhibit 2.5. Scenario, question, and variables

#### Scenario

- A small grocery store carries a number of items, groups of which tend to be purchased together.
- Organizing the grocery store by placing items that are purchased together near one another is believed to increase sales.
- A decision was made to compile data on purchases of items over a period of time.
- Data on 550 purchases was compiled.

#### Question

- How should the items be organized in the grocery store?

#### Variables

- Milk: Milk purchase, in dollars.
- Cheese: Cheese purchase, in dollars.
- Eggs: Eggs purchase, in dollars.
- Sardine: Canned sardine purchase, in dollars.
- Tuna: Canned tuna purchase, in dollars.
- Chicken: Canned chicken purchase, in dollars.
- Chips: Bagged chips purchase, in dollars.
- Candy: Candy purchase, in dollars.
- Soda: Soda purchase, in dollars.
- Pear: Pear purchase, in dollars.
- Banana: Banana purchase, in dollars.
- Apple: Apple purchase, in dollars.
- Plate: Plastic plate purchase, in dollars.
- Spoon: Plastic spoon purchase, in dollars.
- Knife: Plastic knife purchase, in dollars.

Here the MDDA analysis is aimed at finding out how the items should be organized in the grocery store. The variables used in this analysis, which store data about each purchase made at the store, are: Milk (milk purchase, in dollars), Cheese (cheese purchase, in dollars), Eggs (eggs purchase, in dollars), Sardine (canned sardine purchase, in dollars), Tuna (canned tuna purchase, in dollars), Chicken (canned chicken purchase, in dollars), Chips (bagged chips purchase, in dollars), Candy (candy purchase, in dollars), Soda (soda purchase, in dollars), Pear (pear purchase, in dollars), Banana (banana purchase, in dollars), Apple (apple purchase, in dollars), Plate (plastic plate purchase, in dollars), Spoon (plastic spoon purchase, in dollars), and Knife (plastic knife purchase, in dollars).

The importance of this analysis comes from the need for the small grocery store to physically organize the items that it sells, and whose sales are measured by the variables above, by placing the items that are purchased together near one another. This is perceived by the store's management, as likely to increase sales, and thus absolute profits. This belief held by the store's management is based on anecdotal data, namely the observation that customers tend to buy more

## Model-Driven Data Analytics: Applications with WarpPLS

of items that are usually purchased together (e.g., dairy items) if they are placed near one another in the store.

### 2.6.1. Inspect indicator correlations

Figure 2.6.1 shows the correlations among several of the indicators that we intend to group into latent variables in a model. They are available from the menu option “**View or save correlations and descriptive statistics for indicators**”, which is under the “**Data**” menu option. These correlations among indicators are available prior to the model being created, which is necessary for this specific application.

**Figure 2.6.1. Create the model**

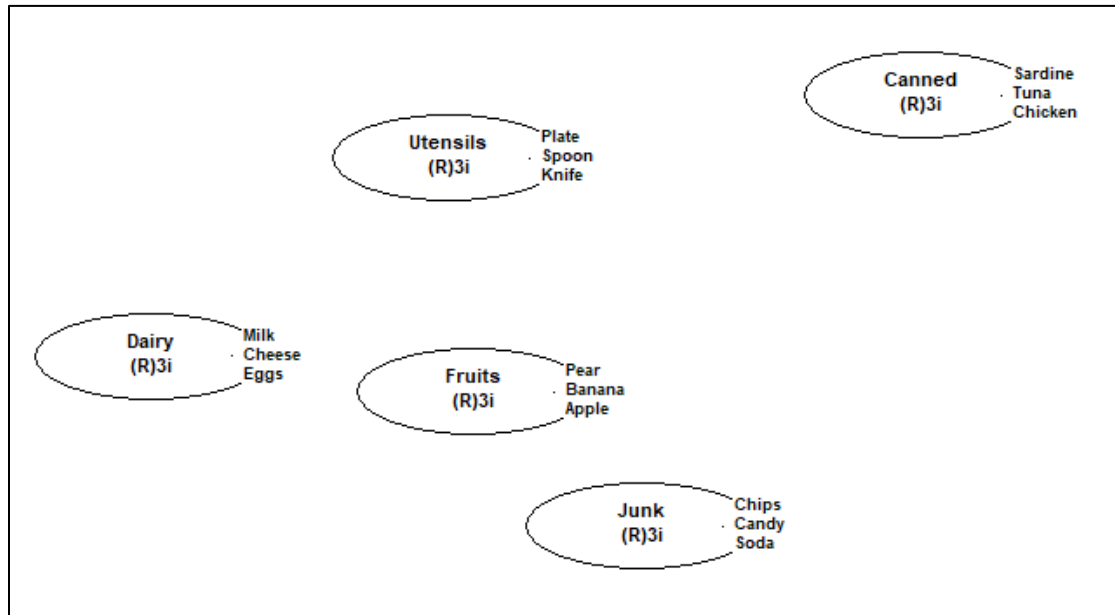
	Milk	Cheese	Eggs	Sardine	Tuna	Chicken	Chips	Candy	Soda	
Milk	1.000	0.625	0.642	0.151	0.127	0.068	-0.206	-0.223	-0.225	0.4
Cheese	0.625	1.000	0.625	0.163	0.151	0.116	-0.236	-0.196	-0.228	0.3
Eggs	0.642	0.625	1.000	0.084	0.112	0.050	-0.254	-0.235	-0.238	0.3
Sardine	0.151	0.163	0.084	1.000	0.610	0.622	-0.089	-0.078	-0.073	0.0
Tuna	0.127	0.151	0.112	0.610	1.000	0.617	-0.098	-0.125	-0.075	0.1
Chicken	0.068	0.116	0.050	0.622	0.617	1.000	-0.084	-0.107	-0.059	0.0
Chips	-0.206	-0.236	-0.254	-0.089	-0.098	-0.084	1.000	0.600	0.603	-0.
Candy	-0.223	-0.196	-0.235	-0.078	-0.125	-0.107	0.600	1.000	0.586	-0.
Soda	-0.225	-0.228	-0.238	-0.073	-0.075	-0.059	0.603	0.586	1.000	-0.

What we will do here is to **select as indicators or each latent variable** those that have **correlations among themselves** that are **equal to or greater than 0.5**. For example, Milk (milk purchase, in dollars), Cheese (cheese purchase, in dollars), and Eggs (eggs purchase, in dollars) have the following correlations among themselves: Milk <> Milk (1.000), Milk <> Cheese (0.625), and Milk <> Cheese (0.642). Note that the correlation of any variable with itself (Milk <> Milk) is 1.000.

## 2.6.2. Create the model

Figure 2.6.2 shows the model that was built to serve as the basis for our analysis. It contains five latent variables, which were created to aggregate the indicators. The aggregation of indicators into latent variables was based on the indicator correlations. As previously noted, indicators that were found to have correlations among themselves that were equal to or greater than 0.5 were aggregated into the same latent variables.

Figure 2.6.2. Create the model

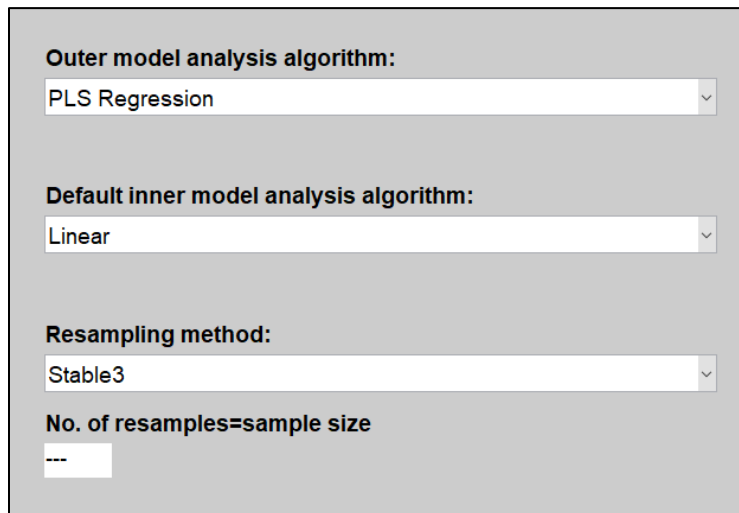


The latent variables created were the following: Dairy (indicators: Milk, Cheese, and Eggs), Canned (indicators: Sardine, Tuna, and Chicken), Junk (indicators: Chips, Candy, and Soda), Fruits (indicators: Pear, Banana, and Apple), and Utensils (indicators: Plate, Spoon, and Knife). In this initial version of our model, the latent variables are not yet placed in any particular order. That will be done later, after inspection of the correlations among latent variables.

### 2.6.3. Choose general settings

The options shown in Figure 2.6.3 were the ones chosen for this analysis. They are common in exploratory analyses that employ one or more latent variables that are measured through multiple indicators, as is the case in our model. The options can be selected through the “**View or change general settings**” menu option.

Figure 2.6.3. Choose general settings



The screenshot shows a settings dialog box with a light gray background. It contains four sections, each with a label and a dropdown menu:

- Outer model analysis algorithm:** The dropdown menu is open, showing "PLS Regression" as the selected option.
- Default inner model analysis algorithm:** The dropdown menu is open, showing "Linear" as the selected option.
- Resampling method:** The dropdown menu is open, showing "Stable3" as the selected option.
- No. of resamples=sample size:** The dropdown menu is open, showing "---" as the selected option.

**PLS Regression** has been the default outer model algorithm since the software’s inception. This algorithm iterates by making the outer model weights directly proportional to the loadings, until the weights become stable. This algorithm does not let the inner model influence the outer model. The **Linear** default inner model analysis algorithm does not perform any warping of relationships; that is, it does not model the relationships as nonlinear. This helps with the interpretation of the results. The **Stable3** method is the default resampling method of the software, because of its high accuracy and robustness to deviations of normality. An advantage of this method is that it does not assume that the data is normally distributed, which is often the case with empirical data. That is, empirical data is typically *not* normally distributed, even though many data analysis techniques assume that it is.

### 2.6.4. Inspect latent variable correlations

The “**View correlations among latent variables and errors**” menu options allow users to view tables containing correlations among latent variables, the P values associated with those correlations, square roots of average variances extracted (AVEs), correlations among latent variable error terms (or residuals), and the VIFs associated with latent variable error terms. The table containing correlations among latent variables and square roots of AVEs (see Figure 2.6.4) is typically the one used here, for this type of application.

**Figure 2.6.4. Inspect latent variable correlations**

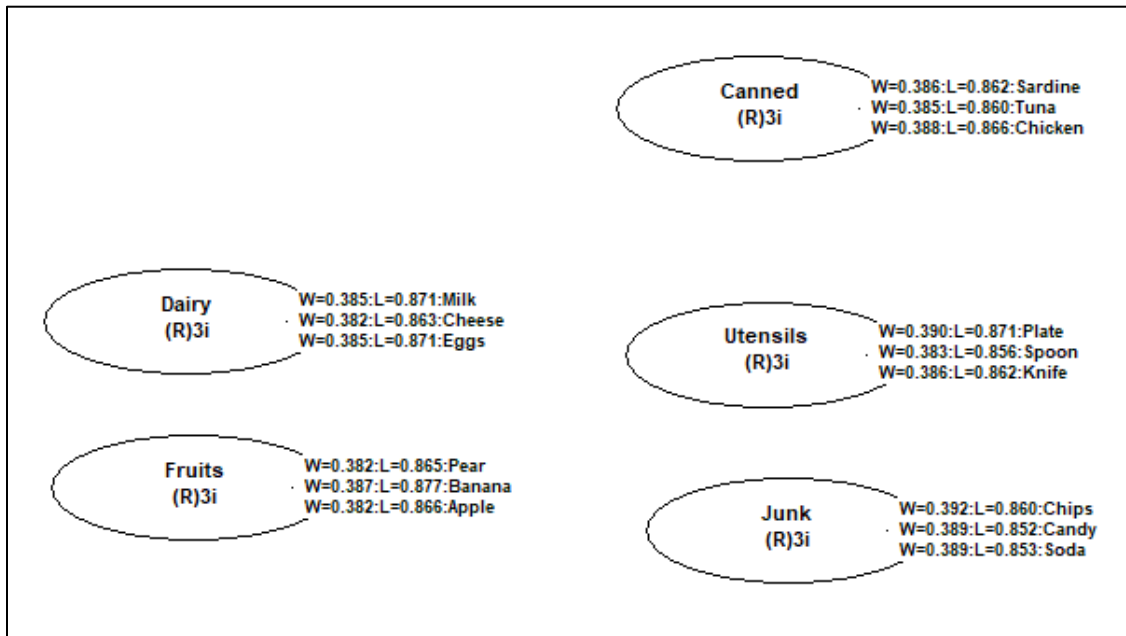
Correlations among l.vs. with sq. rts. of AVEs					
	Dairy	Canned	Junk	Fruits	Utensils
Dairy	(0.868)	0.151	-0.306	0.520	-0.454
Canned	0.151	(0.863)	-0.119	0.085	0.372
Junk	-0.306	-0.119	(0.855)	-0.198	0.415
Fruits	0.520	0.085	-0.198	(0.869)	-0.441
Utensils	-0.454	0.372	0.415	-0.441	(0.863)

What we will do here is to use these correlations to place the latent variables next to one another based on their correlations. Here correlation signs matter. If two items are positively correlated, then tend to be purchased together; if the correlation is negative, they tend not to be purchased together. The most important correlations are the following, selected from each column as we move from left to right: Dairy <> Fruits (0.520), Canned <> Utensils (0.372), Junk <> Utensils (0.415), Fruits <> Dairy (0.520).

### 2.6.5. Adjust the model

In Figure 2.6.5 we show how we placed the latent variables next to one another based on their correlations. Again, here correlation signs matter, and the most important correlations are: Dairy  $\diamond$  Fruits (0.520), Canned  $\diamond$  Utensils (0.372), Junk  $\diamond$  Utensils (0.415), Fruits  $\diamond$  Dairy (0.520). This configuration assumed a particular store configuration. For example, here we assumed that the entrance to the store is near “Junk”, where there is a shelving unit with shelves on both sides, and that the cash register stand is behind “Dairy” and “Fruits”.

Figure 2.6.5. Adjust the model



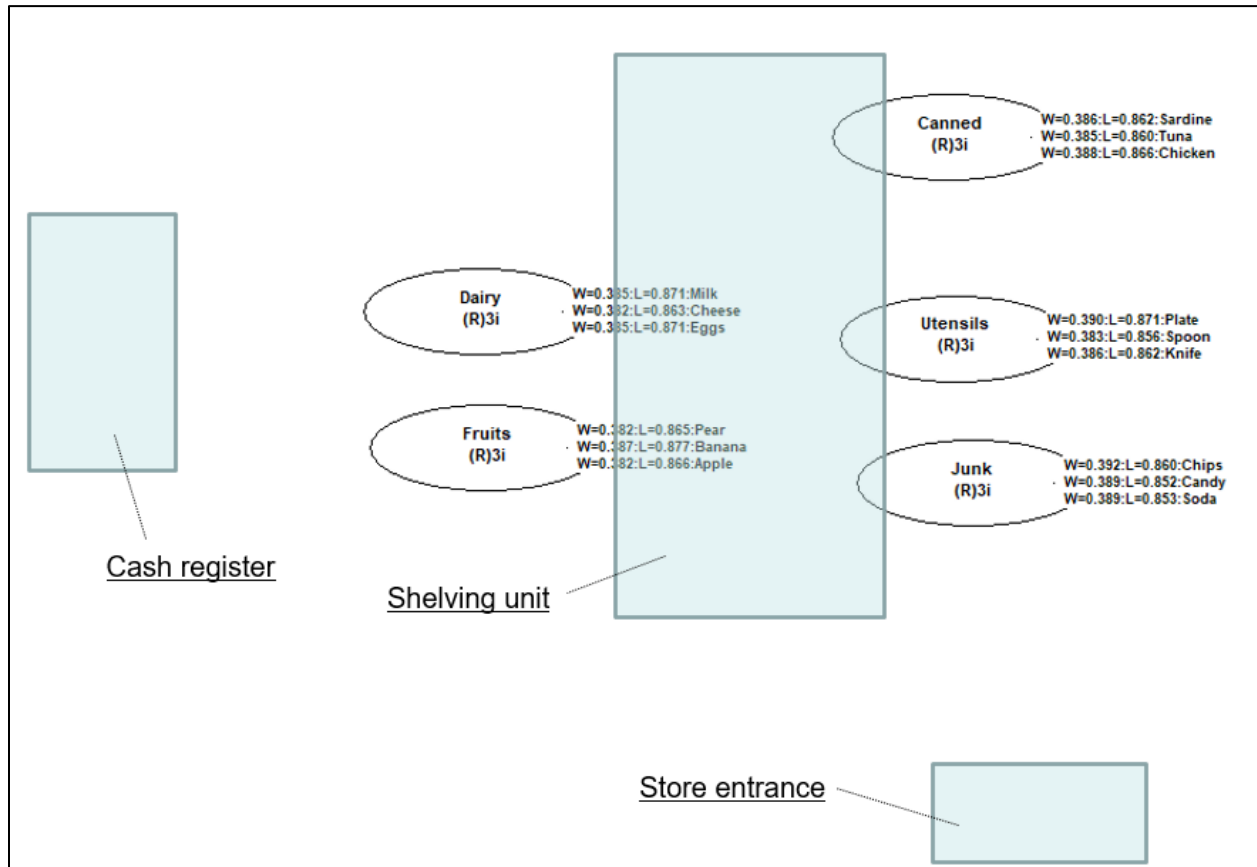
It is also a good idea to check the loadings of the indicators on the latent variables, which we can do by using the show/hide indicators option of the model graph menu. The loadings are listed after the “L=” symbol. Ideally the loadings will all be 0.5 or higher. Lower loadings may suggest that the corresponding indicators should not have been grouped in the way they were.



### 2.6.6. Add contextual information

In Figure 2.6.6 we show how we placed the latent variables next to one another based on their correlations, and also show elements that suggest how the store was configured. This makes it easier for the main customer of the analysis, the store manager, to visualize how the store will look like if it is configured consistently with the results of the analysis.

Figure 2.6.6. Add contextual information



For example, the entrance to the store that is near “Junk” is now more clearly indicated. The same is true for the shelving unit, with shelves on both sides. Finally, the cash register stand location is now more clearly indicated as well, behind “Dairy” and “Fruits”. Ultimately, the physical configuration of the items is the main recommendation that comes from the analysis.

## **Part 3: Concluding remarks**

### **3.1. Applied model-driven data analytics**

This document described the technique of MDDA; which involves the creation of a path model expressing an applied theory, and testing the model using path analysis with latent variables. The latter, path analysis with latent variables, is generally known as SEM.

The applications presented in this document showed how MDDA can be employed in a variety of different contexts, where typically data is collected from organizations with the goal of answering questions that ultimately affect the ability of the organizations to grow their sales and profits.

The applications followed a similar set of steps. There was repetition across applications, of both steps and the text that describes them. This repetition was aimed at helping with the internalization by analysts (i.e., learning) of complex concepts and techniques, while at the same time making each application section fairly self-contained.

### **3.2. Use of simulated data**

Some of the data discussed in this document have been compiled based on publicly available sources, some have been created via Monte Carlo simulations based on empirical studies, and some have been produced as a mix of both approaches. For ethical reasons, and to protect individual privacy, all of the individual-level data have been created via Monte Carlo simulations, based on empirical studies – to mimic what happened with real data.

# Glossary

This glossary includes terms that go beyond those used in the applications discussed in this document. We are including this extended set of terms here because some readers may want to go beyond the features discussed in the applications, and explore other more advanced features that refer to some of the terms in this extended set.

**Adjusted R-squared coefficient.** A measure equivalent to the R-squared coefficient, with the key difference that it corrects for spurious increases in the R-squared coefficient due to predictors that add no explanatory value in each latent variable block. Like R-squared coefficients, adjusted R-squared coefficients can assume negative values. These are rare occurrences that normally suggest problems with the model in which they occur; e.g., severe collinearity or model misspecification.

**Analytic composites.** Analytic composites are weighted aggregations of indicators where the relative weights are set by the user, usually based on an existing theory.

**Average variance extracted (AVE).** A measure associated with a latent variable, which is used in the assessment of the discriminant validity of a measurement instrument. Less commonly, it can also be used for convergent validity assessment.

**Composite reliability coefficient.** This is a measure of reliability associated with a latent variable. Another name for it is Dillon–Goldstein rho coefficient. Unlike the Cronbach’s alpha coefficient, another measure of reliability, the composite reliability coefficient takes indicator loadings into consideration in its calculation. It often is slightly higher than the Cronbach’s alpha coefficient.

**Constrained latent growth.** The constrained latent growth method is essentially the same method as that employed in a full latent growth analysis with the difference that here it is constrained to a sub-sample, typically formed by two groups being compared. This method is normally used in multi-group analyses, whereby the dataset is segmented into various groups, all possible combinations of pairs of groups are generated, and each pair of groups is compared.

**Construct.** A conceptual entity measured through a latent variable. Sometimes it is referred to as “latent construct”. The terms “construct” or “latent construct” are often used interchangeably with the term “latent variable”.

**Convergent validity of a measurement instrument.** Convergent validity is a measure of the quality of a measurement instrument; the instrument itself is typically a set of question-statements. A measurement instrument has good convergent validity if the question-statements (or other measures) associated with each latent variable are understood by the respondents in the same way as they were intended by the designers of the question-statements.

**Cronbach’s alpha coefficient.** This is a measure of reliability associated a latent variable. It usually increases with the number of indicators used, and is often slightly lower than the composite reliability coefficient, another measure of reliability.

**Discriminant validity of a measurement instrument.** Discriminant validity is a measure of the quality of a measurement instrument; the instrument itself is typically a set of question-statements. A measurement instrument has good discriminant validity if the question-statements (or other measures) associated with each latent variable are not confused by the respondents, in terms of their meaning, with the question-statements associated with other latent variables.

**Effect size.** The effect size is a measure of the magnitude of an effect that is independent of the size of the sample analyzed. The effect sizes are calculated by this software as the absolute

values of the individual contributions of the corresponding predictor latent variables to the R-squared coefficients of the criterion latent variable in each latent variable block. With the effect sizes users can ascertain whether the effects indicated by path coefficients are small, medium, or large. The values usually recommended are 0.02, 0.15, and 0.35; respectively. Values below 0.02 suggest effects that are too weak to be considered relevant from a practical point of view, even when the corresponding P values are statistically significant; a situation that may occur with large sample sizes.

**Endogeneity.** The term “endogeneity” refers to a phenomenon that is characterized by the structural error term for an endogenous variable being correlated with any of the variable’s predictors. For example, let us consider a simple population model with the following links  $A > B$  and  $B > C$ . This model presents endogeneity with respect to C, because variation flows from A to C via B, leading to a biased estimation of the path for the link  $B > C$  via ordinary least squares regression. Adding a link from A to C could be argued as “solving the problem”, but in fact it creates the possibility of a type I error, since the link  $A > C$  does not exist at the population level. A more desirable solution to this problem is to create an instrumental variable  $iC$ , incorporating only the variation of A that ends up in C and nothing else, and revise the model so that it has the following links:  $A > B$ ,  $B > C$  and  $iC > C$ . The link  $iC > C$  can be used to test for endogeneity, via its P value and effect size. This link (i.e.,  $iC > C$ ) can also be used to control for endogeneity, thus removing the bias when the path coefficient for the link  $B > C$  is estimated via ordinary least squares regression. Endogeneity may also arise from multilevel effects (Kock, 2020b).

**Endogenous latent variable.** This is a latent variable that is hypothesized to be affected by one or more other latent variables. An endogenous latent variable has one or more arrows pointing at it in the model graph.

**Exogenous latent variable.** This is a latent variable that does not depend on other latent variables, from a SEM analysis perspective. An exogenous latent variable does not have any arrow pointing at it in the model graph.

**Factor score.** A factor score is the same as a latent variable score; see the latter for a definition.

**Formative latent variable.** A formative latent variable is one in which the indicators are expected to measure certain attributes of the latent variable, but the indicators are not expected to be highly correlated with the latent variable score, because they (i.e., the indicators) are not expected to be correlated with one another. For example, let us assume that the latent variable “Satisf” (“satisfaction with a meal”) is measured using the two following question-statements: “I am satisfied with the main course” and “I am satisfied with the dessert”. Here, the meal comprises the main course, say, filet mignon; and a dessert, a fruit salad. Both main course and dessert make up the meal (i.e., they are part of the same meal) but their satisfaction indicators are not expected to be highly correlated with each other. The reason is that some people may like the main course very much, and not like the dessert. Conversely, other people may be vegetarians and hate the main course, but may like the dessert very much.

**Full collinearity VIFs.** Variance inflation factors (VIFs) are measures of the degree of collinearity (or multicollinearity) among variables, including both indicators and latent variables. With latent variables, collinearity can take two main forms: vertical and lateral collinearity. Vertical, or classic, collinearity is predictor-predictor latent variable collinearity in individual latent variable blocks. Lateral collinearity is a term that refers to predictor-criterion latent variable collinearity; a type of collinearity that can lead to particularly misleading results. Full collinearity VIFs allow for the simultaneous assessment of both vertical and lateral collinearity

in a SEM model. They can also be used for common method bias and discriminant validity assessment.

**Full latent growth.** Sometimes the actual inclusion of moderating variables and corresponding links in a model leads to problems; e.g., increases in collinearity levels, and the emergence of instances of Simpson's paradox. By using the full latent growth analysis method, users can completely avoid these problems. This method allows one to estimate the effects of a latent variable or indicator on all of the links in a model (all at once), without actually including any links between the variable and other variables in the model (Kock, 2020a). Moreover, growth in coefficients associated with links among different latent variables and between a latent variable and its indicators, can be estimated; allowing for measurement invariance tests applied to loadings and/or weights. Finally, growth coefficients can be used in the assessment of moderated mediation effects.

**Heterotrait-monotrait (HTMT) ratios.** These ratios, as well as the updated HTMT2 ratios, have been proposed for discriminant validity assessment, particularly in the context of composite-based SEM via classic PLS algorithms; as opposed to factor-based SEM via modern algorithms that estimate factors (which have been available from this software for quite some time now). Discriminant validity is a measure of the quality of a measurement instrument; the instrument itself is typically a set of question-statements. A measurement instrument has good discriminant validity if the question-statements (or other measures) associated with each latent variable are not confused by the respondents, in terms of their meaning, with the question-statements associated with other latent variables.

**Indicator.** The term indicator is frequently used as synonymous with that of manifest variable; a convention that is used here. Thus, see the latter for a definition. More technically though, indicators are manifest variables that are actually used in the measurement model as direct measures of latent variables. As such, technically speaking, there can be manifest variables that are not indicators, if the manifest variables in question are part of the original dataset but not included in the measurement model.

**Inner model.** In a SEM analysis, the inner model is the part of the model that describes the relationships among the latent variables that make up the model. In this sense, the path coefficients are inner model parameter estimates.

**Instrumental variable.** Instrumental variables are variables that selectively share variation with other variables, and only with those variables. Instrumental variables can be used to test and control for endogeneity, and also to estimate reciprocal relationships. Endogeneity may arise from multilevel effects.

**Latent growth.** Generally speaking, latent growth refers to underlying growth in coefficients associated with links among different latent variables and between a latent variable and its indicators. This underlying growth is often reflected in significant moderating and nonlinear effects.

**Latent variable.** A latent variable is a variable that is measured through multiple variables called indicators or manifest variables. For example, "satisfaction with a meal" may be a latent variable measured through two manifest variables that store the answers on a 1 to 7 scale (1=strongly disagree; 7 strongly agree) to the following question-statements: "I am satisfied with this meal", and "After this meal, I feel full".

**Latent variable block.** A latent variable block is a group of latent variables in which one or more predictor latent variables point at one criterion latent variable. In a PLS-based SEM analysis, once latent variable scores are calculated, a series of multiple least squares regressions

are conducted to calculate path coefficients. Each multiple least squares regression is performed on a latent variable block, until all blocks are covered. The term “latent variable block” is also used in the PLS-based SEM literature to refer to a group of manifest variables linked to their assigned latent variable; i.e., a latent variable and its indicators.

**Latent variable score.** Latent variable scores are values calculated based on the indicators defined by the user as associated with the latent variable. They are calculated using one of the outer model analysis algorithms available. These scores may be understood as new columns in the data, with the same number of rows as the original data (unless a range-restricted analysis is conducted), and which generally tend to maximize the loadings and minimize the cross-loadings of a pattern matrix of loadings after an oblique rotation.

**Latent variable error.** An error variable that accounts for the variance in an endogenous latent variable that is not accounted for by the latent variable predictors that point at the endogenous latent variable. The terms “error” and “residual” are used interchangeably in this document. Nevertheless, they refer to subtly different entities. Technically speaking, the term “error” typically refers to the error variable in the true population model, which is assumed to be uncorrelated with latent variables other than the endogenous latent variable to which it is associated. Conversely, the term “residual” typically refers to the corresponding estimated error, the difference between the expected value of the latent variable and its point estimate, which in practice is often correlated with latent variables other than the endogenous latent variable to which it is associated. This is an example of a broader occurrence in multivariate analyses: more often than not sample-specific estimates violate assumptions about the theoretical true values, even if slightly.

**Manifest variable.** A manifest variable is one of several variables that are used to indirectly measure a latent variable. For example, “satisfaction with a meal” may be a latent variable measured through two manifest variables, which assume as values the answers on a 1 to 7 scale (1=strongly disagree; 7 strongly agree) to the following question-statements: “I am satisfied with this meal”, and “After this meal, I feel full”.

**Minimum required sample size.** The minimum required sample size needed for an SEM test to achieve an acceptable level of power (usually .8) depends on the effect size associated with the path coefficient under consideration and the significance level used for hypothesis testing (normally 0.05). The higher is the magnitude of a path coefficient at the population level, the higher is usually its effect size, and the greater is the probability that a true effect will be properly detected with a small sample. Therefore strong path coefficients at the population level, whether they are negative or positive, tend to require very small sample sizes for their proper identification. This software allows users to obtain estimates of the minimum required sample sizes for empirical studies based on the following model elements: the minimum absolute significant path coefficient in the model (e.g., 0.21), the significance level used for hypothesis testing (e.g., 0.05), and the power level required (e.g., 0.80).

**Outer model.** In a SEM analysis, the outer model is the part of the model that describes the relationships among the latent variables that make up the model and their indicators. In this sense, the weights and loadings are outer model parameter estimates.

**Portable document format (PDF).** This is an open standard file format created by Adobe Systems, and widely used for exchanging documents. It is the format used for this software’s documentation.

**Power.** Statistical power, often referred to simply as “power”, is a statistical test’s probability of avoiding type II errors, or false negatives. Power is often estimated for a particular coefficient



of association and sample size, for samples drawn from a population, at a given significance level (usually  $P < .05$ ). For example, let us consider an SEM test employing PLS Mode A and bootstrapping. Let us assume that such a test is able to recognize a path coefficient as statistically significant, where the path coefficient is associated with a “real” effect at the population level of magnitude .2; which would be referred to as the “true” path coefficient. Let us also assume that the test correctly recognizes the path coefficient as significant 83 percent of the time when samples of size 150 are randomly taken from the population. Under these circumstances, we would conclude that the power of the test is 83 percent, or .83.

**Q-squared coefficient.** This measure is also known after its main proponents as the Stone-Geisser Q-squared coefficient. The Q-squared coefficient is a nonparametric measure traditionally calculated via blindfolding. It is used for the assessment of the predictive validity (or relevance) associated with each latent variable block in the model, through the endogenous latent variable that is the criterion variable in the block. The Q-squared coefficient is sometimes referred to as a resampling analog of the R-squared. It is often similar in value to that measure. The Q-squared coefficient can assume negative values.

**Reflective latent variable.** A reflective latent variable is one in which all of the indicators are expected to be highly correlated with the latent variable score, and also highly correlated with one another. For example, the answers to certain question-statements by a group of people, measured on a 1 to 7 scale (1=strongly disagree; 7 strongly agree) and answered after a meal, are expected to be highly correlated with the latent variable “satisfaction with a meal”. The question-statements are: “I am satisfied with this meal”, and “After this meal, I feel full”. Therefore, the latent variable “satisfaction with a meal”, can be said to be reflectively measured through these two indicators. These indicators store answers to the two question-statements. This latent variable could be represented in a model graph as “Satisf”, and the indicators as “Satisf1” and “Satisf2”.

**Reliability of a measurement instrument.** Reliability is a measure of the quality of a measurement instrument; the instrument itself is typically a set of question-statements. A measurement instrument has good reliability if the question-statements (or other measures) associated with each latent variable are understood in the same way by different respondents.

**R-squared coefficient.** This is a measure calculated only for endogenous latent variables, and that reflects the percentage of explained variance for each of those latent variables. The higher the R-squared coefficient, the better is the explanatory power of the predictors of the latent variable in the model, especially if the number of predictors is small. Contrary to popular belief and in spite of what their name implies, R-squared coefficients are not calculated by squaring a correlation-like measure. They can assume negative values, although these are rare occurrences that normally suggest problems with the model in which they occur; e.g., severe collinearity or model misspecification.

**Statistical power.** Statistical power is often referred to simply as “power”; see the latter for a definition.

**Structural equation modeling (SEM).** A general term used to refer to a class of multivariate statistical methods where complex relationships among latent variables and indicators are estimated at once. In a SEM analysis, each latent variable is typically measured through multiple indicators, although there may be cases in which only one indicator is used to measure a latent variable. Key measures of relationships among latent variables are path coefficients (or standardized partial regression coefficients) and corresponding P values. Key measures of

relationships among latent variables and their respective indicators are weights and loadings, and corresponding P values.

**Structural error.** An error variable that accounts for the variance in an endogenous latent variable that is not accounted for by the latent variable predictors that point at the endogenous latent variable. A structural error is the same as a latent variable error; see the latter for an expanded definition.

**Variance inflation factor (VIF).** This is a measure of the degree of collinearity (or multicollinearity) among variables, including both indicators and latent variables. With latent variables, collinearity can take two main forms: vertical and lateral collinearity. Vertical, or classic, collinearity is predictor-predictor latent variable collinearity in individual latent variable blocks. Lateral collinearity is a term that refers to predictor-criterion latent variable collinearity; a type of collinearity that can lead to particularly misleading results. Full collinearity VIFs allow for the simultaneous assessment of both vertical and lateral collinearity in a SEM model.

# Acknowledgements

Revised text and other materials from previously published documents by the author have been used in the development of this book. The author would like to thank those users of WarpPLS who employ the software to inform decision making by organizational leaders. He is grateful to those users for their questions, comments, and suggestions. Some of the data discussed here have been compiled based on publicly available sources, some have been created via Monte Carlo simulations based on empirical studies, and some have been produced as a mix of both approaches. For ethical reasons, and to protect individual privacy, all of the individual-level data have been created via Monte Carlo simulations, based on empirical studies – to mimic what happened with real data.

# Bibliography

- Adelman, I., & Lohmoller, J.-B. (1994). Institutions and development in the nineteenth century: A latent variable regression model. *Structural Change and Economic Dynamics*, 5(2), 329-359.
- Amora, J. T. (2021). Convergent validity assessment in PLS-SEM: A loadings-driven approach. *Data Analysis Perspectives Journal*, 2(3), 1-6.
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality & Social Psychology*, 51(6), 1173-1182.
- Bera, A.K., & Jarque, C.M. (1981). Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo evidence. *Economics Letters*, 7(4), 313-318.
- Berglund, E., Lytsy, P., & Westerling, R. (2012). Adherence to and beliefs in lipid-lowering medical treatments: A structural equation modeling approach including the necessity-concern framework. *Patient Education and Counseling*, 91(1), 105-112.
- Biong, H., & Ulvnes, A.M. (2011). If the supplier's human capital walks away, where would the customer go? *Journal of Business-to-Business Marketing*, 18(3), 223-252.
- Bollen, K.A. (1987). Total, direct, and indirect effects in structural equation models. *Sociological Methodology*, 17(1), 37-69.
- Brewer, T.D., Cinner, J.E., Fisher, R., Green, A., & Wilson, S.K. (2012). Market access, population density, and socioeconomic development explain diversity and functional group biomass of coral reef fish assemblages. *Global Environmental Change*, 22(2), 399-406.
- Chew, L.P. (1989). Constrained Delaunay triangulations. *Algorithmica*, 4(1-4), 97-108.
- Chiquoine, B., & Hjalmarsen, E. (2009). Jackknifing stock return predictions. *Journal of Empirical Finance*, 16(5), 793-803.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.
- Cox, J. (2024). Combining sub-samples for improved statistical power in PLS-SEM: A constrained latent growth approach. *Data Analysis Perspectives Journal*, 5(1), 1-5.
- Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.
- Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39 (1), 1-38.
- Diamantopoulos, A. (1999). Export performance measurement: Reflective versus formative indicators. *International Marketing Review*, 16(6), 444-457.
- Diamantopoulos, A., & Siguaw, J.A. (2006). Formative versus reflective indicators in organizational measure development: A comparison and empirical illustration. *British Journal of Management*, 17(4), 263–282.
- Diamantopoulos, A., & Winklhofer, H. (2001). Index construction with formative indicators: An alternative scale development. *Journal of Marketing Research*, 37(1), 269-177.
- Dillon, W.R., & Goldstein, M. (1984). *Multivariate analysis: Methods and applications*. New York, NY: Wiley.

- Edwards, J.R. (2011). The fallacy of formative measurement. *Organizational Research Methods*, 14(2), 370-388.
- Efron, B., Rogosa, D., & Tibshirani, R. (2004). Resampling methods of estimation. In N.J. Smelser, & P.B. Baltes (Eds.). *International Encyclopedia of the Social & Behavioral Sciences* (pp. 13216-13220). New York, NY: Elsevier.
- Ehremberg, A.S.C., & Goodhart, G.J. (1976). *Factor analysis: Limitations and alternatives*. Cambridge, MA: Marketing Science Institute.
- Enders, C.K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.
- Ferguson, G.A. (1981). *Statistical analysis in psychology and education*. New York, NY: McGraw-Hill.
- Fornell, C., & Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39-50.
- Geisser, S. (1974). A predictive approach to the random effects model. *Biometrika*, 61(1), 101-107.
- Gel, Y.R., & Gastwirth, J.L. (2008). A robust modification of the Jarque–Bera test of normality. *Economics Letters*, 99(1), 30-32.
- Giaquinta, M. (2009). *Mathematical analysis: An introduction to functions of several variables*. New York, NY: Springer.
- Guo, K.H., Yuan, Y., Archer, N.P., & Connelly, C.E. (2011). Understanding nonmalicious security violations in the workplace: A composite behavior model. *Journal of Management Information Systems*, 28(2), 203-236.
- Hubona, G., & Belkhamza, Z. (2021). Testing a moderated mediation in PLS-SEM: A full latent growth approach. *Data Analysis Perspectives Journal*, 2(4), 1-5.
- Jarque, C.M., & Bera, A.K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6(3), 255-259.
- Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3), 187-200.
- Kaiser, H.M. (2010). *Mathematical programming for agricultural, environmental, and resource economics*. Hoboken, NJ: Wiley.
- Ketkar, S., Kock, N., Parente, R., & Verville, J. (2012). The impact of individualism on buyer-supplier relationship norms, trust and market performance: An analysis of data from Brazil and the U.S.A. *International Business Review*, 21(5), 782–793.
- Kim, M.J., Park, C.G., Kim, M., Lee, H., Ahn, Y.-H., Kim, E., Yun, S.-N., & Lee, K.-J. (2012). Quality of nursing doctoral education in Korea: Towards policy development. *Journal of Advanced Nursing*, 68(7), 1494-1503.
- Klaassen, C.A., Mokveld, P.J., & Es, B.V. (2000). Squared skewness minus kurtosis bounded by 186/125 for unimodal distributions. *Statistics & probability letters*, 50(2), 131-135.
- Kline, R.B. (1998). *Principles and practice of structural equation modeling*. New York, NY: The Guilford Press.
- Kock, N. (2010). Using WarpPLS in e-collaboration studies: An overview of five main analysis steps. *International Journal of e-Collaboration*, 6(4), 1-11.
- Kock, N. (2011a). A mathematical analysis of the evolution of human mate choice traits: Implications for evolutionary psychologists. *Journal of Evolutionary Psychology*, 9(3), 219-247.

- Kock, N. (2011b). Using WarpPLS in e-collaboration studies: Mediating effects, control and second order variables, and algorithm choices. *International Journal of e-Collaboration*, 7(3), 1-13.
- Kock, N. (2011c). Using WarpPLS in e-collaboration studies: Descriptive statistics, settings, and key analysis results. *International Journal of e-Collaboration*, 7(2), 1-18.
- Kock, N. (2013). Using WarpPLS in e-collaboration studies: What if I have only one group and one condition? *International Journal of e-Collaboration*, 9(3), 1-12.
- Kock, N. (2014a). Advanced mediating effects tests, multi-group analyses, and measurement model assessments in PLS-based SEM. *International Journal of e-Collaboration*, 10(3), 1-13.
- Kock, N. (2014b). *Stable P value calculation methods in PLS-SEM*. Laredo, TX: ScriptWarp Systems.
- Kock, N. (2014c). *Single missing data imputation in PLS-SEM*. Laredo, TX: ScriptWarp Systems.
- Kock, N. (2014d). Using data labels to discover moderating effects in PLS-based structural equation modeling. *International Journal of e-Collaboration*, 10(4), 1-16.
- Kock, N. (2015a). One-tailed or two-tailed P values in PLS-SEM? *International Journal of e-Collaboration*, 11(2), 1-7.
- Kock, N. (2015b). A note on how to conduct a factor-based PLS-SEM analysis. *International Journal of e-Collaboration*, 11(3), 1-9.
- Kock, N. (2015c). Common method bias in PLS-SEM: A full collinearity assessment approach. *International Journal of e-Collaboration*, 11(4), 1-10.
- Kock, N. (2015d). Wheat flour versus rice consumption and vascular diseases: Evidence from the China Study II data. *Cliodynamics*, 6(2), 130–146.
- Kock, N. (2015e). How likely is Simpson’s paradox in path models? *International Journal of e-Collaboration*, 11(1), 1-7.
- Kock, N. (2016a). Non-normality propagation among latent variables and indicators in PLS-SEM simulations. *Journal of Modern Applied Statistical Methods*, 15(1), 299-315.
- Kock, N. (2016b). Hypothesis testing with confidence intervals and P values in PLS-SEM. *International Journal of e-Collaboration*, 12(3), 1-6.
- Kock, N. (2016c). Advantages of nonlinear over segmentation analyses in path models. *International Journal of e-Collaboration*, 12(4), 1-6.
- Kock, N. (2016d). Visualizing moderating effects in path models with latent variables. *International Journal of e-Collaboration*, 12(1), 1-7.
- Kock, N. (2017). Structural equation modeling with factors and composites: A comparison of four methods. *International Journal of e-Collaboration*, 13(1), 1-9.
- Kock, N. (2018a). Single missing data imputation in PLS-based structural equation modeling. *Journal of Modern Applied Statistical Methods*, 17(1), 1-23.
- Kock, N. (2018b). Should bootstrapping be used in PLS-SEM: Toward stable p-value calculation methods. *Journal of Applied Structural Equation Modeling*, 2(1), 1-12.
- Kock, N. (2019a). From composites to factors: Bridging the gap between PLS and covariance-based structural equation modeling. *Information Systems Journal*, 29(3), 674-706.
- Kock, N. (2019b). Factor-based structural equation modeling with WarpPLS. *Australasian Marketing Journal*, 27(1), 57-63.

- Kock, N. (2019c). Factor-based structural equation modeling: Going beyond PLS and composites. *International Journal of Data Analysis Techniques and Strategies*, 11(1), 1–28.
- Kock, N. (2020a). Full latent growth and its use in PLS-SEM: Testing moderating relationships. *Data Analysis Perspectives Journal*, 1(1), 1-5.
- Kock, N. (2020b). Multilevel analyses in PLS-SEM: An anchor-factorial with variation diffusion approach. *Data Analysis Perspectives Journal*, 1(2), 1-6.
- Kock, N. (2020c). Using indicator correlation fit indices in PLS-SEM: Selecting the algorithm with the best fit. *Data Analysis Perspectives Journal*, 1(4), 1-4.
- Kock, N. (2021a). Common structural variation reduction in PLS-SEM: Replacement analytic composites and the one fourth rule. *Data Analysis Perspectives Journal*, 2(5), 1-6.
- Kock, N. (2021b). Harman’s single factor test in PLS-SEM: Checking for common method bias. *Data Analysis Perspectives Journal*, 2(2), 1-6.
- Kock, N. (2021c). Moderated mediation and J-curve emergence in path models: An information systems research perspective. *Journal of Systems and Information Technology*, 23(3), 303-321.
- Kock, N. (2022a). Testing and controlling for endogeneity in PLS-SEM with stochastic instrumental variables. *Data Analysis Perspectives Journal*, 3(3), 1-6.
- Kock, N. (2022b). Using causality assessment indices in PLS-SEM. *Data Analysis Perspectives Journal*, 3(5), 1-6.
- Kock, N. (2023a). Assessing multiple reciprocal relationships in PLS-SEM. *Data Analysis Perspectives Journal*, 4(3), 1-8.
- Kock, N. (2023b). Using logistic regression in PLS-SEM: Dichotomous endogenous variables. *Data Analysis Perspectives Journal*, 4(4), 1-6.
- Kock, N. (2024a). Combining composites and factors in PLS-SEM models: A multi-algorithm technique. *Data Analysis Perspectives Journal*, 5(4), 1-8.
- Kock, N. (2024b). Conducting a difference-in-differences analysis with PLS-SEM: The classic 2x2 approach. *Data Analysis Perspectives Journal*, 5(5), 1-8.
- Kock, N. (2024c). Will PLS have to become factor-based to survive and thrive? *European Journal of Information Systems*, 33(6), 882-902.
- Kock, N., & Chatelain-Jardón, R. (2011). Four guiding principles for research on evolved information processing traits and technology-mediated task performance. *Journal of the Association for Information Systems*, 12(10), 684-713.
- Kock, N., & Gaskins, L. (2014). The mediating role of voice and accountability in the relationship between Internet diffusion and government corruption in Latin America and Sub-Saharan Africa. *Information Technology for Development*, 20(1), 23-43.
- Kock, N., & Gaskins, L. (2016). Simpson’s paradox, moderation, and the emergence of quadratic relationships in path models: An information systems illustration. *International Journal of Applied Nonlinear Science*, 2(3), 200-234.
- Kock, N., & Hadaya, P. (2018). Minimum sample size estimation in PLS-SEM: The inverse square root and gamma-exponential methods. *Information Systems Journal*, 28(1), 227–261.
- Kock, N., & Lynn, G.S. (2012). Lateral collinearity and misleading results in variance-based SEM: An illustration and recommendations. *Journal of the Association for Information Systems*, 13(7), 546-580.

- Kock, N., & Mayfield, M. (2015). PLS-based SEM algorithms: The good neighbor assumption, collinearity, and nonlinearity. *Information Management and Business Review*, 7(2), 113-130.
- Kock, N., & Moqbel, M. (2016). Statistical power with respect to true sample and true population paths: A PLS-based SEM illustration. *International Journal of Data Analysis Techniques and Strategies*, 8(4), 316-331.
- Kock, N., & Moqbel, M. (2021). Social networking site use, positive emotions, and job performance. *Journal of Computer Information Systems*, 61(2), 163-173.
- Kock, N., & Sexton, S. (2017). Variation sharing: A novel numeric solution to the path bias underestimation problem of PLS-based SEM. *International Journal of Strategic Decision Sciences*, 8(4), 46-68.
- Kock, N., & Verville, J. (2012). Exploring free questionnaire data with anchor variables: An illustration based on a study of IT in healthcare. *International Journal of Healthcare Information Systems and Informatics*, 7(1), 46-63.
- Kock, N., Avison, D., & Malaurent, J. (2017). Positivist information systems action research: Methodological issues. *Journal of Management Information Systems*, 34(3), 754-767.
- Kock, N., Mayfield, M., & Mayfield, J. (2022). Altruistic leadership and job performance: A Darwinian evolutionary perspective. *Revista Interdisciplinar de Ciência Aplicada*, 6(1), 1-10.
- Kock, N., Mayfield, M., Mayfield, J., Sexton, S., & De La Garza, L. (2019). Empathetic leadership: How leader emotional support and understanding influences follower performance. *Journal of Leadership and Organizational Studies*, 26(2), 217-236.
- Kock, N., Moqbel, M., Jung, Y., & Syn, T. (2018). Do older programmers perform as well as young ones? Exploring the intermediate effects of stress and programming experience. *Cognition, Technology & Work*, 20(3), 489-504.
- Lee, D.T., & Schachter, B.J. (1980). Two algorithms for constructing a Delaunay triangulation. *International Journal of Computer & Information Sciences*, 9(3), 219-242.
- Lindell, M., & Whitney, D. (2001). Accounting for common method variance in cross-sectional research designs. *Journal of Applied Psychology*, 86(1), 114-121.
- Lohmöller, J.-B. (1989). *Latent variable path modeling with partial least squares*. Heidelberg, Germany: Physica-Verlag.
- MacKinnon, D.P., Krull, J.L., & Lockwood, C.M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, 1(4), 173-181.
- Melton, B. L., Moqbel, M., Kanaan, S., & Sharma, N. K. (2016). Structural equation model of disability in low back pain. *Spine*, 41(20), 1621-1627.
- Miller, R.B., & Wichern, D.W. (1977). *Intermediate business statistics: Analysis of variance, regression and time series*. New York, NY: Holt, Rinehart and Winston.
- Moqbel, M., Guduru, R., & Harun, A. (2020). Testing mediation via indirect effects in PLS-SEM: A social networking site illustration. *Data Analysis Perspectives Journal*, 1(3), 1-6.
- Morrow, D. L., & Conger, S. (2021). Assessing reciprocal relationships in PLS-SEM: An illustration based on a job crafting study. *Data Analysis Perspectives Journal*, 2(1), 1-5.
- Mueller, R.O. (1996). *Basic principles of structural equation modeling*. New York, NY: Springer.
- Nevitt, J., & Hancock, G.R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling*, 8(3), 353-377.



- Newman, D.A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, 17(4), 372-411.
- Nunnally, J.C. (1978). *Psychometric theory*. New York, NY: McGraw Hill.
- Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.
- Ogasawara, H. (1999). Standard errors for the direct oblimin solution with Kaiser's normalization. *Japanese Journal of Psychology*, 70(4), 333-338.
- Pearl, J. (2009). *Causality: Models, reasoning, and inference*. Cambridge, England: Cambridge University Press.
- Peterson, R.A., & Yeolib, K. (2013). On the relationship between coefficient alpha and composite reliability. *Journal of Applied Psychology*, 98(1), 194-198.
- Petter, S., Straub, D., & Rai, A. (2007). Specifying formative constructs in information systems research. *MIS Quarterly*, 31(4), 623-656.
- Rasoolimanesh, S.M., Jaafar, M., Kock, N. and Ahmad, A. G. (2017). The effects of community factors on residents' perceptions toward World Heritage Site inscription and sustainable tourism development. *Journal of Sustainable Tourism*, 25(2), 198-216.
- Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, 21(2), 173-184.
- Rencher, A.C. (1998). *Multivariate statistical inference and applications*. New York, NY: John Wiley & Sons.
- Robert, C.P., & Casella, G. (2010). *Monte Carlo statistical methods*. New York, NY: Springer.
- Rohatgi, V.K., & Székely, G.J. (1989). Sharp inequalities between skewness and kurtosis. *Statistics & Probability Letters*, 8(4), 297-299.
- Rosenthal, R., & Rosnow, R.L. (1991). *Essentials of behavioral research: Methods and data analysis*. Boston, MA: McGraw Hill.
- Samak, A., Islam, M. R., & Hanke, D. (2024). A comparison of data analyses with WarpPLS and Stata: A study of trust and its role regarding internet use and subjective well-being. *Data Analysis Perspectives Journal*, 5(3), 1-6.
- Schumacker, R.E., & Lomax, R.G. (2004). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum.
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1), 107-120.
- Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causality, prediction and search*. Berlin, Germany: Springer-Verlag.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, 36(1), 111-147.
- Tarkom, A., & Gopal, P. (2024). A comparison of multiple regression analyses in Stata and WarpPLS. *Data Analysis Perspectives Journal*, 5(2), 1-8.
- Temme, D., Kreis, H., & Hildebrandt, L. (2006). *PLS path modeling – A software review*. Berlin, Germany: Institute of Marketing, Humboldt University Berlin.
- Tenenhaus, M., Vinzi, V.E., Chatelin, Y.-M., & Lauro, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis*, 48(1), 159-205.
- Theil, H. (1958). *Economic forecasts and policy*. Amsterdam, Netherlands: North-Holland.
- Wagner, C.H. (1982). Simpson's paradox in real life. *The American Statistician*, 36(1), 46-48.
- Wetzels, M., Odekerken-Schroder, G., & van Oppen, C. (2009). Using PLS path modeling for assessing hierarchical construct models: Guidelines and empirical illustration. *MIS Quarterly*, 33(1), 177-196.

## Model-Driven Data Analytics: Applications with WarpPLS

Wold, H. (1980). Model construction and evaluation when theoretical knowledge is scarce. In J. Kmenta and J. B. Ramsey (Eds.), *Evaluation of econometric models* (pp. 47-74). Waltham, MA: Academic Press.

Wold, S., Trygg, J., Berglund, A., & Antti, H. (2001). Some recent developments in PLS modeling. *Chemometrics and Intelligent Laboratory Systems*, 58(2), 131-150.

Wooldridge, J.M. (1991). A note on computing r-squared and adjusted r-squared for trending and seasonal data. *Economics Letters*, 36(1), 49-54.