



**PATRICIA MENDES DOS SANTOS**

**STRUCTURAL EQUATION MODELS WITH ADAPTIVE  
REGRESSION AND CONSTRUCTION OF AN INDEX TO  
VALIDATE CONSTRUCTS USED TO DISTINGUISH THE  
PROFILES OF SPECIALTY COFFEE CONSUMERS**

**LAVRAS – MG**

**2021**

**PATRICIA MENDES DOS SANTOS**

**STRUCTURAL EQUATION MODELS WITH ADAPTIVE REGRESSION AND  
CONSTRUCTION OF AN INDEX TO VALIDATE CONSTRUCTS USED TO  
DISTINGUISH THE PROFILES OF SPECIALTY COFFEE CONSUMERS**

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Prof. Dr. Marcelo Ângelo Cirillo  
Orientador

**LAVRAS – MG**

**2021**

**Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca  
Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).**

dos Santos, Patricia Mendes.

Structural equation models with adaptive regression and construction of an index to validate constructs used to distinguish the profiles of specialty coffee consumers / Patricia Mendes dos Santos. - 2021.

78 p. : il.

Orientador(a): Marcelo Ângelo Cirillo.

Tese (doutorado) - Universidade Federal de Lavras, 2021.  
Bibliografia.

1. Structural equation models. 2. Adaptive linear regression. 3. Specialty coffees. I. Cirillo, Marcelo Ângelo. II. Título.

**PATRICIA MENDES DOS SANTOS**

**MODELOS DE EQUAÇÕES ESTRUTURAIS COM REGRESSÕES ADAPTATIVAS E  
CONSTRUÇÃO DE UM ÍNDICE PARA VALIDAÇÃO DE CONSTRUTO COM  
APLICAÇÕES NA DISCRIMINAÇÃO DE PERFIS DE CONSUMIDORES DE CAFÉS  
ESPECIAIS**

**STRUCTURAL EQUATION MODELS WITH ADAPTIVE REGRESSION AND  
CONSTRUCTION OF AN INDEX TO VALIDATE CONSTRUCTS USED TO  
DISTINGUISH THE PROFILES OF SPECIALTY COFFEE CONSUMERS**

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

APROVADA em 02 de Agosto de 2021.

Prof. Dr. Eliandro Rodrigues Cirilo UEL  
Prof. Dra. Lúcia Pereira Barroso USP  
Prof. Dr. Ronaldo Rocha Bastos UFJF  
Prof. Dr. Tales Jesus Fernandes UFLA

Prof. Dr. Marcelo Ângelo Cirillo  
Orientador

**LAVRAS – MG  
2021**

*Agradeço a Deus por estar sempre presente em minha vida e ter me sustentado durante essa longa jornada, agradeço aos meus professores, meu orientador, meus familiares e aos meus colegas de curso que me apoiaram para a conclusão dessa tese.*

## **AGRADECIMENTOS**

A Deus, por me guiar, fortalecer e pela presença constante em minha vida.

A minha mãe, Maria Aparecida, pelo apoio incondicional, carinho, compreensão e comprometimento com a minha educação. A você, meu amor e gratidão eternos.

Aos meus irmãos e sobrinhos, pelo carinho, incentivo e por se orgulharem de mim.

Ao meu orientador, prof. Dr. Marcelo Ângelo Cirillo, que aceitou prontamente o convite para me orientar. Agradeço pela paciência, pelo incentivo, pelos ensinamentos, por sua dedicação e apoio para realizar este trabalho. Levo e continuarei levando seu exemplo de profissional e pessoa para minha vida. Muito obrigada por tudo!

Aos amigos do DES, em especial aos meus grandes amigos Vânia e Rafael, Cristian, Rodnei, Luciano, Carlos, Cláudio e Rodrigo por todos os momentos de suporte, alegria, carinho e descontração compartilhados, tornando mais fácil todo o percurso. A amizade de vocês foi fundamental para a realização e conclusão desse trabalho.

Aos colegas de pós-graduação com os quais tive bons momentos durante a fase de estudos. Agradeço pela colaboração, incentivo e apoio, que foram fundamentais para que eu pudesse chegar até aqui.

Aos professores do Departamento de Estatística (DES), pelos conhecimentos transmitidos durante as aulas, pela disponibilidade, paciência e auxílio nos momentos de dificuldades inerentes do curso de doutorado.

Aos demais funcionários do Departamento Estatística (DES), por nos auxiliar nas diversas atividades burocráticas, em especial a Nádia Ferreira, por ser sempre tão prestativa.

À Universidade Federal de Lavras (UFLA) e ao Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, pela oportunidade concedida para realização do curso de doutorado. Em especial, ao coordenador Prof. Dr. Renato Ribeiro Lima, que não mede esforços para a excelência do curso.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa de estudos.

## ACKNOWLEDGEMENTS

I thank God, for His constant guidance, strength, and presence in my life.

I thank my mother, Maria Aparecida, for her unconditional support, love, understanding, and commitment to my education. To you, all my love and gratitude. To my siblings, nieces, and nephews, thanks for your love, your support and for being proud of me.

I thank my advisor, Prof. Dr. Marcelo Ângelo Cirillo, who immediately accepted my invitation to be my advisor. I am grateful for your patience, your support, your teachings, and your dedication to help me do this work. I have taken and will always take you as my example for life. Thank you for everything!

To all my friends from DES, especially Vânia and Rafael, Cristian, Rodnei, Luciano, Carlos, Cláudio, and Rodrigo, thanks for all the moments of support, happiness, love, and fun we shared, making it a lot easier for me. Your friendship has been fundamental to carry out and complete this work.

I would also like to thank my colleagues from graduate school, with whom I shared great moments during this period. Thank you for the collaboration, incentive, and support, which were crucial for me to reach this point.

To the faculty members of the Department of Statistics (DES), thanks for all the knowledge you shared in class, for your availability, patience, and help in the moments of struggle we all go through during doctorate.

Thanks to the employees of the Department of Statistics (DES), for helping us with all the paperwork, especially Nádia Ferreira, for being so helpful.

I also thank Universidade Federal de Lavras (UFLA) and the Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, for the opportunity you gave me to pursue a Doctoral degree. Special thanks to the coordinator, Prof. Dr. Renato Ribeiro Lima, who constantly focuses on bringing excellence to the course.

To the Coordination for the Improvement of Higher Education Personnel (CAPES), thanks for granting the scholarship.

*“Success is not measured by the position one has reached in life, rather by the obstacles one  
overcomes while trying to succeed”  
Abraham Lincoln*



## ABSTRACT

This work consists of presenting a new approach to Adaptive Linear Regression adapted to structural equation models and improving the index related to the Average Variance Extracted (AVE), given a plug-in approach, and replacing the error variances with the factor loadings of the estimated adaptive regressions. To do so, a Monte Carlo simulation study was performed considering scenarios with different numbers of outliers, which were generated by distributions with symmetry deviations and kurtosis excess. Sample sizes were defined as  $n=50$ , 100 and 200. In formative structural models and considering outliers generated either from symmetrical distributions or from multivariate log-normal distributions, the Adaptive Linear Regression modeling was found to be efficient in the different scenarios under analysis. Likewise, for models with specification errors, this method was proven to have low efficiency, as expected. Furthermore, constructs were elaborated with variables that could enable both the characterization and the distinction of individuals among the different groups of Brazilian specialty coffee consumers and that could provide different perspectives on the transition among them. The results made it possible to better distinguish the consumers and better characterize the proposed categories, thus contributing to the improvement and simplification of marketing strategies used by players in this market. In addition, the results also promoted the discussion on which factors stimulate the transition of an individual from an initial construct to another, and we showed that transitioning from regular consumers to enthusiasts is easier than moving from enthusiasts to specialists.

**Keywords:** Structural equation models. Outliers. Adaptive linear regression. Specialty coffees. Consumer behavior.

## RESUMO

Este trabalho consiste na apresentação de uma nova abordagem de Regressão Linear Adaptativa (RLA) adaptada a modelos de equações estruturais e no aprimoramento do índice correspondente a variância média extraída, dada uma abordagem plug-in, substituindo as variâncias dos erros pelas cargas fatoriais das regressões adaptativas estimadas. Para tanto, realizou-se um estudo de simulação Monte Carlo considerando cenários com diferentes concentrações de outliers, gerados por distribuições com desvios de simetria e excesso de curtose e tamanhos amostrais definidos como  $n= 50,100$  e  $200$ . Concluiu-se que, em modelos estruturais formativos, considerando outliers gerados a partir de distribuições simétricas ou da distribuição log-normal multivariada, o método RLA, apresentou boa eficiência para modelos corretamente especificados. Da mesma forma, para modelos com erros de especificação, foi evidenciado baixa eficiência desse método, sendo coerente com o que era esperado. Além disso, elaborou-se construtos cujas variáveis possibilitassem a caracterização e distinção de indivíduos entre os diferentes grupos de consumidores brasileiros de cafés especiais e, fornecessem percepções sobre a transição entre eles. Os resultados permitiram uma melhoria na distinção dos consumidores e caracterização entre as categorias propostas, contribuindo para o aprimoramento e simplificação das estratégias de marketing realizadas pelos atores deste mercado. Além disso, incentivou-se a discussão sobre quais fatores estimulam a transição de um indivíduo de um construto inicial para um subsequente e demonstramos uma maior facilidade de transição dos indivíduos de consumidores regulares para entusiastas, do que de entusiastas para especialistas.

**Palavras-chave:** Modelos de equações estruturais. Outliers. Regressão linear adaptativa. Cafés especiais. Comportamento do consumidor.

## LIST OF FIGURES

Figura 2.1 – Basic elements used to build a path diagr. . . . .	17
Figura 2.2 – Example of a path diagram. . . . .	18
Figura 2.3 – Reflective (a) indicators and Formative (b) indicators. . . . .	20

## CONTENTS

	<b>FIRST PART</b> . . . . .	10
<b>1</b>	<b>INTRODUCTION</b> . . . . .	11
<b>2</b>	<b>THEORETICAL BACKGROUND</b> . . . . .	14
<b>2.1</b>	<b>Structural equation modeling - SEM</b> . . . . .	14
<b>2.1.1</b>	<b>Model specification</b> . . . . .	16
<b>2.1.1.1</b>	<b>Direct, indirect, and total effects</b> . . . . .	19
<b>2.1.1.2</b>	<b>Reflective indicators and formative indicators</b> . . . . .	20
<b>2.1.2</b>	<b>Structural model formulation</b> . . . . .	21
<b>2.1.2.1</b>	<b>Model identification</b> . . . . .	24
<b>2.1.2.2</b>	<b>Model estimation</b> . . . . .	25
<b>2.1.2.3</b>	<b>Considerations on error specifications and interpretation</b> . . . . .	25
<b>2.2</b>	<b>Concepts and preliminary definitions of robust inference</b> . . . . .	26
<b>2.2.1</b>	<b>Breakdown point</b> . . . . .	26
<b>2.2.2</b>	<b>Elementary sets</b> . . . . .	27
<b>2.2.3</b>	<b>Estimator of Least Trimmed of Squares (LTS)</b> . . . . .	28
<b>2.2.4</b>	<b>Adaptive estimator</b> . . . . .	30
<b>2.2.5</b>	<b>Validity</b> . . . . .	31
<b>2.3</b>	<b>Exploratory factor analysis</b> . . . . .	33
<b>2.4</b>	<b>Synthesis of opinions given by panel members on published papers</b> . . . . .	35
	<b>REFERENCES</b> . . . . .	37
	<b>SECOND PART</b> . . . . .	41
	<b>ARTICLE 1 Construction of the average variance extracted index for construct validation in structural equation models with adaptive regressions</b> . . . . .	43
	<b>ARTICLE 2 Specialty coffee in Brazil: transition among consumers' constructs using structural equation modeling.</b> . . . . .	57
	<b>FINAL CONSIDERATIONS</b> . . . . .	76

## **FIRST PART**

## 1 INTRODUCTION

Structural equation models are a family of multivariate statistical models used to model relationships between multiple variables, that is, to estimate parameters simultaneously in an equation system (BOLLEN, 2002). Putting together concepts of multiple regression, path analysis, and factor analysis, these models allow the inclusion of latent variables (constructs), that is, theoretical variables that cannot be directly measured, into a covariance structure given by a theoretical model that explains their linear relationship with observed variables, referred to as indicators.

Regarding the specification of these models, a definition of two factor models is necessary: the structural model which defines the cause-and-effect relationship between the latent variables in the model, and the mensuration model which relates the endogenous and exogenous observed variables to one or more latent variables.

Assuming that a construct is mathematically formed by linear combinations of observed variables, one may wonder about the number of variables to be used in the formation of a construct, so they may provide information enough to characterize a concept in its interpretation. Therefore, to validate constructs, several indexes and coefficients have been proposed, one of which, known as Average Variance Extracted (AVE), has been more commonly used.

Many authors have used AVE index in their research. Kyougoku et al. (2015), for example, stands out by developing the Assessment of Belief Conflict in Relationship-14 (ABCR-14), a new scale which assesses belief conflict in the health area. The authors analyzed the psychometric properties of ABCR-14 in relation to entropy, polyserial correlation coefficient, exploratory factor analysis, confirmatory factor analysis, average variance extracted, Cronbach alpha, Person product-moment correlation coefficient, and multidimensional item response theory (MIRT). Carter (2016) who provides evidence to the internal structure. The aim of the study is the use of statistical techniques to provide evidence for the number of dimensions measures in a scale.

Valentini e Bruno (2017) aimed at discussing the concepts of average variance extracted and composed reliability, refuting the proposal of Fornell e Larcker (1981) to consider AVE as a convergent validity indicator. They also showed that the values for these indexes were altered as a function of the number of observed variables and the homogeneity of factor loadings. Farzandipour et al. (2021) elaborated a psychometric instrument to determine the qualification of computation abilities of nurses working at centers for educational assistance.

Different statistical methods may be used to estimate the parameters of structural equation models; however, to comply with the proposals of this research study, a procedure was used to estimate factor loadings by means of adaptive linear regressions. This procedure resulted in a combination of estimations obtained by the method of least squares, and robust estimations, generated by LTS regression.

When it comes to estimating the AVE index, the formalization occurs as a function of the average sum of variance of errors. However, considering the error of latent variables defined as constructs is controversial, since such errors may be interpreted systemically, that is, a construct may not be correctly specified due to the absence of a given variable (GOSLING; GONÇALVES, 2003).

Another aspect consists of considering an endogenous variable as an error that represents part of the variable that is not taken into consideration by the linear influence of other variables in the system, thus causing the error to be interpreted as random (MACCALLUM et al., 1995)..

Based on these arguments, this study aimed at estimating the AVE index by following a plug-in estimation procedure, that is, replacing the variance of errors with the factor loadings obtained in the adaptive regressions. In this context, after such modification, the indexes proposed by this study and published were named Adaptive AVE indexes.

This work is organized as a paper and is divided into two parts: the first one is composed by a general introduction and the theoretical background needed to support this thesis. The second part consists of two scientific papers. The first paper presents the construction of a new index to validate constructs and assesses it in different scenarios, based on average variance extracted (AVE) with adaptive regressions.

The second paper, using structural equation models, aimed at elaborating constructs with variables that could characterize and distinguish individuals among different consumer groups, and provided perceptions on their transition. To validate the composition of constructs, the AVE index was estimated by adaptive regression, considering a data sample provided by Guimarães et al. (2019). The authors gathered 864 self-appliable online questionnaires, widely disclosed to Brazilian consumers of specialty coffees, between January and February of 2017. Despite its importance and advancements in the understanding of specialty coffee consumers, the work of Guimarães et al. (2019) has considerable limitations, which we seek to overcome in this work, by means of more robust and more adequate statistical methods. Furthermore, further

research on which factors would stimulate the transition of individuals among the categories of consumers identified is necessary.

The final considerations are given at the end of the second part, which is part of this thesis, as well as aspects of this research that are relevant to innovate the use of new statistical methodologies and contribute to marketing strategies.



## 2 THEORETICAL BACKGROUND

### 2.1 Structural equation modeling - SEM

Structural Equation Modeling (SEM) is a multivariate statistical methodology which allows to study cause-and-effect relationships and the correlations between a set of variables simultaneously. This technique has become more and more popular in behavioral, and educational areas, as well as in social research, because it allows several relationships to be studied simultaneously, including variables that are not measured directly, but by their effects (named indicators) (SONG; LEE, 2012).

In this approach, a series of equations is defined to describe hypothetical structures in the relationships among the countless variables added to the model. These equations represent the way the non-observed variables (constructs) interact with the measured observed variables, as well as the way these variables interact with one another.

The development of SEM models was due to the works of Joreskog (1973), Keesling (1972), and Wiley (1973). These authors were able to propose a general model by incorporating to their representations a path diagram which could be applied to different situations. In this moment, the model was seen as two: Measurement Equations and Structural Equations.

At first, this approach was known as JKW model, but then it changed to Linear Structural Relationship model (LISREL) with the development of the first software, LISREL (JORES-KOG; SORBOM, 1986). The development of LISREL helped make these models so popular that the terms Structural Equation Model and LISREL are considered synonymous by many authors. In addition to LISREL, other statistical software programs also stand out, such as Proc Calis and Proc Tcalis, by SAS, AMOS by SPSS, and SEM pack by R.

Regarding the usual estimation techniques, the works of Lawley (1940), Anderson, Rubin et al. (1956), and Jöreskog (1969) also stand out because they helped define tests of hypotheses to be used in the SEM model. Johnson e Creech (1983) and Muthén e Kaplan (1985) suggested analyses to be run when there are categoric variables in the model. In econometrics, the property of structural equation estimators with observed variables was established, for instance, by Goldberger et al. (1964). In psychometrics, Bock e Bargmann (1966) developed an analysis of covariance structure to estimate the variance of latent variables. Jöreskog (1970) proposed maximum likelihood estimators based on multinormal distributions of observed variables. Jöreskog e Goldberger (1972) and Browne (1974) proposed general least square

estimators for better flexibility in the applications. Bentler (1983) suggested estimators to treat products of moments of observed variables.

To sum up, SEM is a consequence of the evolution of three components that are present in the general model: They are the path analysis, the conceptual synthesis of latent variables and measurement models, and the estimation procedures (BOLLEN, 1989). According to Hair et al. (2009), all these components are unified by the main elements of SEM, which are:

**Latent variables or constructs:** hypothetical or theoretical variables that cannot be directly measured, but that can be represented by other indicators, composed by the items in the scale or by the researcher's observation, which, together, will provide a reasonably precise measurement of the attitude. Example: quality, beauty, satisfaction.

**Observed variables:** variables directly measured or observed which are used to measure the latent variable. They are used as indicators of the attitude to be measured. The researcher should use several indicators for each latent variable to obtain a more comprehensive and trustworthy understanding of the construct. Example: temperature, weight, height.

**Exogenous variables:** variables that are not affected by other variables in the model, and act only as a cause of effects on other variables in the theoretical model. They are also known as independent or predictors.

**Endogenous variables:** variables that are influenced by other variables in the model. The researcher will be able to differentiate which independent variables predict each variable dependent, based on the theory and on their own experience. They are also known as dependent variables.

The general structural equation model consists of two sub-models: the structural model, which determines the cause-effect relationships between latent variables and show the effects and the total non-explained variance; and the measurement model, which explains how latent variables are measured as a function of observed variables and describes the measurement properties (validity and reliability) of these variables (PEREIRA, 2014).

Contrary to other multivariate methodologies, SEM is a technique based on theoretical concepts and is proper for confirmatory analyses in which the researcher's knowledge is decisive in almost every step. Therefore, to be applied with the collected dataset, the researcher's experience is used, since it determines which causal relationship should be applied to the data.

This makes it possible to confirm hypotheses, but it also represents a limitation, because the choice of variables and causal relationships made by the researcher influence the predictive

structure of the model. Therefore, the entire model construction should be strongly grounded on theoretical elements.

These models provide plenty of advantages, such as the possibility to analyze several relationships at the same time; to work with latent variables; to model mediator and moderator variables; to model terms of error; to test models with many equations altogether, among others (XIAO et al., 2013).

Thanks to their power of analysis, this methodology may be applied in problem analysis in several areas of knowledge, such as the study of Pilati e Laros (2007) , who discussed the conceptions and applications of structural equation models to producing knowledge in Psychology and related areas. Silva et al. (2018) who used structural equation modeling to assess the effect of mediating satisfaction on the relationship between organizational culture and commitment to work. Ribeiro et al. (2019) proposed a conceptual model using structural equation modeling to analyze the influence of physical and psychological violence against pregnant women and of depression symptoms during pregnancy on post-partum depression. Castro et al. (2020) developed a conceptual model and explored the direct and indirect associations between sociodemographic factors, health, work, social interactions, and the participation in paid jobs of men and women in a representative sample of the national population aged 50 or older. Tavares et al. (2021) used structural equation modeling to assess functional incapacity and its associated factors among the elderly.

### **2.1.1 Model specification**





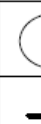
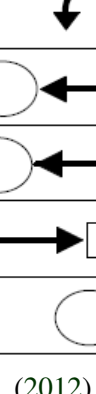


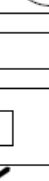
Model specification is the first step in structural equation modeling. It includes the use of all theory, research, relevant information available to develop a theoretical model. In this step, the available information is used to decide which variables should be included in the theoretical model and how these variables are related.

Therefore, one must determine which observed variables and which constructs, as well as their function in the model (exogenous or endogenous variables), which causal relationships exist among the constructs and/or which observed variables should be included or excluded, which (non-causal) associations should be included in or omitted from the model, and which errors or residues should be correlated (MARÔCO, 2010).

All structural relationships between the variables will be represented by a path diagram, which is a graph used to make it easier to understand and visualize the cause-effect relationships between the variables (observed and latent) in the SEM model under analysis.

A path diagram consists of a set of geometric forms and arrows used to evidence the type of variable (observed or latent) and the type of relationship they have. In general, latent variables are represented by ellipses, observed variables by rectangles and their causal relationships (all linear) are represented by arrows. In Figure 2.1 are the conventions used to represent the relationships between variables.

Figure 2.1 – Basic elements used to build a path diagram.

Description	Basic element
Latent variable or construct	
Observed variable or indicator	
Direct or directional causal relationship between two variables	
Indirect or non-directional relationship between two variables	
Correlation between two variables	
Relationship between two latent variables	
Relationship between an observed variable and a latent variable	
Measurement error in the observed variable	
Prediction error in the latent variable	

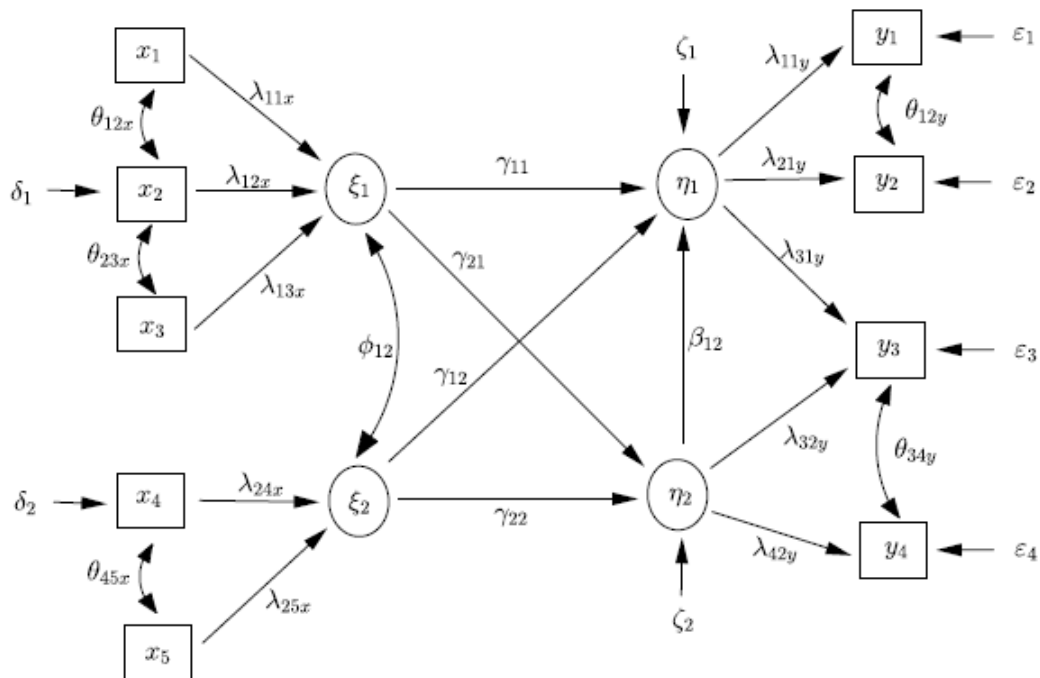
Source: Amorim et al. (2012).

Explanatory relationships between variables are represented by unidirectional arrows. Each of these arrows may be described as a linear regression. According to Hox e Bechger (1998), these relationships may also be understood as factor loadings if they are relative to the confirmatory factor analysis of measurements instruments. Curve arrows and bidirectional arrows represent correlation or covariance between the variables (PILATI; LAROS, 2007).

Based on these definitions it is possible to build a path diagram to organize all the supposed relationships in SEM (to which values will be fixed, or parameters will be adjusted, so called effects). These effects are also represented in the model as the Greek letters added to the indexes. All relationship supposed by the model must be represented in the diagram, even if their adjustment indicates a zero estimation to the corresponding parameter. If a relationship is not specified, it suggests that it is worth zero, for example, that two variables are not correlated.

Therefore, the structural equation model is completely determined by the path diagram, from which it is possible to specify the model mathematically. To better illustrate the concept of a path diagram, refer to the diagram below.

Figure 2.2 – Example of a path diagram.



Source: The author.

In Figure 2.2, the following structure is observed. Variables  $\xi_1$ ,  $\xi_2$ ,  $\eta_1$  and  $\eta_2$  are latent; variables  $x_1$ ,  $x_2$ ,  $x_3$ ,  $x_4$ ,  $x_5$  and  $y_1$ ,  $y_2$ ,  $y_3$ ,  $y_4$  are observed; variables  $\delta_1$ ,  $\delta_2$ ,  $\varepsilon_1$ ,  $\varepsilon_2$ ,  $\varepsilon_3$ ,  $\varepsilon_4$  and  $\zeta_1$ ,  $\zeta_2$  are error variables; variable  $\xi_1$  is caused by  $x_1$ ,  $x_2$  and  $x_3$  and  $\xi_2$  is caused by  $x_4$  and  $x_5$ ; variables  $\xi_1$  e  $\xi_2$  cause  $\eta_1$  and  $\eta_2$ ; variable  $\eta_1$  causes  $y_1$ ,  $y_2$  e  $y_3$  and  $\eta_2$  causes  $y_3$  and  $y_4$ ; variables  $\xi_1$  and  $\xi_2$ ,  $x_1$  and  $x_2$ ,  $x_2$  and  $x_3$ ,  $x_4$  and  $x_5$  are correlated, as well as  $y_1$  and  $y_2$ ,  $y_3$  and  $y_4$ .

### 2.1.1.1 Direct, indirect, and total effects

Based on the path diagram in Figure 2.2., it is possible to observe that a variable may influence another either directly or by means of another variable. This existing influence among variables is called effect. There are three types of effects in a path diagram: direct, indirect, and total.

According to Bollen (1989), a direct effect is the influence of a given variable on another, without the interference of any other variable. An indirect effect is defined as a variable that influences another by means of at least another variable. The total effect is determined by summing the direct effect and all indirect effects.

Bollen (1989) points out that this effect is related to the structure of the model under consideration, that is, if a new structural model is adjusted by altering, for example, the structure of correlations or the set of covariables, the effects are re-estimated.

Using the path diagram in Figure 2.2., we may observe, for example, that:

- a) The direct effect of  $\xi_1$  on  $\eta_1$  is  $\gamma_{11}$ , the indirect effect of  $\xi_1$  on  $\eta_1$  is  $\phi_{12}\gamma_{12} + \gamma_{21}\beta_{12}$ , and the total effect is  $\gamma_{11} + \phi_{12}\gamma_{12} + \gamma_{21}\beta_{12}$ ;
- b) The direct effect of  $\xi_1$  on  $\eta_2$  is  $\gamma_{21}$ , the indirect effect  $\xi_1$  on  $\eta_2$  is  $\phi_{12}\gamma_{22}$ , and the total effect is  $\gamma_{21} + \phi_{12}\gamma_{22}$ ;
- c) The direct effect of  $\xi_2$  on  $\eta_1$  is  $\gamma_{12}$ , the indirect effect of  $\xi_2$  on  $\eta_1$  is  $\phi_{12}\gamma_{11} + \gamma_{22}\beta_{12}$ , and the total effect is  $\gamma_{12} + \phi_{12}\gamma_{11} + \gamma_{22}\beta_{12}$ ;
- d) The direct effect of  $\xi_2$  on  $\eta_2$  is  $\gamma_{22}$ , o efeito indireto  $\xi_2$  on  $\eta_2$  is  $\phi_{12}\gamma_{21}$ , the indirect effect of  $\gamma_{22} + \phi_{12}\gamma_{21}$ .

### 2.1.1.2 Reflective indicators and formative indicators

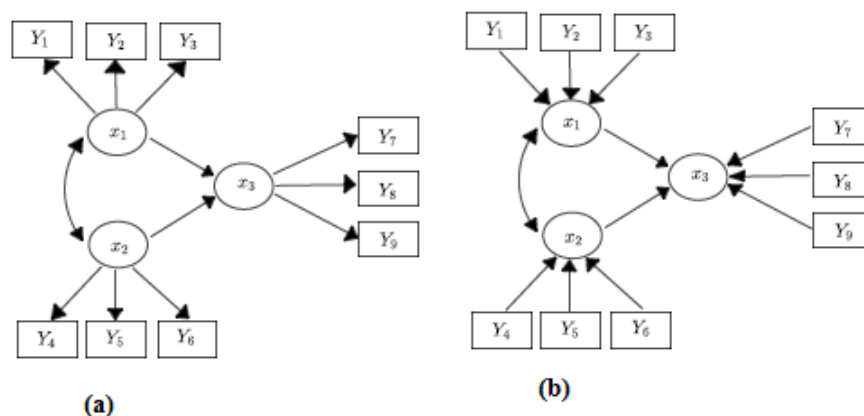
During model development, when there is a causal relationship between a latent variable and an observed variable, the latter is considered an indicator of the former. There are two types of indicators: reflective indicators and formative indicators (BISTAFFA, 2010).

In a reflective model, the observed variable is caused by the latent variable, that is, changes in the construct cause changes in the indicators. In this case, the observed variable is named reflective indicator for that construct. Example: “Grade obtained in a test” is an effect of the latent variable “Acquired knowledge” (PEREIRA, 2014).

Regarding a formative model, the indicator causes the construct, that is, variations in the indicators cause changes in the construct to which they are linked. Then, the observed variable is called formative indicator. Example: “Number of hours dedicated to studying” causes the latent variable “Study habits” (PEREIRA, 2014).

Most models have reflective indicators for the constructs. However, depending on their interpretation, in some cases, it would make more sense for them to be formative (BOLLEN, 1989). When specifying a SEM model, considering an indicator as formative or reflective may completely alter the result of the adjustment, and even influence the effects of other variables. In Figure 2.3 are two path diagrams: one involving reflective indicators, and another involving formative indicators.

Figure 2.3 – Reflective (a) indicators and Formative (b) indicators.



Source: Adapted from Brei e Neto (2006).

### 2.1.2 Structural model formulation

The complete structural equation model consists of a system of structural equations formed by the structural model and the measurement model. According to Bollen (1989), the equation of a structural model can be represented as:

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}. \quad (2.1)$$

in which  $\boldsymbol{\eta}_{m \times 1}$  is a vector of endogenous latent variables,  $\boldsymbol{\xi}_{n \times 1}$  is a vector of exogenous latent variables,  $\mathbf{B}_{m \times m}$  is a matrix of coefficients that relates  $m$  endogenous latent variables to one another,  $\boldsymbol{\Gamma}_{m \times n}$  is a matrix of coefficients that relates  $n$  exogenous latent variables with  $m$  endogenous latent variables and  $\boldsymbol{\zeta}_{m \times 1}$  is an error vector of structural equations.

Equation (2.1) can be rewritten as follows:

$$(\mathbf{I} - \mathbf{B})\boldsymbol{\eta} = \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}. \quad (2.2)$$

assuming that  $(\mathbf{I} - \mathbf{B})$  is not singular, then its inverse form exists. Otherwise, system (2.2) would either not have a solution, or have infinite solutions, which is not of interest. Therefore, equation (2.1) is given by:

$$\boldsymbol{\eta} = (\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}). \quad (2.3)$$

From equation (2.1), the structural model equation has the following matrixial form:

$$\begin{bmatrix} \eta_1 \\ \vdots \\ \eta_m \end{bmatrix} = \begin{bmatrix} 0 & \cdots & \beta_{1m} \\ \vdots & \ddots & \vdots \\ \beta_{m1} & \cdots & 0 \end{bmatrix} \times \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_m \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1n} \\ \vdots & \ddots & \vdots \\ \gamma_{m1} & \cdots & \gamma_{mn} \end{bmatrix} \times \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \vdots \\ \zeta_m \end{bmatrix}.$$

To complete the mathematical representation of the model, it is necessary to make a few suppositions about the parameters and define their correlation structure. Thus, suppose the expectation for the error vector and the latent variables is zero  $E(\boldsymbol{\zeta}) = 0$ ,  $E(\boldsymbol{\eta}) = 0$ ,  $E(\boldsymbol{\xi}) = 0$  and  $\boldsymbol{\zeta}$  is not correlated to  $\boldsymbol{\xi}$ , being the covariance matrix of  $\boldsymbol{\xi}$  given by  $\boldsymbol{\Phi}_{(n \times n)} = E(\boldsymbol{\xi}\boldsymbol{\xi}')$  and the matrix of  $\boldsymbol{\zeta}$  given by  $\boldsymbol{\Psi}_{(m \times m)} = E(\boldsymbol{\zeta}\boldsymbol{\zeta}')$ .



Regarding the measurement model, Bollen (1989) suggests that it may be represented by a set of equations, specified in terms of the exogenous variables or the endogenous variables, respectively, as follows:

$$\mathbf{x} = \mathbf{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta}. \quad (2.4)$$

$$\mathbf{y} = \mathbf{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\varepsilon}. \quad (2.5)$$

in which  $\mathbf{x}_{qx1}$  is a vector with  $q$  observed variables which indicate  $n$  exogenous latent variables,  $\mathbf{y}_{px1}$  is a vector with  $p$  observed variables which indicate  $m$  endogenous latent variables,  $\boldsymbol{\delta}_{qx1}$  is a vector of mensuration errors for  $x$ ,  $\boldsymbol{\varepsilon}_{px1}$  is a vector of mensuration errors for  $y$ ,  $\mathbf{\Lambda}_{x(qxn)}$  is a matrix of regression coefficients which relates  $n$  exogenous latent variables with each of the  $q$  indicators and  $\mathbf{\Lambda}_{y(pxm)}$  is a matrix of regression coefficients which relates the  $m$  endogenous latent variables with each of the  $p$  indicators.

In its matrixial form, the equation for the measurement model specified in terms of the exogenous variables is represented by:

$$\begin{bmatrix} x_1 \\ \vdots \\ x_q \end{bmatrix} = \begin{bmatrix} \lambda_{11x} & \cdots & \lambda_{1nx} \\ \vdots & \ddots & \vdots \\ \lambda_{q1x} & \cdots & \lambda_{qnx} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_q \end{bmatrix}.$$

In its matrixial form, the equation for the measurement model specified in terms of the endogenous variables is represented by:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} \lambda_{11y} & \cdots & \lambda_{1my} \\ \vdots & \ddots & \vdots \\ \lambda_{p1y} & \cdots & \lambda_{pmy} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{bmatrix}.$$

We maintained the model suppositions that the expectations for the error vectors and the latent variables are equal to zero  $E(\boldsymbol{\varepsilon}) = 0$ ,  $E(\boldsymbol{\delta}) = 0$ ,  $E(\boldsymbol{\xi}) = 0$ ,  $E(\boldsymbol{\eta}) = 0$ ,  $\boldsymbol{\varepsilon}$  is not correlated with  $\boldsymbol{\delta}$ ,  $\boldsymbol{\xi}$  and  $\boldsymbol{\eta}$ ,  $\boldsymbol{\delta}$  is not correlated with  $\boldsymbol{\varepsilon}$ ,  $\boldsymbol{\xi}$  and  $\boldsymbol{\eta}$ , and the covariance matrix of  $\boldsymbol{\varepsilon}$  is given by  $\Theta_{\boldsymbol{\varepsilon}(p \times p)} = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')$  and the covariance matrix of  $\boldsymbol{\delta}$  is given by  $\Theta_{\boldsymbol{\delta}(q \times q)} = E(\boldsymbol{\delta}\boldsymbol{\delta}')$ .

Considering the path diagram, as an example for Figure 2.2:

The structural model equations are given by:

$$\begin{aligned}\eta_1 &= \beta_{12}\eta_2 + \gamma_{11}\xi_1 + \gamma_{12}\xi_2 + \zeta_1 \\ \eta_2 &= \gamma_{21}\xi_1 + \gamma_{22}\xi_2 + \zeta_2.\end{aligned}\quad (2.6)$$

Their matrixial form is:

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} 0 & \beta_{12} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}.\quad (2.7)$$

The measurement model equations are given by:

$$\begin{aligned}\xi_1 &= \lambda_{11x}x_1 + \lambda_{12x}x_2 + \lambda_{13x}x_3 + \delta_1 \\ \xi_2 &= \lambda_{24x}x_4 + \lambda_{25x}x_5 + \delta_2.\end{aligned}\quad (2.8)$$

Their matrixial form is:

$$\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} \lambda_{11x} & \lambda_{12x} & \lambda_{13x} & 0 & 0 \\ 0 & 0 & 0 & \lambda_{24x} & \lambda_{25x} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix}.\quad (2.9)$$

The measurement model equations are given by:

$$\begin{aligned}y_1 &= \lambda_{11y}\eta_1 + \varepsilon_1 & y_2 &= \lambda_{21y}\eta_1 + \varepsilon_2 \\ y_3 &= \lambda_{31y}\eta_1 + \lambda_{32y}\eta_2 + \varepsilon_3 & y_4 &= \lambda_{42y}\eta_2 + \varepsilon_4.\end{aligned}\quad (2.10)$$

Their matrixial form is:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} \lambda_{11y} & 0 \\ \lambda_{21y} & 0 \\ \lambda_{31y} & \lambda_{32y} \\ 0 & \lambda_{42y} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}.\quad (2.11)$$

### 2.1.2.1 Model identification

Before estimating the structural equation model, it is necessary to determine whether they are identifiable. Identification aims at demonstrating that the unknown parameters of the model are only functions of identifiable parameters and that these functions have a single solution, that is, they are not improper or undefined. Therefore, if the parameters of the model are not identifiable, they cannot be estimated.

The identifiability of a structural model usually occurs when the number of elements in the covariance matrix among the observed variables ( $\Sigma$ ) is larger than or equal to the number of parameters to be estimated (KLINE, 2004).

Should the parameter be only a function of parameters that can be estimated, and this function present a single solution, then, the parameters is identified. If all parameters in the model are identified, then the model is identified. Should there be multiple solutions, the parameter is considered underidentified. Should there be excessive information to estimate a parameter, then, it is considered overidentified. If at least one parameters in the model is overidentified, the whole model is considered overidentified (PEREIRA, 2014).

In the literature, there are certain rules that check whether a model can be considered identified. In this study, rule t was used.

#### Rule t

Rule t establishes that, so a model may be identified, the number of unknown parameters must be less than or equal to the number of elements in the variance-covariance matrix of the observed variables, that is,

$$t \leq \frac{(p+q)(p+q+1)}{2}, \quad (2.12)$$

in which  $t$  is the number of unknown parameters in the model;  $p$  is the number of observed variables of  $\eta'$ s;  $q$  is the number of observed variables of  $\xi'$ s (BOLLEN; DAVIS, 2009).

### 2.1.2.2 Model estimation

SEM aims at reproducing the populational covariance matrix by means of sample covariances associated to the imposition of parameters determined by the researcher. Therefore, to estimate the parameters of a structural equation model, one assumes the fundamental hypothesis  $\Sigma = \Sigma(\theta)$ , in which  $\Sigma$  is the populational covariance model of  $x$  and  $y$  and  $\Sigma(\theta)$  is the covariance matrix written as a function of the model parameters, given by:

$$\Sigma(\theta) = \begin{bmatrix} \Lambda_y(I-B)^{-1}(\Gamma\Phi\Gamma' + \Psi)(I-B')^{-1}\Lambda_y' + \Theta_\varepsilon & \Lambda_y(I-B)^{-1}\Gamma\Phi\Lambda_x' \\ \Lambda_x\Phi\Gamma'(I-B')^{-1}\Lambda_y' & \Lambda_x\Phi\Lambda_x' + \Theta_\delta \end{bmatrix}. \quad (2.13)$$

Because the populational covariance matrix of the observed variables is unknown, the sample covariance matrix, resulting from the equation system  $\hat{\Sigma} = \Sigma(\hat{\theta})$ , is used (BOLLEN, 1989).

With these specifications and based on the path diagram, usual estimation methods are adapted, including least squares, maximum likelihood, ridge estimation. Next, we present a synthesis of the methods used in this study to adjust the regressions that form the structural models.

### 2.1.2.3 Considerations on error specifications and interpretation

The formalization of indexes that validate constructs, since it considers error variance, may reveal different interpretations, due to the nature of the variable to be observed or latent, as well as the exogenous or endogenous classification.

In this sense, reference is made to the review described by Gosling e Gonçalves (2003), in which they mention that the errors are exogenous variables, therefore, their variances are parameters to be estimated. The authors also suggest that the estimation of each parameter, given as error variance, may be interpreted as relative information, somehow described in indexes that quantify the variability concentrated on the associated endogenous variable, due to the absence of other influences (arrows) from other variables in the model.

Hair et al. (2006) explains that every exogenous variable in the model (whether observable, or latent) will have a variance which will be defined as parameter of the model. Endogenous variables also have variances, but they are not parameters. The variances of endogenous variables are estimated by the other variables and influences in the model, that is, the variance of any

endogenous variable may be algebraically expressed as a function of the variance of exogenous variables, including the errors and parameters associated with the linear influences in the model.

Chou e Bentler (1995) consider that the parameters of the model to be estimated from the data are regression coefficients and the variances and covariances of independent variables. MacCallum et al. (1995) also explains that any directional effect specified in the model is another category of parameters. Such directional effects include the effects of latent variables on other latent variables.

Given this brief discussion, we understand that the research on indices that provide more accuracy in the validation of constructs is still a wide field of research, which begs for new methods on different suppositions to be incorporated into the specification of constructs, thus providing analytical and interpretable results for the researcher who uses the structural model equation approach.

## 2.2 Concepts and preliminary definitions of robust inference

Robust statistics may be described as a generalization of the traditional statistics, which considers the possibility of incorrect specifications in the model and in the distribution of the data under analysis (DAMIÃO, 2007).

### 2.2.1 Breakdown point

There are estimators which can deal with data containing a certain number of outliers. To formalize this aspect, the concept of breakdown point was elaborated.

The breakdown point of an estimator was asymptotically defined by Hampel (1971). However, Rousseeuw e Leroy (1987) approach a version for finite samples, adapted by Donoho e Huber (1983). The authors present an idea of breakdown point as the largest proportion of data contaminated by outliers that an estimator could stand and still provide good estimations.

Consider a dataset,  $Z = \{(x_{11}, x_{12}, \dots, x_{1p}, y_1), \dots, (x_{n1}, x_{n2}, \dots, x_{np}, y_n)\}$ , which relate to each other according to a regression model. Consider  $T$  a regression estimator, so that if we apply  $T$  to sample  $Z$ , we produce a vector of regression coefficients  $\hat{\theta}$ , that is,  $T(Z) = \hat{\theta}$ , in which  $\hat{\theta} = (\theta_1, \dots, \theta_p)$ .

Now, consider all possible contaminated samples  $Z'$ , which are obtained by replacing any  $m'$  observations from the original dataset with outliers. Then, the contamination ratio is

given by  $\varepsilon = \frac{m'}{n}$ . We will call  $b(m'; T, Z)$  the maximum bias that may be caused by any contamination  $\varepsilon$  and that is defined by:

$$b(m'; T, Z) = \sup_{Z'} \left\| T(Z') - T(Z) \right\|, \quad (2.14)$$

in which the supreme is calculated in relation to all possible sets  $Z'$  resulting from the replacement of  $m'$  observations in the data set  $Z$ .

If  $b(m'; T, Z)$  is infinite, then the  $m'$  outliers may produce an arbitrarily large effect on  $T$ , and a “breakdown” may occur for some of them. Thus, the breakdown point of estimator  $T$  is given by:

$$\varepsilon_n^*(T, Z) = \min \left\{ \frac{m'}{n}; b(m'; T, Z) = \infty \right\}, \quad (2.15)$$

that is, it is the lowest contamination ratio that may cause estimator  $T$  to assume values arbitrarily distant from  $T(Z)$ .

Considering a finite sample  $Z$ , for example, for a sample average, we notice that a single outlier may strongly affect the estimation of least squares, because if a single observation assumes a value that tends to infinity, then the sample average will also tend to infinity. Thus, the breakdown point of this estimator is given by:

$$\varepsilon_n^*(\bar{X}, Z) = \frac{1}{n} \quad \text{e} \quad \lim_{n \rightarrow \infty} \frac{1}{n} = 0.$$

Therefore, it is possible to affirm that the method of least squares has a breakdown point of 0%, that is, a single outlier is enough to affect estimator  $T$ .

### 2.2.2 Elementary sets

The resampling algorithm of elementary sets was proposed by Theil (1950). The author introduced this algorithm as a computational method to calculate estimators with high a breakdown point, in which the exact solution of a goal function is computationally complex (MACHADO, 1997).

The resampling algorithm consists of removing, but not replacing, all possible subsets (subsamples) of  $k$  size from a data set. These subsets should have the minimum size enough to

make it possible for the parameters of the model to be estimated (MACHADO, 1997; CHAGAS, 2011).

Consider the data set  $Z_{n \times (k+p)} = (X_{n \times k} Y_{n \times p})$  whose elements relate to each other following a multivariate regression model. Matrix  $Z$  as follows:

$$Z = \begin{pmatrix} X_J & Y_J \\ X_{n-J} & Y_{n-J} \end{pmatrix}. \quad (2.16)$$

in which  $Y_J$  and  $X_J$  are, respectively, the submatrix of size  $k \times p$  of the response variable and the complete submatrix  $k \times k$  of the predicting variables, related to the observations of this set.

Consider  $S = \{\{1, \dots, k\}, \dots, \{j_1, \dots, j_k\}, \dots, \{(n-k+p), \dots, n\}\}$  the set of all possible subsets of indices we may obtain with  $n$  and  $k$  fixated, in which  $J = \{j_1, \dots, j_k\}$  is a subset of indices with  $k$  observations of the original data.

Each subset  $Z_J = \begin{pmatrix} X_J & Y_J \end{pmatrix}$ ,  $J \subset S$ , in which

$$Z_J = \begin{pmatrix} X_{J_1 1} & \cdots & X_{J_1 k} & Y_{J_1 1} & \cdots & Y_{J_1 p} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ X_{J_k 1} & \cdots & X_{J_k k} & Y_{J_k 1} & \cdots & Y_{J_k p} \end{pmatrix}. \quad (2.17)$$

is named an elementary set.

Considering the elementary regression as the adjustment of regression of  $Y_J$ , as a function of  $X_J$ , the estimator of this type of regression is given by

$$\hat{\beta}_J = (X_J' X_J)^{-1} X_J' Y_J. \quad (2.18)$$

Based on the estimation method used, the algorithm calculates the predicted elementary residues for each elementary set  $Z_J$ , and chooses, as final estimator of the regression coefficients, vector  $\hat{\beta}_J$ , whose predicted elementary residues minimize the goal function (MACHADO, 1997; CHAGAS, 2011).

### 2.2.3 Estimator of Least Trimmed of Squares (LTS)

The regression of least trimmed of squares (LTS) is a highly robust method to adjust a linear regression model. Furthermore, it is an alternative to the traditional techniques of least

squares when the error distributions do not meet the conditions of normality (that is, the errors do not fit a normal distribution) or when the data have significant outliers (JUNIOR et al., 2003).

Estimation  $\hat{\beta}_{LTS}$  minimizes the sum of the  $h$  lowest square residues.

$$\min_{\hat{\beta}} \sum_{i=1}^h (\varepsilon^2)_{i:n}. \quad (2.19)$$

in which  $\{\varepsilon^2\}_{1:n} \leq \dots \leq \{\varepsilon^2\}_{n:n}$  are the ordinated square residues.

The cutoff constant is  $h$ ,  $\frac{n}{2} < h \leq n$ , in which  $\frac{h}{n}$  is the ratio of observations used by the LTS estimator. This ratio determines the breakdown point of LTS, since definition (2.19) implies that  $n - h$  observations with the largest residues do not affect the estimator directly (ČÍŽEK, 2013).

The maximum breakdown point is asymptotically equal to  $\frac{1}{2}$  (or 50%) for  $h = \frac{n}{2} + \frac{p+1}{2}$ , in which  $p$  represents the total number of parameters. As for  $h = n$ , the maximum breakdown point is asymptotically equal to 0, which corresponds to the least square estimator. Quantity  $h$  may also depend on a cutoff proportion  $\alpha$ , for example,  $h = [n(1 - \alpha)] + 1$  or  $h = [n(1 - \alpha)] + [\alpha(p + 1)]$ . Thus, the breakdown point will be approximately equal to  $\alpha$ . If  $\alpha$  tends to 50%, we have the estimator of least trimmed of squares, whereas if  $\alpha$  tends to 0%, then we have the estimator of least squares (ROUSSEEUW; LEROY, 1987).

The LTS estimator has as good properties as a robust estimator. The LTS is equivariant, that is, it is an estimator that correctly transforms its estimators in accordance with the changes made in the observations (ROUSSEEUW; LEROY, 1987). The LTS meets the three properties below:

a) Estimator T is an equivariant of Regression if:

$T((x_i; y_i + x_i v); i = 1, \dots, n) = T((x_i; y_i); i = 1, \dots, n) + v$ , in which  $v$  is any vector,  $x_i$  are explanatory variables and  $y_i$  are response variables.

b) Estimator T is an equivariant of Scale if:  $T((x_i; c y_i); i = 1, \dots, n) = c T((x_i; y_i); i = 1, \dots, n)$ ,  $c$  is a constant.

c) Estimator T is an Affine Equivariant if  $T((x_i A; y_i); i = 1, \dots, n) = A^{-1} T((x_i; y_i); i = 1, \dots, n)$   $A$  is a non-singular square matrix.

LTS reaches the usual consistency  $\sqrt{n}$ , that is, LTS converges faster than other robust estimators and displays under normality the asymptotic relative efficiency of 8% (ČÍŽEK, 2013).



The main deficiency of LTS is that its computational complexity is far less understood than that of other robust estimators. The most practical approach might then be heuristics by Rousseeuw e Driessen (2006).

#### 2.2.4 Adaptive estimator

In practical situations, to a given data set, it is often unclear whether the distribution of error is normal. Thus, the choice between ordinary least squares (OLS) and robust estimators remains subjective for the researcher, and the efficiency of the estimator may depend a lot on that choice. In general, the issue of parameter estimation when the errors are not normal is deeply studied, and several authors have suggested countless alternatives.

Arnab e Michael (2008) cite robust regression methods as the procedures of regression  $L_p$ . Thus, choosing  $p$  has become a very important matter. Sposito (1990) proposed the choice of  $p$  based on the kurtosis coefficient,  $k$ , of the error distribution.

According to Arnab e Michael (2008), Sposito (1990) imposes the choice of a single estimator based on the estimated kurtosis. For slightly simple distributions, this may not represent a serious problem, but for heavy error distributions, that is not the same. To remedy the effect of the performance of a kurtosis sample estimator, one must combine the estimators, as weighted mean, instead of considering a single estimator.

Based on this, Arnab e Michael (2008) proposed an estimator which is simply a linear combination of  $\hat{\beta}_{OLS}$  and the robust estimator. Thus, the estimator maintains the efficiency of the robust estimator in the presence of outliers or heavy-tailed error distributions but does not lose much efficiency when the conditions of the estimator of ordinary least squares are truly meet. Considering the robust estimator LTS, the estimator known as adaptive weighted estimator may be represented by the following equation:

$$\hat{\beta}_{ADP} = w_k \hat{\beta}_{OLS} + (1 - w_k) \hat{\beta}_{LTS}. \quad (2.20)$$

in which, weight  $w_k$  must be chosen so it may express the nature of the error distribution. In addition, more weight should be given to OLS for light-tailed distributions, whereas, for heavy-tailed distributions, LTS should have more weight. To assess the nature of the error distribution, the residues of OLS and of the LTS robust estimator are used, and the sample kurtosis is analyzed. Kurtosis is estimated by the average kurtosis of two sets of residues:

$$k = \{ \text{kurtosis (OLS residue)} + \text{kurtosis (LTS residue)} \} / 2. \quad (2.21)$$

In this work, the kurtosis measurement was defined as the ratio between the fourth central moment and the square of the second central moment (CASELLA; BERGER, 2011).

$$\frac{\mu_4}{\mu_2} = \frac{E(X - E(X))^4}{[E(X - E(X))^2]^2}. \quad (2.22)$$

Thus, weight  $w_k$  is chosen from the data and may be represented by

$$w_k = \left\{ \begin{array}{ll} 1 & \text{se } k \leq 3 \\ 3/k & \text{se } k > 3 \end{array} \right\}. \quad (2.23)$$

Moreover, this estimator is directly implemented because only OLS and LTS should be executed. The method is general and easy to use, which makes it effective for a wide range of error distributions.

### 2.2.5 Validity

Validity refers to an instrument measuring exactly what has been proposed, that is, validity assesses, in measurement equations, whether a given indicator is really a measure of whatever is proposed (ROBERTS; PRIEST, 2006; MOKKINK et al., 2010).

**Construct Validity:** construct validity refers to how much an indicator truly reflects the latent construct it measures, that is, it is the extension in which a set of variables truly represents the construct to be measured (HAIR et al., 2009).

This sort of validity is hardly obtained with a single study. Generally, several research studies are done about the theory of the construct to be measured (MARTINS, 2006). Therefore, the more evidence, the more valid the interpretation of results is.

To confirm the validity of a construct, two types of validity are used: convergent validity and discriminating validity.

**Convergent validity:** convergent validity corresponds to the level at which the indicators assigned to measure a given construct are related and convergent. Thus, convergent validity is understood to exist when two different measurements of the same construct confirm the expectation of them being strongly related to each other, that is, there is a high correlation coefficient between both measurements (COSTA, 2011).

To assess the convergent validity, the factor loadings could be assessed. High factor loadings are an indication that they converge onto a common spot, that is, there is convergent validity. The literature states that the factor loadings should be at least 0.5, and ideally higher than that (HAIR et al., 2009). Furthermore, the criterium proposed by Fornell e Larcker (1981), which indicates convergent validation when the Average Variance Extracted is higher than 50%, is used.

The Average Variance Extracted represents the average proportion of the variance of the indicators explained by the latent variable (VALENTINI; BRUNO, 2017).

To calculate this index, Hair et al. (2009) proposes the following equation:

$$AVE = \frac{\sum(\lambda^2)}{\sum(\lambda^2) + \sum(var(\varepsilon))}. \quad (2.24)$$

in which *AVE* is the average variance extracted;  $\lambda^2$  represents the squared factor loadings; therefore  $\sum(\lambda^2)$  means the sum of squared factor loadings, and  $\sum(var(\varepsilon))$  is the sum of variances.

To calculate AVE, Valentini e Bruno (2017) suggest that standard factor loadings should be used. Thus, the AVE index may also be understood simply as the mean of the standard squared factor loadings. Considering that  $\lambda^2 + \varepsilon = 1$ , if the standard factor loadings are used, the denominator of equation (2.24) will be multiplied by the number of indicators. Therefore, only when considering the standard factor loadings can the denominator of equation (2.24) be replaced with the number of indicators, and the equation be simplified as follows:

$$AVE = \frac{\sum(\lambda_p^2)}{p}. \quad (2.25)$$

in which,  $\lambda_p^2$  represents the squared standard factor loading and  $p$  represents the number of indicators.

**Discriminating validity:** discriminating validity is the level at which a construct differs from the others (HAIR et al., 2009). It tests the hypothesis that the measurement under study is not improperly related to different constructs, that is, with variables from which it should differ (POLIT, 2015).

The main way to assess the discriminating validity is by comparing the squared roots of the AVE values of each construct with the correlation values between the latent constructs. A discriminating validity will occur if the latent variables are lower than the AVE squared roots (HAIR et al., 2009).

### 2.3 Exploratory factor analysis

The exploratory factor analysis (EFA) is a statistical method used to identify subjacent relationships among observed variables (RUOTOISTENMÄKI; SEPPÄLÄ, 2007). According to Joseph et al. (2005) by investigating the correlation structure in a set of observed variables, EFA determines the factor(s) that best explains their covariance. Thus, the observed variables belong to the same factor if, and when they share a variance, that is, if they are influenced by the same subjacent construct (BROWN, 2015).

The mathematical model is given by:

$$x_i = \lambda_{i1}\xi_1 + \lambda_{i2}\xi_2 + \dots + \lambda_{in}\xi_n + \delta_i, i = 1, 2, \dots, q, n, q \in N, n \leq q. \quad (2.26)$$

in which  $x_1, x_2, \dots, x_q$  are observed variables (or indicators);  $\xi_1, \xi_2, \dots, \xi_q$  are common factors (or latent variables);  $\lambda_{ij}$  is the estimated coefficient, named factor loading, which represents the correlation between the observed variables and the common factors. Since the observed variables and the common factors are standard variables, the matrix of correlative coefficients is the same as the covariance matrix and  $\delta_i$  is the term of error (FABRIGAR et al., 1999; WANG et al., 2015).

EFA is generally used in the initial phases of a research study to explore data. The exploratory analysis should occur when the researcher has little or no knowledge about the latent structure behind the set of observed variables in the research. In other words, it should occur when there is no prior empirical study or when the theory supporting the phenomenon is initial in that no one knows beforehand how many dimensions may result from it, nor the construct composition. Thus, the exploratory use makes it possible to understand the latent structure and, therefore, the phenomenon (JORESKOG, 2007; PASQUALI, 2012; BIDO; MANTOVANI; COHEN, 2018).

When running EFA, several decisions must be made to obtain an adequate factor structure (COSTELLO; OSBORNE, 2005). Because the results obtained during the EFA are strongly dependent on the decisions made by the researcher, the technique is highly likely to provide wrong and/or unreliable results (PATIL et al., 2008)). Therefore, all decisions made during an EFA must be based on clear, methodological, and theoretical criteria, seeking to obtain adequate factor models (DAMÁSIO, 2012).

After the initial data exploration and verification that the database is adequate, the exploratory factor analysis should be run. First, it is necessary to find the number of factors that best represent the correlation pattern between the variables. An optimal solution would be finding the minimum number of factors that maximize the total explained variance (FILHO; JÚNIOR, 2010). However, there is not only one criterion to determine the number of factors. The main methods found in the literature are eigenvalue, Scree test, and accumulated variance percentage (MATOS; RODRIGUES, 2019).

After defining the number of factors in the model, the next step is to decide which technique will be used to calculate the factor loadings. There are various techniques to extract the factors, such as principal components, principal factors, maximum likelihood, ordinary least squares, general least squares (MATOS; RODRIGUES, 2019).

After extracting the factors, it is possible use the factor loadings to calculate how well the variables adapt to the factors (MATOS; RODRIGUES, 2019). In this sense, the technique of factor rotation is used to maximize high charges between the factors and the variables, and minimize low charges, thus obtaining a better distinction among the factors and making the empirical result more easily interpretable (TABACHNICK; FIDELL, 2007).

According to Field (2009), the rotation choice will depend on whether there is good theoretical reason to suppose the factors are related or independent. There are two types of factor rotation:

- Orthogonal factor rotation: each factor is independent (orthogonal) in relation to the others, that is, there is no correlation between the factors.
- Oblique factor rotation: is calculated so that the extracted factors are correlated (JOSEPH et al., 2005).

These methods differ on how factors are rotated and produce different results. Regarding the available options for each of these types of rotation, software R has four methods of orthogonal rotation (varimax, quartimax, BentlerT, and geominT). Varimax method is the most used one (FIELD; MILES; FIELD, 2012). This method seeks to minimize the number of variables that have high loadings in each factor, which result in groups of more interpretable factors. Software R also has five methods of oblique rotation (oblimin, promax, simplimax, BentlerQ, and geominQ). Promax is the fastest method developed for large databases and one of the most important methods of oblique rotation (MATOS; RODRIGUES, 2019). Promax is the method

which eliminates the assumption that the factors are independent, thus allowing them to rotate freely and have their interpretation simplified (HAIR et al., 2006).

The final step of EFA consists of examining how the variables group, naming the factors, and theoretically justifying how variables and factors relate to each other. All variables belonging to a given factor should be examined, especially those with higher loadings, and the factor that provides the most adequate reflection for the set of variables belonging to it should be named.

It is worth pointing out that the researcher should always be based on a theory and/or prior research because, even though this is an exploratory technique, there should always be some sort of hypothesis about the group of variables.

#### **2.4 Synthesis of opinions given by panel members on published papers**

After the valuable arguments and contributions given at the end of this work, the main points were specifically related to creating and naming an index with the acronym AVE, which reflects an average of the error variances extracted by the construct.

As requested, the use of this name was widely considered, therefore, we were motivated to insert in the introduction its use in relation to an approach via plug-in estimation. Another issue was related to the different interpretations on error variance in the literature, upon the specification of the nature of the variables (observed or latent), which are classified as exogenous or latent. In a practical scenario, there is some controversy using the conventional AVE index.

In this sense, we added Section 2.1.2.3, which allows the researcher to pay attention to the treatment of error variance when elaborating the index; however, in the approach used to publish the papers, the AVE index was considered formalized as a function of the factor loadings, due to the relationship of a plug-in adaptation.

Particularly regarding the second paper, which covers the application of Adaptive AVE index to real data, we agree that the most adequate name for the constructs would be given by expressing feelings that were supposedly explained by the empirical method. Therefore, it makes more sense naming the constructs that explain the transition between specialty coffee consumers, defined by regularity, enthusiasm, and expertise.

Moreover, still on this paper, we were asked in which situation outliers are likely to appear. Given that, we register in this opinion, that outliers are caused simply by divergent observations, and they may be represented by a set of scores lower than the higher responses, for

example, in a likert scale, score 1 diverges from scores 4 and 5, considering the transformation onto a continuous scale of 0-1.

We followed a rigorous writing style, standardizing notes in the entire theoretical background, to minimize possible misunderstandings by other readers and researchers who may have access to this dissertation.

Lastly, we would like to show our gratitude for the debate on the papers and the discussion on the theoretical background, which provided us with valuable viewpoints regarding the analysis of structural equation models.

## REFERENCES

- AMORIM, L. D. A. F. et al. Modelagem com equações estruturais: Princípios básicos e aplicações. 2012.
- ANDERSON, T. W.; RUBIN, H. et al. **Statistical inference in factor analysis**. California: Proceedings of the third Berkeley symposium on mathematical statistics and probability, 1956. v. 5. 111–150 p.
- ARNAB, M.; MICHAEL, S. On adaptive linear regression. **Journal of Applied Statistics**, Taylor & Francis, v. 35, n. 12, p. 1409–1422, 2008.
- BENTLER, P. M. Simultaneous equation systems as moment structure models: With an introduction to latent variable models. **Journal of econometrics**, Elsevier, v. 22, n. 1-2, p. 13–42, 1983.
- BIDO, D. S.; MANTOVANI, D. M. N.; COHEN, E. D. Destruição de escalas de mensuração por meio da análise fatorial exploratória nas pesquisas da área de produção e operações. **Gestão & Produção**, SciELO Brasil, v. 25, n. 2, p. 384–397, 2018.
- BISTAFFA, B. C. **Incorporação de indicadores categóricos ordinais em modelos de equações estruturais**. Tese (Doutorado) — Universidade de São Paulo, 2010.
- BOCK, R. D.; BARGMANN, R. E. Analysis of covariance structures. **Psychometrika**, Springer, v. 31, n. 4, p. 507–534, 1966.
- BOLLEN, K. A. **Structural Equations with Latent Variables**. 1. ed. New York: Wiley, 1989.
- BOLLEN, K. A. Latent variables in psychology and the social sciences. **Annual review of psychology**, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 53, n. 1, p. 605–634, 2002.
- BOLLEN, K. A.; DAVIS, W. R. Causal indicator models: Identification, estimation, and testing. **Structural Equation Modeling: A Multidisciplinary Journal**, Taylor & Francis, v. 16, n. 3, p. 498–522, 2009.
- BREI, V. A.; NETO, G. L. O uso da técnica de modelagem em equações estruturais na área de marketing: um estudo comparativo entre publicações no brasil e no exterior. **Revista de Administração Contemporânea**, SciELO Brasil, v. 10, n. 4, p. 131–151, 2006.
- BROWN, T. A. **Confirmatory factor analysis for applied research**. Boston: Guilford publications, 2015. 2. ed p.
- BROWNE, M. W. Generalized least squares estimators in the analysis of covariance structures. **South African statistical journal**, South African Statistical Association (SASA), v. 8, n. 1, p. 1–24, 1974.
- CARTER, S. R. Using confirmatory factor analysis to manage discriminant validity issues in social pharmacy research. **International Journal of Clinical Pharmacy**, Springer, v. 38, n. 3, p. 731–737, 2016.
- CASELLA, G.; BERGER, R. L. Inferência estatística-tradução da 2ª edição norteamericana. **Centage Learning**, 2011.



- CASTRO, C. M. S. d. et al. Determinantes do trabalho remunerado entre brasileiros mais velhos usando modelagem de equações estruturais: evidências do elsi-brasil. **Cadernos de Saúde Pública**, SciELO Public Health, v. 36, p. e00194619, 2020.
- CHAGAS, E. d. N. Eficiência de estimadores robustos a observações discrepantes em regressão multivariada com aplicação na análise sensorial de café. Universidade Federal de Lavras, 2011.
- CHOU, C.-P.; BENTLER, P. M. Estimates and tests in structural equation modeling. In: **Hoyle Rick H. (ed.) Structural Equation Modeling: Concepts, Issues and Applications**, London: Sage Publications Inc, Cap.3, p. 37–55, 1995.
- ČÍŽEK, P. Reweighted least trimmed squares: an alternative to one-step estimators. **Test**, Springer, v. 22, n. 3, p. 514–533, 2013.
- COSTA, F. d. Mensuração e desenvolvimento de escalas: aplicações em administração. **Rio de Janeiro: Ciência Moderna**, 2011.
- COSTELLO, A. B.; OSBORNE, J. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. **Practical assessment, research, and evaluation**, v. 10, n. 1, p. 7, 2005.
- DAMÁSIO, B. F. Uso da análise fatorial exploratória em psicologia. **Avaliação Psicológica: Interamerican Journal of Psychological Assessment**, Instituto Brasileiro de Avaliação Psicológica (IBAP), v. 11, n. 2, p. 213–228, 2012.
- DAMIÃO, J. **Comparação de carteiras otimizadas segundo o critério média-variância através de estimativas robustas de risco e retorno. 2007. 36 p.** Tese (Doutorado) — Dissertação (Mestrado em Economia)—Faculdade Ibmec São Paulo, São Paulo, SP, 2007.
- DONOHU, D. L.; HUBER, P. J. The notion of breakdown point. **A festschrift for Erich L. Lehmann**, Belmont, Wadsworth, v. 157184, 1983.
- FABRIGAR, L. R. et al. Evaluating the use of exploratory factor analysis in psychological research. **Psychological methods**, American Psychological Association, v. 4, n. 3, p. 272, 1999.
- FARZANDIPOUR, M. et al. Designing a national model for assessment of nursing informatics competency. **BMC medical informatics and decision making**, BioMed Central, v. 21, n. 1, p. 1–12, 2021.
- FIELD, A. **Descobrendo a estatística usando o SPSS-5**. Porto Alegre: Penso Editora, 2009.
- FIELD, A. P.; MILES, J.; FIELD, Z. **Discovering statistics using R/Andy Field, Jeremy Miles, Zoë Field**. England: London; Thousand Oaks, Calif.: Sage, 2012.
- FILHO, D. B. F.; JÚNIOR, J. A. d. S. Visão além do alcance: uma introdução à análise fatorial. **Opinião pública**, SciELO Brasil, v. 16, n. 1, p. 160–185, 2010.
- FORNELL, C.; LARCKER, D. F. Evaluating structural equation models with unobservable variables and measurement error. **Journal of marketing research**, Sage Publications Sage CA: Los Angeles, CA, v. 18, n. 1, p. 39–50, 1981.
- GOLDBERGER, A. S. et al. Econometric theory. **Econometric theory**, New York: John Wiley & Sons., 1964.

GOSLING, M.; GONÇALVES, C. A. Modelagem por equações estruturais: conceitos e aplicações. **Revista de Administração FACES Journal**, v. 2, n. 2, p. 83–95, 2003.

GUIMARÃES, E. R. et al. The brand new brazilian specialty coffee market. **Journal of Food Products Marketing**, Taylor & Francis, v. 25, n. 1, p. 49–71, 2019.

HAIR, J. F. et al. **Multivariate data analysis 6th Edition**. United States: New Jersey: Prentice Hall, 2006.

HAIR, J. F. et al. **Análise multivariada de dados**. 6. ed. Porto Alegre: Bookman Editora, 2009.

HAMPEL, F. R. A general qualitative definition of robustness. **The Annals of Mathematical Statistics**, JSTOR, p. 1887–1896, 1971.

HOX, J. J.; BECHGER, T. M. An introduction to structural equation modeling. **Structural Equation Modeling**, 1998.

JOHNSON, D. R.; CREECH, J. C. Ordinal measures in multiple indicator models: A simulation study of categorization error. **American Sociological Review**, JSTOR, p. 398–407, 1983.

JÖRESKOG, K. G. A general approach to confirmatory maximum likelihood factor analysis. **Psychometrika**, Springer, v. 34, n. 2, p. 183–202, 1969.

JÖRESKOG, K. G. A general method for estimating a linear structural equation system. **ETS Research Bulletin Series**, Wiley Online Library, v. 1970, n. 2, p. i–41, 1970.

JÖRESKOG, K. G. A general method for estimating a linear structural equation system. **Em Structural Equation Models in the Social Sciences**, New York, USA. Academic Press., v. 1970, n. 2, p. 85–112, 1973.

JÖRESKOG, K. G. Factor analysis and its extensions. **Factor analysis at**, v. 100, p. 47–78, 2007.

JÖRESKOG, K. G.; GOLDBERGER, A. S. Factor analysis by generalized least squares. **Psychometrika**, Springer, v. 37, n. 3, p. 243–260, 1972.

JÖRESKOG, K. G.; SORBOM, D. **LISREL VI: Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods**. 1. ed. New York: Scientific Software, 1986.

JOSEPH, F. H. et al. **Análise multivariada de dados (AS Sant'Anna, Trad.)**. Porto Alegre: Bookman, 2005.

JUNIOR, J. R. H. et al. Vegetation greenness impacts on maximum and minimum temperatures in northeast colorado. **Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling**, Wiley Online Library, v. 10, n. 3, p. 203–215, 2003.

KEESLING, J. W. Maximum likelihood approaches to causal analysis. **Ph. D. dissertation. Department of Education: University of Chicago**, 1972.

KLING, R. B. **Principles and practice of structural equation modeling (methodology in the social sciences)**. 3. ed. New York: The Guilford Press, 2004.

KYOGOKU, M. et al. Development of the assessment of belief conflict in relationship-14 (abc-14). **PloS one**, Public Library of Science, v. 10, n. 8, p. e0129349, 2015.

LAWLEY, D. N. Vi.—the estimation of factor loadings by the method of maximum likelihood. **Proceedings of the Royal Society of Edinburgh**, Royal Society of Edinburgh Scotland Foundation, v. 60, n. 1, p. 64–82, 1940.

MACCALLUM, R. C. et al. Alternative strategies for cross-validation of covariance structure models. **Multivariate Behavioral Research**, v. 29, n. 1, p. 1–32, 1995.

MACHADO, H. C. **Detecção de dados atípicos e métodos de regressão com alto ponto de ruptura. Dissertação (Mestrado em Estatística)**. 145 p. Dissertação (Mestrado) — Universidade Estadual de Campinas, Campinas, 1997.

MARÔCO, J. **Análise de equações estruturais: Fundamentos teóricos, software & aplicações**. 1. ed. Pêro Pinheiro: ReportNumber, Lda, 2010.

MARTINS, G. Sobre confiabilidade e validade. **RBGN**, v. 8, n. 20, p. 1–12, 2006.

MATOS, D. A. S.; RODRIGUES, E. C. **Análise fatorial**. Escola Nacional de Administração Pública (Enap), 2019.

MOKKINK, L. B. et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. **Journal of Clinical Epidemiology**, Elsevier, v. 63, n. 7, p. 737–745, 2010.

MUTHÉN, B.; KAPLAN, D. A comparison of some methodologies for the factor analysis of non-normal likert variables. **British Journal of Mathematical and Statistical Psychology**, Wiley Online Library, v. 38, n. 2, p. 171–189, 1985.

PASQUALI, L. **Análise fatorial para pesquisadores. Laboratório de Pesquisa em Avaliação e Medida (LabPAM)-Instituto de Psicologia**. Brasília: Universidade de Brasília, 2012.

PATIL, V. H. et al. Efficient theory development and factor retention criteria: Abandon the ‘eigenvalue greater than one’ criterion. **Journal of Business Research**, Elsevier, v. 61, n. 2, p. 162–170, 2008.

PEREIRA, G. A. **Estimadores ridge generalizados adaptados em modelos de equações estruturais: estudo de simulação e aplicação no perfil de consumidores de café**. Universidade Federal de Lavras, 2014.

PILATI, R.; LAROS, J. A. Modelos de equações estruturais em psicologia: conceitos e aplicações. **Psicologia: teoria e pesquisa**, SciELO Brasil, v. 23, n. 2, p. 205–216, 2007.

POLIT, D. F. Assessing measurement in health: beyond reliability and validity. **International journal of nursing studies**, Elsevier, v. 52, n. 11, p. 1746–1753, 2015.

RIBEIRO, S. V. O. et al. Violência e sintomas de depressão na gestação e materna na coorte brisa: uma abordagem com modelagem de equações estruturais. **Revista Brasileira de Saúde Materno Infantil**, SciELO Brasil, v. 19, p. 173–184, 2019.

ROBERTS, P.; PRIEST, H. Reliability and validity in research. **Nurs Stand**, v. 20, n. 44, p. 41–45, 2006.

ROUSSEEUW, P. J.; DRIESSEN, K. V. Computing l<sub>1</sub> regression for large data sets. **Data mining and knowledge discovery**, Springer, v. 12, n. 1, p. 29–45, 2006.

ROUSSEEUW, P. J.; LEROY, A. M. **Robust regression and outlier detection**. New Jersey: John Wiley & Sons, 1987. v. 589.

RUOTOISTENMÄKI, A.; SEPPÄLÄ, T. Road condition rating based on factor analysis of road condition measurements. **Transport policy**, Elsevier, v. 14, n. 5, p. 410–420, 2007.

SILVA, L. P. d. et al. Comprometimento no trabalho e sua relação com a cultura organizacional mediada pela satisfação. **Revista Brasileira de Gestão de Negócios**, SciELO Brasil, v. 20, p. 401–420, 2018.

SONG, X.-Y.; LEE, S.-Y. A tutorial on the Bayesian approach for analyzing structural equation models. **Journal of Mathematical Psychology**, Elsevier, v. 56, n. 3, p. 135–148, 2012.

SPOSITO, V. Some properties of l<sub>1</sub>-estimation. **Robust Regression: Analysis and Applications**, Marcel Dekker, Inc., 1990.

TABACHNICK, B.; FIDELL, L. **Using Multivariate Statistics**. Needham Heights. United States: Pearson Education Inc./Allyn and Bacon, 2007.

TAVARES, D. M. d. S. et al. Uso da modelagem de equações estruturais na compreensão da incapacidade funcional em idosos. **Revista Latino-Americana de Enfermagem**, SciELO Brasil, v. 29, 2021.

THEIL, H. A rank-invariant method of linear and polynomial regression analysis. **Indagationes Mathematicae**, v. 12, n. 85, p. 173, 1950.

VALENTINI, F.; BRUNO, D. F. Variância média extraída e confiabilidade composta: indicadores de precisão. **Psicologia: Teoria e Pesquisa**, v. 32, n. 2, 2017.

WANG, B. et al. Holiday travel behavior analysis and empirical study under integrated multimodal travel information service. **Transport Policy**, Elsevier, v. 39, p. 21–36, 2015.

WILEY, D. E. The identification problem for structural equation models with unmeasured variables. **Structural equation models in the social sciences**, Academic Press, p. 69–83, 1973.

XIAO, X. et al. Structural equation modeling compared with ordinary least squares in simulations and life insurers data. The University of Texas at Austin, 2013.

## **SECOND PART**

**ARTICLE 1****Construction of the average variance extracted index for construct validation in structural equation models with adaptive regressions**

Published on the paper Communications in Statistics - Simulation and Computation -

Doi:10.1080/03610918.2021.1888122



## Construction of the average variance extracted index for construct validation in structural equation models with adaptive regressions

Patricia Mendes dos Santos & Marcelo Ângelo Cirillo

To cite this article: Patricia Mendes dos Santos & Marcelo Ângelo Cirillo (2021): Construction of the average variance extracted index for construct validation in structural equation models with adaptive regressions, Communications in Statistics - Simulation and Computation, DOI: [10.1080/03610918.2021.1888122](https://doi.org/10.1080/03610918.2021.1888122)

To link to this article: <https://doi.org/10.1080/03610918.2021.1888122>



Published online: 09 Mar 2021.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



## Construction of the average variance extracted index for construct validation in structural equation models with adaptive regressions

Patricia Mendes dos Santos  and Marcelo Ângelo Cirillo 

Statistics Department, Federal University of Lavras, Lavras, Brazil

### ABSTRACT

A range of indicators, such as the average variance extracted (AVE), is commonly used to validate constructs. In statistics, AVE is a measure of the amount of variance that is captured by a construct in relation to the amount of variance due to measurement error. These conventional indices are formed by factor loadings resulting from estimated least squares or maximum likelihood regressions. Thus, a new proposition that provides new factor loadings may result in a more informative AVE index. Consequently, this study consists of the improvement of the index by using adaptive regressions. A Monte Carlo simulation study was performed considering different numbers of outliers generated by distributions with symmetry deviations and excess kurtosis and sample sizes defined as  $n = 50, 100, \text{ and } 200$ . The conclusion was that, in formative structural models, the adaptive linear regression (ALR) method showed good efficiency for correctly specified models. The results obtained from the ALR method for models with specification errors showed low efficiency, as expected.

### ARTICLE HISTORY

Received 26 August 2020  
Accepted 5 February 2021

### KEYWORDS

Adaptive regression;  
Outliers; Structural  
equation model

## 1. Introduction

Among the many advantages of the structural equation modeling (SEM) technique, the flexibility to include latent variables in a covariance structure assumed by a theoretical model that explains their linear relationships with observed variables stands out.

The inclusion of these variables is important in the composition of a construct, resulting from a directly unobservable theoretical concept. However, based on an empirical knowledge represented by the specification of two factorial models that relate the endogenous and exogenous observed variables to one or more latent variables. These models are defined as measurement models, in which a structural model defines the relationship between latent variables.

Assuming that a construct is mathematically formed by linear combinations of observed variables, it is important to analyze the number of variables to be used in its formation. This provides enough information for a concept to be characterized in its interpretation.

Several indices and coefficients have been proposed to validate the constructs; for example, the average variance extracted (AVE), which is the average amount of variation that a latent construct is able to explain in the observed variables to which it is theoretically related.

Of late, the use of this indicator has been increasing in scientific literature. Farrell (2010) discussed the procedures for establishing discriminant validity and the representation for this type of validity. In addition, the author presented an example of an inaccurate assessment of

**CONTACT** Patricia Mendes dos Santos  [patymendesdossantos@hotmail.com](mailto:patymendesdossantos@hotmail.com)  Statistics Department, Federal University of Lavras, 37200-000 Lavras, Brazil.

© 2021 Taylor & Francis Group, LLC



discriminant validity based on the study of Bove et al. (2009). Niclasen et al. (2013) examined the factor structure of the Strengths and Difficulties Questionnaire using an analytical approach of the structural confirmation factor. Henseler, Ringle, and Sarstedt (2015) proposed the heterotrait-monotrait ratio of correlations based on the multitrait-multimethod matrix. Their method was compared with the Fornell and Larcker (1981) criterion, based on AVE and its squared correlations, and with the evaluation of (partial) cross loadings to assess discriminant validity. This comparison was performed using Monte Carlo simulations. Valentini and Bruno (2016) problematized the concepts of AVE and compound reliability, refuting Fornell and Larcker (1981) proposal regarding the use of AVE as an indicator of convergent validity.

Arnab and Michael (2008) propose an adaptive estimator given by a linear combination of the estimates obtained by both the method of least squares (OLS) and robust (LAD). In the case of kernel-based estimators, Zhao et al. (2016) use an adaptive estimation procedure with the purpose of estimating Varying-coefficient models functions, which are extensions of the classic linear models. However, each coefficient  $\beta_j(\cdot)$  ( $j = 1, \dots, p$ ) is estimated by functions and/or smoothers, which, in certain situations, may show leaps in discontinuity. Thus, an alternative to circumvent this problem, proposed by the authors, is the use of an adaptive procedure called adaptive jump-preserving (AJP) that considers nonparametric estimation based on local linear smoothing and jump-preserving regression techniques.

Regarding the asymptotic properties of the AJP estimator, Zhao and Lin (2019) conduct a study of asymptotic properties both for Varying-coefficient models and under some mild conditions. The proposed estimators are established not only in the continuous regions of coefficient functions, but also in the neighborhoods of the jump-preserving points. Simulations and empirical examples have been presented to validate the use of these estimators (Zhao et al. 2017).

Following the motivation provided by the mentioned applications, and aiming at an innovation in the use of new statistical methodologies, this study aimed to improve the index corresponding to the AVE constructed by adaptive regressions, considering that the conventional indices are formed by factor loadings resulting from estimated least square or maximum likelihood regressions. In this respect, a new proposition that provides new factor loadings could result in a more informative AVE index. We ought to use adaptive regressions to combine ordinary least squares estimates with estimates obtained by robust methods in situations that involve the presence of outliers or heavy-tailed error distributions. The aim is not to lose much efficiency when the conditions of the ordinary least squares estimator are met.

## 2. Material and methods

The methodology used in this study considered the theoretical structural equation model illustrated in Figure 1.

The topics of this section are structured as follows: 2.1. Definition of the structural equation model and the parameters used in the simulation process; 2.2. Generation of samples contaminated by outliers; 2.3. Regression estimators incorporated in the structural equation model; 2.4. Adaptive regression estimators and adaptive index for the average variance extracted; 2.5. Average variance index extracted considering model adjustments with specification errors.

### 2.1. Definition of the structural equation model and the parameters used in the simulation process

Following the definition of the structural equation model by Cassel, Hackl, and Westlund (1999) in the Monte Carlo simulation process, the structural model (Figure 1) is represented by the following equations:

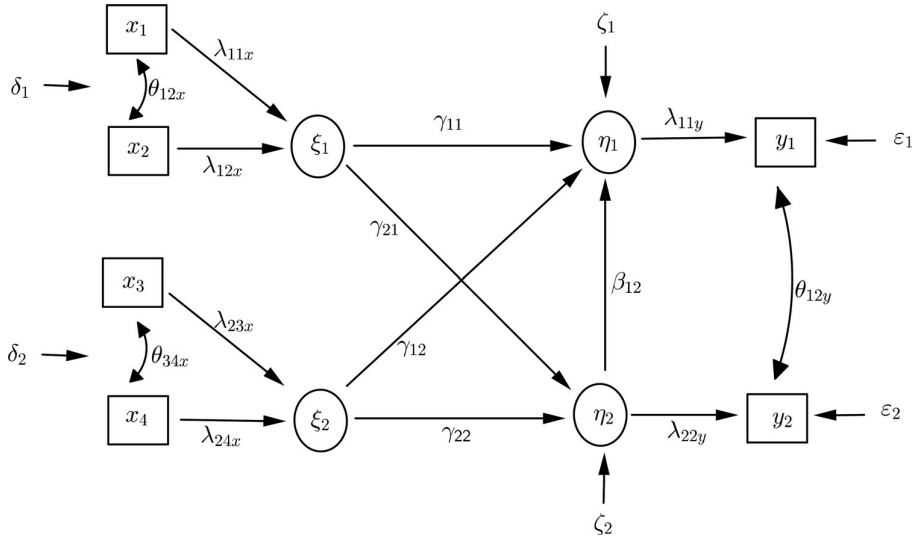


Figure 1. Theoretical structural equation model used in the Monte simulation.

$$\eta_1 = \beta_{12}\eta_2 + \gamma_{11}\xi_1 + \gamma_{12}\xi_2 + \zeta_1, \tag{1a}$$

$$\eta_2 = \gamma_{21}\xi_1 + \gamma_{22}\xi_2 + \zeta_2, \tag{1b}$$

in which  $\beta_{12}, \gamma_{ii}$  for  $i=1, 2$  correspond to the regression coefficients and  $\zeta_1$  and  $\zeta_2$  corresponds to the structural errors.

The observed variables are defined by  $x_q$  ( $q=1, \dots, 4$ ), so that each equation of the  $x$  measurement model is composed by (2), considering  $\delta_j$  ( $j=1, 2$ ) as the measurement error.

$$\xi_1 = \lambda_{11x}x_1 + \lambda_{12x}x_2 + \delta_1, \tag{2a}$$

$$\xi_2 = \lambda_{23x}x_3 + \lambda_{24x}x_4 + \delta_2, \tag{2b}$$

Regarding the equations of the measurement model for  $\eta$ , the equations are given by

$$y_1 = \lambda_{11y}\eta_1 + \epsilon_1, \tag{3a}$$

$$y_2 = \lambda_{22y}\eta_2 + \epsilon_2. \tag{3b}$$

The assumptions of the structural model were maintained, in which the likelihoods of error vectors and latent variables are zero,  $\zeta_k$  and  $\xi_n$  ( $k, n=1, 2$ ) is not correlated, and  $\epsilon_i$  ( $i=1, 2$ ) is not correlated with  $\eta_m$  ( $m=1, 2$ ), and  $\xi_n$  and  $\delta_j$  are not correlated with  $\xi_n, \eta_m$ , and  $\epsilon_i$ .

Following the procedure proposed by McDonald and Hartmann (1992), the structural equation model (Figure 1) was simulated with respect to the relation  $v = Av + u$  in the matrix form below.

$$\begin{bmatrix} y_1 \\ y_2 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ \eta_1 \\ \eta_2 \\ \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.3 & 0.7 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.6 & 0.8 \\ 0 & 0 & 0.2 & -0.2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.6 & 0.6 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ \eta_1 \\ \eta_2 \\ \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} 0.75 \\ 0.84 \\ 0 \\ 0 \\ 0 \\ 0 \\ \zeta_1 \\ \zeta_2 \\ 0.96 \\ 0.64 \end{bmatrix}$$

where  $v$  corresponds to the vector formed by the latent and observed variables. The regression coefficient matrix with the parametric values assumed in the Monte Carlo simulation is defined in  $A$ . The average error vectors formed by the structural errors are provided by (1), the measurement errors are provided by the submodels in (2), and (3) formed the vector  $u$ .

Considering the number of variables defined in  $v$  and  $k = p + q$ , the number of observed variables, for which the covariance matrix imposed by the structural model (Figure 1) was completely specified, the following matrices were defined:

$$P_{axa} = \begin{bmatrix} 1 & 0.4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -0.2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$J_{kxa} = \begin{bmatrix} I_k \\ 0 \end{bmatrix}$$

where  $P$  represents the covariance matrix of vector  $u$ , assuming the parametric values used in the Monte Carlo simulation.  $J$  is a matrix in which the first  $k$  columns form the  $I_k$  identity matrix.

With these specifications, the covariance matrix parameter imposed by the structural model was formed using (4).

$$\Sigma_\theta = J(I_a - A)^{-1}P[(I_a - A)^{-1}]^t J^t \quad (4)$$

After obtaining the covariance matrix (Eq. (4)), multivariate samples were generated in relation to the structural model (Figure 1), such that the data matrix was generated according to  $G \sim N_k(0, \Sigma_\theta)$ . Thus,  $\hat{\Sigma}_\theta$  was obtained for each Monte Carlo simulation.

## 2.2. Generation of samples contaminated by outliers

Considering the definitions of the parametric values relevant to the structural and sample covariance matrices specified in the previous section, and arbitrarily assuming  $\mu_\sim = 0$ , the simulation of samples contaminated with distribution outliers that present symmetry and kurtosis, respectively, were performed using the following distributions, according to Johnson (1987) and Cirillo and Barroso (2017): Uniform, t-Student, and multivariate log-normal.

The multidimensional random variables derived from these distributions were represented by  $G_{1i}$ ,  $G_{2i}$ , and  $G_{3i}$  ( $i = 1, \dots, n$ ), given an arbitrary mix of probabilities set to:  $\alpha = 0.05$  and  $0.25$ , and sample sizes set to  $n = 50, 100$ , and  $200$ .

$$G_{1i} = \alpha N_k(0, \Sigma_\theta) + (1 - \alpha)U_k(i = 1, \dots, k = p + q) \quad (5)$$

such that each component  $U_i$  is defined as

$$U_i = \frac{x_i}{(x_1 + \dots + x_k)^{\frac{1}{k}}}(i = 1, \dots, k = p + q), \text{ where } x_i \sim N(0, 1) \quad (6)$$

$$G_{2i} = \alpha N_k(0, \Sigma_\theta) + (1 - \alpha)t_k(\Sigma_\theta, \nu = 5)(i = 1, \dots, k)$$

such that  $\nu$  represents the number of degrees of freedom in the t-Student distribution and

$$G_{3i} = \alpha N_k(0, \Sigma_\theta) + (1 - \alpha)W_k \quad (7)$$

where  $W_k = [\exp(Z_1), \exp(Z_2), \dots, \exp(Z_k)]$ , when  $Z_i \sim N_k(0, \Sigma_\theta)$ .

### 2.3. Regression estimators incorporated in the structural equation model

The least trimmed squares (LTS) estimators applied in SEM were incorporated considering the structural model (1) and the measurement models (2) and (3). For each sample size set at  $n = 50, 100, \text{ and } 200$ , the estimates were obtained by minimizing Eqs. (8a)–(8c) (Leroy and Rousseeuw 1987). Thus, except for the structural model, each equation of the measurement models was specified by the index  $s = 1, \dots, Q$ , where  $Q$  is the total number of equations, considering the estimates of the parameters of each equation obtained by the least squares method.

$$\min \left\{ \sum_{r=1}^h \zeta_r^2, \dots, \zeta_h^2 \right\}, \quad (8a)$$

$$\min \left\{ \sum_{r=1}^h e_{sr}^2, \dots, e_{sh}^2 \right\}, s = 1, \dots, Q, \quad \text{para } Q = 2, \quad (8b)$$

$$\min \left\{ \sum_{r=1}^h (\delta_{sr}^2, \dots, \delta_{sh}^2) \right\}, s = 1, \dots, Q, \quad \text{para } Q = 2. \quad (8c)$$

### 2.4. Adaptive regression estimators and adaptive index for the average variance extracted

Considering the estimates of latent variables obtained by ordinary squares  $\hat{\beta}_{OLS}$  and least trimmed squares  $\hat{\beta}_{LTS}$ , the estimates of the structural equation parameters using adaptive regressions are given by:

$$\hat{\beta}_{ALR}(\eta_1) = \bar{k}_1 \hat{\beta}_{OLS}(\eta_1) + (1 - \bar{k}_2) \hat{\beta}_{LTS}(\eta_1), \quad (9a)$$

$$\hat{\beta}_{ALR}(\eta_2) = \bar{k}_1 \hat{\beta}_{OLS}(\eta_2) + (1 - \bar{k}_2) \hat{\beta}_{LTS}(\eta_2), \quad (9b)$$

$$\hat{\beta}_{ALR}(\xi_1) = \bar{k}_1 \hat{\beta}_{OLS}(\xi_1) + (1 - \bar{k}_2) \hat{\beta}_{LTS}(\xi_1), \quad (9c)$$

$$\hat{\beta}_{ALR}(\xi_2) = \bar{k}_1 \hat{\beta}_{OLS}(\xi_2) + (1 - \bar{k}_2) \hat{\beta}_{LTS}(\xi_2). \quad (9d)$$

In which  $(\bar{k}_1 \text{ and } \bar{k}_2)$ , refer to the averages of the kurtosis coefficients, obtained through the empirical distributions of the factor loads generated in 1000 Monte Carlo realizations, were referred to the variables of each construct (9a)–(9d), which are estimated respectively by the OLS and LTS methods.

The attribution of these coefficients as weights to be used in the adaptive regressions is justified by the maintenance of the assumption of multivariate normality and by the attribution of the parametric values given in an interval  $[-1.1]$ . As a consequence, the average shortness estimate  $\bar{k}_1 \approx 0$  will be close to zero, resulting in a mesokurtic distribution.

Possible misrepresentations will be plausible to occur, in the kurtosis estimates, through the degree of contamination of outliers, specified in  $\alpha = 0.05$  and  $0.25$ . However, it should be noted that given the breaking point defined by  $h = \frac{(n+p+1)}{2}$ , where  $n$  is the sample size and  $p$  is the number of variables in each regression. As a result, the percentages of mixture of outliers ( $\alpha$ ), being lower than the breaking point  $h$ , leads us to state that the estimates of the factorial loads obtained by the LTS method will be close to the estimates achieved in the OLS method, implying that,  $\bar{k}_1$  is closest to  $\bar{k}_2$ .

The asymptotic behavior of adaptive regressions for all adaptive constructs, whose loads are represented  $\hat{\beta}_{ALR}(\cdot)$  (9a)–(9d), can be understood by considering  $k_{ALR}(\cdot)$  kurtosis defined both by the ratio of the fourth and second order squared moments Joanes and Gill (1998) adapted to adaptive estimates, as expressed (10).

$$k_{ALR}(\cdot) = \frac{E\left(\hat{\beta}_{ALR}(\cdot) - E\left(\hat{\beta}_{ALR}(\cdot)\right)\right)^4}{\left(E\left(\hat{\beta}_{ALR}(\cdot) - E\left(\hat{\beta}_{ALR}(\cdot)\right)\right)^2\right)^2} \quad (10)$$

Therefore, we have the classification defined by (11)

$$k_{ALR}(\cdot) = \begin{cases} 1; & \text{se } k_{ALR}(\cdot) \leq 3 \\ \frac{3}{k_{ALR}(\cdot)}; & \text{se } k_{ALR}(\cdot) > 3 \end{cases} \quad (11)$$

Rewriting adaptive regressions using the  $k_{ALR}(\cdot)$  coefficient.

$$\hat{\beta}_{ALR}(\cdot) = k_{ALR}(\cdot)\hat{\beta}_{OLS}(\cdot) + (1 - k_{ALR}(\cdot))\hat{\beta}_{LTS}(\cdot). \quad (12)$$

As a conclusion, we pose that  $k_{ALR}(\cdot) \leq 3$  then  $\hat{\beta}_{ALR}(\cdot) = 1\hat{\beta}_{OLS}(\cdot)$ . Otherwise,  $\hat{\beta}_{ALR}(\cdot) = \hat{\beta}_{LTS}(\cdot)$ . In this way, we can state the following results:

$$k_{ALR}(\cdot) > (1 - k_{ALR}(\cdot)) \text{ então } \hat{\beta}_{ALR}(\cdot) \Rightarrow \hat{\beta}_{OLS}(\cdot);$$

$$k_{ALR}(\cdot) < (1 - k_{ALR}(\cdot)) \text{ então } \hat{\beta}_{ALR}(\cdot) \Rightarrow \hat{\beta}_{LTS}(\cdot).$$

Following the hypothesis that  $\lambda^2 + \epsilon = 1$  for exogenous constructs and  $\lambda^2 + \delta = 1$  for endogenous constructs, adaptive regression indices to explain the average variance extracted were obtained by:

$$AVE_{ALR(\eta_1)} = \frac{\sum \hat{\beta}_{ALR}^2}{p}, \quad (13a)$$

$$AVE_{ALR(\eta_2)} = \frac{\sum \hat{\beta}_{ALR}^2}{p}, \quad (13b)$$

$$AVE_{ALR(\xi_1)} = \frac{\sum \hat{\beta}_{ALR}^2}{q}, \quad (13c)$$

$$AVE_{ALR(\xi_2)} = \frac{\sum \hat{\beta}_{ALR}^2}{q}, \quad (13d)$$

where p and q correspond to the number of observed variables used in the composition of each construct. For comparative purposes,  $AVE_{ALR}$  was calculated, replacing  $\hat{\beta}_{ALR}$  to  $\hat{\beta}_{OLS}$  and  $\hat{\beta}_{LTS}$ .

### 2.5. Average variance extracted index performance on fit of models with specification errors between constructs

Based on the simulated theoretical model (Figure 1), with the purpose of studying the performance of the  $AVE_{ALR}$  (10a)–(10d) index in situations, the adjustment was made considering some errors caused by the absence of any relationship between constructs, which are represented by the dashed lines in Figures 2 and 3.

All analyses and graphing were performed using the R (R Core Team 2019) software.

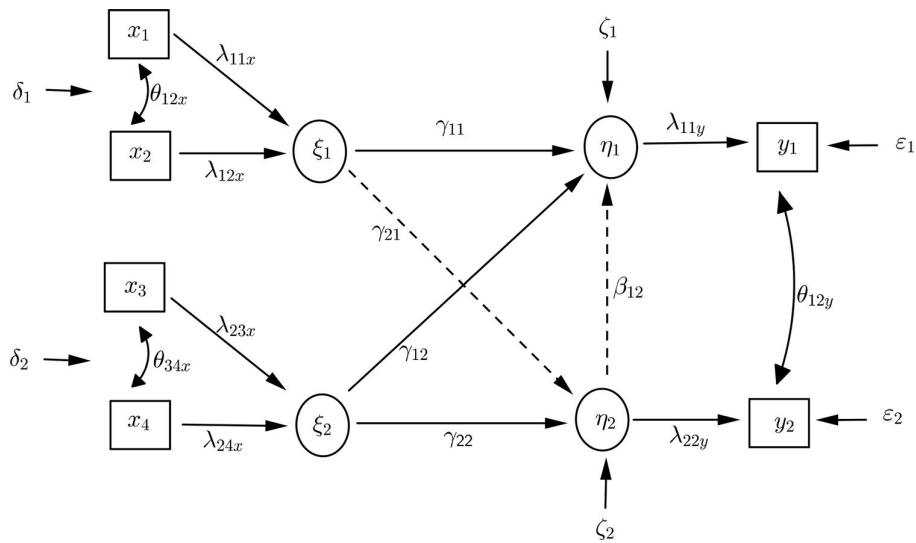


Figure 2. Structurally poorly specified model, omitting the path of  $\zeta_1$  for the variable  $\eta_2$  and  $\eta_2$  for  $\eta_1$ .

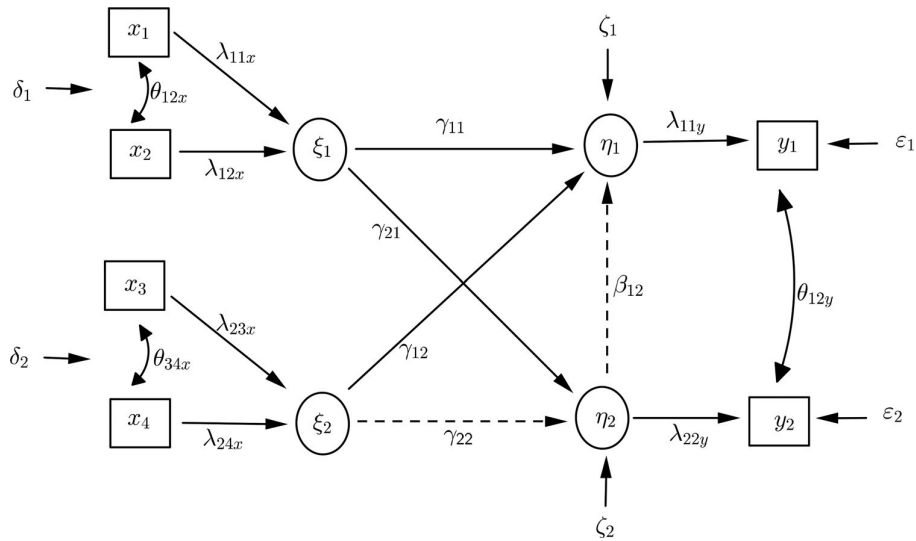


Figure 3. Structurally poorly specified model, omitting the path of  $\zeta_2$  for the variable  $\eta_2$  and  $\eta_2$  for  $\eta_1$ .

### 3. Results and discussion

Figure 4 shows the estimates of the AVE index, in which the factor loadings were obtained by OLS, least trimmed squares (LTS), and combined by adaptive linear regression (ALR).

Considering the results illustrated in Figure 4, the AVE values obtained from the ALR method for the endogenous ( $\eta_1$  and  $\eta_2$ ) and exogenous ( $\xi_1$  and  $\xi_2$ ) constructs were overestimated in relation to the unit value.

Similarly, the same behavior was observed when considering the calculation of AVE for the endogenous constructs ( $\eta_1$  and  $\eta_2$ ), with factor loads obtained by the LTS method. However, for

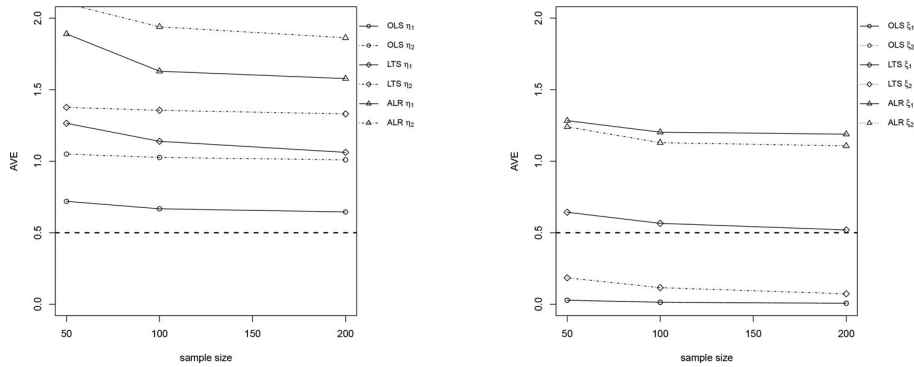


Figure 4. AVE values as a function of different sample sizes  $n$ , considering the normal distribution, using the OLS, LTS, and ALR methods.

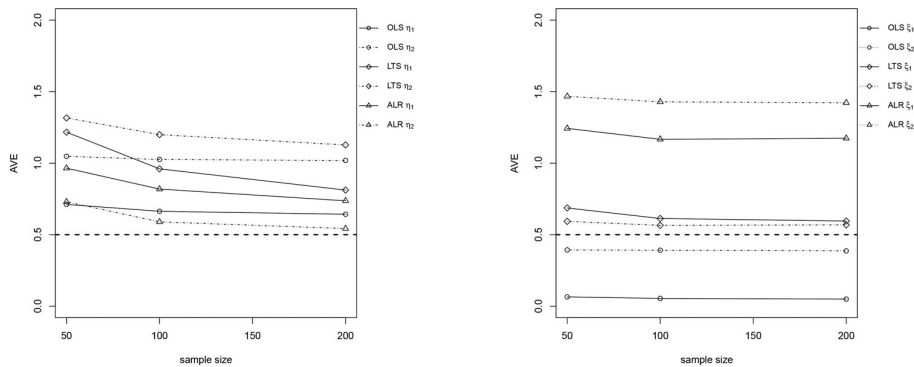


Figure 5. AVE values as a function of different sample sizes  $n$ , considering a mix rate set at  $\alpha = 0.05$  of log-normal distribution, using the OLS, LTS, and ALR methods.

the exogenous constructs ( $\xi_1$  and  $\xi_2$ ), the values remained close to 0.50, which is consistent with the minimum value specified by Fornell and Larcker (1981) to validate the construct.

When using the factor loadings resulting from the OLS method for the endogenous constructs ( $\eta_1$  and  $\eta_2$ ), the AVE values apparently did not show a reduction trend, remaining constant. In addition, their values were more efficient ( $0.50 < AVE \leq 1.00$ ) in relation to the other methods as the sample size increased.

As for the exogenous constructs ( $\xi_1$  and  $\xi_2$ ), these results were null as the sample size increased, but informative in the context that variables that form the constructs were insufficient to allow any practical interpretation. These results were noticeable by the approximation of the result to the value of the reference unit, represented by the dashed line.

In the situation where the data presented symmetry deviation with excess kurtosis characterized by the presence of outliers generated by the multivariate lognormal distribution, maintaining a low concentration of outliers ( $\alpha = 5\%$ ), it was observed that for endogenous constructs ( $\eta_1$  and  $\eta_2$ ), AVE values obtained from the ALR method (Figure 5) ranged from 0.50 to 1.00 as the sample size increased. Regarding the effect of the sample size, for  $n > 100$ , AVE apparently did not show a downward trend, remaining constant as the sample size increased. Thus, it can be stated that the ALR estimator was more efficient, since the other methods overestimated the AVE values for some of these constructs.

However, for the exogenous constructs ( $\xi_1$  and  $\xi_2$ ), the AVE values obtained from the ALR method were overestimated. In addition, it was observed that the AVE values obtained from the

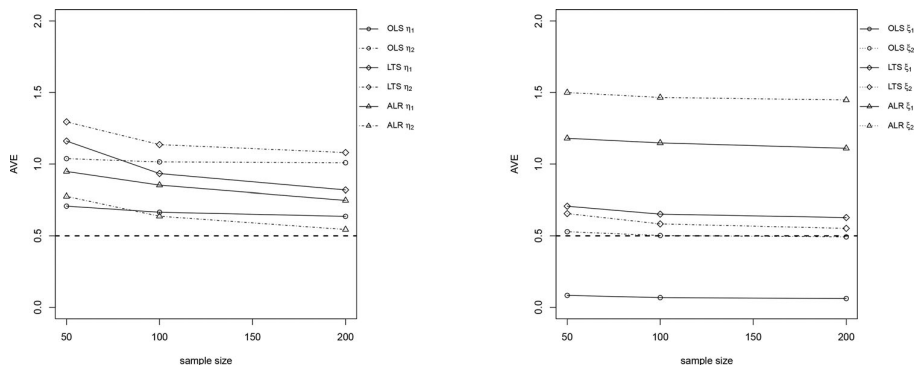


Figure 6. AVE values as a function of different sample sizes  $n$ , considering a mix rate fixed at  $\alpha = 0.25$  of log-normal distribution, using the OLS, LTS, and ALR methods.

Table 1. AVE values obtained by the ALR method in sample sizes  $n = 50$  and  $n = 100$ , contaminated with a minimum percentage of outliers ( $\alpha = 5\%$ ).

Distribution	$\alpha = 0.05$	Constructs	$n = 50$	$n = 100$
			$AVE_{ALR}$	$AVE_{ALR}$
t-Student		$\eta_1$	*0.9724	*0.8320
		$\eta_2$	*0.7349	*0.6168
		$\zeta_1$	1.2261	1.2034
		$\zeta_2$	1.4238	1.3977
		Uniform	$\eta_1$	*0.9809
	$\eta_2$	*0.7671	*0.5778	
	$\zeta_1$	1.2209	1.2119	
	$\zeta_2$	1.4096	1.4322	

\*The AVE values correspond to estimates of the indices that justify the selection of variables that will compose the constructs, it is not about the estimates of the model parameters, therefore, there is no significance probability ( $p$ -values).

Table 2. AVE values obtained by the ALR method in sample sizes  $n = 50$  and  $n = 100$ , contaminated with a maximum percentage of outliers ( $\alpha = 25\%$ ).

Distribution	$\alpha = 0.25$	Constructs	$n = 50$	$n = 100$
			$AVE_{ALR}$	$AVE_{ALR}$
t-Student		$\eta_1$	*0.9750	*0.8237
		$\eta_2$	*0.7624	*0.6217
		$\zeta_1$	1.2010	1.2372
		$\zeta_2$	1.3155	1.3424
		Uniform	$\eta_1$	*0.9689
	$\eta_2$	*0.7365	*0.5573	
	$\zeta_1$	1.2214	1.1269	
	$\zeta_2$	1.3887	1.3898	

\*The AVE values correspond to estimates of the indices that justify the selection of variables that will compose the constructs, it is not about the estimates of the model parameters, therefore, there is no significance probability ( $p$ -values).

LTS method remained higher than 0.50, and that the same values obtained from the OLS method were lower than the standard value. Thus, it can be stated that there is evidence that the combination of OLS and LTS may not be a good alternative, suggesting the need to use robust estimates.

In the same situation, maintaining a high concentration of outliers ( $\alpha = 25\%$ ), it was observed that for larger samples ( $n \geq 100$ ), the AVE estimates obtained from the OLS, LTS, and ALR methods presented results similar to those obtained in the situation with a low concentration of outliers (Figure 6).

For situations where outliers were generated by symmetric multivariate distributions, in the scenarios with the lowest ( $\alpha = 0.05$ ) and the highest level of contamination ( $\alpha = 0.25$ ), the AVE estimates are described for samples  $n = 50$  and  $n = 100$  in Tables 1 and 2, with emphasis on the performance of  $AVE_{ALR}$ .



**Table 3.** AVE values obtained by the ALR method in sample sizes  $n = 50$  and  $n = 100$ , contaminated with a minimum percentage of outliers ( $\alpha = 5\%$ ), considering the model of Figure 2.

Distribution	$\alpha = 0.05$	Constructs	n = 50	n = 100
			$AVE_{ALR}$	$AVE_{ALR}$
Log-normal		$\eta_1$	*0.5207	*0.4229
		$\eta_2$	*0.3623	*0.3038
		$\xi_1$	1.2335	1.1842
		$\xi_2$	1.4741	1.4713
t-Student		$\eta_1$	*0.5312	*0.4331
		$\eta_2$	*0.3598	*0.3062
		$\xi_1$	1.2213	1.1575
		$\xi_2$	1.4221	1.4183
Uniform		$\eta_1$	*0.5480	*0.4241
		$\eta_2$	*0.3607	*0.3179
		$\xi_1$	1.2164	1.1671
		$\xi_2$	1.4730	1.4559

\*The AVE values correspond to estimates of the indices that justify the selection of variables that will compose the constructs, it is not about the estimates of the model parameters, therefore, there is no significance probability ( $p$ -values).

**Table 4.** AVE values obtained by the ALR method in sample sizes  $n = 50$  and  $n = 100$ , contaminated with a minimum percentage of outliers ( $\alpha = 5\%$ ), considering the model in Figure 3.

Distribution	$\alpha = 0.05$	Constructs	n = 50	n = 100
			$AVE_{ALR}$	$AVE_{ALR}$
Log-normal		$\eta_1$	*0.5213	*0.4325
		$\eta_2$	*0.4305	*0.3552
		$\xi_1$	1.2263	1.1928
		$\xi_2$	1.4341	1.4357
t-Student		$\eta_1$	*0.5007	*0.4335
		$\eta_2$	*0.3841	*0.3464
		$\xi_1$	1.1970	1.1733
		$\xi_2$	1.4116	1.4105
Uniform		$\eta_1$	*0.5588	*0.4444
		$\eta_2$	*0.4293	*0.3343
		$\xi_1$	1.2288	1.2038
		$\xi_2$	1.4693	1.4514

\*The AVE values correspond to estimates of the indices that justify the selection of variables that will compose the constructs, it is not about the estimates of the model parameters, therefore, there is no significance probability ( $p$ -values).

Based on the results described in Table 1 and given the effect of these observations from the t-Student and uniform distributions, it was observed that when using the factor loadings resulting from the ALR method, the AVE values for the endogenous constructs ( $\eta_1$  and  $\eta_2$ ) ranged from 0.50 to 1.00 as the sample size increased. These results showed that the ALR method presented good efficiency. In the same situation, the AVE values for the exogenous constructs ( $\xi_1$  and  $\xi_2$ ), obtained by the ALR method, were overestimated.

In order to increase the contamination rate ( $\alpha = 0.25$ ) of outliers and given the effect of these observations from the t-Student and uniform distributions, the results regarding AVE were similar to those obtained in the situation with the lowest contamination level (Table 2).

### 3.1. Average variance index extracted considering model adjustments with specification errors

For situations where outliers were generated by multivariate lognormal distribution and symmetric multivariate distributions, in the scenarios with the lowest ( $\alpha = 0.05$ ) and the highest level of contamination ( $\alpha = 0.25$ ), the estimates of the AVE indices are described for samples  $n = 50$  and  $n = 100$  in Tables 3 and 4 with an emphasis on the performance of  $AVE_{ALR}$ .

**Table 5.** AVE values obtained by the ALR method in sample sizes  $n = 50$  and  $n = 100$ , contaminated with a maximum percentage of outliers ( $\alpha = 25\%$ ), considering the model in Figure 2.

Distribution	$\alpha = 0.25$	Constructs	$n = 50$	$n = 100$
			$AVE_{ALR}$	$AVE_{ALR}$
Log-normal		$\eta_1$	*0.5386	*0.4384
		$\eta_2$	*0.4074	*0.3664
		$\xi_1$	1.1544	1.1230
		$\xi_2$	1.5319	1.5206
t-Student		$\eta_1$	*0.4975	*0.4427
		$\eta_2$	*0.3807	*0.3147
		$\xi_1$	1.1751	1.1976
		$\xi_2$	1.3601	1.3493
Uniform		$\eta_1$	*0.5275	*0.4481
		$\eta_2$	*0.3567	*0.3372
		$\xi_1$	1.2061	1.1388
		$\xi_2$	1.4363	1.3553

\*The AVE values correspond to estimates of the indices that justify the selection of variables that will compose the constructs, it is not about the estimates of the model parameters, therefore, there is no significance probability ( $p$ -values).

**Table 6.** AVE values obtained by the ALR method in sample sizes  $n = 50$  and  $n = 100$ , contaminated with a maximum percentage of outliers ( $\alpha = 25\%$ ), considering the model in Figure 3.

Distribution	$\alpha = 0.25$	Constructs	$n = 50$	$n = 100$
			$AVE_{ALR}$	$AVE_{ALR}$
Log-normal		$\eta_1$	*0.5489	*0.4485
		$\eta_2$	*0.4975	*0.4250
		$\xi_1$	1.1989	1.0960
		$\xi_2$	1.5506	1.4730
t-Student		$\eta_1$	*0.5181	*0.4360
		$\eta_2$	*0.4044	*0.3671
		$\xi_1$	1.1960	1.1740
		$\xi_2$	1.4078	1.3255
Uniform		$\eta_1$	*0.6027	*0.4227
		$\eta_2$	*0.4274	*0.3493
		$\xi_1$	1.2289	1.1292
		$\xi_2$	1.4228	1.3793

\*The AVE values correspond to estimates of the indices that justify the selection of variables that will compose the constructs, it is not about the estimates of the model parameters, therefore, there is no significance probability ( $p$ -values).

Keeping a low concentration of outliers ( $\alpha = 0.05$ ) and considering the models in Figures 2 and 3, it was observed that, for samples of size  $n = 50$ , the estimates of AVE obtained from the ALR method for the endogenous constructs ( $\eta_1$  and  $\eta_2$ ) were less than or close to 0.50. For larger samples ( $n = 100$ ), the AVE index values were lower than the specified default values. These results are expected when the template is incorrectly specified. In this situation, there is evidence to affirm that the ALR method presented low efficiency. In addition, for the exogenous constructs ( $\xi_1$  and  $\xi_2$ ), the ALR method overestimated the AVE values (Tables 3 and 4).

In situations where a high concentration of outliers was considered ( $\alpha = 0.25$ ), the estimates of AVE obtained from the ALR method presented similar results to those obtained in a situation with a low concentration of outliers (Tables 5 and 6).

#### 4. Conclusions

In formative structural models, the ALR method was efficient for correctly specified models, considering that the discrepant values were generated from symmetrical distributions or a multivariate lognormal distribution. Similarly, the results obtained from the ALR method for models with specification errors showed low efficiency, as expected.

## ORCID

Patricia Mendes dos Santos  <http://orcid.org/0000-0002-6989-7982>

Marcelo Ângelo Cirillo  <http://orcid.org/0000-0003-2026-6802>

## References

- Arnab, M., and S. Michael. 2008. On adaptive linear regression. *Journal of Applied Statistics* 35 (12):1409–22.
- Bove, L. L., S. J. Pervan, S. E. Beatty, and E. Shiu. 2009. Service worker role in encouraging customer organizational citizenship behaviors. *Journal of Business Research* 62 (7):698–705. doi:10.1016/j.jbusres.2008.07.003.
- Cassel, C., P. Hackl, and A. H. Westlund. 1999. Robustness of partial least-squares method for estimating latent variable quality structures. *Journal of Applied Statistics* 26 (4):435–46. doi:10.1080/02664769922322.
- Cirillo, M. A., and L. P. Barroso. 2017. Effect of outliers on the GFI quality adjustment index in structural equation model and proposal of alternative indices. *Communications in Statistics—Simulation and Computation* 46 (3): 1895–905. doi:10.1080/03610918.2015.1018998.
- Farrell, A. M. 2010. Insufficient discriminant validity: A comment on Bove, Pervan, Beatty, and Shiu (2009). *Journal of Business Research* 63 (3):324–27. doi:10.1016/j.jbusres.2009.05.003.
- Fornell, C., and D. F. Larcker. 1981. Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research* 18 (1):39–50. doi:10.2307/3151312.
- Henseler, J., C. M. Ringle, and M. Sarstedt. 2015. A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science* 43 (1):115–35. doi:10.1007/s11747-014-0403-8.
- Joanes, D., and C. Gill. 1998. Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (the Statistician)* 47 (1):183–89. doi:10.1111/1467-9884.00122.
- Johnson, M. E. 1987. *Multivariate statistical simulation* (1st ed.). New York: Wiley.
- Leroy, A. M., and P. J. Rousseeuw. 1987. *Robust regression and outlier detection*. *Wiley series in probability and mathematical statistics*. New York: Wiley.
- McDonald, R. P., and W. M. Hartmann. 1992. A procedure for obtaining initial values of parameters in the RAM model. *Multivariate Behavioral Research* 27 (1):57–76. doi:10.1207/s15327906mbr2701\_5.
- Niclasen, J., A. M. Skovgaard, A.-M. N. Andersen, M. J. Sømshovd, and C. Obel. 2013. A confirmatory approach to examining the factor structure of the strengths and difficulties questionnaire (SDQ): A large scale cohort study. *Journal of Abnormal Child Psychology* 41 (3):355–65. doi:10.1007/s10802-012-9683-y.
- R Core Team. 2019. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.
- Valentini, F., and D. F. Bruno. 2016. Variância média extraída e confiabilidade composta: Indicadores de precisão. *Psicologia: Teoria e Pesquisa* 32 (2):1–7. doi:10.1590/0102-3772e322225.
- Zhao, Y.-Y., and J.-G. Lin. 2019. Estimation and test of jump discontinuities in varying coefficient models with empirical applications. *Computational Statistics & Data Analysis* 139:145–63. doi:10.1016/j.csda.2019.05.003.
- Zhao, Y.-Y., J.-G. Lin, X.-F. Huang, and H.-X. Wang. 2016. Adaptive jump-preserving estimates in varying-coefficient models. *Journal of Multivariate Analysis* 149:65–80. doi:10.1016/j.jmva.2016.03.005.
- Zhao, Y.-Y., J.-G. Lin, H.-X. Wang, and X.-F. Huang. 2017. Jump-detection-based estimation in time-varying coefficient models and empirical applications. *Test* 26 (3):574–99. doi:10.1007/s11749-017-0525-7.

**ARTICLE 2**

**Specialty coffee in Brazil: transition among consumers' constructs using structural equation modeling.**

Published on the paper British Food Journal - Doi:<https://doi.org/10.1108/BFJ-06-2020-0537>

# Specialty coffee in Brazil: transition among consumers' constructs using structural equation modeling

Patricia Mendes dos Santos, Marcelo Ângelo Cirillo and  
Elisa Reis Guimarães  
*UFLA, Lavras, Brazil*

Specialty  
coffee in Brazil

Received 23 June 2020  
Revised 9 December 2020  
Accepted 14 December 2020

## Abstract

**Purpose** – Building on Guimarães *et al.* (2019) study and using the modeling of structural equations, the objective of this paper was to elaborate constructs whose variables would enable the characterization and distinction of individuals among these different groups of consumers and to provide insights into their transition between them.

**Design/methodology/approach** – The constructs were validated by the average variance extracted adaptive (AVEADP) index. The transition between consumer groups is explained and encouraged by advances in their conceptual and perceptual knowledge. Thus, regular consumers should be addressed with messages aimed primarily for the social aspect of consumption; enthusiasts, by reinforcing simple to moderate aspects commonly used as product purchase criteria and experts, attracted by the emphasis on complex criteria related to specialty coffee's conceptual and perceptual knowledge, highlighting their influence on the beverage's sensory profile.

**Findings** – Those results enabled a better understanding of these consumers and can guide the marketing strategies of different actors in this market.

**Originality/value** – Important attempts to understand and characterize Brazilian specialty coffee consumers were conducted by Guimarães *et al.* (2019) and Ramírez-Correa *et al.* (2020). However, further studies are needed to differentiate different specialty coffee consumer groups and enhance the market applicability of those studies results. In addition, despite its importance, there is a paucity of public domain studies about the national consumption of specialty coffees, being the results of this work important for the wide dissemination of such information.

**Keywords** Coffee waves, Consumer behavior, Consumption motivations, Marketing, Purchasing criteria

**Paper type** Research paper

## 1. Introduction

Brazil is the world's major coffee producer and exporter as well as the second largest consumer of the beverage, second only to the United States. Moreover, the country is increasingly recognized for its specialty coffee industry (Guimarães *et al.*, 2019). The demand for specialty coffee in Brazil has gained momentum in the last 15 years (Costa, 2020) and should reach 1,063 million bags in 2021 (Proença, 2017). In addition, “the Brazilian coffee chain has strengthened, producers and consumers have become closer, and the specialty coffee market has developed” (Tales and Behrens, 2020, p. 267); its estimated traded value in retail was expected to reach R\$ 3.9bn by 2020 [1] (Proença, 2017).

The development of the global specialty coffee market, usually referred to as “coffee waves”, represented significant changes in the coffee bean production and processing methods, the product differentiation features and quality evaluation criteria and the goals and philosophies of its consumption (Boaventura *et al.*, 2018; Guimarães *et al.*, 2019). Recently,

The authors would like to thank the Coordination for the Improvement of Higher Education Personnel (CAPES) and National Council for Scientific and Technological Development (CNPq) for financial support. Process: 305.599/2018-6.



---

a new type of consumers, often called “coffee lovers” and characterized as “third-wavers”, has emerged (Quintão, 2015) and has been addressed in different studies (e.g. Manzo, 2010; Quintão *et al.*, 2017a, b; Ufer *et al.*, 2019; Urwin *et al.*, 2019, among others).

However, the specificities of the Brazilian coffee market, especially regarding the proximity among producers and consumers, and a well-established and traditional coffee consumption culture, might result in different consumption behaviors (Guimarães *et al.*, 2019). Thus, important attempts to understand and characterize Brazilian specialty coffee consumers were conducted by Guimarães *et al.* (2019) and Ramírez-Correa *et al.* (2020). However, further studies are necessary to differentiate specialty coffee consumer groups and enhance the applicability of those results in the market. Despite their importance, public domain studies about the national consumption of specialty coffees are scarce, which is why this work’s results are important to for the wide spread of this information.

This paper builds on the work by Guimarães *et al.* (2019), using the database provided by the authors to fill those gaps. Using the modeling of structural equations, our purpose was to elaborate constructs whose variables would allow the characterization and distinction of individuals among these different groups of consumers and to provide insights into their transition between them.

As highlighted by Ufer *et al.* (2019), “consumers have become increasingly aware of the ethical implications of their food, as well as the production processes and people behind their food” (p. 1) and the increasing competitiveness in the marketplace demands that “food producers and retailers understand the major determinants of consumer decision-making processes to ensure they both fully capitalize on consumer demand for their products and meet consumer expectations” (p. 2).

This work aims at assisting professionals in the Brazilian specialty coffee market in their strategies to gather and engage coffee consumers, stimulating their conceptual or perceptual knowledge (LaTour and LaTour, 2010; LaTour *et al.*, 2011) on the product and, consequently, encouraging their transition to higher involvement categories. Should these professionals focus on the variables with the greatest influence on specialty coffee consumers’ behavior, they would also be able to optimize the resources and results of the strategies designed for this purpose.

## 2. Theoretical framework

### 2.1 Coffee waves and the emergence of specialty coffee consumers

Since the mid-nineteenth century, the international coffee market has undergone important changes regarding coffee bean production and processing methods, differentiation features and quality evaluation criteria and the goals and philosophies of consumption. These changes are usually explained by the concept of “coffee waves”, coined in 2003 by Trish Skeie (Andrade *et al.*, 2015; Boaventura *et al.*, 2018; Guimarães, 2016; Guimarães *et al.*, 2016, 2019; Lima *et al.*, 2020; Skeie, 2003). Each wave is subject to several ideological, technological and social factors (Teles and Behrens, 2020). As highlighted by Teles and Behrens (2020, p. 260),

At the end of the 19th century, the postindustrial revolution world experienced an abandonment of handicrafts to the detriment of industrialized products, mainly due to urban growth and social changes arising from the economic system. In fact, the behavior of consumers changed [...]. So, coffee changed from a colonial good to an industrial product.

The “first coffee wave” occurred mainly due to the industrial revolution and the advances in the product processing, packaging and marketing, which facilitated storage, increased its useful life and allowed its distribution over long distances and its commercialization in several formats, such as soluble, thus considerably increasing coffee consumption worldwide (Guimarães *et al.*, 2019; Lima *et al.*, 2020; Manzo, 2010; Teles and Behrens, 2020). Coffee

consumption was predominantly domestic, motivated by utilitarian issues, such as physical disposition and improved intellectual performance provided by caffeine (Andrade *et al.*, 2015; Guimarães *et al.*, 2016; Lima *et al.*, 2020; Skeie, 2003, Teixeira and Nunes, 2016, Teixeira, 2020), in addition to its convenience and affordable price (Teles and Behrens, 2020).

The second wave started around the 1960s [2], when coffee became a high-quality artisanal food (Tales and Behrens, 2020), and new concepts of origin and roasting were introduced to reach different beverage-consuming profiles (Guimarães *et al.*, 2019; Ramírez-Correa *et al.*, 2020; Skeie, 2003). In this way, the specialty coffee market was created. Quality perception was influenced by the origin of the beans, still at a national level, stimulating curiosity about the environmental, social and economic conditions of these places and leading to an important interest in sustainable coffees (Guimarães *et al.*, 2019; Ramírez-Correa *et al.*, 2020).

New modes of coffee consumption, which focused on quality, differentiation and added value of coffee, created a segment of the specialty coffee market that has grown considerably since its inception (Roseberry, 1996). Thus, coffee shop chains (e.g. Starbucks and similar establishments) have become important places of consumption (Boaventura *et al.*, 2018; Borrella *et al.*, 2015; Lima *et al.*, 2020; Teixeira and Nunes, 2016; Teixeira, 2020). However, with their rapid expansion, they had to standardize their product through strategies like the adoption of dark roast profiles, masked by the addition of chocolate, whipped cream, syrup, among others, sacrificing some of their product's quality and differentiation. Therefore, these companies shifted their focus to the social characteristic of consumption (Andrade *et al.*, 2015; Guimarães, 2016; Guimarães *et al.*, 2016).

The third coffee wave came in response to the massification and high standardization of late second wave coffees, seeking greater quality and differentiation by adopting deeply artisanal processes (Borrella *et al.*, 2015; Guimarães, 2016; Tales and Behrens, 2020). The consumption of third wave coffees occurs mainly in high-end independent coffee shops (HEICS) or in the domestic environment. Coffee consumers have demanded more quality, engaged in preparation and consumption practices, sought for more knowledge by participating in communities and courses on coffee and required greater sustainability across the productive chain (Andrade *et al.*, 2015; Guimarães, 2016; Guimarães *et al.*, 2016; Quintão, 2015; Quintão *et al.*, 2017a). In this wave, each coffee specificity is valued and the beverage is savored considering several sensory characteristics (Teixeira, 2020). As described by Tales and Behrens (2020, p. 266),

Part of the third wave also involves the growth of network communities of home baristas participating in the growing availability of artisanal roasters, green coffee retailers, equipment suppliers (often under the same roof as the cafés themselves), and targeted courses such as tasting and preparation of coffees, that provide support and assistance to consumers who seek new experiences.

Because of these market transformations, among other factors, a new type of consumer often referred to as “coffee evangelists, coffee snobs, coffee lovers, coffee aficionados” and others (Quintão, 2015, pp. 25–26) has emerged. Those are highly demanding and market-engaged individuals, usually considered the main responsible for the performance and growth of this market (Guimarães *et al.*, 2016; Proença, 2017; Quintão *et al.*, 2017a).

As demonstrated by Quintão *et al.* (2017a), they “adopt a highly differentiated system to understand, evaluate, and appreciate the beverage, expressing it through their consumption practices” (Guimarães *et al.*, 2019, p. 50). As a part of a broader trend in consumer behavior (Angus, 2020; Hoşafçı, 2018), those individuals are “highly concerned about product quality, origin, and social, environmental, and economic sustainability” (Guimarães *et al.*, 2019, p. 52).

In a large qualitative study conducted in the United States and Canada, Quintão and Brito (2016) and Quintão *et al.* (2017a, b) demonstrated the existence of heterogeneous communities

---

of specialty coffee consumers in those countries. The authors subdivided these consumers into three categories, associated with their level of expertise on conceptual and perceptual knowledge about the product (LaTour and LaTour, 2010; LaTour *et al.*, 2011): regular consumers; amateurs (connoisseurs) and professionals (in that study context, represented by baristas). In sum, regular consumers are those immersed in mass consumption in large coffee shop networks; amateurs are those who started developing subcultures composed of performance standards, values and rules and professionals are baristas with a high subcultural capital, playing a role in defining and creating it, later sharing it with the specialty coffee community (Quintão *et al.*, 2017a).

According to Quintão *et al.* (2017a), the transition of individuals between the categories of regular consumers and amateurs would take place through “taste transformation rituals”. However, the authors do not address the rituals or motivations for the transition of individuals from amateurs (connoisseurs) to professionals.

In turn, Guimarães *et al.* (2019) pointed out that the specificities of the Brazilian specialty coffee market – that is, the country’s position as the world’s largest coffee producer and exporter and, simultaneously, second largest consumer of the beverage, with a still nascent market for specialty coffees – can translate into different consumption behaviors and result in a new dynamic of consumer communities.

Thus, in a recent study with a large sample of respondents, Guimarães *et al.* (2019) proved the existence of different groups/categories of Brazilian specialty coffee consumers. Similarly to Quintão and Brito (2016), Quintão *et al.* (2017a, b) and considering the reflections of LaTour and LaTour (2010) and LaTour *et al.* (2011), Guimarães *et al.* (2019) proposed their classification in (1) regular consumers; (2) enthusiasts and (3) experts. Despite its importance and advances in understanding specialty coffee consumers, the work by Guimarães *et al.* (2019) presents important limitations, which we seek to overcome in this work by using more robust and adequate statistical methods. First, due to the different categories of specialty coffee consumers in Brazil, their study presented a wide number of variables (*i.e.* consumption motivations and product purchasing criteria) to influence these individuals’ behavior, some of which can also be considered similar among the identified categories. Therefore, the application of their findings to the development of marketing strategies by the actors in this market can still be considered complex. In addition, there is a need for further study of what factors would encourage the transition of individuals from one consumer category to another.

### 3. Methods

#### 3.1 Database characterization

This paper follows the work by Guimarães *et al.* (2019), using the database they provided for adoption of the proposed methodology, as presented below. Through a non-probabilistic convenience sampling, based on the “snowball method” (Baltar and Brunet, 2012), Guimarães *et al.* (2019) collected 864 self-administered online questionnaires – whose studied variables are displayed in Table 1 – widely spread among Brazilian specialty coffee consumers between January and February 2017. The demographic characterization of their sample is presented in Table 2.

Over 60% of the respondents consumed specialty coffees daily, and many showed a tendency to reduce or even quit consuming commodity coffees. Their average consumption of and monthly expenditure with specialty coffees showed great variation in the sample, but 74% of respondents stated willingness to increase consumption of this product. In addition, over 71% would increase their expenditure if more product-related information were available. Such information was usually sought on the product package (67.9%), online (63.5%), directly with a barista or other professionals (51.8%), on specialized journals (33.3%)



and with family or friends (24.8%). Only 2.3% of the respondents stated that they did not seek more information about those coffees (Guimarães *et al.*, 2019).

The respondents mostly acquired specialty coffees from coffee shops (59%), through direct purchase from the producer (40.5%), online (32.7%) and in supermarkets (30.7%). Coffee consumption occurred mainly in households (92.3%) and coffee shops (79%), followed by professional environments (42.4%) and homes of relatives and/or friends (25.9%) (Guimarães *et al.*, 2019). This profile is similar to the one described by Tales and Behrens (2020).

Guimarães *et al.* (2019) identified three different groups/categories of specialty coffee consumers (Table 3), each with a different set of motivations for products consumption (Table 4) and its purchasing criteria (Table 5): (1) regular consumers; (2) enthusiasts and (3) experts, as explained in the introductory section.

Among the identified categories, Guimarães *et al.* (2019) highlight regular consumers as the least involved in this consumption practice. Their consumption motivations are essentially based on the taste and aroma and pleasure of its consumption. In addition, they are practically indifferent to supporting sustainable and socially responsible initiatives, which explains their low consumption of sustainable coffees, when compared to other categories of consumers. Most of these individuals are not professionals in the coffee market

Variables

1. Demographic criteria (sex, income, age and schooling)
2. Professional exercise in the coffee market (production, processing/industry, retail, research or teaching)
3. Commodity and specialty coffees consumption frequency
4. Specialty coffee average consumption and expenditure
5. Willingness in increasing average specialty coffee consumption
6. Prior acquisition and performed investment in accessories or differentiated extraction methods
7. Willingness in increasing expenditure in specialty coffees if more product-related information are disclosed
8. Sustainable coffee consumption

Source(s): Guimarães *et al.* (2019, p. 66)

**Table 1.**  
Variables used by Guimarães *et al.* (2019) in their survey

Sex	Male (65.1%)
Age	21–35 years old (59.4%)
Monthly family income	Over five minimum wages (about US\$1.450)
Education level	College or postgraduate degrees (80.1%)
Professional association with coffee	Yes (54%)

Source(s): Adapted from Guimarães *et al.* (2019)

**Table 2.**  
Guimarães *et al.* (2019) sample characterization

Coffee professionals	27.1%	33.0%	61.4%
Commodity coffee daily consumption	45.8%	39.6%	24.0%
Specialty coffee daily consumption	42.4%	46.0%	77.3%
Specialty coffee monthly average consumption	Up to 250 g	Up to 500 g	Over 500 g
Disposition in increasing specialty coffee consumption	66.1%	70.0%	79.0%
Equipment/extraction methods acquisition	Did not buy	Up to R\$250	From R\$250 to R\$500
Expenditure increasing disposition	54.2%	58.3%	85.9%
Sustainable coffee consumption	59.3%	76.0%	89.6%

Source(s): Guimarães *et al.* (2019, p. 66)

**Table 3.**  
Comparative between Guimarães *et al.* (2019) identified clusters' features

	Variable	Regular consumers	Enthusiasts	Experts
Consumption motivations	Pleasure in consumption	Totally agree	Totally agree	Totally agree
	Beverage flavor and aroma	Totally agree	Totally agree	Totally agree
	Beans' history or origin	Partially agree	Partially agree	Totally agree
	Support for sustainable initiatives	Indifferent	Partially agree	Totally agree
	Learning or professionalization desire	Indifferent	Partially agree	Totally agree
	Coffee professionals' influence	Partially disagree	Indifferent	Partially agree
	Energy and disposition	Indifferent	Indifferent	Partially agree
	Improvement			
	Family or friends influence	Indifferent	Indifferent	Indifferent
Family habit or tradition	Indifferent	Indifferent	Indifferent	

**Table 4.** Summarization of the discussed clusters' consumption motivations

Source(s): Guimarães *et al.* (2019, p. 66)

	Variable	Regular consumers	Enthusiasts	Experts
Purchasing criteria	Roasting intensity/color	Little relevance	High relevance	High relevance
	Roasting date	Little relevance	High relevance	High relevance
	Origin	Little relevance	Medium relevance	High relevance
	Certification	Little relevance	Medium relevance	High relevance
	SCA scoring	Little relevance	Medium relevance	High relevance
	Processing method	Little relevance	Medium relevance	High relevance
	Variety	Little relevance	Medium relevance	High relevance
	Altitude	Irrelevant	Medium relevance	High relevance
	Package	Little relevance	Medium relevance	Medium relevance
	Price	Medium relevance	Medium relevance	Medium relevance
	Brand	Medium relevance	Medium relevance	Medium relevance

**Table 5.** Summarization of the discussed clusters' purchasing criteria

Source(s): Guimarães *et al.* (2019, p. 67)

and give little or no importance to recognized specialty coffee purchasing criteria, except for price and brand, to which they attribute medium relevance, and which are generally associated with the consumption of commodity coffee (ABIC, 2010). Among the three consumer groups identified by the authors, regular consumers give the highest importance to the price criterion.

Enthusiasts consume specialty coffees in greater volume and frequency when compared to regular consumers. Similarly to regular consumers, however, most enthusiasts are not yet coffee professionals. They are willing to increase their expenses with and consumption of specialty coffees if more information about its characteristics and processing or preparation methods is made available. Their consumption of specialty coffee is highly motivated by the beverage taste and aroma, the pleasure in its consumption and the knowledge about the history and origin of the beans. Moreover, they support sustainable and socially responsible initiatives. Enthusiasts also attach great importance to the roasting date and intensity/color as specialty coffee purchasing criteria and demonstrate a desire to learn and/or become professionals in the specialty coffee market (Guimarães *et al.*, 2019).

Experts are consumers who have the highest level of involvement with this consumption practice. Although they consume specialty coffees on a daily basis, being the largest consumption among the identified groups, they are still willing to increase their consumption. Their motivation is based on the beverage taste and aroma, the pleasure in its consumption, the satisfaction in knowing the history and origin of the beans and the support for sustainable and socially responsible initiatives, which explains their impressive consumption of sustainable coffees. As for their specialty coffees purchasing criteria, roasting date and intensity/color stand out, followed by processing methods, origin, SCA scoring [3], production altitude, variety and the presence of certification seals (Guimarães *et al.*, 2019). Furthermore, they have been highly motivated by the desire to learn or become professionals in this market, which explains the fact that experts are mostly professionals in the coffee market, whether in specialty coffees or in commodities (Guimarães *et al.*, 2019). The authors also pointed out the need to work on a broader concept of “coffee professionals” in the Brazilian context, which concentrates, in the same nation, all stages from coffee production to consumption.

### 3.2 The structured equation model

In this work, the specialty coffee consumption motivations and purchasing criteria, as proposed by Guimarães *et al.* (2019), were considered observed variables which, hypothetically combined, formed the different constructs/latent variables (that is, consumer groups) that make up the structural equation model. In this way, we seek to advance their studies by proposing a more accurate and easy-to-understand interpretation of which variables differentiate consumers within the different proposed categories. This estimation can help the actors of coffee market, especially small producers, micro roasters and independent coffee shops, translate this knowledge into more effective marketing strategies.

With that purpose, we initially carried out an exploratory factor analysis to find the underlying structure in the data matrix and determine the number and nature of the latent variables (factors) that best represent the set of observed variables. In other words, we sought the best distribution of the observed variables into factors, which occur when they share a common variance (Brown, 2006). This exploratory three-factor analysis – to reflect the different groups of consumers proposed by Guimarães *et al.* (2019) – was performed using the Bartlett method and Promax rotation.

With the formulated model, we proceeded with the estimation of the factor loads and numerical validation of the formalized constructs, according to the combinations of variables. Following the theoretical model description, the representation of the equations whose factorial loads were estimated is presented in Table 6.

To estimate the factor loads of the proposed model, the ordinary least square (OLS) method was used and, in the hypothesis that the variables had outliers, the model was also adjusted by least trimmed squares (LTS), following the specifications given by Cirillo and Barroso (2012).

Using the parameter estimates and to validate the composition of the constructs  $\xi_1$ ,  $\xi_2$  e  $\xi_3$  in relation to the number of variables considered, the index defined by the average variance extracted (AVE) was estimated, resulting from the combination of the factorial loads obtained by the two estimation methods mentioned above, according to the following equations:

$$AVE_{ADP(\xi_i)} = \frac{\sum_{i=1}^p \hat{\lambda}_{ADP}^2(\xi_i)}{p}$$

with  $i = 1, 2$  e  $3$  being,

$$\hat{\lambda}_{ADP}(\xi_i) = k_1 \hat{Z}_{OLS}(\xi_i) + (1 - k_2) \hat{Z}_{LTS}(\xi_i)$$

where  $p$  referred to the number of variables observed in each construct; and  $k_1$  and  $k_2$  corresponded to the kurtosis coefficient, considering the vector of the factor load estimates, represented by  $\hat{Z}_{OLS}$  e  $\hat{Z}_{LTS}$ , respectively, obtained by applying the OLS and LTS methods.

$AVE_{ADP}$  is the average amount of variation that a latent construct is able to explain in the observed variables to which it is theoretically related. It also consists of an adapted combination of the OLS and LTS methods, being a proposition that provides new factor loads and may result in a more informative index (AVE) with the use of adaptive regressions. This technique combines the estimates of ordinary least squares with estimates obtained by robust methods, in situations involving the presence of outliers or error distributions with heavy tails, so that it will not lose much efficiency when the conditions of the OLS estimator are really satisfied (Arnab and Michael, 2008). Finally, we used *R* software to estimate and build the adaptive index.

#### 4. Results and discussion

In the proposed model (Figure 1), groups of consumers, as classified by Guimarães *et al.* (2019) –  $\xi_1$ , regular consumers;  $\xi_2$ , enthusiasts and  $\xi_3$ , experts – named the latent variables represented by the constructs. The description of the observed variables used in the composition of the constructs is presented in Table 7.

The idealized structural equation model is empirical, with linear assumptions regarding the cause-effect relationship between the variables, naturally identified by the arrows. Method adequacy is based on the interpretation of the model according to the data. In this work, the numerical justification for formalizing the constructs, represented by the profiles as a function of the observed variables, is confirmed by the  $AVE_{ADP}$  index, as described in the methodology. Based on the satisfactory result, statistics evidence that there is no need to insert new observed variables to represent the construct.

The observed variables may present values diagnosed as outliers. Therefore, adaptive indexes were formalized to make a robust index to these observations since estimates of least squares regressions and robust regressions (LTS) are combined in the formalization of the statistical criterion that confirms the adequacy of results, using the proposed model and adds confidence to the validation of the construct.

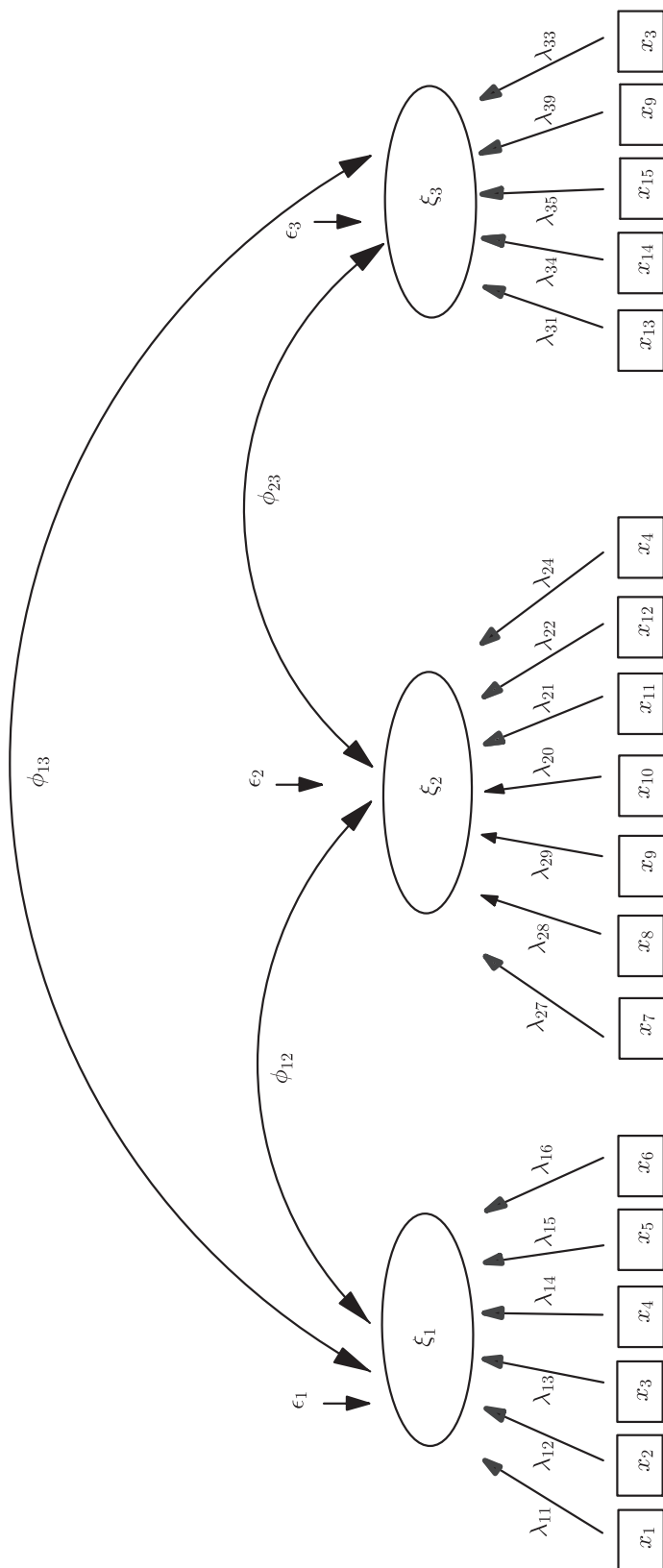
The characterization of the model and its constructs are presented in the following subsections.

##### 4.1 Regular consumers ( $\xi_1$ )

Regular consumers ( $\xi_1$ ) are essentially characterized by their consumption motivations, with emphasis on the influence exercised by family members, friends and coffee professionals, as well as the family habit or tradition of coffee consumption. As highlighted by Sabioa and Spers (2020, p. 308), “social factors can also contribute to experimentation, being influenced by someone who is part of the conviviality of the potential consumer”. Therefore, these social

**Table 6.**  
Equations with the coefficients representing the estimated factor loads in the structural model

$$\begin{aligned} \xi_1 &= \lambda_{11}X_1 + \lambda_{12}X_2 + \lambda_{13}X_3 + \lambda_{14}X_4 + \lambda_{15}X_5 + \lambda_{16}X_6 + \varepsilon_1 \\ \xi_2 &= \lambda_{27}X_7 + \lambda_{28}X_8 + \lambda_{29}X_9 + \lambda_{20}X_{10} + \lambda_{21}X_{11} + \lambda_{22}X_{12} + \lambda_{24}X_{14} + \varepsilon_2 \\ \xi_3 &= \lambda_{31}X_{13} + \lambda_{34}X_{14} + \lambda_{35}X_{15} + \lambda_{39}X_9 + \lambda_{33}X_3 + \varepsilon_3 \\ \text{corr}(\xi_1, \xi_2) &= \emptyset_{12} \\ \text{corr}(\xi_2, \xi_3) &= \emptyset_{23} \\ \text{corr}(\xi_1, \xi_3) &= \emptyset_{13} \end{aligned}$$



**Figure 1.** Theoretical model to define the variables that characterize the specialty coffee's consumer groups and to evaluate their transition between constructs

Latent variables	Observed variables
$\xi_1$ – Regular consumers	$x_1$ = Consumption motivation: energy and disposition improvement $x_2$ = Consumption motivation: family habit or tradition $x_3$ = Consumption motivation: family or friends influence $x_4$ = Consumption motivation: support for sustainable initiatives $x_5$ = Consumption motivation: coffee professionals' influence $x_6$ = Consumption motivation: learning or professionalization desire
$\xi_2$ – Enthusiasts	$x_7$ = Purchasing criteria: price $x_8$ = Purchasing criteria: brand $x_9$ = Purchasing criteria: SCA scoring $x_{10}$ = Purchasing criteria: roasting intensity/color $x_{11}$ = Purchasing criteria: packaging type or design $x_{12}$ = Consumption motivation: beans history or origin $x_4$ = Consumption motivation: support for sustainable initiatives
$\xi_3$ – Experts	$x_{13}$ = Purchasing criteria: origin $x_{14}$ = Purchasing criteria: altitude $x_{15}$ = Purchasing criteria: processing methods $x_9$ = Purchasing criteria: SCA scoring $x_3$ = Consumption motivation: family or friends influence

**Table 7.**  
Description of the variables observed in the composition of the proposed model constructs

factors could explain why consumers are willing to experiment specialty coffees. This can also be explained by the still-restricted conceptual knowledge (LaTour and LaTour, 2010; LaTour *et al.*, 2011) regular consumers have about specialty coffees. Among the identified consumer groups, regular consumers demonstrate the lowest level of involvement with the product (Guimarães *et al.*, 2019) and immersion in that market, making them more susceptible to influences by other individuals.

The energy and disposition improvement, a characteristic typically associated with consumers of the first coffee wave (Andrade *et al.*, 2015; Guimarães *et al.*, 2016; Skeie, 2003), highlights the transition process of regular consumers ( $\xi_1$ ) from the consumption of traditional/commodity coffees to specialty coffees. In other words, this might be considered as a determinant feature of their consumption because these individuals are still transitioning from commodity to specialty coffee consumers. However, the feeling of supporting sustainable and socially responsible initiatives heavily influences their consumption, indicating changes on their perception toward coffee, which starts to be considered as a truly differentiated and special product. Consumers also value such non-organoleptic traits in coffee, especially due to their willingness to pay premium prices for sustainable or ethical labels (Sabioa and Spers, 2020; Tales and Behrens, 2020; Ufer *et al.*, 2019). As highlighted by Tales and Behrens (2020, p. 269),

Concern about reliability, traceability, sustainability, and ethics is also growing in Brazil, with an increasing number of consumers who are aware of the preservation of environmental resources, both in terms of environmental impact, animal abuse, and fair trade.

Regular consumers ( $\xi_1$ ) are also characterized by their desire to learn more about specialty coffees and/or become an expert in that market. Even though this characteristic is noticeable in all consumer groups presented by Guimarães *et al.* (2019), we assumed that it stood out among regular consumers because of its importance guaranteeing that individuals effectively transition from the consumption of traditional/commodity to specialty coffees. This characteristic may also be considered essential because, as they acquire conceptual knowledge about the product and increase their engagement with consumption, these individuals begin to transition to the categories of enthusiasts ( $\xi_2$ ) and experts ( $\xi_3$ ).

---

#### 4.2 Enthusiasts ( $\xi_2$ )

Enthusiasts ( $\xi_2$ ) are essentially characterized by their specialty coffee purchasing criteria, namely: price, brand, packaging type or design, roasting intensity/color and SCA scoring. This may indicate that these consumers have acquired more conceptual knowledge about the product and now feel more confident selecting and purchasing those coffees that please them the most, regardless of the influence of other individuals in their decision.

However, considering brands as proxy for quality and differentiated packaging, with greater availability of information about the product (Guimarães *et al.*, 2019), higher prices are low complexity purchasing criteria and require reduced conceptual and perceptual knowledge about the product. Thus, we consider that enthusiasts would be limiting the influence of other individuals in their product acquisition by starting to look for extrinsic signs of product quality by themselves.

When compared to the other specialty coffee purchasing criteria, the “SCA scoring” and “roasting intensity/and color” demand greater conceptual and/or perceptual knowledge about the product from enthusiasts. SCA scoring is an international standard still used in a restricted way in Brazil. It is mostly known by professionals in the area and competes nationally with other methodologies for grading and evaluating the quality of coffee, such as the methodology of the Brazilian Association of the Coffee Industry (ABIC, 2020) [4].

Roasting intensity/color is an important distinguishing feature between commodity and specialty coffees since third wave members have adopted lighter roasts to highlight the beans’ distinctive features and obtain maximum quality (Skeie, 2003; Guimarães *et al.*, 2016, 2019; among others). For them, darker roasts were commonly adopted by second wavers to conceal product defects and impurities, in addition to allowing its standardization on a large scale, contradicting attempts to differentiate specialty coffees by quality.

As for the enthusiasts’ ( $\xi_2$ ) characteristic consumption motivations, the support for sustainable initiatives and the search for knowledge about the product history and/or origin stand out. This is consistent with the results given by Sabioa and Spers (2020, p. 312), which indicate that “origin influences coffee choice as a source of information about the product, which would reduce risk aversion, and add a differential to flavor, thus, attracting the consumer”. Such differentiation might also be influenced by the consumer’s level of involvement with coffee and higher knowledge about the product (Sabioa and Spers, 2020). These issues are pointed out by different researchers (*e.g.* Guimarães *et al.*, 2016; Skeie, 2003) as outstanding among consumers of the second and, mainly, the third wave of coffee movements that marked the progressive differentiation of the product, both for its quality and for the search for sustainability in the productive chain, resulting in the specialty coffee category.

#### 4.3 Experts ( $\xi_3$ )

Finally, expert consumers ( $\xi_3$ ) stand out for the adoption of advanced and complex criteria, in terms of conceptual and perceptual knowledge, for evaluation and acquisition of specialty coffees. These features, that is, bean production, origin and altitude, processing methods and SCA scoring, influence the beverage sensory profile in distinct ways. In other words, not only do they affect the overall beverage quality but they also give it distinct nuances, which make each coffee special, according to the perspective adopted by consumers of the third wave of coffee (Guimarães *et al.*, 2016, 2019; Quintão *et al.*, 2017a, b).

We can also say that consumption motivation based on “origin”, as depicted by enthusiasts ( $\xi_2$ ), becomes an even more relevant criteria for experts ( $\xi_3$ ), when purchasing specialty coffee, showing their deep and consolidated conceptual and perceptual knowledge about the product.

Similarly to regular consumers ( $\xi_1$ ), experts ( $\xi_3$ ) find the influence of family and friends significant as a specialty coffee consumption motivation, which can be explained by the still

recent, although significant growth of this market in the country. Thus, this consumer community (Quintão *et al.*, 2017b) would grow organically with its members influencing one another and attracting individuals related to their immediate social circles.

#### 4.4 General discussion

The characterization of the different Brazilian specialty coffee consumer groups, presented in this study, resembles the one proposed by Guimarães *et al.* (2019). The main differences refer to the regular consumers ( $\xi_1$ ): unlike those authors, we identified that such consumers are also highly motivated by the desire to support sustainable and socially responsible initiatives associated with the product. Besides, they are highly influenced by their closest social circles and often rely on professionals in the field to indicate the best options. As they expand their conceptual knowledge on specialty coffees, regular consumers tend to migrate to the category of enthusiastic consumers ( $\xi_2$ ) and then, start to adopt more specific purchase criteria, possibly by feeling more confident about their coffee choices.

This work advances in determining which variables would indicate the best group of consumers for each individual. This reduces the ambiguities Guimarães *et al.* (2019) found in their profiles and better guides the marketing strategies of players in this market. Based on this information, we reason that they should focus on different aspects, according to the group of consumers which is considered their priority target audience. Regular consumers ( $\xi_1$ ) could be addressed with messages aimed primarily at the social aspect of consumption, due to the strong influence exercised by their social circle and by professionals in the field. Enthusiasts ( $\xi_2$ ) would be better addressed with messages reinforcing simple-to-moderate aspects they commonly use as product purchase criteria. Experts ( $\xi_3$ ), in turn, would be attracted by complex criteria related to the conceptual and perceptual knowledge about specialty coffee that highlight aspects with distinct influence on the beverage sensory profile and that add a strong subjective component to their assessment.

Nevertheless, we should point out that consumers usually evaluate quality in food and beverages in terms of sensory attributes, healthiness, convenience and process characteristics. Plus, their quality perception “often weights more heavily on their individual expectations and past experience than on intrinsic product characteristics”. (Giancalone *et al.*, 2016, p. 2463). Moreover, “though few consumers are indeed coffee experts, the consumers’ own perception of their knowledge, or level of “subjective knowledge”, can influence their purchase and consumption decisions” (Ufer *et al.*, 2019, p. 7). These reflections should be considered when developing marketing strategies for the specialty coffee market.

#### 4.5 Transition among constructs

The results described in Table 8 show that, when using the adaptive AVE index, the interpretation of constructs corresponding to the classification of specialty coffee consumers is corroborated. These results reflect a measure of validity, in which the closer to the unit value, the closer the correspondence between the measure and the construct.

In Table 9, we demonstrate the parameter estimates and the correlations between the constructs of the theoretical model used to define the variables that characterize the classification of specialty coffee consumers.

Construct	AVE <sub>ADP</sub>
$\xi_1$	1.541
$\xi_2$	1.372
$\xi_3$	1.174

**Table 8.**

Adaptive AVE values for the constructs used in the proposed model



Parameter	Estimates	Parameters	Estimates
$\lambda_{11}$	0.292	$\lambda_{20}$	0.189
$\lambda_{12}$	0.312	$\lambda_{21}$	0.301
$\lambda_{13}$	0.307	$\lambda_{22}$	0.198
$\lambda_{14}$	0.553	$\lambda_{24}$	0.126
$\lambda_{15}$	0.272	$\lambda_{31}$	0.124
$\lambda_{16}$	0.384	$\lambda_{34}$	0.227
$\lambda_{27}$	0.143	$\lambda_{35}$	0.187
$\lambda_{28}$	0.517	$\lambda_{39}$	0.379
$\lambda_{29}$	0.311	$\lambda_{33}$	0.114
Correlations	$\phi_{12} = 0.204$	$\phi_{13} = -0.307$	$\phi_{23} = -0.499$

**Table 9.**  
Estimated parameters  
for the proposed model

Regarding the estimates of the model parameters and considering the adaptive regressions, we found a slightly positive correlation, ( $\phi_{12} = 0.204$ ) between the constructs of regular consumers ( $\xi_1$ ) and enthusiasts ( $\xi_2$ ). Thus, we identified a certain alignment between these individuals' specialty coffees consumption motivations and purchasing criteria, with the purchase criteria adopted by enthusiasts ( $\xi_2$ ) being perceived as a natural evolution or deepening of the consumption motivations depicted by regular consumers ( $\xi_1$ ). The easy deepening and consolidation of conceptual knowledge (LaTour *et al.*, 2011) of simple or moderate complexity by consumers, to migrate from regular consumers ( $\xi_1$ ) to enthusiasts ( $\xi_2$ ), would justify a certain ease (as in less effort or intensity of product involvement) for individuals to transit between these constructs.

The negative correlation identified among the constructs of enthusiasts ( $\xi_2$ ) and experts ( $\xi_3$ ) - ( $\phi_{23} = -0.499$ ) - and between regular consumers ( $\xi_1$ ) and experts ( $\xi_3$ ) - ( $\phi_{13} = -0.307$ ) - indicates distinct behaviors regarding specialty coffee consumption motivations and purchase criteria, which reflect different levels of knowledge about and involvement with the product (Guimarães *et al.*, 2019). These results may indicate a significant change in the consumers' perception of what constitutes quality and what criteria they use to evaluate it, as well as differences in consumption motivations.

We assume, therefore, that an individual's transition from enthusiast ( $\xi_2$ ) to expert ( $\xi_3$ ) is more complex than the migration of a regular consumer ( $\xi_1$ ) to the enthusiasts ( $\xi_2$ ) category, demanding greater engagement from individuals and an important reformulation of their conception about the product.

As highlighted by Teixeira (2020, p. 287), "specialty coffee culture is not for everyone, for a refinement of taste, one needs to perceive subtle nuances of aroma, flavor, and body that each method of coffee preparation can provide. Thus, only a restrict number of individuals are willing and able to transition to the enthusiast, and later to expert, consumer groups. In other words, "average consumers may not have access or experience with finer, high-end examples of specific products and hence may lack the necessary frame of reference to judge food quality", but "given such experience, consumers learn to recognize the appeal of higher-end products and acquire new preferences for these" (Giancalone *et al.*, 2016, p. 2463).

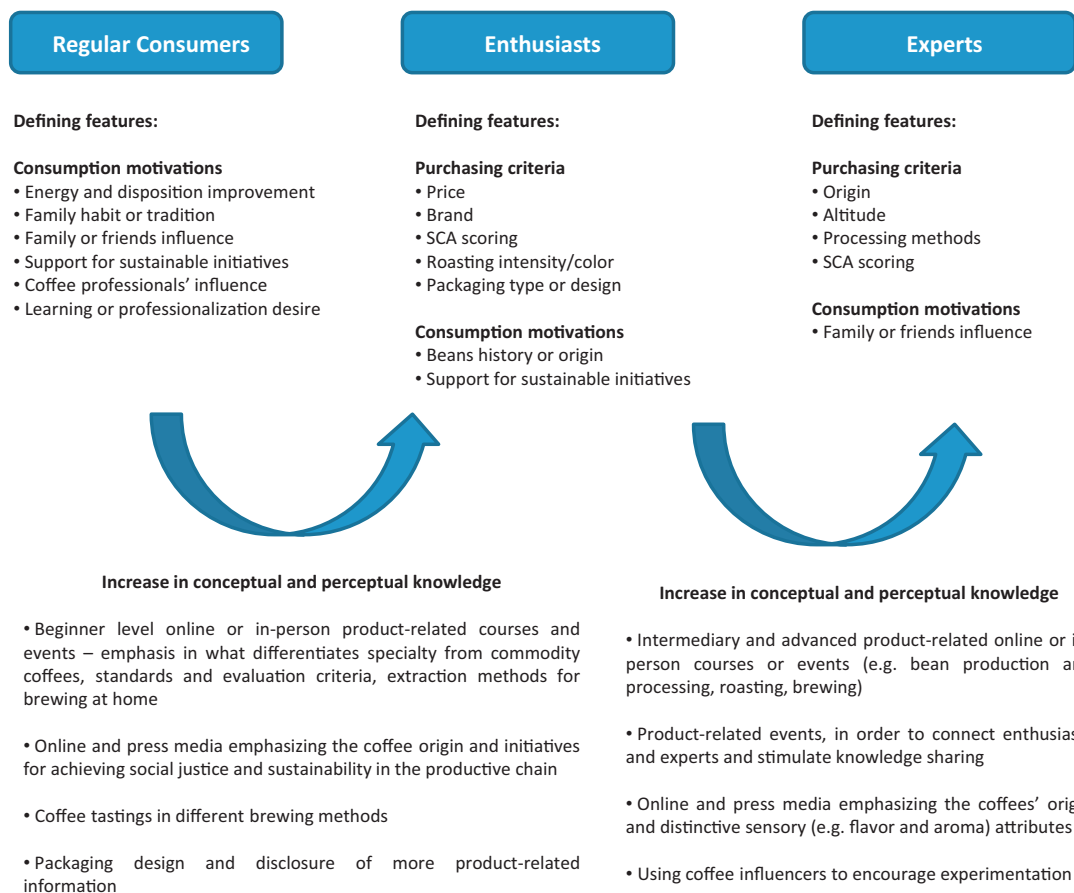
Thus, on one hand, we consider that the transition of individuals between the categories of regular consumers ( $\xi_1$ ) and enthusiasts ( $\xi_2$ ) reflects a moderate deepening and consolidation of certain conceptual knowledge considered simpler (see previous section). On the other hand, the transition from enthusiasts ( $\xi_2$ ) to experts ( $\xi_3$ ) would require both the deepening and consolidation of complex conceptual knowledge, as well as the deconstruction of previous perceptual knowledge and the construction of new and complex forms of perceptual evaluation of the product, which are commonly developed collaboratively with other members of the consumer community (Quintão *et al.*, 2017a).

As demonstrated by Quintão *et al.* (2017b), “connoisseurs”, or experts, would seek to distinguish themselves from other consumers through ritualistic practices, in a process the authors name the “taste transformation ritual”. However, considering our results, we reason that the taste transformation ritual comprises, at different intensities, the individuals’ entire transition process between the specialty coffee consumer categories and is not restricted to the transformation of regular consumers into “connoisseurs”. In Figure 2, we emphasize possible strategies specialty coffee market agents can adopt to stimulate consumers’ transition from lower to higher-engagement constructs.

According to Lammigan (2020), it is important to address both online and in-person consumer education strategies since “online platforms cannot effectively assist in changing tastes, but only in reiterating what interactions have already occurred” (p. 6). The author also adds that such strategies are especially adequate to experience-based industries anchored to cognitive and evaluative complex frameworks, such as the specialty coffee industry. In addition, educating consumers as to what differentiates specialty from traditional coffees is complex and demands “intricate communication and coordination between interlocutors and consumers in the form of sensory experience and a shared vocabulary to understand these differences” (p. 7).

As for the transition of individuals between constructs, two important questions remain to be addressed in future research:

- (1) Is it possible to directly migrate regular consumers ( $\xi_1$ ) to the experts ( $\xi_3$ ) category? This would mean that the level of conceptual and perceptual knowledge on the



**Figure 2.** Infographic with concrete examples of marketing strategies that can be implemented to improve effectiveness of the transitions between the three different profiles

product (LaTour and LaTour, 2010; LaTour *et al.*, 2011), as well as the individual's involvement in its consumption, advanced abruptly, which would indicate a consumer "passion" for this product or an intense desire for professionalization in this market. This would require greater susceptibility to the influence of other individuals and a great learning desire, as explored in the previous section. In this sense and according to the researchers' experience, professionals from different areas have abandoned their original professions in search of a career in the specialty coffee market. To the best of our knowledge, there are no studies demonstrating what would lead such professionals to make this decision.

- (2) Is it possible for individuals to reversely migrate from one construct to another, that is, "returning" from experts ( $\xi_3$ ) to enthusiasts ( $\xi_2$ ) and, finally, to regular consumers ( $\xi_1$ )? In this case, an abrupt reduction in these consumers' level of involvement with the product would be necessary, which, over time, would result in their low conceptual and/or perceptual knowledge about specialty coffees, based on the principle that general knowledge about this product category will continue to evolve.

## 5. Conclusions

Important empirical advances have been made, contributing to the better comprehension of the Brazilian specialty coffee consumers. The structural equation modeling enabled the construction of a model composed of three constructs (*i.e.* groups of specialty coffee consumers), which that represent different levels of involvement with and knowledge about the product. It is also composed of observed predictive variables about which category best represents each consumer. The results allowed an improvement in the consumers' distinction and characterization among the categories proposed by Guimarães *et al.* (2019), contributing to the enhancement and simplification of the marketing strategies carried out by the players in this market.

We encouraged the discussion about which factors stimulate an individual's transition from an initial to a subsequent construct and demonstrated a greater ease of individuals' transition from regular consumers to enthusiasts, than from enthusiasts to experts. This could be explained by the need to perform "taste transformation rituals" (Quintão *et al.*, 2017a, b) so that the second transition may occur. However, we considered that such rituals comprise the individual's entire transition process between such categories, whether to a greater or lesser extent. In this aspect, we recommend further studies to verify the possibility of direct transition from regular consumers to experts and, if so, the factors that stimulate it; in addition to the possibility of transition between constructs in reverse, that is, from experts ( $\xi_3$ ) to enthusiasts ( $\xi_2$ ) and finally, regular consumers ( $\xi_1$ ). This knowledge allows different players in the national coffee chain – e.g. coffee growers, roasters, coffee shops, associations, among others – to adopt strategies aimed at encouraging the migration of consumers to more engaged constructs, thus promoting the expansion and qualification of the Brazilian specialty coffee market. Important effects of the expansion of this market would be the greater internal appreciation of the product and the consequent higher prices earned by coffee growers, also reducing their dependence on exports.

This study aimed not at exhausting the studied topic but providing a basis for broadening the understanding of the Brazilian consumer of specialty coffees, to be achieved in future studies.

Regarding limitations, we highlight the exclusive use of criteria related to conceptual knowledge (LaTour *et al.*, 2011) as predictive variables for the establishment of constructs. In future work, we recommend carrying out qualitative and experimental research, addressing aspects related to both conceptual and perceptual knowledge (LaTour *et al.*, 2011) to support

the determination of predictive variables to compose the different consumer constructs. In addition, assessing coffee quality might be highly complex and subjective, so it is interesting to verify whether consumers are truly able to recognize quality differences in coffee (Sabioa and Spers, 2020) and whether these differences correspond to industry standards (Giancalone et al., 2016). Furthermore, we identified different variables that characterize the consumers' groups, but, at first, it is not possible to determine the each one's importance in the constructs, which is an issue that requires further study.

Finally, there are limitations associated with the original database used in this study, whose data might be biased and highly homogeneous. Therefore, it is not possible to generalize the results for the entire coffee consuming population in the country. Further studies are also necessary to verify whether the proposed model is highly place-based or if part of the results can be applied in other countries or cultural contexts.

### Notes

1. At the time of the estimates, the possible effects of the Covid-19 pandemic on the national economy were not considered. These figures may soon be revised.
2. The emergence and adherence to the coffee waves occurred in distinct periods in different countries, being largely influenced by cultural factors. Thus, the dates mentioned refer especially to countries considered mature for the beverage's consumption, like the USA (Guimarães et al., 2016).
3. The Specialty Coffee Association (SCA) quality standard is the most widely accepted by coffee professionals worldwide. Through a cupping protocol, based on 11 criteria, those coffees that achieve a scoring above 80 points, on a scale from zero to 100, are considered specialty (Associação de Cafés Especiais, 2020).
4. According to ABIC's methodology, coffees are classified according to their global quality score (GQ), being categorized as traditional/extra strong ( $GQ \geq 4.5$  and  $\leq 5.9$ ), superior/higher coffees ( $GQ \geq 6.0$  and  $\leq 7, 2$ ) or gourmet ( $HQ \geq 7.3$  and  $\leq 10$ ). Although these methodologies use distinct product evaluation criteria, different consumers wrongly associate ABIC's gourmet coffees with SCA's specialty coffees.

### References

- Andrade, H.C.C., Alcântara, V.C., Aldano, A.P.M. and Santos, A.C. (2015), "Atribuição de sentidos e agregação de valor: Insumos para o turismo rural em regiões cafeicultoras", *Revista Brasileira De Ecoturismo*, Vol. 8 No. 12, pp. 333-346.
- Angus, A. (2020), "Top 10 global consumer trends for 2018 - emerging forces shaping consumer behaviour", *Euromonitor International*, available at: <http://go.euromonitor.com/white-paper-economies-consumers-2018-global-consumer-trends-EN.html> (accessed 21 May 2020).
- Arnab, M. and Michael, S. (2008), "On adaptive linear regression", *Journal of Applied Statistics*, Taylor & Francis, Vol. 35 No. 12, pp. 1409-1422.
- Associação Brasileira da Indústria do Café (2010), "Tendências de Consumo de Café – VIII", available at: <http://abic.com.br/src/uploads/2017/10/Pesquisa-Tendencias-de-Consumo-VIII-2010-Final.pdf> (accessed 21 May 2020).
- Associação Brasileira da Indústria do Café (2020), "Categorias de qualidade", available at: <https://www.abic.com.br/certificacao/qualidade/categorias-de-qualidade/> (accessed 21 May 2020).
- Associação de Cafés Especiais (2020), "Protocolos e práticas recomendadas", available at: <https://sca.coffee/research/protocols-best-practices> (accessed 21 May 2020).
- Baltar, F. and Brunet, I. (2012), "Social research 2.0: virtual snowball sampling method using Facebook", *Internet Research*, Vol. 22 No. 1, pp. 57-74, available at: <https://doi.org/10.1108/10662241211199960> (accessed 22 June 2020).

- 
- Boaventura, P.S.M., Abdalla, C.C., Araújo, C.L. and Arakelian, J.S. (2018), "Value co-creation in the specialty coffee value chain: the third-wave coffee movement", *Revista de Administração de Empresas (Journal of Business Management)*, Vol. 58 No. 3, pp. 254-266.
- Borrella, I., Mataix, C. and Carrasco-Gallego, R. (2015), "Smallholder farmers in the speciality coffee industry: opportunities, constraints and the businesses that are making it possible", *IDS Bulletin*, Vol. 46 No. 3, pp. 29-44, available at: <https://doi:10.1111/idsb.2015.46.issue-3> (accessed 22 June 2020).
- Brown, T.A. (2006), *Confirmatory Factor Analysis for Applied Research*, The Guilford Press, New York.
- Cirillo, M.A. and Barroso, L.P. (2012), "Robust regression estimates in the prediction of latent variables in structural equation models", *Journal of Modern Applied Statistical Methods*, Vol. 11, pp. 42-53.
- Costa, B.R. (2020), "Brazilian specialty coffee scenario", in Almeida, L.F. and Spers, E.E. (Eds), *Coffee Consumption and Industry Strategies in Brazil*, Woodhead Publishing, Elsevier.
- Giacalone, D., Fosgaard, T.R., Steen, I. and Münchow, M. (2016), "Quality does not sell itself: divergence between 'objective' product quality and preference for coffee in naïve consumers", *British Food Journal*, Vol. 118 No. 10, pp. 2462-2474, doi: [10.1108/BFJ-03-2016-0127](https://doi.org/10.1108/BFJ-03-2016-0127).
- Guimarães, E.R. (2016), *Terceira Onda do Café: Base Conceitual e Aplicações*, Masters dissertation, Universidade Federal de Lavras, Lavras, available at: <http://repositorio.ufla.br/jspui/handle/110972> (accessed 22 June 2020).
- Guimarães, E.R., Castro Júnior, L.G. and Andrade, H.C.C. (2016), "A terceira onda do café em Minas Gerais", *Organizações Rurais E Agroindustriais*, Vol. 18 No. 3, pp. 214-227, available at: <https://doi:10.21714/2238-68902016v18n3p214> (accessed 22 June 2020).
- Guimarães, E.R., Leme, P.H.M.V., Rezende, D.C., Pereira, S.P. and Santos, A.C. (2019), "The brand new Brazilian specialty coffee market", *Journal of Food Products Marketing*, Vol. 25 No. 1, pp. 49-71, doi: [10.1080/10454446.2018.1478757](https://doi.org/10.1080/10454446.2018.1478757) (accessed 22 June 2020).
- Hoşafçı, P. (2018), "8 food trends for 2018", *Euromonitor International*, available at: <https://blog.euromonitor.com/2018/03/8-food-trends-2018-2.html> (accessed 22 June 2020).
- Lannigan, J. (2020), "Making a space for taste: context and discourse in the specialty coffee scene", *International Journal of Information Management*, Vol. 51, 101987.
- LaTour, K.A. and LaTour, M.S. (2010), "Bridging aficionados' perceptual and conceptual knowledge to enhance how they learn from experience", *Journal of Consumer Research*, Vol. 37 No. 4, pp. 688-697, doi: [10.1086/655014](https://doi.org/10.1086/655014) (accessed 22 June 2020).
- LaTour, K.A., LaTour, M.S. and Feinstein, A.H. (2011), "The effects of perceptual and conceptual training on novice wine drinkers' development", *Cornell Hospitality Quarterly*, Vol. 52 No. 4, pp. 445-457, doi: [10.1177/1938965511420695](https://doi.org/10.1177/1938965511420695) (accessed 22 June 2020).
- Lima, L.M., Elias, L.P., Silva, M.M.D., Silva, K.V. and Pacheco, A.S.V. (2020), "Behavioral aspects of the coffee consumer in different countries: the case of Brazil", in Almeida, L.F. and Spers, E.E. (Eds), *Coffee Consumption and Industry Strategies in Brazil*, Woodhead Publishing, Elsevier, 2020.
- Manzo, J. (2010), "Coffee, connoisseurship, and an ethnomethodologically-informed sociology of taste", *Human Studies*, Dordrecht, Vol. 33 Nos 2-3, pp. 141-155.
- Proença, M. (2017), "Pesquisa revela crescimento de 18,1% no consumo brasileiro de cafés especiais", *Revista Espresso*, available at: <http://revistaespresso.com.br/2017/11/28/pesquisa-revela-crescimento-de-181-no-cosumo-brasileiro-de-cafes-especiais/> (accessed 22 June 2020).
- Quintão, R.T. (2015), *The Rite of Passage from Regular to Connoisseur Consumer: The Role of the Taste Transformation Ritual in the Specialty Coffee Context*, Doctoral dissertation, Fundação Getúlio Vargas, São Paulo, available at: <http://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/13614/Dissertation%20Ritual%20v.24%20entregue%20secretaria%20CMCD.pdf?sequence=2&isAllowed=y> (accessed 22 June 2020).
- Quintão, R.T. and Brito, E.P.Z. (2016), "Connoisseurship consumption and market evolution: an institutional theory perspective on the growth of specialty coffee consumption in the USA", *Brazilian Journal of Marketing*, Vol. 15 No. 1, pp. 1-15.

- 
- Quintão, R.T., Brito, E.P.Z. and Belk, R.W. (2017a), "Connoisseurship consumption community and its dynamics", *Review of Business Management*, Vol. 19 No. 63, pp. 1-17, doi: [10.7819/rbgn.v0i0.2982](https://doi.org/10.7819/rbgn.v0i0.2982) (accessed 22 June 2020).
- Quintão, R.T., Brito, E.P.Z. and Belk, R.W. (2017b), "The taste transformation ritual in the specialty coffee market", *Revista De Administração De Empresas*, Vol. 57 No. 5, pp. 483-494, available at: <https://doi.org/10.1590/s0034-759020170506> (accessed 22 June 2020).
- Ramírez-Correa, P., Rondán-Cataluña, F.J., Moulaz, M.T. and Arenas-Gaitán, J. (2020), "Purchase Intention of specialty coffee", *Sustainability*, Vol. 12, 1329.
- Roseberry, W. (1996), "The rise of yuppie coffees and the reimagination of class in the United States", *American Anthropologist*, Vol. 98 No. 4, pp. 762-775.
- Sabioa, R.P. and Spers, E.E. (2020), "Does coffee origin matter? An analysis of consumer behavior based on regional and national origin", in Almeida, L.F. and Spers, E.E. (Eds), *Coffee Consumption and Industry Strategies in Brazil*, Woodhead Publishing, Elsevier.
- Skeie, T.R. (2003), "Norway and coffee", available at: [https://web.archive.org/web/20031011091223/http://roastersguild.org/052003\\_norway.Shtml](https://web.archive.org/web/20031011091223/http://roastersguild.org/052003_norway.Shtml) (accessed 22 June 2020).
- Teixeira, L.V. (2020), "The consumption of experiences in specialty coffee shops", in Almeida, L.F. and Spers, E.E. (Eds), *Coffee Consumption and Industry Strategies in Brazil*, Woodhead Publishing, Elsevier.
- Teixeira, L.V. and Nunes, M.R.F. (2016), "Café e cenas culturais na cidade de São Paulo: consumo, memória e ambiências comunicacionais", *Razón y Palabra*, Vol. 20 Nos 3-94, pp. 347-383.
- Teles, C.R.A. and Behrens, J.H. (2020), "The waves of coffee and the emergence of the new Brazilian consumer", in Almeida, L.F. and Spers, E.E. (Eds), *Coffee Consumption and Industry Strategies in Brazil*, Woodhead Publishing, Elsevier.
- Ufer, D., Lin, W. and Ortega, D.L. (2019), "Personality traits and preference for specialty coffee: results from a coffee shop field experiment", *Food Research International*, Vol. 125, pp. 1-9.
- Urwin, R., Kesa, H. and São João, E. (2019), "The rise of specialty coffee: an investigation into the consumers of specialty coffee in Gauteng", *African Journal of Hospitality, Tourism and Leisure*, Vol. 8 No. 5, pp. 1-17.

**Corresponding author**

Patricia Mendes dos Santos can be contacted at: [patymendesdossantos@hotmail.com](mailto:patymendesdossantos@hotmail.com)

## FINAL CONSIDERATIONS

This study introduced a new approach on Adaptive Linear Regression, adapted to structural equation models, and offered a new study perspective for areas which use this kind of modeling. In addition, seeking to innovate the use of new statistical methodologies, we proposed an improvement to the average variance extracted (AVE) index, given a plug-in approach, by replacing error variances with the factor loadings of estimated adaptive regressions. In this sense, a proposal providing new factor loadings may result in a more informative index (AVE) with the use of adaptive regressions. The results from the first paper show that the adaptive linear regression method in formative structural models, considering that outliers originated from symmetrical distributions or from a multivariate log-normal distribution, was effective for correctly specified models. Likewise, for models with specification errors, this method was not as effective, which was expected. The second paper aimed at deepening the studies of Guimarães et al. (2019), who proposed a more precise and more easily understandable interpretation of which variables place the Brazilian specialty coffee consumers into different groups. The results made it possible to improve the differentiation and separation of consumers into the categories proposed by Guimarães et al. (2019), thus contributing with the enhancement and simplification of the marketing strategies used by players in this market. Moreover, we discussed which factors stimulate the transition of an individual from an initial construct to a second one and showed that transitioning from regular consumers to enthusiasts is easier than from enthusiasts to specialists. Thus, we provided material to better understand the Brazilian specialty coffee consumer. For future research, the AVE index may be compared to other construct validation indices and a study could also be done to define a limit to determine the adequacy of this index.