University of Brasília
Institute of Psychology
Post-Graduate Program of Social, Work, and Organizational Psychology – PSTO

Ph.D. Dissertation

Beyond Psychometric Assumptions: How to Develop New Psychological Measures

Para Além de Pressupostos Psicométricos: Como Desenvolver Novas Medidas Psicológicas

Víthor Rosa Franco

Brasília – DF, 4th December 2019

Beyond Psychometric Assumptions: How to Develop New Psychological Measures

Para Além de Pressupostos Psicométricos: Como Desenvolver Novas Medidas Psicológicas

Víthor Rosa Franco

Doctoral dissertation elaborated under the supervision of Prof. Ph.D. Jacob Arie Laros, and presented to the Post-Graduate Program of Social, Work, and Organizational Psychology of the University of Brasília, as partial requirement for the degree of Doctor in Social, Work, and Organizational Psychology.

Supervisor: Prof. Ph.D. Jacob Arie Laros

Examining Committee:

Prof. Ph.D. Felipe Valentini
Programa de Pós-Graduação em Psicologia
Universidade São Francisco - USF
Membro externo à UnB

Prof. Ph.D. Ricardo José de Moura
Programa de Pós-graduação em Ciências do Comportamento
Universidade de Brasília - UnB
Membro externo ao PSTO

Prof. Ph.D. Josemberg Moura de Andrade
Programa de Pós-Graduação em Psicologia Social, do Trabalho e das Organizações
Universidade de Brasília - UnB
Membro interno ao PSTO

Prof. Ph.D. Elaine Rabelo Neiva
Programa de Pós-Graduação em Psicologia Social, do Trabalho e das Organizações
Universidade de Brasília - UnB
Membro suplente

Dedicatory

Oh, life is good... As good as you wish!

“For me, it is far better to grasp the Universe as it really is than to persist in delusion, however satisfying and reassuring.”

Carl Sagan

“*Give a boy a hammer and chisel; show him how to use them; at once he begins to hack the doorposts, to take off the corners of shutter and window frames, until you teach him a better use for them, and how to keep his activity within bounds.*”

Charles Reade and Dion Boucicault

Acknowledgements

I thank my family for all the support and personal sacrifice they made so that I could get here. Starting with my grandparents—Teó, Arino, Cecilia, and Hélio—and my parents—Cynthia da Silva Rosa, and Hélio José Franco Júnior—; I am so grateful that they always reinforced my nerdy side and taught me to love to study. Thanks also to my siblings—Lázaro Rosa Franco, Sofia Maria Rosa Franco, and Thomas Rosa Franco—for being nerds and the best friends of all time. I also thank all of my extended family—cousins, uncles, aunts—for always supporting me in all my decisions. I love you all.

I thank my supervisor Jacob Arie Laros for accepting to accompany me on this endeavor. From the first request, not even sure which topic I would like to study, to the last-minute decisions, remote supervisions and supervisions on weekends. Laros is a great advisor, teacher, and person, and I believe all of his support was instrumental in getting the job done in the quality it did. I also thank the examining committee—Elaine Rabelo Neiva, Felipe Valentini, Josemberg Moura de Andrade, and Ricardo José de Moura—for agreeing to participate in this moment and for all the fundamental contributions.

I also thank the professors/teachers Antonio Abreu, Antonio Pedro Mello, Bartholomeu Tôrres Tróccoli, Carla Antloga, Cristiane Faiad, Dida Mendes, Edna Akemi Ueda, Gerson Américo Janczura, Hudson Golino, Jorge Mendes, Luciano Grüdtner Buratto, Ricardo Primi, Timothy Mulholland, Wanessa Loureiro, and Zorg Ribeiro da Costa. They are all excellent masters and have always reinforced in me the desire to pursue academic life and to be increasingly nerdy, hoping one day to be as capable (and nerdy) as all of them. In particular I would also like to thank professor Fabio Iglesias. I started my history in science under his guidance, and the history follows with your partnership and personal friendship, also of great value to me. Thank you for all the teachings and guidance in my life.

To my colleagues from laboratory, course or profession Ana Luiza Marinho, André Paiva, André Rabelo, Angelica Oliveira, Beatriz Cavendish, Camila Gastal, Carlos Manoel, Daniel Barbieri, Douglas Piasson, Elena Pinheiro, Elis Ramos, Elis Martins, Filipe Lima, Filipe Gabriela Campelo, Gabriela Macedo, Gabriela Ribeiro, Giordana Bruna, Hannah Hämmer, Isangelo Souza, Izabella Melo, Jazon Torres, Jessica Farias, Jessica Riechelmann, João Modesto, Jonathan Jones, Julia Gisler, Laura Andrade, Letícia Ferreira, Lorena Andreoli, Luana Veiga, Lucas Heiki, Lude Marieta, Luiz Victorino, Marcos Pimenta, Marcos Lima, Mariana Santos, Marina Caricatti, Martina Mazzoleni, Mauricio Sarmet, Raiane Nunes Nogueira, Patrícia Santos, Paula Gabriela, Raquel Hoersting, Raquel Loewenhaupt, Renan Benigno, Stela de Lemos, Teresa Clara, Tiago França, Vitória Lima, and Victor de Souza. Thank you for all the discussions, teachings and mutual support.

To the friends I met someplace in life Adler Adriel, Adrielli Nazario, Alexandre Barba Ruiva, Amanda do Couto, Anna Thais, Britt Bayesian Jane, Diach Selch, Diux Ronan, Erick di Serio, Fernando Alexandre, Filipe Cardoso, Frederico Bicalho, Gabriel Mosna, Henrique Simas, Hugo Sousa, João Roberto, Juliana Simas, Maitê Assis, Marcella Pantarotto, Mari Junqueira, Mari Sá, Naty Sá, Paula Souza, Paulo Victor, Pp Martins, Rafael Marks, Raquel Simas, Raul Marques, Sarah Goulart, Stefano Mosna, Talitha Pumar, Thais Staudt, Thiago Fernando, Thiago Pereira, Vitor Guimarães, and Victor Keller. Thank you for all the hours of laughter and happiness that come with me. Also, many of you are nerds too. Special thanks to Alexandre Gomide, Guilherme Gonçalves and Isabela Lima; they know why.

Till alla vänner jag träffade från Sverige Anders Lundquist, Angel Angelov, Anita Lindmark, David Källberg, Emma Persson, Flavia Raschini, Gabriel Wallin (extra tack för den svenska fixen!), Guilherme Barros, Ingeborg Waernbaum, Ingela Klinga, Jenny Häggström, Jessica Fahlen, Johan Svensson, Kadri Meister, Katarina Kempe, Kreske Ecker, Lina Schelin, Marco Doretti, Maria Karlsson, Massimo Maresca, Minna Genbäck, Niloofar

Moosavi, Pär “Rossi” Sehlström, Pryiantha Wijayatunga, Sajad Mortazavi, Sandra Behren, Simone Mellquist, Svante Klinga Tanya Gorbach, och Xavier de Luna. Tack så mycket för att ni hjälpte mig att anpassa mig så enkelt till detta underbara land och fick mig att känna mig som hemma från början. Särskilt tack till Leo Nazar för hans så stora hjärta som bara förlorar till hans stora nördhjärna. Från att hjälpa till att hantera ångesten av att vara på en ny plats, till att hjälpa med att hitta i staden, busskort, kasta bort tiden på fredag eller helgen, kort sagt, genom din mycket speciella vänskap. Särskilt tack till mina goda vänner Anastasia Potehina, Daria Nikitina, David Kvist och Yulia Ryanova. Torsdagskvällar har inte varit så bullriga utan er, jag saknar det. Ett särskilt tack också till professor Marie Eriksson för allt stöd, arbete och personligt, för de snabba sex månaderna jag var på din underbara enhet. Ett extra speciellt tack till min andra handledare Marie Wiberg. Inte bara hjälpte du mig att förverkliga en dröm, utan det var också viktigt för mig att få det här jobbet gjort. Slutligen, framför allt, tack vare denna erfarenhet, ändrade jag fullständigt min åsikt om hur det akademiska livet kan vara, på ett mycket positivt sätt. Tack så mycket.

To my beautiful friends Aline Fernandes, Juliana Almeida, Ligia Abreu, Lucas Caldas, and Raissa Damasceno. For all the hours of support, lunch, dinner, fun, crying, smiling, studying, and chatting away; none of this can be represented in any kind of currency but the currency of love and friendship. Special thanks to my lab friend Talita Alves. Her strength to overcome all her difficulties is extremely inspiring; you help me to be better!

Finally, Gabriela Yukari Iwama. Three years that feel like forever. We’ve spent much of the last year apart from each other, but I couldn’t feel closer to you than I feel now. I am so happy for everything you accomplished and I am very happy you are here with me in this accomplishment of mine. You are so amazing; you do so many things that can only make me think that I am the luckiest person in the world. I hope we share many more adventures, many scientific papers, and many more years of happiness. I love you so much, Mozinha.

Summary

List of Tables

List of Figures

List of Abbreviations
(in alphabetic order)

1PLM/1PL	One-Parameter Logistic Model
2PLM	Two-Parameter Logistic Model
Acc	Accuracy
AIC	Akaike Information Criterion
Bayesian-SOFA	Bayesian implementation of a SOFA model
BIC	Bayesian Information Criterion
CDF	Cumulative Density Function
CFT	Common Factor Theory
CG	Chain Graph
CGMM	Correlation Gaussian Mixture Model
CIRM/CM	Conditional Item Response Model
CMT	Conjoint Measurement Theory
Coop-Comp scale	Cooperation and Competition attitudes' scale
CPM	Cognitive Psychometric Modeling
CTT	Classical Test Theory
DAG	Directed Acyclic Graph
DGP	Data Generating Process
DIC	Deviance Information Criterion
DP	Dirichlet Process
EAP	Expected a Posteriori
EBIC	Extended BIC
EFA	Exploratory Factor Analysis
EGA	Exploratory Graph Analysis
ENEM	Exame Nacional do Ensino Médio
GAM	Generalized Additive Models
GAM-SFA	GAM regression model of SFA
GARI	Graph Adjusted Rand Index
GES	Greedy Equivalence Search algorithm
HC	Hill-Climbing algorithm
HitND	Percentage of hits of the number of dimensions
IRF	Item Response Function

IRM	Item Response Model
IRT	Item Response Theory
ISE	Integrated Squared Error
Kernel-SFA	Kernel smooth regression model of SFA
LASSO	Least Absolute Shrinkage and Selection Operator
Loess-SFA	Locally estimated scatterplot smoothing model of SFA
LR	Likelihood Ratio
LVT	Latent Variable Theory
MAE	Mean Absolute Error
MCMC	Markov Chain Monte Carlo
MLE	Maximum Likelihood estimation
MMHC	Min-Max Hill-Climbing algorithm
MSA	Mokken Scale Analysis
NIRM	Nonparametric Item Response Model
NMI	Normalized Mutual Information
OS-IRM/OM	Optimal-score procedure
PA	Parallel Analysis
PA-EFA	Combined model using PA and EFA
PCG	Power Chain Graph
PG	Power Graph
PPV	Positive Predictive Value
RM	Rasch model
RMSE	Root-Mean Squared Error
SEM	Structural Equation Model
SFA	Stochastic Frontier Analysis
SOFA	Situational Optimization Function Analysis
SS	Sum Scores
TIRM	Tree-Based Item Response Model
TNR	True Negative Rate
TPR	True Positive Rate
UG	Undirected Graph
VI	Variation of Information

GENERAL ABSTRACT

What defines a good measurement? In the present dissertation we argue, and show, that defining a good measurement can be much more complex than simply performing a factor analysis or an analysis using item response theory. The overall objective of this dissertation is to present three principal assumptions of psychometric measurement, and to develop alternatives for traditional psychological measurement. The dissertation is divided in four studies. The first one is a theoretical study in which three central assumptions common to psychometric theory and psychometric practice are presented, and in which is shown how alternatives to traditional psychometric approaches can be used to improve psychological measurement. These alternatives were developed by adapting each of these three assumptions: (1) the assumption of structural validity; (2) the process assumption; and, (3) the construct assumption. The structural validity assumption relates to the implementation of mathematical models. The process assumption implies that a specific underlying process is generating the observed data. The construct assumption infers that the observed data on its own do not constitute a measurement, but the measure are the latent variables that originate the observed data. Several examples of already existing alternative psychometric approaches are presented in the first study. The second study relates to the structural validity assumption and aimed to develop two new item response models for polytomous and binary items that do not assume a normal distribution of the true scores. The first model that was developed, the Conditional Item Response Model (CIRM), assumes a beta-binomial distribution. The second new model is a Bayesian implementation of the optimal score procedure (OS-IRM). Both new models were compared with the traditional Rasch model: the results indicate that the two developed models improve various aspects of the Rasch model. The third study was derived from the process assumption and had three objectives. First, to develop a Bayesian implementation of the situational optimization function analysis (SOFA) framework. Second, to compare this Bayesian implementation of SOFA with three other Maximum Likelihood-based models that are used to estimate true scores. The third objective was to show how joint modeling can be used for validity research. One of the main advantages of the SOFA framework compared to the traditional psychometric approach is that SOFA relies on experimental data, improving the validity of the measures. The fourth and final study was derived from the construct assumption and its main objective was to develop a procedure of structure learning of power chain graphs (PCGs). A PCG is a type of graph that represents causal relations between groups of variables. It can be thought as a full exploratory version of structural equation modeling, as well as a psychometric model that is not dependent on latent variables. These four studies intend to show that psychometric modeling should not be restricted to the use of traditional measurement models, but should also consider adapting these traditional models in accordance with the intended use and theoretical processes that originate the observed measures.

Keywords: psychometrics; quantitative modeling; formal theorizing; Bayesian modeling; measurement theory.

RESUMO GERAL

O que define uma boa medida? Na presente tese, argumentamos e mostramos que definir uma boa medida pode ser muito mais complexo do que simplesmente executar uma análise fatorial ou uma análise usando a teoria da resposta ao item. O objetivo geral desta dissertação é apresentar três principais pressupostos da medida psicométrica e desenvolver alternativas para a medida psicológica tradicional. A tese está dividida em quatro estudos. O primeiro é um estudo teórico no qual são apresentados três pressupostos centrais comuns à teoria psicométrica e à prática psicométrica, e no qual é mostrado como alternativas às abordagens psicométricas tradicionais podem ser usadas para melhorar a medição psicológica. Essas alternativas foram desenvolvidas adaptando cada um desses três pressupostos: (1) o pressuposto de validade estrutural; (2) o pressuposto do processo; e (3) o pressuposto de construto. O pressuposto de validade estrutural refere-se à implementação de modelos matemáticos. O pressuposto de processo implica que um processo subjacente específico está gerando os dados observados. O pressuposto de construto infere que os dados observados por si só não constituem uma medida, mas que as medidas são as variáveis latentes que originam os dados observados. Vários exemplos de abordagens psicométricas alternativas já existentes são apresentados no primeiro estudo. O segundo estudo se refere ao pressuposto de validade estrutural e teve como objetivo desenvolver dois novos modelos de resposta aos itens para itens politômicos e binários que não assumem uma distribuição normal dos escores verdadeiros. O primeiro modelo desenvolvido, o Modelo de resposta ao item condicional (CIRM), assume uma distribuição beta-binomial. O segundo novo modelo é uma implementação Bayesiana do procedimento de escore ótimo (OS-IRM). Ambos os novos modelos foram comparados com o modelo tradicional de Rasch: os resultados indicam que os dois modelos desenvolvidos melhoram vários aspectos do modelo de Rasch. O terceiro estudo foi derivado do pressuposto do processo e tinha três objetivos. Primeiro, desenvolver uma implementação Bayesiana do *framework* de análise da função de otimização situacional (SOFA). Segundo, comparar essa implementação Bayesiana do SOFA com outros três modelos baseados em Máxima Verossimilhança, usados para estimar escores verdadeiros. O terceiro objetivo foi mostrar como a modelagem conjunta pode ser usada para pesquisas de validade. Uma das principais vantagens do *framework* SOFA em comparação com a abordagem psicométrica tradicional é que o SOFA depende de dados experimentais, melhorando a validade das medidas. O quarto e último estudo foi derivado do pressuposto de construto e seu principal objetivo era desenvolver um procedimento de aprendizado de estrutura de gráficos de cadeia de potência (PCGs). Um PCG é um tipo de gráfico que representa relações causais entre grupos de variáveis. Pode ser pensado como uma versão exploratória completa da modelagem de equações estruturais, bem como um modelo psicométrico que não depende de variáveis latentes. Esses quatro estudos pretendem mostrar que a modelagem psicométrica não deve se restringir ao uso de modelos tradicionais de mensuração, mas também deve considerar a adaptação desses modelos tradicionais de acordo com o uso pretendido e os processos teóricos que originam as medidas observadas.

Palavras-chave: psicometria; modelagem quantitativa; teorização formal; modelagem Bayesiana; teoria de medida.

PRESENTATION

What does it mean to measure something? In this dissertation central aspects related to traditional psychometric practices, such as Factor Analysis and Item Response Theory are discussed and alternatives for these traditional practices are proposed. In this context, it is necessary to understand first which changes can be realized in psychometrics to differentiate it substantially from what is usually already done by researchers and by people who depend on psychometric tools for their work. Therefore, the overall aim of this dissertation is to present the assumptions of contemporary psychometrics and to show how models derived from these assumptions can be modified in order to develop meaningful measurement in psychology (Sijtsma, 2012).

It is necessary to emphasize that the use of psychological measurement tools developed in this dissertation is not intended to be the default practice in psychometrics. On the contrary, the new developed models are intended to inspire other psychometricians and researchers to seek new tools that may better suit their specific contexts. However, whether such tools are appropriate to the context depends, obviously, on empirical evidence of adequacy. Although psychometrics forms the foundation for a large amount of psychological studies, especially those using tests and scales, is considered as one of the areas that spread various misconceptions (Flake & Fried, 2019). This is not necessarily due to lack of ethical principles, but principally related to the complexity of the topics covered in the psychology area as a whole and the small number of psychometric models used in these studies.

In the psychometric literature, the difficulty and mathematical complexity of theoretical models are sometimes presented as the main factors of misuses of psychometric techniques (Borsboom, 2006). On the other hand, it is the responsibility of the writer to seek clearer ways of conveying the proposed message, as it is critical that the targeted audience understands the message (Silvia, 2007). Therefore, despite consisting mostly of

methodological papers, a more accessible writing approach was used in all four studies of this dissertation.

The dissertation consists of four manuscripts, one of them being theoretical and the other three empirical/methodological. The first manuscript is a theoretical study in which theoretical issues inherent to psychometrics and the concept of measurement are discussed. The objectives of this first manuscript were: (1) to present and discuss three basic assumptions in psychometric literature, and (2) develop new measurement models in psychology by adapting these assumptions. In the second manuscript, the Conditional Item Response Model is proposed, along with a Bayesian implementation of optimal scores (Ramsay & Wiberg, 2017), as alternatives for the traditional Rasch model. In the third manuscript, an analytical and methodological framework for measuring dispositions with experimental data, named situational optimization function analysis, is presented and tested with simulated and empirical data. In the fourth manuscript, an extension of power graphs (Royer, Reimann, Andreopoulos, & Schroeder, 2008), which we called power chain graphs, is presented as an alternative to structural equation modeling, and other psychometric models, when causal relations between groups of variables are to be estimated. The final considerations section outlines research agendas based on the three empirical studies, as well as the overall contribution of the present dissertation.

References

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3), 425-470.

Flake, J. K., & Fried, E. I. (2019). *Measurement schmeasurement: Questionable measurement practices and how to avoid them*. Retrieved from https://doi.org/10.31234/osf.io/hs7wm

Ramsay, J. O., & Wiberg, M. (2017). A strategy for replacing sum scoring. *Journal of Educational and Behavioral Statistics, 42*(3), 282-307.

Royer, L., Reimann, M., Andreopoulos, B., & Schroeder, M. (2008). Unraveling protein networks with power graph analysis. *PLoS Computational Biology, 4*(7), 1-17.

Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory & Psychology, 22*(6), 786-809.

Silvia, P. J. (2007). *How to write a lot: A practical guide to productive academic writing*. New York: American Psychological Association.

I

How to Think Straight About Psychometrics:

Measurement Theories and Practice in Psychology

Vithor Rosa Franco[1], Jacob Arie Laros[1] and Marie Wiberg[2]

Affiliations

[1]Post-graduate program of Social, Work and Organizational Psychology, Institute of Psychology, University of Brasília, Brasília, Brazil;

[2]Department of Statistics, Umeå School of Business, Economics and Statistics, Umeå University, Umeå, Sweden;

Abstract

The aim of the current study is to present three assumptions common to psychometric theory and psychometric practice, and to show how alternatives to traditional psychometrical approaches can be used to improve psychological measurement. These alternatives are developed by adapting each of these three assumptions. The assumption of structural validity relates to the implementation of mathematical models. The process assumption which is underlying process generates the observed data. The construct assumption implies that the observed data on its own do not constitute a measurement, but the latent variable that originates the observed data. Nonparametric item response modeling and cognitive psychometric modeling are presented as alternatives for relaxing the first two assumptions, respectively. Network psychometrics and measurement theory are alternatives for relaxing the third assumption. Final remarks sum up the most important conclusions of the study.

Keywords: Psychological measurement; item response theory; measurement theory; network psychometrics.

How to Think Straight About Psychometrics:

Measurement Theories and Practice in Psychology

Is it possible to measure psychological entities? This question, albeit less troublesome for most current psychology researchers (Borsboom, 2005; Stanovich, 2012) were a main concern for scientists in the beginning of the XXth century. Campbell (1928) and others argued bluntly that psychological entities cannot be properly concatenated. Therefore, measurement in psychology must be impossible and scientific psychology as well. This, of course, was not well received by most psychologists at the time (Hull, 1943). One of the most influential theories from this period was Stevens' (1946) operational view on measurement. This theory popularized the measurement levels, which allowed psychologists to define their variables as a different type of measures than those from hard sciences.

An older field, known as psychometrics, was being developed since the beginning of the same century (Jones & Thissen, 2006). From the classical test theory to the item response theory, several models to measure psychological constructs were developed (van der Linden & Hambleton, 2013), allowing for measurement instruments to be constructed as well (Furr, 2011). This development was not without controversy. Trendler (2009), for instance, says that measurement in psychology, as defined by psychometric theory, is not scientific. Michell (1997) agrees in some degree, stating that psychometrical methods do not allow for true quantitative measures to be attained. A more balanced view is sustained by Sijtsma (2012), who affirms that the two measurement approaches proposed to psychology—the statistical (i.e., the psychometrical approach) and the physical (i.e., measurement theory)—can be useful. However, they are usually not as useful as they could be, as they disregard meaningful psychological theory.

The aim of the current study is to present three assumptions common to psychometric theory and practice. Focusing mainly on the statistical approach to measurement, we also

present how alternatives to traditional psychometrical approaches can be used to improve measurement in psychology. The rest of this paper is structured as follows. In the next section, we explain how quantitative and qualitative reasoning impacts theorizing in psychology, originating the latent framework in psychometrics. We then present the latent framework as the basis for the three most popular theories in psychometrics and list the three assumptions regarding these theories. The next three sections discuss each of the three assumptions, presenting how research in psychological measurement can better explore each of these assumptions. The paper ends with a number of concluding remarks.

Quantitative and qualitative thinking in psychology

To assure a scientific status, researchers in the field of psychology have preferred to use quantitative practices for data analysis (Mertens, 2014). This happened because, in the beginning of the XXth century, to be considered a science, any field of study should rely on mathematics and formal logic (Price, 1986). Nevertheless, the theorization in psychology is still, and increasingly (Myung & Pitt, 2001; Towsend, 2008), done on basis of natural language, meaning that relations between variables are not objectively defined. This, on itself, is not a problem, given that qualitative thinking can be beneficial for science. Nevertheless, methods and theorization should suit the research question, not the other way around.

In methodological textbooks for undergraduate and graduate students (e.g., Shaughnessy, Zechmeister, & Zechmeister, 2014), much is said about how methods and data analysis should properly be selected to answer each type of research question. For instance, Kish (2004) proposes that every research can be of one—or a combination of—design category: realistic; representative; and randomized. Realistic research designs are those centered on being profound about a singular subject, usually using qualitative research

techniques and are meant, mainly, to be of a descriptive nature. Representative research designs should be used when one wants to know if a characteristic is generalizable to a population, as in survey research. Finally, randomization designs are, basically, experiments: “randomization” is used to express the random group assignment and are defined as the type of design that should be used to infer causal relations.

Despite all the different types of methodological designs that exist, they only help to answer an already posed research question. The problem for the development of psychological science is, therefore, not only dependent on the research design, but also on how the research question was posed (Shaughnessy et al, 2014). This is a considerably less discussed topic in scientific psychological literature. Most textbooks and tutorial papers will focus mainly on where research questions come from, rather than the procedures used to derive them (e.g., Sandberg & Alvesson, 2011). For instance, Shaughnessy et al (2014) suggests there are two important sources for scientific theorization: past research or our personal experiences. Provided that both are further and critically evaluated, using a proper method, they are valid sources for theorizing. However, this does not answer the question of what theorizing is and how to properly do it. For instance, given a prior scientific result, how does one create new hypotheses or proposes modifications to a given theory?

Theorizing is certainly not an effortless endeavor (Thabane, Thomas, Ye, & Paul, 2009), and also defining what is proper theorizing is not a straight forward effort. Nevertheless, some authors propose some alternatives. As in many things in science, there is no unique way for theorizing, but it can be categorized in, at least, two types, depending on the amount of formalization used to describe the phenomena of interest (Myung & Pitt, 2001). Formalization is used here to define the use of mathematical, logical or any objective language (Shoenfield, 2018) in contrast to the natural language, such as English, Portuguese, Swedish, and many others (Manning, Manning, & Schütze, 1999). Therefore, the first type of

theorizing, which is also the most common in psychology and other sciences (Townsend, 2008), is the natural language theorization. As the name suggests, this type of theorization is done by simply stating, in natural language, what, how and why empirical data is how it is. It also involves a lot of rationalization over past empirical results. A simple example can be given by the theory of cognitive dissonance, classically defined as the mental discomfort experienced by a person who simultaneously holds two or more contradictory beliefs, ideas, or values (Festinger, 1962). This definition, despite being clear and meaningful for most individuals who understand the English language, does not explain, for instance, how mental discomfort is caused by holding contradictory beliefs. Most psychological theories and hypothesis have this format. They vaguely state some expected relation between variables, without acknowledging the process that originates this expected relation.

Even theories and hypothesis that are more preoccupied with the process and with more complex relations between variables do it by using natural language, meaning they will, for the nature of natural languages, lack precision. For instance, the multicomponent model of working memory by Baddeley and Hitch (1974) states that three components are necessary for working memory: the central executive, the phonological loop, and the visuospatial sketchpad with the central executive functioning. For the current presentation, it is not necessary to describe these components. It is necessary only to know that the authors stated that all of the components are necessary to prevent decay of relevant memory information (Baddeley & Hitch, 1974). However, they do not explain how the decaying process works. When does it begin? Is the decaying rate constant or variable? Does the decaying process have some limit of "data exclusion" or the information can be completely lost? For illustrative purposes, Figure 1 shows four possible decaying rate models that could all be true, given the definition used by the authors.

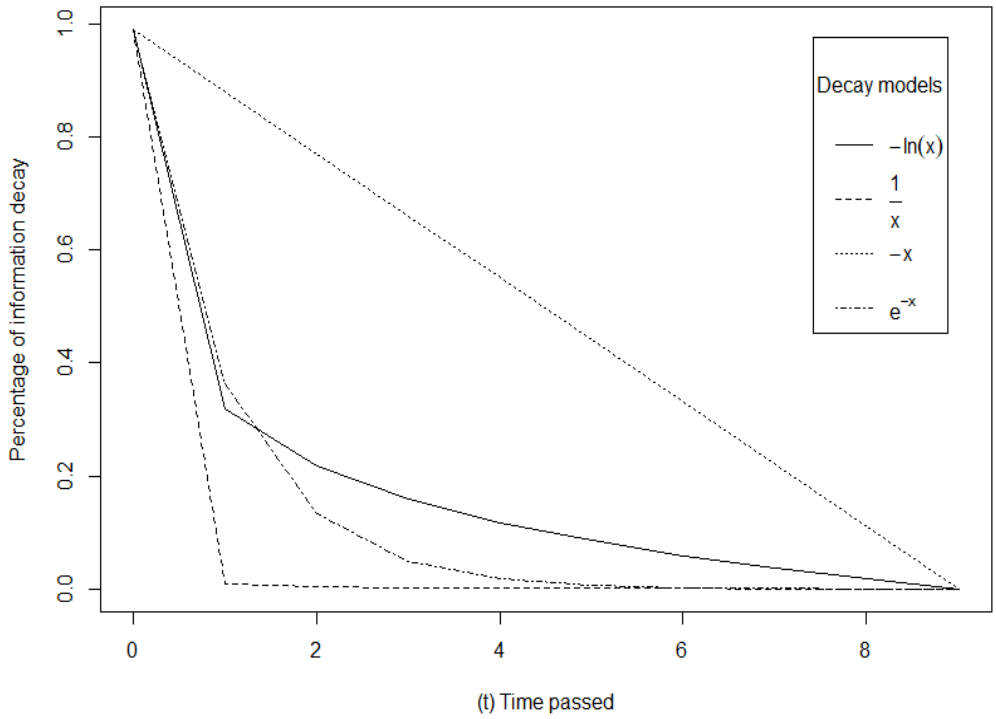


Figure 1. *Four possible models for decaying rate in memory tasks*.

Are we trying to imply that natural language does not have its place in scientific theorization? Certainly not. The true message here is that when describing quantities and patterns, a more appropriate symbolical tool should be used; which is mathematics. This has been the practice of what is known as mathematical psychology (Coombs, Dawes, & Tversky, 1970): an approach to psychological research that is based on mathematical modeling and on the establishment of psychological rules of quantifiable psychological processes. Again, psychology researchers seem to prefer to use quantitative practices for data analysis, but quantitative reasoning and quantitative theorizing seems still to be lacking (Towsend, 2008). One example of use of quantitative theorizing is the latent variable theory on psychometrics (McDonald, 2013).

Psychometrics and its three assumptions

Psychometrics is a field in psychology concerned with the theory and technique of psychological measurement (McDonald, 2013). Three major theories were developed to

explain the relations between observed data and psychological traits: classical test theory (CTT); common factor theory (CFT); and item response theory (IRT). Despite the fact that many authors defend, for instance, that IRT is superior to CTT (Borsboom, 2005; Hays, Morales, & Reise, 2000; Reise, Ainsworth, & Haviland, 2005), all these psychometric theories can be understood as different applications of the same general Latent Variable Theory (LVT; McDonald, 2013). LVT holds that psychological variables (or constructs) are explanatory variables which are not directly observable, but inferred from their effects on human behavior.

The following equations, that can respectively be used to express CTT, CFT, and IRT, clarify why all these theories are related to LVT:

$$X = T + \varepsilon, \tag{1}$$

$$X = \lambda T + \varepsilon, \tag{2}$$

$$g(X) = f(T), \tag{3}$$

where X stands for the observables, T for the true, latent, score, λ for the factor loadings (or factor weights) and ε for the random error. All equations can be understood as an extension of the previous, with the same monotonic relation between observed and latent variables. The functions $g(X)$ and $f(T)$ are usually represented, respectively, by probability mass functions (e.g., binomial or categorical distributions) and link functions (e.g., logistic function or cumulative normal function) to properly scale the observed and latent variables.

It is possible to perceive that all functions generate similar inferences about relations of true and observed scores. The differences in the results rely, mostly, on the methods used to estimate the parameters of these equations (McDonald, 2013), meaning that, from a more computational perspective, these methods are rather different. The true score in Equation (1) is normally estimated using sum or average scores, Equation (2) is tested by means of confirmatory or exploratory factor analysis (Thompson, 2004), and Equation (3) will be

tested by some item response model (IRM; van der Linden & Hambleton, 2013). All these methods harness some pragmatic assumptions, being some of them testable (e.g., unidimensionality assumption; Stout, 1987), but some others are untestable (Michell, 2000). For instance, some variation of the logistic function is generally assumed as the item response function (IRF) for $f(T)$ in Equation (3). Regardless of being usually taken as an obvious assumption, due to a traditional psychophysical empirical finding on tone and loudness perception (Fechner, 1860), there is no direct test for the validity of this assumption. This is to say that, given that the true score is latent and not directly experimentally controllable, it is impossible to test if the logistic function is really the function that relates latent to observed variables (Levine, 2003). However, it is possible to test if the observed IRF is monotonic in relation to the latent IRF (Junker & Sijtsma, 2000).

At this point it should be noted that, despite harsh, this criticism does not necessarily invalidate the general LVT approach for psychometrics. This general psychometric approach can be defined as a statistical approach to measurement (Sijtsma, 2012). A statistical approach is characterized by accepting the assumptions of some statistical or mathematical measurement model, using them to establish "quality" thresholds on data. For instance, when using Factor Analysis, items with low factor loadings are usually suggested to be discarded (Thompson, 2004). On the other hand, factor analysis is based upon linear regression, meaning that, if the true process is quadratic or relies on any other non-monotonic function, the estimates of factor loadings are probably biased (McDonald, 1965). For problems like that, a second approach for dealing with psychological data could be the physical approach (Sijtsma, 2012). The physical approach consists of testing if a particular mathematical structure is true for the data. For instance, utility theory uses a number of axioms to define rational behavior (von Neumann & Morgenstern, 1944). Nevertheless, people not always behave rationally, meaning these models will not always properly fit the data (Allais, 1953).

The usual interpretation is that people are not rational. Kahneman and Tversky (1979), on the other hand, proposed the prospect theory, changing the assumptions of the utility theory, with a consequence to changing the interpretation on humans' decision-making behavior.

The statistical and physical approaches to measurement provide researchers with two different mindsets, respectively: rejecting the data if the model has a poor fit; or rejecting the model if it has a poor fit to data. None is, for itself, the best approach for proper inferences. Sometimes it is better to reject the data after a low fit of the model, given that there may be some bias on the data collecting process (Shaughnessy et al, 2014). On the other hand, if data are properly collected and the model used for testing it systematically shows a pattern of bad fit, maybe an alternative model should be tested. Nevertheless, most researchers are not even aware of the existence of the physical approach to measurement (Michell, 2017), as it demands more knowledge on mathematics and experimental design. Both of these requirements make research and developing quality measurements more difficult for several areas in psychology.

Three major critiques can then be elaborated about the traditional psychometric practice. Each critique is related to one of three assumptions regarding psychological measurement, as we derived from the LVT and the statistical approach to measurement. First, psychological measurement is based on using pre-conceived models that, sometimes, are non-testable and have higher priority than the empirical data. We call this the structural validity assumption. Second, psychological measurement is based on models that, sometimes, do not mirror the psychological phenomena or processes they are intended to represent. We call this the process assumption. Finally, traditional psychological measurement depends heavily on constructs, which are not observable and can be sometimes difficult to define; therefore, difficult to give a proper operationalization. We call this the construct assumption. These assumptions do not need to have a specific hierarchy of complexity or necessity.

As with several other problems within science, relaxation or thorough testing of the assumptions can help science to improve (Kanazawa, 1998). Nonparametric Item Response Modeling (NIRM; Sijtsma & Molenaar, 2002) can be used to test or to relax the structural validity assumption, but it still relies on the process and construct assumptions. Cognitive Psychometric Modeling (Embretson, 2010) can be used to test or relax both structural validity and process assumptions, but is still relying on the construct assumption. Finally, network modeling (Epskamp & Fried, 2018), based on the statistical approach, and measurement theory (Roberts, 1979), based on the physical approach, can be used to relax all three assumptions.

Structural validity assumption and Nonparametric Item Response Modeling

The structural validity assumption is the exact mathematical implementation of psychometric models. For instance, traditional factor analysis assumes latent variables to be linearly related to observed variables (Thompson, 2004). The error of measurement has an expectation of zero and, for adequate fitting, a normal distribution is generally used for modeling the error. The structural validity assumptions can be thought of as being the least related to a particular psychological theory, but the most related with the statistical, mathematical, or computational feasibility of the implementation of a model (e.g., Griffith & Akio, 1995). A good historical example of changing a structural validity assumption occurred in the case of the transition between initial IRMs and the logistic and Rasch models. Initially, the normal cumulative density function was used as the IRF for the binary IRMs (Lord, 1953). Nevertheless, at the time of the development of the first IRMs, computing this IRF was computationally extensive. For this reason, some authors proposed changing from the normal to the logistic cumulative density function as the IRF for IRMs (Birnbaum, 1968). Mathematically, and for

modern computers, this difference makes little to no difference, but at the time it was necessary so using IRMs was feasible (Rasch, 1960).

Apart from changing mathematical characteristic of the models, most of the structural assumptions in IRMs can also be relaxed using NIRMs (Sijtsma & Molenaar, 2002). NIRM is not only a different class of item response models, but also a whole different approach to modeling response patterns. For both NIRM and Parametric IRMs, there are three main assumptions about the relations between the observed scores and the latent trait. All these assumptions are specificities of our structural validity assumption. The first is that of unidimensionality, which simply means that the observed scores have only a single latent cause (Stout, 1987)—or a single more relevant cause.

While there is much theoretical support for multidimensionality in psychological measurements (Knol & Berger, 1991; Reckase, 2009), given the complexity of psychological phenomena, many authors defend that unidimensional measurements should be preferred (Nunnally, 1978; Sijtsma & Molenaar, 2002). The main argument standing the latter can be clearer stated with an analogy. If you would use a scale to measure your weight and your height at the same time, what a score of 104 would mean? Supposing there was no standard unit for both measures, this score would be meaningless to making conclusions about those magnitudes apart. Therefore, multidimensional scores, despite being probably more representative of psychological phenomena, should be avoided so meaningful measures can be achieved (Heene, Kyngdon, & Sckopke, 2016).

The second assumption is that of local independence, which states that the observed score of individual *i* on item *k* does not depend on the response he gave in any other *j* item, conditioned on the latent trait (Zhang & Stout, 1999). Finally, the third assumption is monotonicity, which states that the probability of getting an item right (or of endorsing a higher score in a Likert scale) augments with increasing latent trait (Junker & Sijtsma, 2000).

Another assumption, not common to all IRMs, however, is that of nonintersecting item response functions (IRFs; Rosenbaum, 1987). This assumption is used, for instance, by the Rasch Model and by the One-Parameter Logistic Model (1PLM; van der Linden & Hambleton, 2013). For ordering items by their difficulties' estimates, these assumptions need to hold, simply because when IRFs intersect there is an interaction effect between individuals' levels of the latent trait and items difficulties (Sijtsma & Molenaar, 2002). This means that the ordering of item difficulty is not the same for all individuals, but depends on their latent traits. Obviously, this is not a desirable property when you want to create a standard test.

NIRM begins to differ from parametric IRM when the operationalization of these assumptions takes place. For instance, a common monotonic function relating the latent trait with the observed scores is the logistic function:

$$\Pr(X = x|\theta, \delta) = \frac{e^{\theta-\delta}}{1 + e^{\theta-\delta}} \quad (5)$$

where θ stands for the level of the individual's latent trait and δ for the level of item's difficulty. All the lines depicted in Figure 2 are representations of this function, with different values for the subtraction $\theta - \delta$.

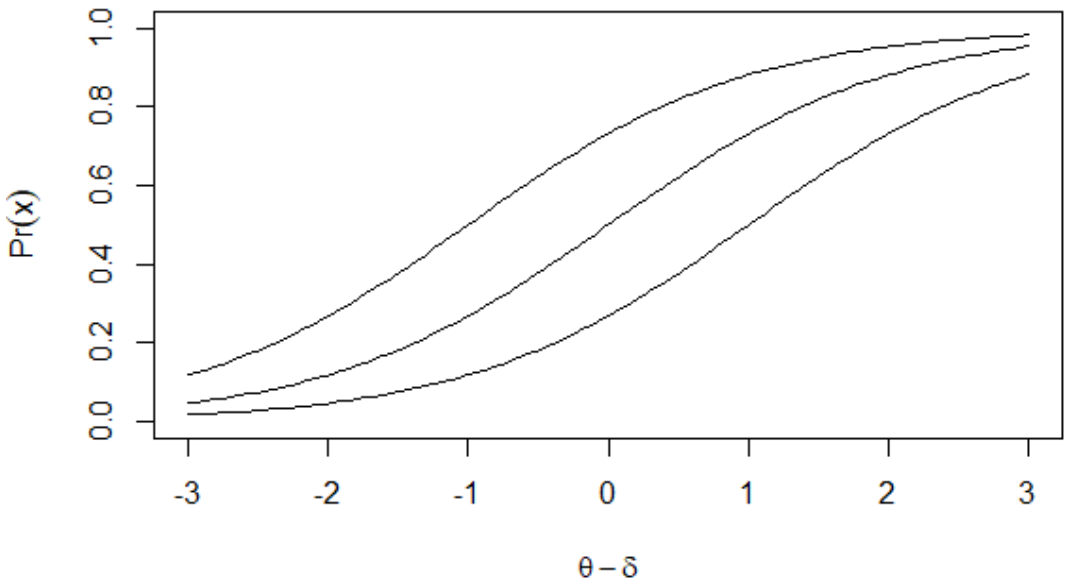


Figure 2. *Depictions of logistic functions*.

It is possible to see that the IRFs will always be “S” shaped. For NIRMs, however, any function can be used, since it does not disregard the monotonicity. The general formulation of IRFs for NIRMs is the following

$$P_i(\theta_a) \leq P_i(\theta_b) \quad (6)$$

which implies that, provided that the function is a nondecreasing function of θ, any function can be used to relate the latent trait with the observed response. Figure 3 depicts five different functions, all which can be used as IRFs in a NIRM perspective.

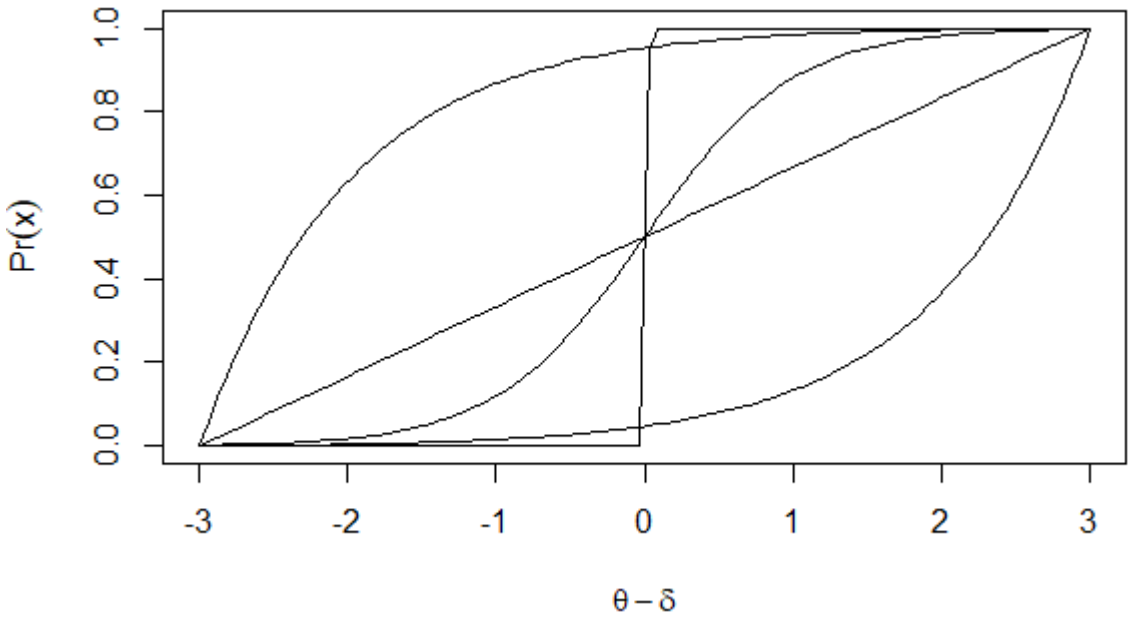


Figure 3. *Depictions of valid functions IRFs in a NIRM perspective*.

As there are, in principle, an infinite number of functions that can conform to the less restrictive forms of NIRMs, it would be computationally extensive (or impossible) to test, for instance, exactly what function the IRF follows (Ferraty & Vieu, 2006). Therefore, in a NIRM perspective, instead of making complex estimates of latent variables, the fit of the model is given by the capacity of the data to follow the relaxed assumptions. This means that, for example, the data could give a result like that in Figure 3, when using NIRM, but, if fitted with a 1PLM, could result in what was shown in Figure 2. This would probably happen because while NIRMs will test the assumptions it makes, parametric IRMs forces data into its functional form.

The most well know models of NIRMs are those from Mokken Scale Analysis (MSA; Mokken, 1971). A Mokken scale is a non-parametric, probabilistic version of the Guttman scale (Mokken & Lewis, 1982). Both Guttman and Mokken scaling assumes that items have a hierarchical order, meaning that respondents who answered a difficult question correctly should also answer an easy question correctly. The main difference between Guttman and Mokken scales is that Guttman scaling assumes that respondents who answered a difficult question correctly will necessarily answer an easier question correctly. When that does not happen, it is said to be a Guttman error (Meijer, 1994). Nevertheless, in real evaluation scenarios usually people do not respond deterministically, but accordingly with some stochastic process, better modeled with MSA or a parametric IRM.

It is important to note that there are many other non-parametric and semi-parametric models that can be found on the literature. Semi-parametric models are the ones that alleviate just some, not all, of the assumptions made by parametric models (Dey, Ghosh, & Mallick, 2000). Many Bayesian models of such kind have been created (e.g., Miyazaki & Hoshino, 2009; Wang, Chang, & Douglas, 2013). There are also tests and models for inferential analysis that are centered on testing the common assumptions for parametric item response models (e.g., Stout, 1987; Straat, van der Ark, & Sijtsma, 2013). All these different techniques make it possible to test and model a larger range of items, without being needed to limit test and scale construction to what is only permitted by traditional IRMs.

Nevertheless, all those different models share an important limitation, as stated by Sijtsma (2012). Using those techniques, the measurement is dependent on the questionnaire—or test, or scale—data, meaning that they all assume that the observed scores are caused by a generic latent variable related to the items and to the respondents. This structural validity assumption makes it easier to mistake the prescriptive structure of a statistical measurement model with the theoretical structure of the attribute of interest.

Nevertheless, meaningful measurement is possible only if enough is known about the attribute so as to justify its logical operationalization into prescriptions from which a measurement instrument can be developed (Michell, 2017; Sijtsma, 2012; Trendler, 2009). Despite theories about attributes in psychology often not being precise enough to justify a logical operationalization, the emerging field of cognitive psychometric modeling has been presented as an interesting alternative (Embretson, 2010).

Process assumption and Cognitive Psychometric Modeling

The process assumption is the definition on how basic constructs, or latent variables, relate to each other to compose a particular psychological model or theory. For instance, the IRT assumes that the probability of correct answering a question is a function of the respondent's latent trait and the difficulty of the items. The signal detection theory, on the other hand, assumes that the probability of correct answering a question is a function of the criterion and discrimination of the respondent (Stanislaw & Todorov, 1999). In this case, it is possible to distinguish the measurement made by these theories as, for the signal detection theory, it makes a difference if correct responses were a hit or a correct rejection, and if the incorrect responses were a false alarm or a miss. For IRT, it usually matters only if the response was correct or incorrect. This entails in the fact that both theories make different assumptions on the underlying process controlling response patterns on the proposed quantitative model.

A quantitative model is a representation of a phenomenon using techniques and procedures due to mathematics and statistics (Edwards & Hamson, 2007). Therefore, a cognitive model is a representation of cognitive phenomena using the same class of mathematical and statistical techniques and procedures (Lee & Wagenmakers, 2014). Lewandowsky and Farrell (2010) describe three different classes of quantitative models. The first class contains models of data description. As the name suggests, they only describe

relations between variables. They are explicitly devoid of psychological content, although the modeled function constrains possible psychological mechanism to the phenomena. An example is linear regression models (Faraway, 2016). The second class is the one of process characterization. These models postulate and measure distinct cognitive components. Yet, they are neutral about how specific instantiations underpinning the cognitive components work. An example is the multinomial processing trees model (Erdfelder et al, 2009). Finally, the third class is the one with models of process explanation. Like characterization models, their advantage stands on hypothetical cognitive constructs. However, they provide detailed explanation about those constructs and how are they related. An example is the generalized context model (Nosofsky, 1986).

Traditional psychometric models can be thought as descriptive models, given that they can be described themselves as only linear or generalized linear regressions with latent predictors (Bock, 1997). Considering psychometric models from this perspective enables to perceive that, despite being important tools to psychological research, traditional IRMs lack explanatory meaning. Therefore, several aspects of psychological phenomena are not taken into account. Sijtsma (2012) states that IRT leaves psychology out of the equation when proposing psychological models of measurement, resulting in fruitless insights for psychological phenomena. Nevertheless, it is important to note that it is not the use of a statistical framework based on latent variables that is the strongest limitation of IRT. The strongest limitation is to use exclusively descriptive models (i.e., a strong process validity assumption) for developing measurement models, which have no concern for the processes that generated the observed data structure.

The cognitive modeling approach can then be used to enrich IRMs and give more significance to the measurement process in psychology. Despite not formally defined as such, the Cognitive Psychometric Models (CPMs) approach has been used as a good alternative for

traditional IRMs (Embretson, 2010). One prominent CPM is Tree Based IRMs (TIRMs; LaHuis, Blackmore, Bryant-Lees, & Delgado, 2018). This kind of model helps to understand in which order latent variables influence each other to cause the observed response patterns. Empirical comparison of different TIRMs can provide proper evidences of validity for a measure, when compared, for instance, with simply correlating expected–to–be–related measures (Borsboom, Mellenbergh, & van Heerden, 2004). Figure 4 illustrates two hypothetical competing TIRMs for the measurement of personality data. Both models state that people have different propensities to act aggressively or peacefully. However, the model on the left states that the probability of a respondent giving an aggressive or a peaceful response depends only on the parameter α, which can be thought as his propensity to act in a given way (e.g., his disposition). On the other hand, the model on the right decomposes the process, stating that self-control, measured by the parameter β, has an important role on regulating individual's actions. Traditional analysis in psychometrics and psychology would only test correlations between these measures (e.g., Kim, Namkoong, Ku, & Kim, 2008), providing only descriptive relations for the constructs. The use of the TIRMS makes it possible to conclude what process is more likely to have originated the data at hand. In the example, individuals with more peaceful personality do not need to have self-control, given that they simply act as is socially expected. This kind of conclusion would be difficult, or even impossible, to be drawn by traditional psychometrical analysis.

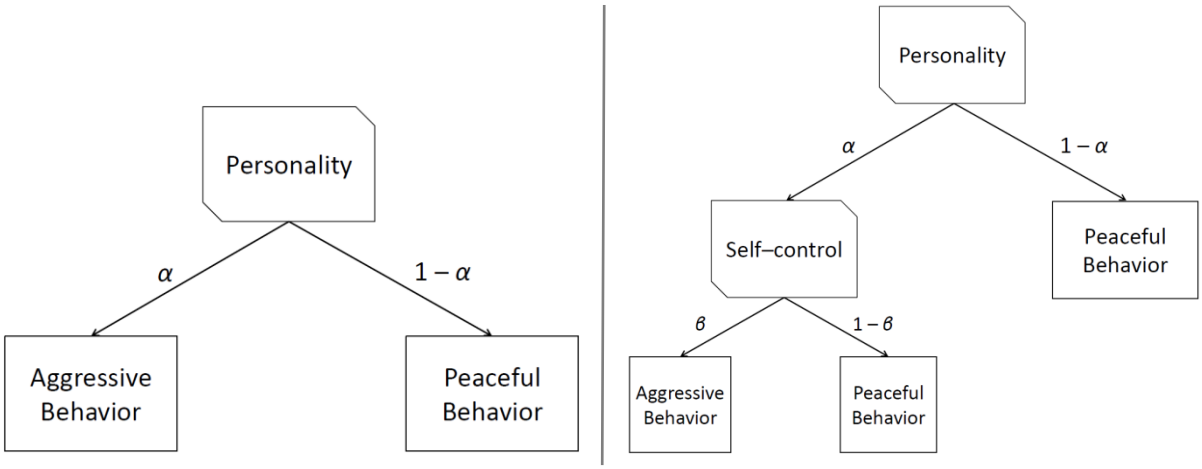


Figure 4. *Two hypothetical competing TIRMs for the measurement of personality data related to observed behavior*.

Following what is proposed by Sijtsma (2012), a measurement occurs only when the relations between the observed measure and the attribute of interest are logically and formally established, making it possible to operationalize the phenomena of interest. Characterization and explanation's CPMs can be used for this end, as illustrated by the example of Figure 4. Certainly, on the other hand, for the proper use of CPMs one cannot only rely on questionnaire or test data. This can be a difficult task. Kagan (2005) criticized the common psychometric practice on relying almost exclusively on the semantic structures activated when participants give their responses to questionnaires. The author also proposes the use of alternative data sources, such as motor activity, distress to unfamiliar visual, auditory, and olfactory stimuli, and others. Sijtsma (2012) states that these alternative data types are rarely used in psychometrical research. Cognitive psychology researchers who use cognitive modeling approaches, on the other hand, are more used with searching for different data sources for their models (Lee, 2011).

One way of increasing the number of data sources and, therefore, making a more robust and valid measurement model, is collecting data in experimental sets (Sijtsma, 2012). This is a good approach for two main nested reasons. First, because using experimental sets reduce the noise and random variance in the data (Kish, 2004). As many sources of confounding

effects are controlled in experiments, each new relevant measurement reduces the amount of unexplained variance. The repeated measurement design is the design that reduces the noise the most (Cook, Campbell, & Shadish, 2002). The second reason, which is nested in the first, is because experimental sets in psychology usually control for the external influences in individual's behavior. Therefore, the variance that is left can be due only to noise or individual differences (Bacon, 2004).

Using CPMs and experimental sets are still a novel approach to psychological measurement, which has the potential to approximate the meaning of measurement in psychology with that of physics (Sijtsma, 2012). Nevertheless, for some authors (e.g., Gould, 1996), measurement in psychology will always be impossible, given that IRMs, and even CPMs, define measurement as the estimates for some latent variable. This means that, for some authors, it is not enough to relax or test the structural and process validity assumptions, but is also necessary to directly observe the measured property or feature. Despite this being problematic for psychology as a whole (Borsboom, Mellenbergh, & van Heerden, 2003; Trendler, 2013), measurement in fact does not need to be defined in terms of latent variables. Network psychometric modelling and the realist measurement theory (Michell, 2005) can be used to this end.

Construct assumption, network modeling, and realist measurement theory

The construct assumption is the definition of a measure as the latent variable that explains the variance of observed variables. All models presented so far rely on this assumption, as the estimates of the magnitude of the latent variables are of central interest, and not the observed variables per se (Borsboom, 2005). This is an old trend in mainstream psychology to attribute mentalist causes to human behavior (Stanovich, 2012). Despite the success of this approach in many areas in psychology (Sijtsma, 2012), philosophical and mathematical critics are

made to accept something that cannot be assessed or, if ever, only indirectly assessed (Michell, 1990; 1997; 2005; 2008). Two work-arounds from this assumption are found in both the recent psychometrical literature, in the form of network psychometric modeling (Constantini et al, 2019), as well as in the traditional measurement theory literature (Roberts, 1979), largely overlooked in the current psychometric field (Michell, 2000).

Probabilistic graphical modeling (Lauritzen, 1996) is a statistical approach, derived from the mathematical graph theory, used to model multivariate conditional dependencies between variables. In this sense, factor analysis and item response models can be considered as special cases of probabilistic graphical models, where dependencies between observed variables are conditioned on latent variables (Kruis & Maris, 2016). For the probabilistic graphical models proposed by Lauritzen (1996), however, no latent variables are considered. Instead, dependencies between any two variables are explained by their relations to a third variable. One of such models is the partial correlation graphical model, also known as partial correlation network model (Epskamp & Fried, 2018). This model has been of growing interest in the field of psychometric construct analysis, such as mental health related ones (Borsboom, 2017), where the existence of a common latent cause is a controversial issue.

The probabilistic graphical modelling approach to psychometrics, also named network psychometrics (Epskamp, Rhemtulla, & Borsboom, 2017), is considered as a part of the statistical approach to measurement because it is not particularly interested in the measurement level of the observed variables. For instance, partial correlations can be calculated from polychoric correlations, combined with regularized regressions, if the observed variables are ordinal (Golino & Epskamp, 2017). It is generally not the objective to estimate interval or ratio measures to predict observed variables, nor are the values of the observed variables transformed to interval or ratio measures. However, some network psychometric models, such as the network Ising model, have been shown to be equivalent to

traditional latent common cause models, such as factor analytical models and item response models (Marsman et al, 2018). This means that not always network psychometrics will present a true alternative to traditional psychometrics, only when it avoids latent variables (Kruis & Maris, 2016; Lauritzen, 1996).

One good example of using network psychometrics, and abandoning latent variables to measure psychological entities, is the model of general intelligence proposed by van der Maas et al (2006), known as the mutualism model of intelligence. Traditionally, the study of human intelligence was concomitantly developed with the psychometric factor analytic model (Buckhalt, 2002). This is represented by the fact that one of the main discussions in the study of intelligence is not about the existence of a true latent variable, but about how many dimensions describe this assumed latent variable the best (e.g., Golino & Demetriou, 2017). The most traditional model, and maybe one of the best corroborated, is the *g*-factor model of general intelligence (Canivez & Watkins, 2010). In this model, a single general latent variable is used to explain the variance of all observed variables. Some extensions (Canivez, 2016) involve using this *g*-factor as the cause of other latent variables (named specific factors), or with other latent variables explaining residual correlations, after conditioning the *g*-factor out. These are known as the second-order and bifactor models, respectively. The traditional, second-order, and bifactor models of intelligence are respectively illustrated in Figure 5, where *g* represents the *g*-factor, *f*1, *f*i, and *f*n represent the possible specific factors, and *V*1, *V*i, and *V*n represent the possible sets of observed variables.

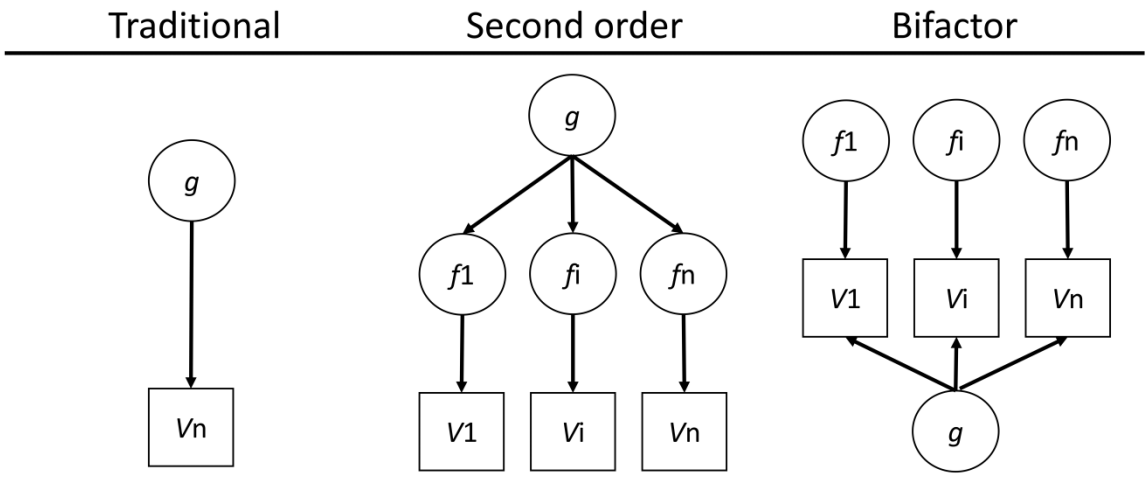


Figure 5. Graphical illustrative example of the traditional (left), second-order (middle), and bifactor (right) models of intelligence.

The network model proposed by van der Maas et al (2006), on the other hand, suggests that the dynamics of intelligence are better described by a network of reciprocal causal effects. The idea of the mutualism model of general intelligence is that such reciprocal causal effects also occur during development, originating the observed correlations, for instance, on responses to items on intelligence tests (van der Maas, Kan, Marsman, & Stevenson, 2017). From an empirical point of view, van der Maas et al (2006) showed that phenomena such as he hierarchical factor structure of intelligence, the low predictability of intelligence from early childhood performance, the integration/differentiation effect, and the increase in heritability of *g* can all be explained by the mutualism model. Despite this fact, the mutualism model has been criticized for not accounting some effects, such as distinguishing between genetic and environmental effects (Nisbett et al, 2012), that are already well studied with traditional psychometrical models. Van Der Maas et al (2017), however, argue and present a new model that can better address most of the criticism of the mutualism model.

Apart from finding conditional dependencies between observed variables, one may intend to test if a given set of observed variables can be measured in a particular

measurement level. To this end, from the physical approach to measurement, there are two steps for creating relevant numerical representations (Scott, 1964). First, one needs to identify the inherent structure of the objects or events, which reveals the property to be measured. Second, one needs to find a method that can properly assign (real) numbers that have an exact correspondence between the property to be measured and the numbers that represents this property. Succeeding in finding the method where the numerical representation correctly represents the measure property means that the researcher has achieved an isomorphic system (Coombs et al, 1970) better known as a "measure".

The first step, which establishes the conditions under which various types of scales can be constructed, is called the measurement theory. The second step, which is the process of assigning numbers to properties, is called scaling. Both are important aspects of measurement and need to be analyzed apart. Nevertheless, the terms measurement and scaling have been used interchangeably in the literature (Coombs et al, 1970), originating what Sijtsma (2012) called "the statistical perspective" of measurement in psychology. Sijtsma (2012) also argued that this practice originated mistakes in the psychometric practice of confusing the prescriptive structure of a statistical measurement model with the theoretical structure of the attribute of interest, therefore, confusing scaling procedures with development of measurement theories.

From this distinction, it is easy to perceive that virtually every scale, test or questionnaire ever created in psychology follows the measurement theories that inspired Spearman (1904), Rasch (1960), and others, which are also similar between them. Most psychometric procedures establish that there are two attributes to measurements (van der Linden & Hambleton, 2013): a generic respondent latent trait and a generic latent characteristic from the test (or item). These components are usually assumed to interact in an additive way. This is certainly a problematic set of assumptions (Trendler, 2009): is it

reasonable to assume that the process that generates observed scores in intelligence tests is the same as the process that generates observed scores in attitude scales?

Some authors propose that this statistical perspective for measurement, despite helping the advancement of science, cannot establish real scientific measures (Michell, 2000; Trendler, 2009). Michell (1990; 1997), instead, proposes that the only way for psychology to perform real scientific measures is to adopt a realist view of measurement. Traditionally, measurement in psychology is based on an operationalist view of measurement (Stevens, 1946): it is accomplished every time a numerical assignment is done in an informed manner. This is to say that one just needs to order objects, or scale magnitudes, following any specific rule and to establish some agreeable observable property as the definition of the phenomena of interest to create a good measurement.

This operationalist view of measurement, despite heavily used in psychology, does not conform to the traditional scientific definition of measurement. Traditionally defined, a measurement is the estimation of the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind (Emerson, 2008). This definition encompasses the concatenation process, which is the operation of joining objects to observe their interactive effect on some measurement scale (Suppes, 1951). Campbell (1928) states that, for a realist view of measurement, units of measurement can be defined only by the concatenation process, which means that scientific measurement in psychology is therefore impossible.

A realist view, nevertheless, can still be applied to psychological measurements. Hölder (1901) proposed a valid mathematical formalization of the concatenation process, providing the beginning of the formal study of measurement. From this work, Luce and Tukey (1964) extended Hölder's (1901) theory for objects or events that are not feasible to be concatenated, proposing the Conjoint Measurement Theory (CMT). For a measurement to be established via CMT, it is necessary to have at least two natural attributes that non-interactively relate to

a third attribute (Krantz, 1964). Via specific relations between the levels of this third attribute, it can be established that all the attributes are continuous quantities, even if initially observed only as ordinal relations.

Despite its importance to guarantee interval or even ratio scales in psychology, the use of CMT has been very limited and virtually absent in psychometric practice and research (Michell, 2017). Probably one of the most famous examples of application of the CMT is the cumulative prospect theory (Kahneman & Tversky, 1979). In the cumulative prospect theory, the utility of a gamble is given by the additive combination of uncertainty-weighted marginal utilities of positive and log-negative outcomes (Wakker & Tversky, 1993). This theory can be illustrated with the reversal of preference effect (Kyngdon, 2013). Suppose a person has to choose between game A, consisting of an 80% chance of winning $4,000, and game B, consisting of a 100% chance of winning $3,000. Then, suppose that the same person has to choose between game C, consisting of a 20% chance of winning $4,000, and game D, consisting of a 25% chance of winning $3,000. It is possible to see that games C and D are simply weighted versions of games A and B, such that C = .25(A) and D = .25(B); $A = 4{,}000 \times .80$ and $B = 3{,}000 \times 1$. Therefore, if participants prefer game B to A, they should prefer game D to C as well. Kahneman and Tversky (1979) found that 80% of their test participants preferred B over A, but 65% preferred C over D. From the prospect theory, this result is explained by the fact that losing have a larger weight on the decision than winning, modeled using a logarithmic function. Decisions are then made choosing the alternative that minimizes loses. In Figure 6 we show a simplified calculation derived from the cumulative prospect theory. Values per row on the "Sum" column are calculated by $(\ln(\text{Loose}) \times \text{Value}) + (\text{Win} \times \text{Value})$. The exception is game B (with value 3,000 and win 1), as it has no probability of losing, the calculation is simply $(\text{Win} \times \text{Value})$. Real modeling using cumulative prospect theory would have an extra step, a transforming function that transforms

the “Sum” values to observed probabilities on choosing each alternative. However, this step is not relevant for understanding the example.

Games A/B			
Value	Loose	Win	Sum
3,000	–	1	**3,000**
4,000	ln(.20)	.80	2,307

Games C/D			
Value	Loose	Win	Sum
3,000	ln(.75)	.25	–113
4,000	ln(.80)	.20	**–92**

Figure 6. Illustrative example on how weighted utilities are calculated from cumulative prospect theory.

Another issue for researchers on psychological measurements is the fact that CMT is not the only measurement theory relevant to psychology. Polynomial Conjoint Measurement (Tversky, 1967) and *n*-component Conjoint Measurement (Krantz, 1968) can be applied when the measurement structure is polynomial and when there are more than three attributes, respectively. Krantz, Luce, Suppes and Tversky (1971), Suppes, Krantz, Luce and Tversky (1989) and Luce, Krantz, Suppes and Tversky (1990) present an exhaustive list of measurement theories that can be used to form ordinal, interval and ratio scales, providing new ways to answer if (and which) IRMs, or any other psychometric model, can really provide quantitative measures of psychological entities.

Discussion

The aim of the current study was to present three assumptions common to psychometric theory and practice. These are the assumptions of structural validity, the process assumption and the construct assumption. We presented the basic idea on measurement model development by adapting each of these assumptions. In a non-exhaustive manner, we also presented how alternatives to traditional psychometrical approaches can be used to improve measurement in psychology. The take home message is similar to that of Sijtsma (2012): only rigorous development of attribute theories can lead to meaningful measurement in psychology. Introducing and stressing the three assumptions in psychometrics not only

allows understanding why measurement in psychometrics is as it is, but also how to change its models so to achieve meaningful measurements.

If latent variables are acceptable and the process is assumed to be known, traditional psychometric models, such as factor analytical and item response models, can be improved by simply changing some of its mathematical functions (structural validity assumptions). We showed as the main examples the change between normal and logistic item response functions to the Rasch model, and the use of nonparametric item response modeling. These procedures change only minor or none of the aspects of the underlying theory about the response or cognitive/behavioral processes. In this case, developing models are mainly related to improving the fit to the data, rather than improving the theoretical description of the underlying psychological process. However, studies on validity can still be used in this approach to improve the understanding of the underlying psychological processes (Cronbach & Meehl, 1955).

If latent variables are acceptable but the underlying psychological process is to be explored, then one should develop models that try to characterize or to explain this process (process assumption). We showed as the main examples the difference between item response theory and signal detection theory, as well as the tree item response models (as a type of characterization models). It is important to note as well that the signal detection theory also provides a descriptive model on respondents' response patterns, similarly to item response theory. Nevertheless, differences on predictions made by both theories allow for designing studies to compare which model makes the most sense, given the experimental apparatus (Trendler, 2009).

Finally, latent variables can be abandoned and measurement in psychology will still be possible (construct assumption). This can be achieved by a statistical approach to measurement, or by means of network psychometrics, as in a physics approach to

measurement, or by means of measurement theory. The main difference between these three approaches is that network psychometrics is similar to other psychometric models: despite the measurement level of the observed variables, the focus is to model the data with multivariate statistics. Measurement theory, on the other hand, focuses on assessing whether a property is quantitative and, if so, what its magnitude is (Roberts, 1979). Depending on the objective, the researcher can use any of these alternatives to propose more meaningful measurements in psychology.

References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école américaine [The rational man's behavior in the face of risk: Critical of postulates and axioms of the American school]. *Econometrica: Journal of the Econometric Society*, 503-546.

Bacon, D. (2004). The contributions of reliability and pretests to effective assessment. *Practical Assessment, Research & Evaluation*, *9*(3), 1-8.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. A. Bower (Ed.), *The psychology of learning and motivation* (pp. 47-90). New York: Academic Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick. (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, PA: Addison-Wesley.

Bock, R. D. (1997). A brief history of item theory response. *Educational Measurement: Issues and Practice*, *16*(4), 21-33.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5-13.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203-219.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061-1071.

Buckhalt, J. A. (2002). A short history of g: Psychometrics' most enduring and controversial construct. *Learning and Individual Differences*, *13*(2), 101-114.

Campbell, N. R. (1928). *An account of the principles of measurement and calculation*. London: Longsman, Green.

Canivez, G. L. (2016). Bifactor modeling in construct validation of multifactored tests: Implications for multidimensionality and test interpretation. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction: Standards and recent advancements* (pp. 247–271). Göttingen, Germany: Hogrefe.

Canivez, G. L., & Watkins, M. W. (2010). Investigation of the factor structure of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS–IV): Exploratory and higher order factor analyses. *Psychological Assessment*, *22*(4), 827-836.

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Costantini, G., Richetin, J., Preti, E., Casini, E., Epskamp, S., & Perugini, M. (2019). Stability and variability of personality networks. A tutorial on recent developments in network psychometrics. *Personality and Individual Differences*, *136*, 68-78.

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, NJ: Prentice-Hall.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Dey, D. K., Ghosh, S. K., & Mallick, B. K. (2000). *Generalized linear models: A Bayesian perspective*. Boca Raton, FL: CRC Press.

Edwards, D., & Hamson, M. (2007). *Guide to mathematical modelling*. New York: Industrial Press.

Embretson, S. E. (2010). *Measuring psychological constructs: Advances in model-based approaches*. New York: American Psychological Association.

Emerson, W. H. (2008). On quantity calculus and units of measurement. *Metrologia, 45*(2), 134-138.

Erdfelder, E., Auer, T. S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Journal of Psychology, 217*(3), 108-124.

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods, 23*(4), 617-634.

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika, 82*(4), 904-927.

Faraway, J. J. (2016). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. Boca Raton, FL: CRC Press.

Fechner, G. T. (1860). *Elemente der psychophysik* [Elements of psychophysics]. Leipzig: Breitkopf & Hartel.

Ferraty, F., & Vieu, P. (2006). *Nonparametric functional data analysis: Theory and practice*. New York: Springer.

Festinger, L. (1962). *A theory of cognitive dissonance*. Palo Alto, CA: Stanford University Press.

Furr, M. (2011). *Scale construction and psychometrics for social and personality psychology*. New York: Sage.

Golino, H. F., & Demetriou, A. (2017). Estimating the dimensionality of intelligence like data using Exploratory Graph Analysis. *Intelligence, 62*, 54-70.

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PLOS One, 12*(6), 1-26.

Gould, S. J. (1996). *The mismeasure of man*. New York, NY: Norton.

Griffith, D. A., & Akio, S. (1995). Trade-offs associated with normalizing constant computational simplifications for estimating spatial statistical models. *Journal of Statistical Computation and Simulation*, *51*(2-4), 165-183.

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care*, *38*(9 Suppl), 1128-1142.

Heene, M., Kyngdon, A., & Sckopke, P. (2016). Detecting violations of unidimensionality by order-restricted inference methods. *Frontiers in Applied Mathematics and Statistics*, *2*, 1-13.

Hölder, O. (1901). Die axiome der quantität und die lehre vom mass. *Reports on the negotiations of the Royal Saxon Society of Sciences to Leipzig Mathematical-Physical Class, 53*, 1–46.

Hull, C. (1943). *Principles of behavior*. New York: Appleton-Century-Crofts.

Jones, L. V., & Thissen, D. (2006). A history and overview of psychometrics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Volume 26* (pp. 1-27). New York: Elsevier.

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, *24*(1), 65-81.

Kagan, J. (2005). A time for specificity. *Journal of Personality Assessment, 85*, 125–127.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 262-292.

Kanazawa, S. (1998). In defense of unrealistic assumptions. *Sociological Theory, 16*(2), 193-204.

Kim, E. J., Namkoong, K., Ku, T., & Kim, S. J. (2008). The relationship between online game addiction and aggression, self-control and narcissistic personality traits. *European Psychiatry*, *23*(3), 212-218.

Kish, L. (2004). *Statistical design for research*. New York: John Wiley & Sons.

Knol, D. L., & Berger, M. P. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research*, *26*(3), 457-477.

Krantz, D. H. (1964). Conjoint measurement: The Luce–Tukey axiomatization and some extensions. *Journal of Mathematical Psychology. 1*(2), 248–277.

Krantz, D. H. (1968). A survey of measurement theory. In G.B. Danzig, & A.F. Veinott, (Eds.), *Mathematics of the decision sciences: Part 2* (pp. 314-250). Providence, Rhode Island: American Mathematical Society.

Krantz, D. H., Luce, R.D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, Vol. I: Additive and polynomial representations*. New York: Academic Press.

Kruis, J., & Maris, G. (2016). Three representations of the Ising model. *Scientific Reports, 6*, 1-11.

Kyngdon, A. (2013). Descriptive theories of behaviour may allow for the scientific measurement of psychological attributes. *Theory & Psychology, 23*(2), 227-250.

LaHuis, D. M., Blackmore, C. E., Bryant-Lees, K. B., & Delgado, K. (2018). Applying Item Response Trees to personality data in the selection context. *Organizational Research Methods, Online first*, 1-12.

Lauritzen, S. L. (1996). *Graphical models*. Oxford: Clarendon Press.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology, 55*(1), 1-7.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Levine, M. V. (2003). Dimension in latent variable models. *Journal of Mathematical Psychology*, *47*(4), 450-466.

Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. New York: Sage.

Lord, F. M. (1953). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika, 18*(1), 57-76.

Luce, R.D., Krantz, D. H., Suppes, P., & Tversky, A. (1990). *Foundations of measurement, Vol. III: Representation, axiomatization, and invariance*. New York: Academic Press.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*(1), 1-27.

Manning, C. D., Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L. J., van der Maas, & Maris, G. (2018). An introduction to network psychometrics: Relating Ising network models to item response theory models. *Multivariate Behavioral Research, 53*(1), 15-35.

McDonald, R. P. (1965). Difficulty factors and nonlinear factor analysis. *British Journal of Mathematical and Statistical Psychology, 18*(1), 11-23.

McDonald R. P. (2013). *Test theory: A unified treatment*. New York: Psychology Press.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, *18*(4), 311-314.

Mertens, D. M. (2014). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods*. New York: Sage.

Michell, J. (1990). *An introduction to the logic of psychological measurement.* Hillsdale, NJ: Erlbaum.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*(3), 355-383.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology, 10*, 639–667.

Michell, J. (2005). The logic of measurement: A realist overview. *Measurement*, *38*(4), 285-294.

Michell, J. (2008). Conjoint measurement and the Rasch paradox: A response to Kyngdon. *Theory & Psychology*, *18*, 119-124.

Michell, J. (2017). On substandard substantive theory and axing axioms of measurement: A response to Humphry. *Theory & Psychology, 27*(3), 419-425.

Miyazaki, K., & Hoshino, T. (2009). A Bayesian semiparametric item response model with Dirichlet process priors. *Psychometrika, 74*(3), 375-393.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton/Berlin: De Gruyter.

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*, 417-430.

Myung, I. J., & Pitt, M. A. (2001). Mathematical modeling. In J. Wixted (Ed.), *Stevens' handbook of experimental psychology. Vol. 4: Methodology* (pp. 429-459). New York: Wiley.

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence: New findings and theoretical developments. *American Psychologist*, *67*(2), 130-159.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39-57.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Price, D. J. S. (1986). *Little science, big science... and beyond*. New York: Columbia University Press.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Educational Research Institute.

Reckase, M. (2009). *Multidimensional item response theory*. New York: Springer.

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory: Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science, 14*(2), 95-101.

Roberts, F. S. (1979). *Measurement theory with applications to decision making, utility, and the social sciences*. Reading, MA: Addison-Wesley Publishing Company.

Rosenbaum, P. R. (1987). Comparing item characteristic curves. *Psychometrika*, *52*(2), 217-233.

Sandberg, J., & Alvesson, M. (2011). Ways of constructing research questions: Gap-spotting or problematization? *Organization*, *18*(1), 23-44.

Scott, D. (1964). Measurement structures and linear inequalities. *Journal of Mathematical Psychology*, *1*(2), 233-247.

Shaughnessy, J. J., Zechmeister, E. B., & Zechmeister, J. S. (2014). *Research methods in psychology*. New York: Alfred A. Knopf.

Shoenfield, J. R. (2018). *Mathematical logic*. Boca Raton, FL: CRC Press.

Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory & Psychology*, *22*(6), 786-809.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. New York: Sage.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137-149.

Stanovich, K. E. (2012). *How to think straight about psychology*. New York: Pearson.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 667-680.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*(4), 589-617.

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification*, *30*(1), 75-99.

Suppes, P. (1951). A set of independent axioms for extensive quantities. *Portugaliae Mathematica, 10*, 163-172.

Suppes, P., Krantz, D. H., Luce, R.D., & Tversky, A. (1989). *Foundations of measurement, Vol. II: Geometrical, threshold, and probabilistic representations*. New York: Academic Press.

Thabane, L., Thomas, T., Ye, C., & Paul, J. (2009). Posing the research question: Not so simple. *Canadian Journal of Anesthesia*, *56*(1), 71-79.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. New York: American Psychological Association.

Townsend, J. T. (2008). Mathematical psychology: Prospects for the 21st century: A guest editorial. *Journal of Mathematical Psychology*, *52*(5), 269-280.

Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory & Psychology, 19*(5), 579-599.

Trendler, G. (2013). Measurement in psychology: A case of ignoramus et ignorabimus? A rejoinder. *Theory & Psychology, 23*(5), 591-615.

Tversky, A. (1967). A general theory of polynomial conjoint measurement. *Journal of Mathematical Psychology. 4*, 1–20.

Van der Linden, W. J., & Hambleton, R. K. (2013). *Handbook of modern item response theory*. New York: Springer.

Van der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review, 113*(4), 842-861.

Van der Maas, H., Kan, K. J., Marsman, M., & Stevenson, C. E. (2017). Network models for cognitive development and intelligence. *Journal of Intelligence, 5*(2), 1-17.

Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton: Princeton University Press.

Wakker, P., & Tversky, A. (1993). An axiomatization of cumulative prospect theory. *Journal of Risk and Uncertainty, 7*(2), 147-175.

Wang, C., Chang, H. H., & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology, 66*(1), 144-168.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*(2), 213-249.

II

Conditional Item Response Model and Optimal scores: Alternatives to the Rasch model

Vithor Rosa Franco[1], Jacob Arie Laros[1], and Marie Wiberg[2]

Affiliations

[1]Post-graduate program of Social, Work and Organizational Psychology, Institute of Psychology, University of Brasília, Brasília, Brazil;

[2]Department of Statistics, Umeå School of Business, Economics and Statistics, Umeå University, Umeå, Sweden;

Abstract

The objective of the present study was to develop two new item response models for polytomous and binary items that do not assume a normal distribution of the true scores. The first model that was developed, the Conditional Item Response Model (CIRM), assumes a beta-binomial distribution. The second model is a Bayesian implementation of the optimal score procedure (OS-IRM). Two studies were conducted with the new developed models: the first was a Monte Carlo simulation comparing the effectiveness of the two new models with the Rasch model. The second study compared the practical equivalence of the two new developed models and the Rasch model empirically. Overall, results show that the CIRM produces the least biased estimates of the true scores while the OS-IRM is the procedure that best recovers the true score distribution. In the discussion a number of limitations of the present study are pointed out and suggestions are given for future studies.

Keywords: Item response theory; Bayesian nonparametric modeling; Monte Carlo simulation.

Conditional Item Response Model and Optimal scores: Alternatives to the Rasch model

Sum scoring is a very common procedure for estimating the true scores of respondents to psychological tests and questionnaires (Borsboom, 2006). This method is regarded by leading modern psychometricians (e.g., Borsboom, 2006; Ramsay & Wiberg, 2017) as inappropriate for two main reasons. First, because it is not possible to assure that the observed item scores are measured on an additive (interval) scale. Secondly, because in the sum scoring procedure psychometric characteristics of the items are not taking into consideration. This way, low quality items receive the same weight as high-quality items.

Contemporary psychometrics deals with these limitations by means of item response models (IRMs), such as the Rasch model (Rasch, 1960). Nevertheless, parametric IRMs also fall short given the fact that they assume the true score to have a normal distribution, which has an unbounded interval. This means that, in principle, respondents can have any score between $-\infty$ and $+\infty$, which seems theoretically implausible and also generates scoring scales that are less intuitive (Ramsay & Wiberg, 2017).

Wright and Panchapakesan (1969) developed a scoring procedure which assumes that binary responses can be modeled by a Bernoulli distribution. Therefore, estimates of respondents' true scores can be given by the probability parameter of this distribution, which is bounded between 0 and 1. This method, however, is limited as it can only be applied to binary response variables and, like the sum scoring procedure, does not take the psychometric characteristics of the items into account. Another model that does not assume a normal distribution of the true scores, was developed by Ramsay and Wiberg (2017) and is known as optimal scoring. It is a nonparametric IRM, which uses the sum scores as initial estimates for the true scores, and that optimizes the likelihood of responses for each item using a B-spline estimate for the item response functions (IRFs).

The present study aimed at achieving three objectives. First to expand Wright and Panchapakesan's (1969) model to polytomous items and use this expansion as the building block to develop two new item response models (IRMs) that do not assume a normal distribution of the true scores. The secondary aim is to compare the effectiveness of these newly developed IRMs to the Rasch model using a Monte Carlo simulation study. The last aim is to compare the practical equivalence of the three models using empirical data.

The rest of this paper is structured as follows. In the next section, we explain the Bernoulli scoring procedure and extend it for, what we call, the binomial scoring procedure. We then present the IRM, called the Conditional item response model (CIRM), which uses beta distributions for the true scores. The fourth section is dedicated to show how to fit the CIRM, the Rasch model and the optimal scoring using Bayesian statistics. The fifth and the sixth sections are dedicated to a simulation study and an example with real data, respectively. The paper ends with a discussion and some concluding remarks.

The binomial scoring procedure

If items are coded as simple binary variables, response patterns (X) can be assumed to be stochastic and dependent on a true score θ: $P(X = 1|\theta)$. Departing from a set of k items, the response pattern of a single respondent can be modeled by a Bernoulli distribution:

$$P(X = 1|\theta) = \theta^{k}(1 - \theta)^{1-k}. \quad (1)$$

This procedure for scoring respondents was proposed by Wright and Panchapakesan (1969) in order to avoid using sum scores. Despite the fact that sum scores are actually good approximations for this kind of procedure (Rosenbaum, 1987), with maximum likelihood estimates (MLE) also standard errors associated with the θ estimates are provided, which is not the case with sum scores (Tarone, 1979).

One disadvantage of the Bernoulli procedure is its limitation to binary data. Relying on the fact that the Bernoulli distribution is a special case of the binomial distribution (Marshall & Olkin, 1985), it is straightforward to expand the procedure so it can give estimates also for polytomous data. Departing from a test with *k* items, but now with polytomous items with *l* response levels, one can model the response patterns by

$$P(X = l|\theta, n) = \binom{n}{S} \theta^S (1-\theta)^{n-S}, \quad (2)$$

where *S* is the sum score and *n=(l-1)k* is a fixed parameter.

Estimation for the θ parameter can be done either with MLE (Kleinbaum & Klein, 2010) or with a Bayesian method such as Markov Chain Monte Carlo (MCMC; Everson & Bradlow, 2002). For the MLE method, the computations can often be simplified by maximizing the loglikelihood (*LL*) function, given by

$$LL(\theta; S) = k + S \ln\theta + (n-S)\ln(1-\theta). \quad (3)$$

The advantage of using the *LL* function is the fact that its optimization is computationally less expensive than the optimization of the original binomial function (for more details, see Edwards, 1984).

To estimate the θ parameter using Bayesian statistics, one can use a beta-binomial model with uninformative priors: $\alpha = \beta = 1$ (Lee & Wagenmakers, 2014). This approach begins with the same binomial distribution, which can be simply rewritten as

$$X \sim \mathrm{Binom}(n, \theta), \quad (4)$$

which means that the response pattern *X* (for a person or for an item) is distributed as a Binomial distribution with parameters *n* and θ. The next step is to define the beta distribution, $\mathrm{Beta}(\alpha, \beta)$, for θ, which is given by

$$f(\theta; \alpha, \beta) = \frac{1}{\mathrm{B}(\alpha, \beta)} S^{\alpha-1}(1-S)^{\beta-1}, \quad (5)$$

where $\mathrm{B}(\alpha, \beta)$ is the beta function. The actualization steps for the MCMC algorithm are given by the mean of the compound distribution:

$$f(S|n, \alpha, \beta) = \binom{n}{S} \frac{\mathrm{Beta}(S+\alpha, n-S+\beta)}{\mathrm{Beta}(\alpha, \beta)}. \quad (6)$$

Furthermore, the beta-binomial distribution can be extended to its semiparametric form (Liu, 1996). Using Bayesian non-parametric modeling via Dirichlet process prior (Blackwell, 1973) for the mixing distribution, this can be accomplished by

$$\begin{aligned} X &\sim \mathrm{Binom}(n, \theta) \\ \theta &\sim F \\ F &\sim \mathrm{DP}(M, F_0) \\ F_0 &\sim \mathrm{Beta}(\alpha, \beta) \end{aligned} \quad (7)$$

where the random probability measure *F* replaces the beta prior of the parametric model in Eq. (5) and DP stands for the Dirichlet process prior. As a base prior distribution, F_0 can also assume a Beta(α, β) with $\alpha = \beta = 1$.

Both the Bernoulli and the beta-binomial scoring procedures have as a limitation the fact that they assume different items to not have different effects on the response pattern, just like with sum scores. However, departing from these procedures it is possible to formulate an IRM that exceeds this limitation. We propose to use the conditional probability function (Ross, 2014) as the (linking) IRF (van der Linden & Hambleton, 2013) when true scores are assumed to come from a beta distribution.

Bounded support and the Conditional Item Response Model

One advantage of using the beta distribution for estimating the true scores is the fact that the distribution of latent variables can be set to be any arbitrary distribution (Ramsay & Wiberg, 2017). Therefore, using distributions with bounded support can improve both interpretability of the estimates and the accuracy of the estimates. Here, items and respondents' scores are

assumed to follow a beta distribution, which is bounded between 0 and 1, thus an initial IRF could simply be the expected value of a beta distribution, given by

$$P(X_{ij} = x_{ij}|\theta_j, \delta_i) = \mathrm{E}[\theta_j, \delta_i] = \frac{\theta_j}{\theta_j + \delta_i}, \tag{8}$$

where θ_j is the latent true score of the j^{th} respondent and δ_i is the difficulty for the i^{th} item. A nice property of Eq. 8 is that it can be directly related to both Rasch and one-parameter logistic (1PL) models by their shared additive property of the estimated scores (Perline, Wright, & Wainer, 1979). This relation can be made evident by evaluating the log transformation of Eq. 8:

$$\begin{aligned} \mathrm{logit}\big(P(X = x|\theta, \delta)\big) &= \log(\theta) - \log(\delta), \\ P(X = x|\theta, \delta) &= \frac{1}{1 + \exp[\log(\theta) - \log(\delta)]}, \end{aligned} \tag{9}$$

given that θ and δ follow beta distributions and, therefore, can only assume nonnegative real values, upper bounded by 1.

An interesting property of other IRMs, such as the Rasch and 1PL models, is whenever θ_j and δ_i are equal, $P(X_{ij} = x_{ij}|\theta_j, \delta_i)$ = .50 (van der Linden & Hambleton, 2013). Nevertheless, if bias for an end of the scale is expected to influence respondents in a sample equally, such as with scales influenced by social desirability (Edwards, 1957), the value of $P(X_{ij} = x_{ij}|\theta_j, \delta_i)$ when θ_j and δ_i are equal should be different from .50. Therefore, we can assume .50 to represent a prior probability *c*. Conditioning $P(X_{ij} = x_{ij})$ on this prior probability will give the CIRM:

$$P(X_{ij} = x_{ij}|\theta_j, \delta_i, c) = \frac{c\theta_j}{c\theta_j + (1 - c)\delta_i}. \tag{10}$$

Because the CIRM assumes a beta distribution for both respondents and items' parameters, it can be considered as a parametric version of the optimal scoring procedure (OS-IRM; Ramsay & Wiberg, 2017). OS-IRM maximizes the likelihood of each respondent's

response pattern, estimating θ as weighted sum scores, therefore, defining a closed interval to θ, that can also be modeled by a beta distribution. This objective is achieved by estimating the logit (or inverse logistic) function $W_i(\theta)$ instead of $P_i(\theta)$, defined as

$$W_i(\theta) = \log\left(\frac{P_i(\theta)}{1 - P_i(\theta)}\right). \quad (11)$$

An efficient nonparametric estimation approach for $W_i(\theta)$ is to use B-spline basis function expansions with Q knot sequences:

$$W_i(\theta) = \sum_{q}^{Q} \gamma_{iq}\phi_{iq}(\theta), \quad (12)$$

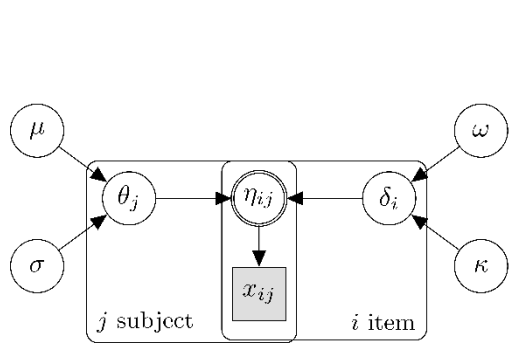
where γ_{iq} is the coefficient of B-spline basis function ϕ_{iq} in the basis function expansion of the i^{th} item's IRF. This approach is preferable because linear combinations do not respect bounds restricted to the variables and $P_i(\theta)$ can only assume values between 0 and 1. The function $W_i(\theta)$, however, can assume any value between $-\infty$ and $+\infty$, while a value of 0 of $W_i(\theta)$ is equivalent to a value of .50, or 50%, in $P_i(\theta)$.

Fitting the CIRM and the OS-IRM

Both MLE and Bayesian methods can be used to fit the CIRM and the OS-IRM. For the present study, a Bayesian method is used as it can simultaneously estimate the parameters of respondents and of items (Gelman, Carlin, Stern, Dunson, Vehtari, & Rubin, 2014), as well as be used directly to assess models' differences (Kruschke, 2015). For the CIRM, responses of each respondent to each item are modelled according to a binomial distribution with parameters $\eta_{ij} = P(X_{ij} = x_{ij}|\theta_j, \delta_i, c)$ and n. The θ_j and δ_i parameters are sampled from the reparametrized beta distributions (Kruschke, 2015) so their distributions approximate a normal distribution, bounded between 0 and 1, but with means μ and ω drawn from a

noninformative beta distribution—Beta(1,1)—and standard deviations σ and κ drawn from a noninformative gamma distribution—Gamma (.001, .001).

The Bayesian procedure to fit the CIRM is shown in the net representation in Figure 1, following Lee's (2008) graphical standards. The observed variables are represented by shaded nodes and the unobserved variables are represented by unshaded nodes. Discrete variables are represented by square nodes, while continuous variables are represented by circular nodes. Stochastic variables are represented by single-bordered nodes, and deterministic variables are represented by double-bordered nodes. Finally, encompassing plates are used to denote independent replications of the graph structure within the model.



$$\mu \sim \text{Beta}(1,\ 1)$$
$$\sigma \sim \Gamma(.001,\ .001)$$
$$\theta_j \sim \text{Beta}(\mu\sigma + 1,\ \sigma(1-\mu) + 1)$$

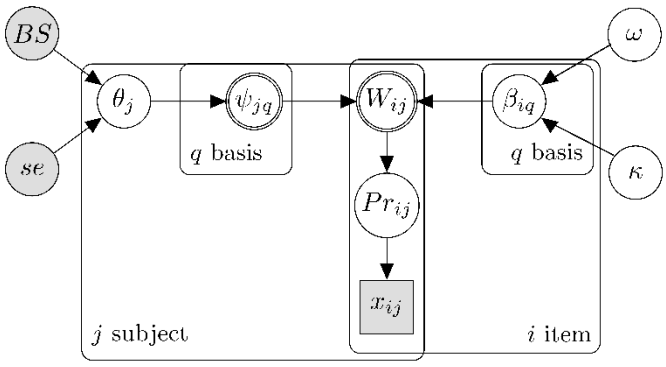
$$\omega \sim \text{Beta}(1,\ 1)$$
$$\kappa \sim \Gamma(.001,\ .001)$$
$$\delta_i \sim \text{Beta}(\omega\kappa + 1,\ \kappa(1-\omega) + 1)$$

$$\eta_{ij} \leftarrow \frac{c\theta_j}{(c\theta_j)+(c\delta_i)}$$
$$x_{ij} \sim \text{Binom}(\eta_{ij},\text{n})$$

Figure 1. Bayesian representation of the CIRM.

In the MLE approach proposed by Wiberg, Ramsay and Li (2019) for fitting the OS-IRM, the initial estimates for the true scores are based on the sum scores. To follow the same reasoning, we propose the use of beta-binomial scoring procedure to get initial estimates (*BS*) for the true scores and also to obtain standard errors estimates (*se*), which can be used to set a prior of normal reparametrized beta distributions for the scores. Wiberg et al (2019) also used a B-spline for estimating W_{ij}. Given that the estimates of true scores are bounded to 0 and 1, and that W_{ij} can have negative values, we propose the use of a Rademacher basis. The Rademacher distribution is a discrete probability distribution which has equal chance for either 1 or –1 (Seth & Príncipe, 2008). Our proposed Rademacher basis is a simple rule: if the value of the θ_j estimate is below or equal to G_q, then the basis equals –1. If the value is above

G_q, then the basis equals 1. The G_q knots were set by equidistance points in the 0 to 1 scale. Finally, we apply a DP to the logistic transform of W_{ij}, following a procedure similar to that of Duncan and MacEachern (2008). This proposed model is presented in Figure 2.



$$\theta_j \sim \text{Beta}(\tfrac{BS}{se}+1,\ \tfrac{(1-BS)}{se}+1)$$
$$\omega \sim N(0,\ .01)$$
$$\kappa \sim \Gamma(.01,\ .01)$$

$$\alpha \sim N(0,\ 1)$$
$$\nu \sim \text{Unif}(.01,\ 100)$$
$$\beta_{iq} \sim N(\alpha,\ \nu)$$

$$\psi_{jq} \leftarrow \begin{cases} -1, & \theta_j \leq G_q \\ +1, & \theta_j > G_q \end{cases}$$

$$W_{ij} \leftarrow \sum_{q=1}^{Q} \beta_{iq}\psi_{jq}$$

$$Pr_{ij} \sim \text{DP}(M_i, \text{ilogit}(W_{ij}))$$
$$x_{ij} \sim \text{Binom}(Pr_{ij}, n)$$

Figure 2. Bayesian implementation of the OS-IRM.

Simulation study

Method

We used both the Rasch model and the CIRM as the true IRFs for the data generation process (DGP). When the data were generated using the Rasch model, the θ parameter was drawn from a truncated normal distribution with mean 0, standard deviation 1 and -3 and 3 as lower and upper bound, respectively. When the data were generated using the CIRM, the θ parameter was drawn from a beta distribution with both parameters α and β equal to 1. These distributions were chosen because they represent traditionally defined non-informative prior distributions (Kruschke, 2015). Difficulties were drawn following the same distributions, according to the IRF. Random draws from sample sizes of 100, 250 and 500 simulated respondents and from test sizes of 10, 20 and 50 items were each iterated 100 times. This

resulted in 9 crossed conditions and 18 total conditions, taking into account both DGPs (Rasch and CIRM).

The Rasch model, the CIRM and the OS-IRM were compared using five measures of effectiveness. The first measure is related to accuracy and was measured by Spearman's correlation between the estimates and the known true scores. To assess accuracy throughout the range of true scores' estimates, we measured bias by the residuals of an additive regression between estimated and true scores. As an average measure of accuracy, we used the mean absolute error (MAE), which is not affected by the variance of the distribution of error magnitudes (Willmott & Matsuura, 2005). MAE is calculated as

$$\text{MAE} = \frac{1}{N}\sum_{j=1}^{N}\left|\theta_j - \hat{\theta}_j\right|, \tag{13}$$

where *N* is the number of respondents, θ_j is the true score and $\hat{\theta}_j$ is an estimate of θ_j, calculated using the expected values of an additive regression between estimated scores and true scores.

Integrated square error (ISE; Shirahata & Chu, 1992) was the fourth measure of effectiveness, used to access how similar were the true and the estimated distribution of θ. The ISE was computed by

$$\text{ISE}(\hat{g}) = \int \{\hat{g}(\theta) - g(\theta)\}^2 d\theta, \tag{14}$$

where $\hat{g}(\theta)$ is the density of the feature-scaled distribution of estimates of the true score θ and $g(\theta)$ is the density of the real distribution of feature-scaled true scores. Both $\hat{g}(\theta)$ and $g(\theta)$ were calculated using kernel density estimates with the number of equally spaced points equal to the sample size and the bandwidth chosen adaptively using Sheather and Jones (1991) method.

For comparing model fit we calculated the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Linde, 2014). The DIC is a hierarchical modeling

generalization of the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). As in the case of BIC and AIC, models with smaller DIC should be preferred over models with larger DIC. The main difference between DIC and both AIC and BIC is the fact that the DIC has larger penalties for the quantity of parameters in the model.

The R (R Core Team, 2019) code with the full simulation is available from the corresponding author upon request. The Bayesian models were all implemented in the JAGS software (Plummer, 2003). The JAGS software was interfaced with R through the R2jags package (Su & Yajima, 2012). We used the gam function from the mgcv package (Wood, 2012) to estimate the residuals of the additive regressions used for measuring the error for the bias. Finally, the density integrate.xy function from the sfsmisc package (Meachler, 2018) was used to calculate the ISE.

Results

In Table 1 the results for data generated by using the Rasch model are displayed. The best performances for each effectiveness measure on each condition are bolded. The average performance shows that all the procedures produce very similar rank correlations with the true score. In terms of accuracy, as measured by the MAE, the CIRM is the best procedure, followed by the Rasch model, and by the OS-IRM procedure in the last place. In terms of properly recovering the density distribution of the true scores, measured with the ISE, the OS-IRM procedure performs the best, closely followed by the Rasch model, and by the CIRM. In terms of DIC, the OS-IRM is the model that loses the less information, followed by the Rasch model and then the CIRM.

Evaluating the different sample sizes conditions, it is possible to see that the average pattern almost does not change. The rank correlations are similar in all conditions. The MAE follows the same patterns as the average performance. The ISE is best for the Rasch model

when sample size is equal to 100 and 250. When the sample size is large (500 cases), OS-IRM has the best ISE, followed by the Rasch model, and then the CIRM. Measures of model fit are always smaller for OS-IRM, followed by the Rasch model and then the CIRM. On the other hand, when evaluating the test size (i.e., the number of variables), different patterns were found for both ISE and DIC. With 10 and 20 items, ISE was smaller for the Rasch model, followed by the OS-IRM and then the CIRM. With 50 items, the ISE was smaller for the OS-IRM, while the DIC was smaller for the Rasch model, followed by the CIRM and then the OS-IRM. Rank correlations and MAE have similar patterns as found before.

Table 1

Comparing accuracy, similarity with the true score distribution, and model fit of the three models (Rasch, CIRM and OS-IRM) for data generated by the Rasch model.

		Rasch	**CIRM**	**OS-IRM**
	Effectiveness measures	**Average performance**		
	Accuracy 1 ($\rho_{\text{true score}}$)	**.862**	.861	.861
	Accuracy 2 (MAE)	.320	**.064**	1.496
	Similarity distributions (ISE)	.250	1.279	**.167**
	Model fit (DIC)	7,050.05	7,103.59	**6,935.74**
Sample size				
100	Accuracy 1 ($\rho_{\text{true score}}$)	**.834**	.831	.833
	Accuracy 2 (MAE)	.322	**.082**	1.261
	Similarity distributions (ISE)	**.056**	.511	.127
	Model fit (DIC)	2,319.89	2,337.51	**2,298.72**
250	Accuracy 1 ($\rho_{\text{true score}}$)	.867	**.869**	.866
	Accuracy 2 (MAE)	.332	**.060**	1.587
	Similarity distributions (ISE)	**.077**	1.377	.110
	Model fit (DIC)	7,036.94	7,091.88	**6,931.86**
500	Accuracy 1 ($\rho_{\text{true score}}$)	.856	**.857**	**.857**
	Accuracy 2 (MAE)	.333	**.055**	1.469
	Similarity distributions (ISE)	0,617	1,948	**0,262**
	Model fit (DIC)	11,800.22	11,888.14	**11,473.24**
Test size				
10	Accuracy 1 ($\rho_{\text{true score}}$)	**.789**	.788	**.789**
	Accuracy 2 (MAE)	.364	**.089**	1.015
	Similarity distributions (ISE)	.607	1.365	**.189**
	Model fit (DIC)	2,540.43	2,554.13	**2,351.10**
20	Accuracy 1 ($\rho_{\text{true score}}$)	**.882**	.879	.880
	Accuracy 2 (MAE)	.332	**.058**	1.491

	Similarity distributions (ISE)	**.079**	.636	.167
	Model fit (DIC)	6,058.37	6,130.63	**5,929.21**
50	Accuracy 1 ($\rho_{\text{true score}}$)	.941	**.942**	.940
	Accuracy 2 (MAE)	.238	**.042**	2.150
	Similarity distributions (ISE)	**.064**	1.835	.144
	Model fit (DIC)	**12,544.46**	12,619.23	12,630.29

Note. $\rho_{\text{true score}}$ = the Spearman correlation of the estimated scores with the true scores. MAE = mean absolute error. ISE = integrated squared error. DIC = deviance information criterion.

In Table 2, the results for data generated by using the CIRM are displayed. The results are similar to those found when the true DGP was based on the Rasch model. Overall, the rank correlations are all similar between the three models, but substantially lower than the correlations found in Table 1. The second measure of accuracy, the mean absolute error (MAE) was always smaller for the CIRM meaning that the CIRM showed better accuracy. The dissimilarity of the score distribution with the true score distribution (TSD) as measured by the Integrate Square Error (ISE) was smaller for the OS-IRMT in most conditions, with exception for test sizes of 20 and 50 items and for a sample size equal to 100. In these cases, the ISE was smaller for the CIRM. The model misfit as measured by the Deviance Information Criterion (DIC) was less for OS-IRM in half of the conitions (sample sizes of 100 and 250 and test size of 50 items), while in the other conditions the CIRM showed better model fit (sample size of 500 and test sizes of 10 and 20 items).

Table 2
Comparing accuracy, similarity with the true score distribution, and model fit of the three models (Rasch, CIRM and OS-IRM) for data generated by the CIRM.

		Rasch	**CIRM**	**OS-IRM**
	Effectiveness measures	**Average performance**		
	Accuracy 1 ($\rho_{\text{true score}}$)	**.795**	.794	.794
	Accuracy 2 (MAE)	.352	**.104**	1.844
	Similarity distributions (ISE)	1.045	.523	**.297**
	Model fit (DIC)	7,338.65	7,258.84	**7,241.78**
Sample size				
100	Accuracy 1 ($\rho_{\text{true score}}$)	**.767**	.765	.765
	Accuracy 2 (MAE)	.329	**.104**	1.518

	Similarity distributions (ISE)	.527	**.147**	.311
	Model fit (DIC)	2,411.93	**2,370.56**	2,390.30
250	Accuracy 1 ($\rho_{\text{true score}}$)	**.805**	.803	.803
	Accuracy 2 (MAE)	.369	**.105**	1.947
	Similarity distributions (ISE)	1.018	.551	**.278**
	Model fit (DIC)	7,344.34	**7,271.23**	7,273.01
500	Accuracy 1 ($\rho_{\text{true score}}$)	.779	.780	**.781**
	Accuracy 2 (MAE)	.373	**.109**	1.784
	Similarity distributions (ISE)	1.591	.870	**.302**
	Model fit (DIC)	12,278.43	12,124.95	**11,955.64**
Test size				
10	Accuracy 1 ($\rho_{\text{true score}}$)	**.712**	.710	.711
	Accuracy 2 (MAE)	.352	**.114**	1.117
	Similarity distributions (ISE)	2.101	1.366	**.204**
	Model fit (DIC)	2,653.93	2,567.32	**2,449.54**
20	Accuracy 1 ($\rho_{\text{true score}}$)	**.813**	.811	.812
	Accuracy 2 (MAE)	.365	**.107**	1.735
	Similarity distributions (ISE)	.376	**.137**	.204
	Model fit (DIC)	6,352.29	6,281.28	**6,216.37**
50	Accuracy 1 ($\rho_{\text{true score}}$)	**.896**	**.896**	.894
	Accuracy 2 (MAE)	.325	**.085**	2.965
	Similarity distributions (ISE)	.659	**.065**	.483
	Model fit (DIC)	12,991.01	**12,937.71**	13,165.84

Note. $\rho_{\text{true score}}$ = the Spearman correlation of the estimated scores with the true scores. MAE = mean absolute error. ISE = integrated squared error. DIC = deviance information criterion.

In Table 3, the results of both conditions of data generation processes were averaged. Once more, the rank correlations are all similar. MAE was always smaller for the CIRM. ISE was always smaller for the OS-IRM. DIC was always smaller for the OS-IRM but when the test size was equal to 50, then the smallest value was due to the Rasch model.

Table 3
Comparing accuracy, similarity with the true score distribution, and model fit of the three models (Rasch, CIRM and OS-IRM) for the average of data generated by both models.

		Rasch	CIRM	OS-IRM
	Effectiveness measures	**Average performance**		
	Accuracy 1 ($\rho_{\text{true score}}$)	**.828**	**.828**	**.828**
	Accuracy 2 (MAE)	.336	**.084**	1.670
	Similarity distributions (ISE)	.648	.901	**.232**
	Model fit (DIC)	7,194.35	7,181.22	**7,088.76**
Sample size				
100	Accuracy 1 ($\rho_{\text{true score}}$)	**.801**	.798	.799
	Accuracy 2 (MAE)	.326	**.093**	1.390
	Similarity distributions (ISE)	.292	.329	**.219**
	Model fit (DIC)	2,365.91	2,354.04	**2,344.51**
250	Accuracy 1 ($\rho_{\text{true score}}$)	**.836**	**.836**	.835
	Accuracy 2 (MAE)	.351	**.083**	1.767
	Similarity distributions (ISE)	.548	.964	**.194**
	Model fit (DIC)	7,190.64	7,181.56	**7,102.44**
500	Accuracy 1 ($\rho_{\text{true score}}$)	.818	**.819**	**.819**
	Accuracy 2 (MAE)	.353	**.082**	1.627
	Similarity distributions (ISE)	1.104	1.409	**.282**
	Model fit (DIC)	12,039.32	12,006.54	**11,714.44**
Test size				
10	Accuracy 1 ($\rho_{\text{true score}}$)	**.751**	.749	.750
	Accuracy 2 (MAE)	.358	**.102**	1.066
	Similarity distributions (ISE)	1.354	1.366	**.197**
	Model fit (DIC)	2,597.18	2,560.73	**2,400.32**
20	Accuracy 1 ($\rho_{\text{true score}}$)	**.848**	.845	.846
	Accuracy 2 (MAE)	.349	**.083**	1.613
	Similarity distributions (ISE)	.228	.387	**.186**
	Model fit (DIC)	6,205.33	6,205.96	**6,072.79**
50	Accuracy 1 ($\rho_{\text{true score}}$)	**.919**	**.919**	.917
	Accuracy 2 (MAE)	.282	**.064**	2.558
	Similarity distributions (ISE)	.362	.950	**.314**
	Model fit (DIC)	**12,767.73**	12,778.47	12,898.06

Note. $\rho_{\text{true score}}$ = the Spearman correlation of the estimated scores with the true scores. MAE = mean absolute error. ISE = integrated squared error. DIC = deviance information criterion.

Empirical example

Aiming at illustrating the differences between the procedures, we used a sample of data of 5,000 respondents from an administration of the Brazilian National High School Exam

(ENEM). The ENEM is a yearly national exam, non-mandatory, which both evaluates high school students in Brazil and can be used as an admission test for enrollment in Brazilian and Portuguese colleges. It consists of four subtests: languages; human sciences; natural sciences; and mathematics. Each subtest is analyzed separately and contains 45 items. In this example, we used the languages subtest of the ENEM. The languages subtest was chosen because it is less skewed and presents less outliers in its sum scores than the other three subtests. The distribution of the languages sum scores can be seen in Figure 3.

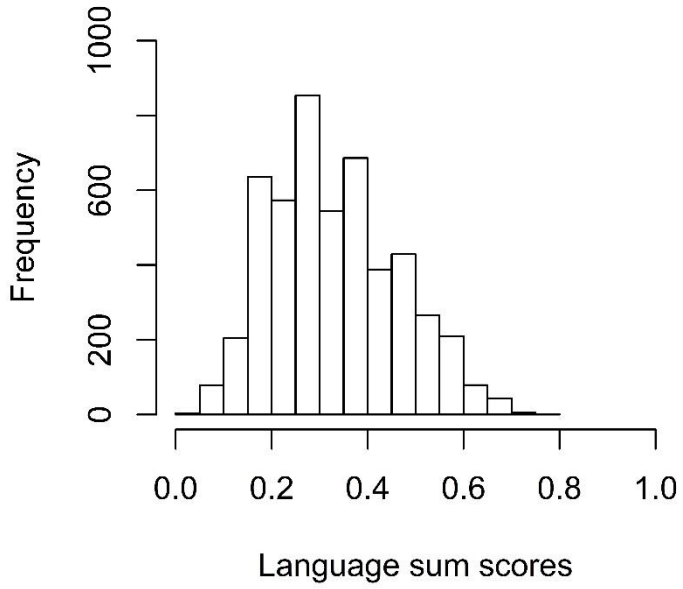


Figure 3. The distribution of the sum scores of ENEM's languages subtest.

Respondents are scored on each subtest using expected a priori estimation (EAP; Bock & Aitkin, 1981) of latent trait scores of the three-parameter logistic model (Andriola, 2011). A test score is then calculated as a composite score of each subtest and is used in the selection process for higher education in Brazil and for some colleges in Portugal. Therefore, differences in the order, or ranking, of the respondents can result in different people having access to higher education. The ENEM scores on the languages' subtest were compared to the sum scores using three different measures. First, the densities of the IRMs' estimates were compared to the sum scores' density using the ISE. Next, Kolmogorov-Smirnov *d* statistic was calculated for the distributions of scores estimated using sum scores and the

IRMs. The *d* statistic simply represents the largest distance (in absolute value) between the cumulative distribution functions of the target distribution (i.e., a normal distribution) and the distribution of the estimated scores. Finally, Spearman's correlation was used to compare how similar the scores rank respondents, using the whole sample and the top 5% and 1% performers on the sum scores of the languages' subtest.

Results

By evaluating the *d* statistic from Table 4 it becomes evident that the Rasch estimates are more normally distributed than the other estimates. Inspection of the Integrate Square Error (ISE) shows that the OS-IRM distribution was the most similar to the sum score distribution.

Table 4

Distributional properties of the estimated scores in terms of distance to a normal distribution (d) and difference from the sum score's distribution (ISE).

Measure	Sum score ENEM	Rasch	CIRM	OS-IRM
d	.516	**.219**	.518	.953
ISE	—	.375	.124	**.080**

The values of *d* and ISE can also be reflected by the densities represented in Figure 4. The OS-IRM has almost the same positively skewed distribution as the sum score's. Rasch estimates closely follow a normal distribution and CIRM scores have a positively skewed distribution, with stronger asymmetry and a different peak when compared to the sum score and OS-IRM distributions.

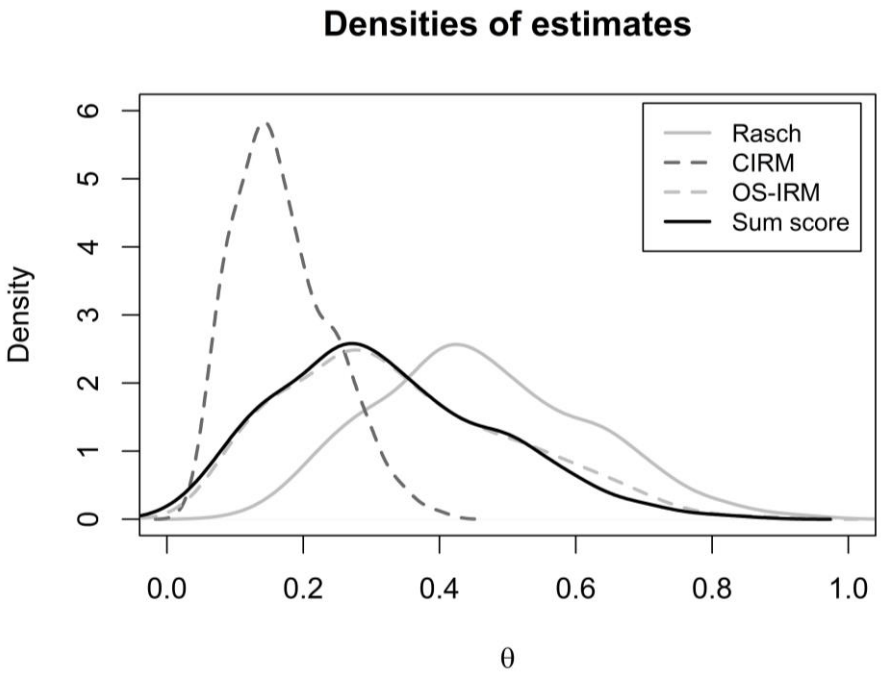
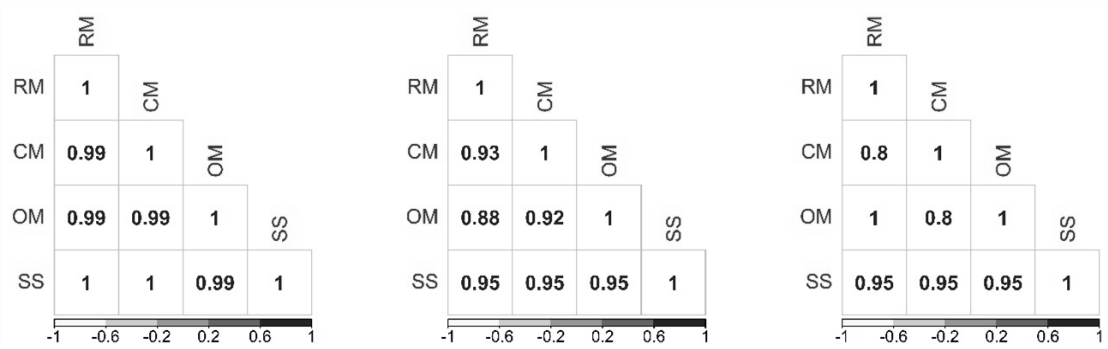


Figure 4. Densities of the estimated scores.

The Spearman correlations in Figure 5 show how similarly the different procedures rank the respondents. When the whole sample is used, all procedures gave almost the same ranking of participants, with the smaller Spearman correlation equals to .99. The rank correlations between the sum scores (SS) and the other procedures are virtually the same (due to rounding). The correlations between the CIRM (CM) and the other two IRMs decrease with the changes in the sample for the top 5% and top1% performers. The correlation between the Rasch model (RM) and the OS-IRM (OM) decreases when considering only the top 5% performers, but increases when considering only the top 1% performers.



Whole sample Top 5% performers Top 1% performers

Figure 5. Correlation between scores given the whole sample, the top 1% and the top 5% performers.

Discussion

The present study aimed at achieving three objectives: (1) to develop two new item response models (IRMs) that do not assume a normal distribution of the true scores; (2) to compare the effectiveness of these newly developed IRMs to the Rasch model using simulated data, and (3) to compare the practical equivalence of the three models using empirical data. The two developed models were named the Conditional Item Response Model (CIRM) and the Optimal Score Item Response Model (OS-IRM). The CIRM is a one parametric IRM for polytomous and binary items that assumes a beta-binomial distribution, and the OS-IRM is Bayesian implementation of the optimal score procedure.

The overall results of the simulation study based on the mean absolute error (MAE) showed that the CIRM produced less biased scores of the true scores than the OS-IRM and the Rasch model. Compared to the Rasch model the OS-IRM yielded more biased estimations of the true scores. This result is probably due to setting the priors on the true scores in the OS-IRM procedure as a beta distribution based on the binomial score of the simulated respondents, while the Dirichlet Prior was only used to estimate the item response functions. This setting probably caused the estimation of true scores being less flexible in comparison with the tilted scaled beta distribution proposed by Ramsay and Wiberg (2017). The OS-IRM, however, almost always produced the best ISE—a measure of similarity with

the distribution of the true sore—, and the best DIC—a measure of model fit—especially when the conditions of data generation process were averaged. These results indicate that the OS-IRM is the procedure that better recovers the distribution of the true scores and better explains the variance of the observed scores. This is especially true as, when working with real data, we do not know what the true data generation process is. In terms of correlation, all the procedures were always quite similar. Finally, in terms of practical equivalence, all the procedures ranked the participants similarly, with the CIRM having the smallest correlation with the other procedures.

On understanding the results, it is important to restate the fact, presented in Equation 9, that the CIRM can be considered just as an alternative form of the Rasch model. Establishing an interval of the scores bounded between 0 and 1, not only improves the interpretation of the scores, but also the statistical estimation of the parameters (Wiberg et al, 2019). However, there are limitations that must be taken into account, especially in relation to the implementation of the OS-IRM. The numbers of knots, for instance, were always equal to 10, increasing the complexity of the model when there were a larger number of items. Also, the Rademacher basis was proposed in the present study and its inferential properties need to be further evaluated (e.g., Claeskens, Krivobokova, & Opsomer, 2009). Setting the priors to be equal to the binomial scores allowed for using the beta distribution as the bounded distribution for estimating the true score with OS-IRM. On the other hand, however, the beta distribution has heavy tails biasing the scores to be closer to the average. More flexible distributions should be tested, such as the tilted scaled beta distribution (Ramsay & Wiberg, 2017) or a truncated t distribution (Kim, 2008).

Future studies should focus on how to extend the CIRM to more complex parametric models, similarly to how the Rasch model was extended to the family of logistic IRMs (Maris, & Bechger, 2009). Also, in the present study, we did not investigate how the *c*

parameter, which represents the probability of getting a correct response given that difficulty and true scores are equal, can influence model estimation and possible inferences from it. Therefore, future studies should be aimed on comparing how parametric extensions of the CIRM, and its nonparametric version, the OS-IRM, can improve over other traditional IRMs. Particularly regarding the CIRM, it can be simply considered as an alternative form of the Rasch model. But even without adding more parameters, the CIRM reduces the bias on the measurement of true scores when compared with the Rasch model. This can be interpreted as evidence that, for improving psychometrical measurement, it may be not necessary to use very complex models, but rather models with better mathematical properties.

References

Andriola, W. B. (2011). Doze motivos favoráveis à adoção do Exame Nacional do Ensino Médio (ENEM) pelas Instituições Federais de Ensino Superior (IFES) [Twelve reasons for the adoption of the National Examination for Secondary Education (ENEM) by Federal Institutions of Higher Education (IFES)]. *Ensaio: Avaliação e Politicas Públicas em Educação*, *19*(70), 107-125.

Blackwell, D. (1973). Discreteness of Ferguson selections. *The Annals of Statistics*, *1*(2), 356-358.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 371-425.

Claeskens, G., Krivobokova, T., & Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, *96*(3), 529-544.

Duncan, K. A., & MacEachern, S. N. (2008). Nonparametric Bayesian modelling for item response. *Statistical Modelling*, *8*(1), 41-66.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. New York: Praeger.

Edwards, A. W. F. (1984). *Likelihood*. Cambridge: Cambridge University Press.

Everson, P. J., & Bradlow, E. T. (2002). Bayesian inference for the beta-binomial distribution via polynomial expansions. *Journal of Computational and Graphical Statistics*, *11*(1), 202-207.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis*. Boca Raton, FL: CRC Press.

Griffiths, D. A. (1973). Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics, 29*(4), 637-648.

Kim, H. J. (2008). Moments of truncated Student-t distribution. *Journal of the Korean Statistical Society*, *37*(1), 81-87.

Kleinbaum, D. G., & Klein, M. (2010). Maximum likelihood techniques: An overview. In *Logistic Regression* (pp. 103-127). New York: Springer.

Kruschke, J. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Cambridge: Academic Press.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*(1), 1-15.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Liu, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputations. *The Annals of Statistics, 24*(3), 911-930.

Maris, G., & Bechger, T. (2009). On interpreting the model parameters for the three parameter logistic model. *Measurement, 7*(2), 75-88.

Marshall, A. W., & Olkin, I. (1985). A family of bivariate distributions generated by the bivariate Bernoulli distribution. *Journal of the American Statistical Association, 80*(390), 332-338.

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement, 3*(2), 237-255.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, No. 125.10).

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ramsay, J. O., & Wiberg, M. (2017). A strategy for replacing sum scoring. *Journal of Educational and Behavioral Statistics*, *42*(3), 282-307.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Oxford, England: Nielsen & Lydiche.

Rosenbaum, P. R. (1987). Comparing item characteristic curves. *Psychometrika*, *52*(2), 217-233.

Ross, S. M. (2014). *Introduction to probability models*. Cambridge: Academic Press.

Seth, S., & Príncipe, J. C. (2008, March). Compressed signal reconstruction using the correntropy induced metric. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 3845-3848). IEEE.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(3), 485-493.

Su, Y. S., & Yajima, M. (2012). R2jags: A Package for Running jags from R. *R package version 0.03-08, URL http://CRAN. R-project. org/package= R2jags*.

Tarone, R. E. (1979). Testing the goodness of fit of the binomial distribution. *Biometrika*, *66*(3), 585-590.

van der Linden, W. J., & Hambleton, R. K. (2013). *Handbook of modern item response theory*. New York: Springer.

Wood, S. (2012). mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. R Package retrieved from *https://cran.r-project.org/web/packages/mgcv/index.html*.

Wiberg, M., Ramsay, J. O., & Li, J. (2019). Optimal scores: An alternative to parametric Item Response Theory and sum scores. *Psychometrika*, *84*(1), 310-322.

Wright, B., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, *29*(1), 23-48.

III

An operationalization of Lewin's Equation:

The situational optimization function analysis

Vithor Rosa Franco[1], Marie Wiberg[2], and Jacob Arie Laros[1]

Affiliations

[1]Post-graduate program of Social, Work and Organizational Psychology, Institute of Psychology, University of Brasília, Brasília, Brazil;

[2]Department of Statistics, Umeå School of Business, Economics and Statistics, Umeå University, Umeå, Sweden;

Abstract

This study presents situational optimization function analysis (SOFA) and has three aims. First, to develop a Bayesian implementation of SOFA. Second, to compare this implementation with three other Maximum Likelihood-based models in their accuracy to estimate true scores. The third aim is to show how joint modeling can be used for validity research. A simulation study was used to test the second aim, while an empirical example was used to illustrate the third aim. The simulation study used three data generating process, with varying degrees of deviation from linear models and with different sample sizes. Results of the simulation study showed that the Bayesian implementation supersedes the other models. In the empirical example, data collected on 63 participants using an iterated prisoner dilemma and a scale on cooperation-competition attitudes was used. Results showed that joint modeling is the best fitting model, also increasing the correlation between the true scores of both measures (deviations from the iterated prisoner dilemma and the scale). Implications, limitations and future studies are discussed.

Key-words: Measurement; SOFA approach; Bayesian modeling; joint modeling.

An operationalization of Lewin's Equation:

The situational optimization function analysis

Psychometrics is the field in psychology dedicated to theory and practice of measuring psychological constructs (Borsboom, 2005). For most measurement models, responses are considered to be a function of the effect of a dispositional trait and particular characteristics of the items used as stimuli in a test or questionnaire (McDonald, 1999). For models developed in the Item Response Theory framework (IRT; van der Linden & Hambleton, 2013; De Ayala, 2009), respondents' dispositional trait θ (i.e., intelligence, attitudes, beliefs, personality, and so on) has an additive interaction with items' δ to score responses in questionnaires and tests. In psychology, on the other hand, the existence of sources of data that are not suitable to be analyzed with an IRM is not uncommon (Eid & Diener, 2006). Psychometric models are generally applied to correlational data (van der Linden & Hambleton, 2013). A common justification for this is the fact that experimental manipulating of psychological variables is either very difficult, impossible or unethical (Meehl, 1967). However, alternatives have been proposed by cognitive psychologists, who developed a number of measurement models that rely on information gathered in experimental settings and theoretical models different from those derived from IRT (Farrell & Lewandowsky, 2018; Lee & Wagenmakers, 2014).

The present study has as its first and main aim to develop a Bayesian implementation of situational optimization function analysis (SOFA), a general modeling framework based on Lewin's equation. The second aim is to compare the Bayesian implementation of SOFA with three other Maximum Likelihood-based approaches in its efficiency to estimate dispositional traits from simulated data. Finally, our third aim is to use real data to illustrate how construct validity analysis can be conducted in this framework by means of joint modeling (Turner et al, 2013).

Lewin's equation: Disposition versus Situation

From the field theory proposed by Lewin (1936), it is a somewhat agreed principle that a number of different and competing influences combine to result in a particular behavior expressed by individuals (Furr & Funder, in press). Stemming from this principle, two areas of research were developed: social psychology and personality psychology (McAdams, 1997). While social psychology studies mainly situational causes of behavior, personality psychology seeks to identify dispositional traits that influence behavior. In practical terms of conducting research, traditionally, social psychology is rooted in experimental designs while personality psychology relies more on psychometrics (Furr & Funder, in press).

Despite constituting seemingly different approaches, personality and social psychology are combined in several studies, as researchers in these areas understand that both are necessary for better explaining human behavior (Furr & Funder, in press). The complementarity of both approaches was stated long ago by Lewin's (Lewin, 1936) equation, where behavior (*B*) is defined as a function of the person (*P*; the dispositional traits) and the environment (*E*; the situational variables):

$$B = f(P, E). \tag{2}$$

This equation, however, was presented only as a heuristic formula. This means that the function relating the person and the environment to behavior may vary on a case-by-case basis (Furr & Funder, in press).

A general practical framework for Lewin's equation can be proposed by combining experimental research design with latent variable mixture modeling (Muthén, 2001). Figure 1 represents the fundamental problem faced by a social psychologist. In this case, one is interested in testing the effect of a situational variable *E* on the behavior *B*. The node *Z* represents all the other possible causes for both *E* and *B*, not directly measured in the study.

Therefore, *Z* represents both dispositional variables as well as other situational variables, less relevant for the study.

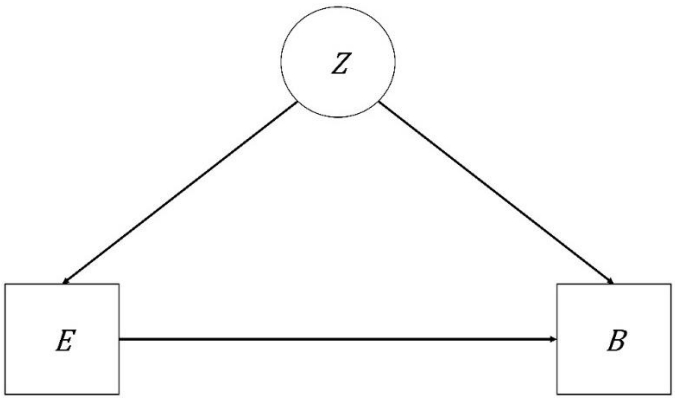


Figure 1. Representation of the fundamental problem.

This type of problem is a common, and somewhat simple, example studied in the causal inference literature (Pearl, 2000). It is a known fact that, to measure the effects of *E* on *B* for this type of problem, it is sufficient and necessary to control for *E* (Pearl, Glymour, & Jewell, 2016, p. 55). Controlling for *E* is a statistical term that can also be interpreted as experimentally manipulating *E*. The procedure of experimentally controlling for *E* causes the arrow between *Z* and *E* to disappear, meaning that both nodes are independent (Pearl, 2000). This is represented in Figure 2. To fully determine the SOFA framework, Z must then be decomposed into two mixture components: a stochastic error (*v*); and the dispositional trait (*P*). A mixture model represents the presence of subpopulations within a sample (Muthén, 2001). In the present case, it more specifically represents the two subpopulations of effects of unobserved variables on behavior. The framework can then be represented by the following equation

$$B = f(E) + F(v, P), \tag{3}$$

where $F(\cdot)$ is the mixture model relating both the effects of the stochastic error and of the dispositional variables on the behavior.

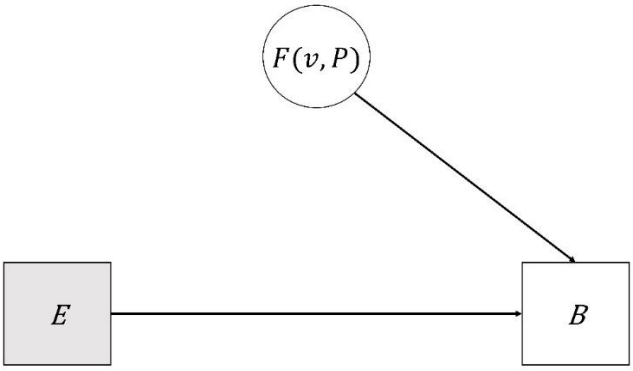


Figure 2. Representation of the SOFA framework.

Assessing dispositions with Stochastic Frontier Analysis

A similar reasoning as the one that originates Eq. 3 was previously employed in economic modeling (e.g., Aigner, Lovell, & Schmidt, 1977). One particular method, known as stochastic frontier analysis (SFA; Aigner et al, 1977), is already recognized as a robust analytical tool to assess the efficiency of firms. The SFA model can be thought as a conjoint measurement model (Luce & Tukey, 1964), as it simultaneously estimates a production function, as well as the efficiency of firms. A production function can be interpreted as a threshold of maximal possible production of outputs, given certain levels of inputs. Efficiency is the distance of particular firms to the production function. Firms above the production function are relatively efficient. Firms below the production function are relatively inefficient (i.e., produce less output than they should, considering the amount of inputs). However, measurement is considered to be imperfect and it can, therefore, with expectancy equals to zero, stochastically vary around the production function.

The SFA model, with output *y* and input *x*, can be represented as

$$y = f(x; \beta) + v - u, \tag{4}$$

where β is the regression coefficient for *x*, *v* is a stochastic component, similar to the error term in a regression model, and *u* is the non-negative inefficiency component. It is possible to see that Eq. 4 is simply a generic regression representation, where the error term is

decomposed in the additive effect $v - u$. For Eq. 4 to be identifiable, independence between x, v and u is assumed. It is also necessary to specify distributional assumptions of the error components (Greene 1990). For the stochastic component v, generally a normal distribution with a mean equaling 0 is used. For the inefficiency component, truncated normal, exponential, or gamma distributions are common choices.

In terms of interpretation, when SFA is extended to SOFA, both assumptions (independence and error terms' distributions) also lead to a reasonable model to implement. As illustrated in Figure 2, experimentally controlling for the situational variable, which corresponds to the x variable in Eq. 4, guarantees that, at least, both error terms are independent from the situational effect. This means that experimentally controlling for the situational variable is necessary for the interpretability of the model estimates. The mixture modeling procedure with a 0 mean normal distribution for the stochastic error and a non-negative component for the dispositional variable is the same as stating that the "optimal" behavior is determined by situational contingencies. For instance, imagine a psychophysical research where the behavior is measured as the average number of errors and the situational variable is the ratio of noise over signal of a series of stimuli. In this case, the dispositional variable, which could be defined as the visual acuity, could not improve the behavior beyond the "optimal" frontier given by the situational contingencies. This is important to note as assuming different distributions for the dispositional variable change the interpretation of the process studied, as well as resulting in different challenges for model identification. In the present study, we are focusing only on the more traditional implementation of the SFA, where dispositional variables are assumed to follow a non-negative distribution (i.e., all values are below the regression function).

In terms of application, Eq. 4 can be extended for different types of SFA models. Traditional linear and log-linear are likely to be the most commonly used SFA models

(Griffin & Steel, 2007). Nevertheless, semi- and nonparametric estimation techniques (Fan, Li, & Weersink, 1996), as well as time-varying inefficiency techniques (Cornwell, Schmidt, & Sickles, 1990; Kumbhakar, 1990) have been considered. In the present study, we focus exclusively on semiparametric models. The reason for this is twofold. First, because linear models are the default in many analyses used in psychology and, therefore, non-linear and non-monotonic relations are usually overlooked (Beller & Baier, 2013). The second reason is that even in traditional IRT literature, some authors have defended that semi- and nonparametric should become the default analyses instead of the more used parametric models, or at least as initial diagnosing tools (e.g., Ramsay, 1991; Sijtsma & van der Ark, 2017).

Fitting an SFA model

In the implementation process of a semiparametric SFA model, it is common to use a two-step approach, as first proposed by Fan et al (1996). In the first step, a semiparametric or nonparametric regression technique is used to estimate the conditional expectation. Previous studies (Ferrara & Vidoli, 2017) applied general additive models (GAMs), kernel, and locally estimated scatterplot smoothing regression (Loess) techniques for estimating this first step. In the second step the stochastic error and the inefficiencies are estimated by maximizing a pseudo-likelihood function (Fan et al, 1996). Figure 3 represents this approach, where the figure to the left represents the result after applying the first step and the figure to the right represents the frontier (line) estimated after the second step.

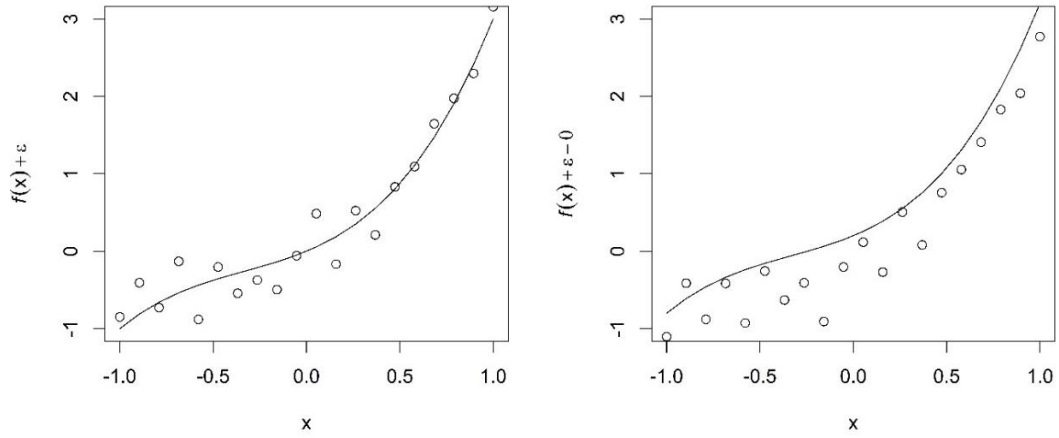


Figure 3. Two step approach for estimating semiparametric SFA models.

In the present study, we developed a Bayesian semiparametric SFA model, following similar approaches as did Griffin and Steel (2004) and Tsionas and Mallick (2019). The model starts with observed values of continuous experimental manipulations x_j and observed values of continuous behavioral responses y_j. The values of the experimental manipulations are then transformed using B-spline like basis. To do so, k knots are defined using the range of possible values of x_j, resulting in values m_k for each knot. These knots are subtracted from x_j exponentiated to $\phi_{0k} + 1$, with this difference divided by the standard deviation of x, σ_x. The k knots are weighted by their regression coefficients β_k and summed over, as well as with the intercept, β_0, and the estimate of the dispositional trait, u_j. This sum represents the expected average response, μ_{yi}, for the y_j response, with an error equaling v.

In terms of distributions, y_j is assumed to follow a normal distribution with mean μ_{yi} and standard deviation v, which is assumed a priori to come from a gamma distribution with shape and rate equals to .001. The dispositional trait u is assumed to follow an exponential distribution with rate parameter equals to λ, which a priori follows a gamma distribution with shape and rate equals to .001. The intercept of the regression model is assumed a priori to come from a normal distribution with mean 0 and standard deviation 1. The regression coefficients are assumed a priori to come from a Laplace distribution with mean 0 and scale 1. This Laplace prior was used since it can be considered as a Bayesian implementation of the

LASSO regression (Park & Casella, 2008). The exponent for the basis, ϕ_{0k}, is drawn from a binomial distribution with bias equals to ζ and maximum possible degree equals to p. The bias is draw from a beta 1-1 distribution, while p is set by the researcher. Traditional B-spline commonly fixes the exponent for the basis to be equal to 3 (De Boor, 1972; Eilers & Marx, 1996). In our Bayesian approach, the model seeks to estimate what should be the best exponent for the given data. As we use both a B-spline like basis and the Laplace prior for the regression coefficients, our Bayesian model can be considered to be a combined GAM/LASSO regression, with an adaptive step for the basis' exponent.

The Bayesian model in our SOFA approach is shown in the net representation in Figure 4, following Lee's (2008) graphical standards. The observed variables are represented by shaded nodes and the unobserved variables are represented by unshaded nodes. Discrete variables are represented by square nodes, while continuous variables are represented by circular nodes. Stochastic variables are represented by single-bordered nodes, and deterministic variables are represented by double-bordered nodes. Finally, encompassing plates are used to denote independent replications of the graph structure within the model.

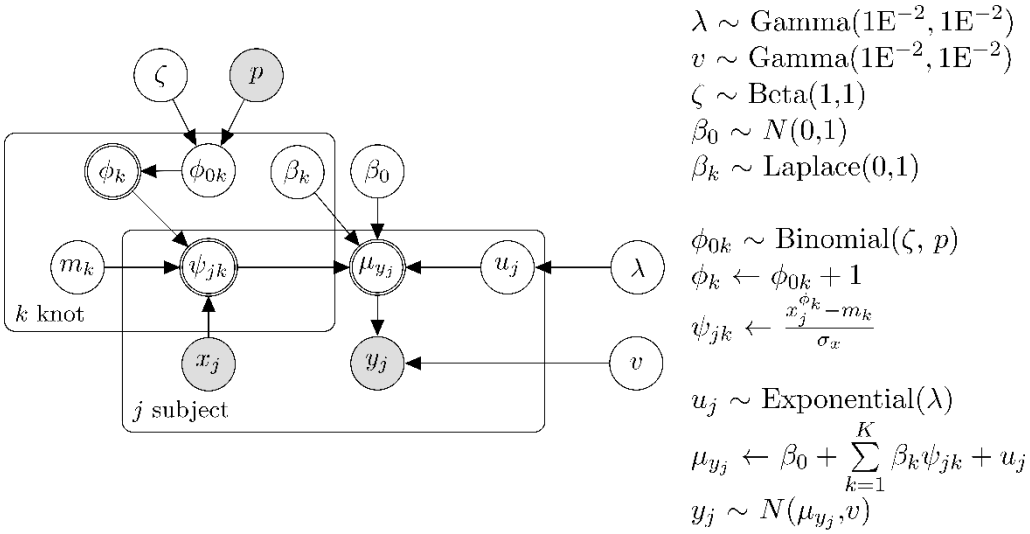


Figure 4. Bayesian implementation of a situational optimization function analysis (SOFA)

Construct validity by joint modeling

A relevant matter in psychometric research is how to demonstrate that a given psychometric tool is really measuring what it aims to measure; this is known as validity research (Borsboom, Mellenbergh, & van Heerden, 2004; Cronbach & Meehl, 1955). In traditional psychometrics, the content of test items and the correlations between tests with similar content and aim are known as content validity and convergent validity, respectively (Messick, 1989). In the SOFA approach, in terms of content, there is little general interpretation one can give to a specific experimentally controlled situation. For instance, if an experimental setting of a game favors competition over cooperation, then it is reasonable to assume that individuals who cooperate more have some dispositional trait that leads them to behave against the situational pressure (e.g., Goto, 1996). Nevertheless, what exactly is this dispositional trait? Answering this question is the aim of validity research on the SOFA approach, as well as in traditional psychometric approach.

One possible alternative for assessing validity in this context would be to use joint modeling (Turner et al, 2013). Joining different types of modeling in a single hierarchical framework allows for data with different sources to influence each other parameters. Turner et al (2013) used joint modeling for constraining parameters of a model from behavioral data from a neuroimaging model. In the present application, the idea is to combine both the SOFA approach with traditional psychometrics. For instance, the data from a questionnaire could be modeled by a two-parameter logistic item response model (2PLM; Swaminathan & Gifford, 1985) and the data from the experiment could be modeled by our Bayesian implementation. Then, estimates of the dispositional traits from both models could be correlated, as it is similarly done in traditional psychometric analysis of convergent validity (Carlson & Herdman, 2012).

In traditional psychometric analysis, convergent validity can be explored best using structural equation modeling (Raines-Eudy, 2000). This approach is normally preferred over using correction for attenuation of correlating scores estimated from two different tests or questionnaires (Osborn, 2003). This is so as structural equation modeling has a natural correction for the measurement error, giving more reliable estimates of the correlations and factor loadings (Bagozzi, 1981). In this sense, joint modeling can be thought of an extension of structural equation modeling, but applying different measurement models than the factor analysis model commonly assumes in this context (Pilati & Laros, 2007). This allows for much more complex types of models and theories to be estimated and tested.

Turner et al (2013) proposed using the multivariate normal distribution to joint modeling parameters of interest. This approach needs no additional modifications if each parameter is indeed better described by a normal distribution. In the present application, if questionnaire/test data are to be used to be joint modeled with our Bayesian SOFA model, then the normal distribution fits well this type of data and model (i.e., the 2PLM). For the Bayesian SOFA model, however, the dispositional trait is better modeled using a non-negative distribution; we proposed the use of the exponential distribution. To avoid this limitation, one can use what is called copula dependence (Genest & MacKay, 1986). A copula is a multivariate cumulative distribution function (CDF). Cumulative distribution functions are always uniformly distributed and, therefore, the marginals of a copula can be modeled using a uniform distribution (Embrechts & Hofert, 2013). From this fact, any type of dependence between continuous variables can be modeled using copulas and then be converted to the appropriate distribution using its marginals (Colonius, 2016). One easy way for fitting copulas is to use the multivariate normal distribution to draw random values for the variables or parameters of interest, calculate the marginals' CDFs and then apply the quantile

function to convert the values to the most appropriate targeted distribution for each variable or parameter (Meyer, 2013).

Simulation study

Method

Simulated data were created by manipulating the sample sizes (100, 250 or 500 data points) for 100 iterations. These data points were generated using one of these three functions (quadratic function, power function and a trigonometric function, respectively) or data generating processes (DGPs):

$$y = -x^2 + x + v + u; \tag{5}$$

$$y = 1.5^x + v + u; \tag{6}$$

$$y = x\sin(\pi x) + x\cos(\pi x) + v + u. \tag{7}$$

The variables *x*, *v* and *u* were drawn from a multivariate normal distribution with mean zero, variances equal to 1 and covariances equal to 0. This was done to guarantee that these values would be uncorrelated, as assumed by the SFA and SOFA models. Copula CDF transformation was used to convert the distribution of the dispositional trait, *u*, to a standardized truncated normal distribution, with lower bound equals to 0. Equations 5, 6 and 7, with *v* and *u* excluded, are represented from left to right in Figure 5.

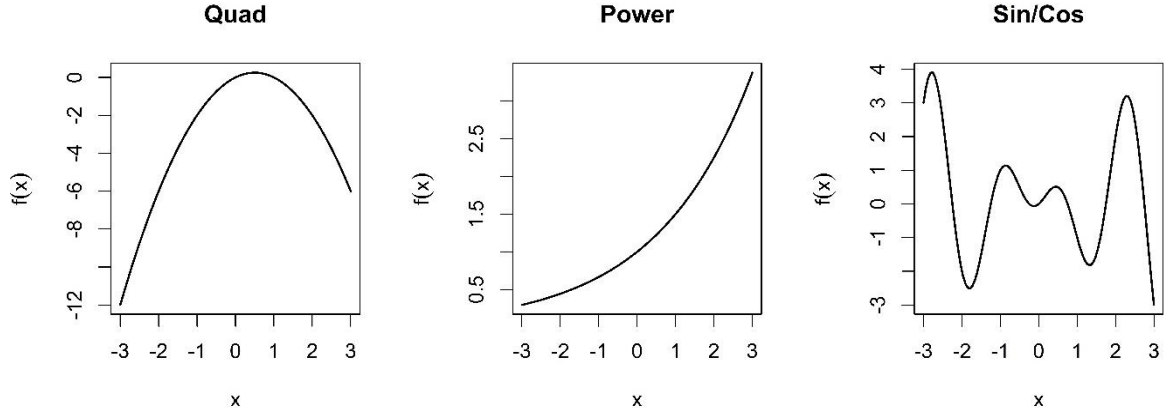


Figure 5. DGPs' functions used for testing the models' performance.

The simulated data were analyzed with four models: GAM-SFA; Kernel-SFA; Loess-SFA; and Bayesian-SOFA. To compare the performance of the models, three measures of accuracy were used. The first measure was the Spearman correlation between the true values of *u* and its estimated values. This measure was used since it is expected that the latent variables and their estimates are monotonic related (Junker & Sijtsma, 2000). The next two measures used were the mean absolute error (MAE) and the root-mean squared error (RMSE) between the true values of *u* and their estimated values. When the values of MAE and RMSE are larger than zero and equal, it means that the bias is equal in all *u*'s of the scale. If the bias is larger for some values of *u* in comparison to others, then RMSE will be larger than MAE. The most effective model will be the one with correlations closest to 1, and MAE and RMSE the most similar and closest to 0.

All the simulations and data analyses were conducted using the R software (R Core Team, 2019) and are available from the corresponding author upon request. To fit the GAM-SFA, the Kernel-SFA and the Loess-SFA the semsfa package was used (Ferrara & Vidoli, 2018). To fit the Bayesian-SOFA, the model was implemented using the JAGS language and software (Plummer, 2003), interfacing with R by means of the jagsUI package (Kellner, 2019). We used the gam function from the mgcv package (Wood, 2012) to estimate the residuals of the additive regressions used to measure the error of the bias.

Results

We first compare the overall performance of all the methods, displayed in Table 1. Per column, the best performances are in bold. Table 1 shows that Bayesian-SOFA and Kernel-SFA models have the same level of correlations with the true scores, followed by the GAM-SFA and the Loess-SFA model. In terms of the mean absolute error (MAE), the Bayesian-SOFA model has the smallest value, followed by the Kernel-SFA model. The GAM-SFA

and the Loess-SFA model have very similar values of MAE. In terms of the root-mean squared error (RMSE), the same pattern found with MAE is followed.

Table 1
Overall performances of each method, measured by Spearman correlations, MAE and RMSE.

Method	Correlation	MAE	RMSE
General Additive Models (GAM-SFA)	.915	.875	1.107
Kernel smoothing (Kernel-SFA)	**.924**	.843	1.063
Locally estimated scatterplot smoothing (Loess-SFA)	.911	.876	1.109
Bayesian Implementation (Bayesian-SOFA)	**.924**	**.474**	**.647**

Note. MAE = mean absolute error. RMSE = root mean-squared error.

The effect of sample size on the accuracy of the models can be observed in Table 2. This time, bolded numbers are used to evaluate the effect of sample size within each method. For the three frequentist methods (GAM, Kernel and Loess), the same pattern was found: Spearman correlations with the true scores increase as the sample size increase. MAE and RMSE, however, are best with a medium sample size (250 data points) than when compared with the large sample size (500 data points). For the Bayesian model, it is possible to see that increasing the sample size improves over all the accuracy measures.

Table 2
Performances of each method, measured by Spearman correlations, MAE and RMSE, compared by sample size.

Method	Sample size	Correlation	MAE	RMSE
General Additive Models (GAM-SFA)	100	.857	1.036	1.299
	250	.933	**.786**	**.994**
	500	**.954**	.804	1.030
Kernel smoothing (Kernel-SFA)	100	.884	.931	1.147
	250	.933	**.785**	**.992**
	500	**.954**	.813	1.050
Locally estimated scatterplot smoothing (Loess-SFA)	100	.854	1.000	1.240
	250	.929	**.805**	**1.023**
	500	**.952**	.824	1.064
Bayesian Implementation (Bayesian-SOFA)	100	.892	.451	.594
	250	.929	.513	.705
	500	**.950**	**.457**	**.643**

Note. MAE = mean absolute error. RMSE = root mean-squared error.

The effect of the data generating process (DGP) on the accuracy of the models can be observed in Table 3. As was the case in Table 2, bolded numbers are used to evaluate the effect of the conditions (in this case, DGP) within each method. This time, the same pattern was observed for all the methods: Spearman correlations, MAE and RMSE, were best for the monotonic condition (Eq 6), followed by the quadratic condition (Eq 5), and worse for the trigonometric function.

Table 3
Performances of each method, measured by Spearman correlations, MAE and RMSE, compared by DGP.

Method	DGP	Correlation	MAE	RMSE
General Additive Models (GAM-SFA)	Equation 5	.914	.878	1.113
	Equation 6	**.927**	**.809**	**1.005**
	Equation 7	.902	.939	1.205
Kernel smoothing (Kernel-SFA)	Equation 5	.927	.826	1.041
	Equation 6	**.932**	**.791**	**.985**
	Equation 7	.912	.913	1.163
Locally estimated scatterplot smoothing (Loess-SFA)	Equation 5	.919	.842	1.061
	Equation 6	**.920**	**.840**	**1.052**
	Equation 7	.894	.949	1.223
Bayesian Implementation (Bayesian-SOFA)	Equation 5	.927	.500	.695
	Equation 6	**.941**	**.298**	**.399**
	Equation 7	.903	.623	.847

Note. MAE = mean absolute error. RMSE = root mean-squared error.

Empirical example

For illustration purposes, the present empirical example is using the data of a study that was carried out to measure cooperation/competition dispositions. In experimental social psychology, social dilemmas are games used to evaluate how people choose and are influenced by the context to behave in a particular manner (Van Lange, Joireman, Parks, & Van Dijk, 2013). One of the most famous games is the iterative prisoner dilemma. This game

was chosen as the optimal behavior is to always compete. Therefore, there is a clear expected effect of the situation that can be modeled by the SOFA approach.

Method

Participants. Data from 63 participants, with an average age of 22.7 years and 50.8% of them woman, were collected. No other sociodemographic data were collected.

Instruments. Two instruments were used: an iterative prisoner dilemma's game, and the cooperation-competition scale (Coop-Comp Scale; Teixeira, Iglesias, & Castro, 2010). In the iterative prisoner dilemma two individuals decide to compete or to cooperate in 10 rounds. For the present implementation, the second player was a simulated participant. The probability of choosing between cooperation or competition was drawn from a standard uniform distribution. The behavior of the simulated player was the experimental condition. The Coop-Comp Scale is composed by 11 phrases to which the participants have to indicate to what extent they agree, ranging from 1(totally disagree) to 5 (totally agree). Competition-related items were inversed so higher scores in all items represented a more cooperative attitude.

Procedures. All the participants were invited from mailing lists and social networks' groups of large universities. The study started with a brief explanation on its objectives, as well as a declaration of intention of the participant to really participate in the study. After declaring being interested in participating, the participant had to read a brief description of the iterative prisoner dilemma and to start the game. The game existed of 10 rounds: after each round the participant had to choose between cooperating or competing. After finishing the game, the participant responded the Coop-Comp Scale and finally answered the sociodemographic questions.

Data analysis. For initial steps of data cleansing, we kept only participants who responded all the questions and also excluded individuals with a biased response pattern on the Coop-Comp scale (e.g., responding everything with 1). A second step of cleaning was to reverse negative items and apply item factor analysis (Wirth & Edwards, 2007). This analysis showed that only item 11 from the Coop-Comp Scale had a factor loading below .30 and, therefore, it was deleted from the rest of the analyses. After the process of data cleansing, we used the joint modeling approach. Data from the cooperation-competition scale was modeled using a Bayesian 2PLM, with discriminations constrained to be positive. Data from the iterative prisoner dilemma's game was modeled with the Bayesian-SOFA approach, with higher values of dispositional trait representing the tendency to compete. We used four different analyses for assessing the quality of the scale. First, we estimated the parameters of the Bayesian-SOFA model and of the 2PLM separately. Then, the aptitude estimate of the 2PLM was correlated with the dispositional estimate of the Bayesian-SOFA model. For the second model, both models were modeled jointly, with a multivariate normal copula used for coupling the aptitude estimates of both the 2PLM and the Bayesian-SOFA model. The last two analyses were conducted forcing the estimates to be completely uncorrelated and perfectly correlated, respectively.

Five fit indices were used to compare the analyses. The fit index was the correlation, and the corresponding highest density intervals, between the estimate made with the 2PLM and the estimate made with the SOFA Bayesian model. Then, the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Linde, 2014) was used to measure to quality of the fit of the joint models; smaller values are best. We also used a similar procedure as the one proposed by Lewandowsky and Farrell (2010) to calculate the DIC difference, likelihood ratio (LR) and weight (*w*) of the models. The DIC difference represents how many units the compared model is larger than the best model; closer to 0 is better. The LR represents how

many times the compared model is as good as the best model; closer to 1 is better. Finally, *w* is the posterior probability of the models; values closer to 1 are better than values closer to 0.

Results

The results from the empirical example are presented in Table 4. The negative correlations found in the first two models, which allowed the correlation to be freely estimated, show that both measurements give similar results. Higher scores in the coop-comp scale characterize a person with more cooperative attitudes, while scores closer to 1 in the Bayesian SOFA model characterize someone with a competitive disposition. Nevertheless, the results show that the joint model has a stronger estimate when compared to the correlations made with separately estimated parameters. The strong negative correlation in the joint model (–.85 [–.99, –.53]) indicates that the coop-comp attitudes is what makes people choose not to compete in a competitive inducing situation, in a similar fashion to traditional convergence validity.

It is also possible to see that the model that best fits the data, according to the deviance information criterion (DIC), is the joint modeling of coop-comp scale and the SOFA approach (1,558.93). Evaluating the LR, it is possible to see that even that second best model is considerably worse (5.77E–11) than the joint model. The *w* shows that the joint model is almost the exclusive model in the capacity of explaining the data at hand, with a probability basically equals to 1. Therefore, the models that forces no correlation and perfect correlation between the dispositional estimates of the coop-comp scale and the iterative prisoner dilemma do not provide a relatively good fit to data, when compared with the joint model.

Table 4
Different procedures for estimating construct validity.

	Correlation	**DIC**	**ΔDIC**	**LR**	***w***
Separately estimated	–.44 [–.65, –.20]	-	-	-	-
Joint modeling	–.85 [–.99, –.53]	**1,558.93**	0	1.000	1.000

No correlation	0	1,606.08	47.15	5.77E–11	5.77E–11
Perfect correlation	1	1,622.63	63.70	1.47E–14	1.47E–14

Notes. DIC = deviance information criterion. **Δ**DIC = variation of DIC. LR = likelihood ratio. w = posterior probability of the models.

Discussion

The present study had three aims: (1) to develop a Bayesian implementation of situational optimization function analysis (SOFA); (2) compare the Bayesian implementation of SOFA with three other Maximum Likelihood-based approaches; and (3) use real data to illustrate how construct validity can be conducted in the SOFA framework using joint modelling. We showed that the SOFA framework, at least in the current Bayesian implementation, is ideal for situations where an expected optimal or ideal behavior is expected, given situational contingencies. Research paradigms such as the ideal observer (Kuss, Jäkel, & Wichmann, 2005) or games with Nash equilibrium (Kalai & Lehrer, 1993) are good examples of possible designs where SOFA can be used to test hypotheses.

The results of our simulation study indicate that the Bayesian implementation of the SOFA approach can outperform more traditional Maximum-Likelihood based SFA models in terms of mean absolute error (MAE) and root-mean squared error (RMSE) between the true scores and their estimated values. It has to be acknowledged, however, that the correlation between the true scores and their estimated values are quite similar. Because MAE does not increase with the variance of bias but RMSE does (Willmott & Matsuura, 2005), our results allow to conclude that more extreme values of the dispositional trait are especially biased when estimated with the Maximum-Likelihood based models. The better performance by the Bayesian implementation of the SOFA framework is probably due to the adaptive GAM/LASSO regression characteristic of the non-parametric regression in the model. Nevertheless, caution should be taken in the sense that the great degree of flexibility of the adaptive GAM/LASSO regression can lead to overfitting (Wood, 2004).

From our empirical example, we have two major outcomes. First that, as in traditional structural equation modeling, the joint modeling of two measurement models increase the correlation between the latent variables, due to the correction for attenuation of the correlations (Bagozzi, 1981; Osborne, 2003). The second outcome may be of special interest to researchers in social dilemmas, as our results present evidence for a strong effect of cooperation-competition attitudes behavior. Of course, our study was only an empirical example and did not test possible correlations with time perspective and values (Milfont & Gouveia, 2006), social norms (Thøgersen, 2008), conversation (Sally, 1995), or other factors important for the study of cooperation-competition behaviors in social dilemmas. Nevertheless, our model and results allow to review hypotheses in this area from a new perspective.

Despite the positive results of our study, one major limitation should be pointed out. This limitation is the fact that, for identifiability reasons, all models tested in the present study assume that the dispositional trait can only have positive values. In other words, it can only explain deviations below the estimated trend function, while deviations above the trend function are considered to be stochastic errors. If only situations with optimal behavior are taken into account, this can be considered as a reasonable assumption. However, researchers are interested not only in studies where an optimal behavior can be clearly defined. Using combined moving average and Gaussian mixture models may be a way of surmounting this limitation (e.g., Yu, Chen, Mori, & Rashid, 2013).

As a final note, it is important to explicitly locate the SOFA approach in the psychometric literature. The success of the factor analytical and item response paradigms (Rust & Golombok, 2014) are related to their easiness of understanding and generality of application. For instance, exploratory factor analysis is widely used for validation research on any type of data form a questionnaire or test, no matter if it is an intelligence or a personality

test, for instance (Thompson & Daniel, 1996). This can be seen as both a weakness and a strength of the factor analytical procedure, as it is a consequence of a lack of theoretical assumptions on the underlying psychological process (Sijtsma, 2012). The SOFA approach is also pretty atheoretical. Nevertheless, because models developed from it need to be applied to experimental data, more objective interpretations and straightforward tests of theories are easier to make.

References

Aigner, D., Lovell, C. K., & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, *6*(1), 21-37.

Bagozzi, R. P. (1981). Evaluating structural equation models with unobservable variables and measurement error: A comment. *Journal of Marketing Research, 18*, 375-381.

Beller, J., & Baier, D. (2013). Differential effects: Are the effects studied by psychologists really linear and homogeneous? *Europe's Journal of Psychology, 9*(2), 378-384.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061-1071.

Carlson, K. D., & Herdman, A. O. (2012). Understanding the impact of convergent validity on research results. *Organizational Research Methods*, *15*(1), 17-32.

Colonius, H. (2016). An invitation to coupling and copulas: With applications to multisensory modeling. *Journal of Mathematical Psychology, 74*, 2-10.

Cornwell, C., Schmidt, P., & Sickles, R. C. (1990). Production frontiers with cross-sectional and time-series variation in efficiency levels. *Journal of Econometrics, 46*(1-2), 185-200.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302.

De Boor, C. (1972). On calculating with B-splines. *Journal of Approximation Theory, 6*(1), 50-62.

Eid, M. E., & Diener, E. E. (2006). *Handbook of multimethod measurement in psychology*. New York: American Psychological Association.

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with *B*-splines and penalties. *Statistical Science, 11*(2), 89-102.

Embrechts, P., & Hofert, M. (2013). A note on generalized inverses. *Mathematical Methods of Operations Research*, *77*(3), 423-432.

Fan, Y., Li, Q., & Weersink, A. (1996). Semiparametric estimation of stochastic production frontier models. *Journal of Business & Economic Statistics, 14*(4), 460-468.

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.

Ferrara, G., & Vidoli, F. (2017). Semiparametric stochastic frontier models: A generalized additive model approach. *European Journal of Operational Research*, *258*(2), 761-777.

Ferrara, G., & Vidoli, F. (2018). semsfa: Semiparametric estimation of stochastic frontier models. R Package retrieved from *https://CRAN.R-project.org/package=semsfa*.

Furr, R. M., & Funder, D. C. (in press). Persons, situations, and person-situation interactions. In O. P. John & R. W. Robins (Eds.), *Handbook of personality: Theory and research* (pp. 1-42). New York: Guilford.

Genest, C., & MacKay, J. (1986). The joy of copulas: bivariate distributions with uniform marginals. *The American Statistician*, *40*(4), 280-283.

Goto, S. G. (1996). To trust or not to trust: Situational and dispositional determinants. *Social Behavior and Personality: An International Journal*, *24*(2), 119-131.

Greene, W. H. (1990). A gamma-distributed stochastic frontier model. *Journal of Econometrics*, *46*(1-2), 141-163.

Griffin, J. E., & Steel, M. F. (2004). Semiparametric Bayesian inference for stochastic frontier models. *Journal of Econometrics*, *123*(1), 121-152.

Griffin, J. E., & Steel, M. F. (2007). Bayesian stochastic frontier analysis using WinBUGS. *Journal of Productivity Analysis*, *27*(3), 163-176.

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, *24*(1), 65-81.

Kalai, E., & Lehrer, E. (1993). Rational learning leads to Nash equilibrium. *Econometrica: Journal of the Econometric Society*, 1019-1045.

Kellner, K. (2019). jagsUI: A Wrapper Around 'rjags' to Streamline 'JAGS' Analyses. R Package retrieved from *https://CRAN.R-project.org/package=jagsUI*.

Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *Journal of Vision*, *5*(5), 478-492.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*(1), 1-15.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. New York: Sage.

Lewin, K. (1936). *Principles of topological psychology*. New York: McGraw-Hill.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*(1), 1-27.

McAdams, D. P. (1997). A conceptual history of personality psychology. In R. Hogan, J. Johnson & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 3-39). New York: Academic Press.

McDonald, R.P. (1999). *Test theory: A unified treatment*. New York: Psychology Press.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*(2), 103-115.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Meyer, C. (2013). The bivariate normal copula. *Communications in Statistics-Theory and Methods*, *42*(13), 2402-2422.

Milfont, T. L., & Gouveia, V. V. (2006). Time perspective and values: An exploratory study of their relations to environmental attitudes. *Journal of Environmental Psychology*, *26*(1), 72-82.

Muthén, B. O. (2001). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 21-54). New York: Psychology Press.

Osborne, J. W. (2003). Effect sizes and the disattenuation of correlation and regression coefficients: lessons from educational psychology. *Practical Assessment, Research & Evaluation*, *8*(11), 1-5.

Park, T., & Casella, G. (2008). The Bayesian LASSO. *Journal of the American Statistical Association*, *103*(482), 681-686.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge: MIT press.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. New York: John Wiley & Sons.

Pilati, R., & Laros, J. A. (2007). Structural Equation Modeling in psychology: Concepts and applications. *Psicologia: Teoria e Pesquisa*, *23*(2), 205-216.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, No. 125.10).

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Raines-Eudy, R. (2000). Using structural equation modeling to test for differential reliability and validity: An empirical demonstration. *Structural Equation Modeling*, *7*(1), 124-141.

Rust, J., & Golombok, S. (2014). *Modern psychometrics: The science of psychological assessment*. London: Routledge.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*(4), 611-630.

Sally, D. (1995). Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society*, *7*(1), 58-92.

Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory & Psychology*, *22*(6), 786-809.

Sijtsma, K., & van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 137-158.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(3), 485-493.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, *50*(3), 349-364.

Thøgersen, J. (2008). Social norms and cooperation in real-life social dilemmas. *Journal of Economic Psychology*, *29*(4), 458-472.

Teixeira, L. A. G., Iglesias, F., & Castro, R. (2010). *Mensurando atitudes em dilemas sociais: Versão brasileira da Escala de Cooperação e Competitividade [Measuring attitudes in social dilemmas: Brazilian version of the Cooperation and Competitiveness Scale]*. Unpublished manuscript.

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement*, *56*(2), 197-208.

Tsionas, M. G., & Mallick, S. K. (2019). A Bayesian semiparametric approach to stochastic frontiers and productivity. *European Journal of Operational Research*, *274*(1), 391-402.

Turner, B. M., Forstmann, B. U., Wagenmakers, E. J., Brown, S. D., Sederberg, P. B., & Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, *72*, 193-206.

Van der Linden, W. J., & Hambleton, R. K. (2013). *Handbook of modern item response theory*. New York: Springer.

Van Lange, P. A., Joireman, J., Parks, C. D., & Van Dijk, E. (2013). The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes*, *120*(2), 125-141.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79-82.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological Methods*, *12*(1), 58-79.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*(467), 673-686.

Wood, S. N. (2012). mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. R Package retrieved from *https://CRAN.R-project.org/package=mgcv*.

Yu, J., Chen, K., Mori, J., & Rashid, M. M. (2013). A Gaussian mixture copula model based localized Gaussian process regression approach for long-term wind speed prediction. *Energy*, *61*, 673-686.

IV

A Structure Learning Procedure for Power Chain Graphs

Vithor Rosa Franco[1], Marie Wiberg[2] and Jacob Arie Laros[1]

Affiliations

[1]Post-graduate program of Social, Work and Organizational Psychology, Institute of Psychology, University of Brasília, Brasília, Brazil;

[2]Department of Statistics, Umeå School of Business, Economics and Statistics, Umeå University, Umeå, Sweden;

Abstract

Structural equation modeling is a psychometric analysis technique relying on the presence of a previous latent variable model (measurement model) and a causal model (structural model). In network psychometrics, such as partial correlation network, no latent variable model or pre-specified causal model are necessary. Nevertheless, current procedures analyzing the structure of multidimensional data with both causal and non-causal relations cannot properly deal with complex data patterns encountered in the field of psychology. The main aim of the present study is to develop a procedure of structure learning of power chain graphs (PCGs). The secondary aim of the study is to compare clustering algorithms and causal discovering algorithms designed to learn the structure of the PCGs. This comparison of algorithms is carried out with simulated and with real data. In a number of conditions, we show that our clustering procedure outperform traditional clustering procedures used in psychometrics. The paper ends with a discussion in which practical implications of this study are reviewed and suggestions for future studies are given.

Keywords: Psychometrics, network modeling, power chain graph, Monte Carlo simulation.

A Structure Learning Procedure for Power Chain Graphs

Structural equation modeling (SEM; Mair, 2018) is frequently used in psychometrical research as an analytical tool to evaluate criterion validity or the causal and predictive relations between two or more latent variables. SEM relies on previous knowledge on both the measurement and the structural model; therefore, it is seldom used as a causal model in exploratory research (Fried & Cramer, 2017; Hevey, 2018). The interest in probabilistic graph models, that can be considered as an alternative to SEM, is growing in psychology (Epskamp, Borsboom, & Fried, 2018; Epskamp, Rhemtulla, & Borsboom, 2017). The advantage of probabilistic graph models is that they can be used to study data dependencies without relying on latent variables and a previous theoretical structure. There are at least three types of probabilistic graphical models: undirected graphs (UGs); directed acyclic graphs (DAGs); and chain graphs (CGs; Lauritzen, 1996).

UGs are graphical representations of multivariate data with associational relations, usually measured by correlations or partial correlations between variables (Epskamp & Fried, 2018). Each variable is represented by a node and each relation is represented by an edge: —. DAGs are used when the intent is to represent multivariate data with causal relations, usually measured by regression coefficients (Pearl, 2009). In DAGs, relations are represented by arrows: ← or →. Finally, CGs are used when representing multivariate data in which both associational and causal relations are present (Peña, 2018). Due to recent developments on the study of statistical dependence and graph theory (e.g., Koller, Friedman, & Bach, 2009; Lauritzen, 1996; Pearl, 2009), several procedures for automatically learning the structure of graphs have been proposed and some procedures exist for reducing high dimensional graphs (i.e., graphs with a very high number of variables).

Power graphs (PGs; Royer, Reimann, Andreopoulos, & Schroeder, 2008), as represented in Figure 1, can be used to reduce high dimensional UGs. This is accomplished

by extending the notion of nodes and edges to that of power nodes and power edges, respectively. Power nodes are sets of similar nodes, where similar means that variables are highly correlated and that these variables correlate to the same other variables. Power edges are edges used to connect two power nodes, implying that all nodes contained in the first power node are connected to all the nodes contained in the second power node. Royer et al (2008) proposed an algorithm of structure learning for PGs that first find potential power nodes and then maximizes the number of power edges over normal edges.

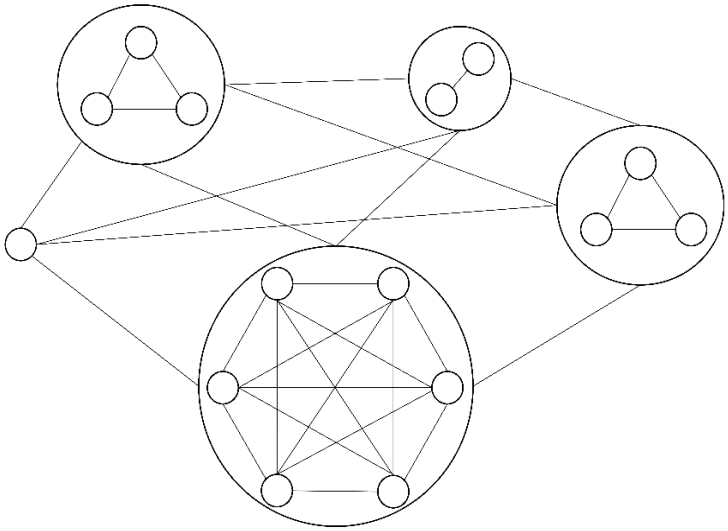


Figure 1. An example of a power graph.

For graphs with causal relations, we are not aware of any other study that has tried to find directed edges on PGs, despite the fact that there are other studies proposing structure learning procedures for CGs (e.g., Drton & Perlman, 2008; Ma, Xie, & Geng, 2008; Peña, Sonntag, & Nielsen, 2014). These procedures for structure learning procedures, nevertheless, do not apply in the present case once they rely on labeled block ordering (Roverato & Rocca, 2006), permitting directed edges within blocks, or not incorporating the community detection step. Therefore, combining both CGs and power graphs, we propose power chain graphs (PCGs) for this end. The main aim of the present study is to develop a procedure of structure

learning of PCGs. The secondary aim of the study is to compare clustering algorithms and causal discovering algorithms that can be used for learning the structure of the PCGs.

The rest of this paper is structured as follows. In the first section, probabilistic graph theory is presented, focusing on characteristics of PGs and CGs. We then present the formal definition of PCGs and a procedure for learning its structure. Next, we present two clustering procedures traditionally used in psychology, which are used in the present study as benchmarks for the proposed procedure. The fourth section is presenting the theory behind algorithms that aim to find causal relations in observational data, and ways to tune our procedure further for the causal relations in the data. The fifth and the sixth sections are dedicated to simulation studies and examples with real data, respectively. The paper ends with a discussion and some concluding remarks.

Probabilistic graph theory, PGs and CGs

A probabilistic graphical model (Koller, Friedman, & Bach, 2009) is a graphical representation of a set of distributions that satisfies a set of conditional independence relations. Probabilistic graphical models are composed of vertexes (V), or nodes, and at least one type of connection between the vertexes: edges (E) or arrows (A). When data are high dimensional, the graph $G(V, E)$ can be simplified as a power graph $P(V', E')$, where the vertexes V are summarized as power vertexes V', using cluster or network motif analysis, and the edges E are summarized as power edges E' (Royer et al, 2008). For an adequate transition from $G(V, E)$ to $P(V', E')$, two conditions should be met. First, the power node hierarchy condition, which establishes that any two power nodes are either disjoint, or one is included in the other. The second is the power edge disjointness condition, which states that each edge of the original graph is represented by one and only one power edge (Nenov & Nikolov, 2015).

Groupings in psychological instruments are sometimes interpreted as evidence of a common latent cause (Bagozzi, 2007; Golino & Epskamp, 2017). Nevertheless, Lauritzen and Richardson (2002) used a simple example to show that, when arrows are also present in a graph, edges should only be interpreted as associational relations. For representing this type of structure, CGs, defined as *G*(*V*, *B*), where *B* represents a set of both edges and arrows, should be preferred in a number of cases. For illustration of this point, at the top of Figure 2 we present an observed CG, with the following factorization of the joint density of variables *a*, *b*, *c*, and *d*:

$$a \perp b, \quad a \perp d | \{b, c\}, \quad b \perp c | \{a, d\}. \tag{1}$$

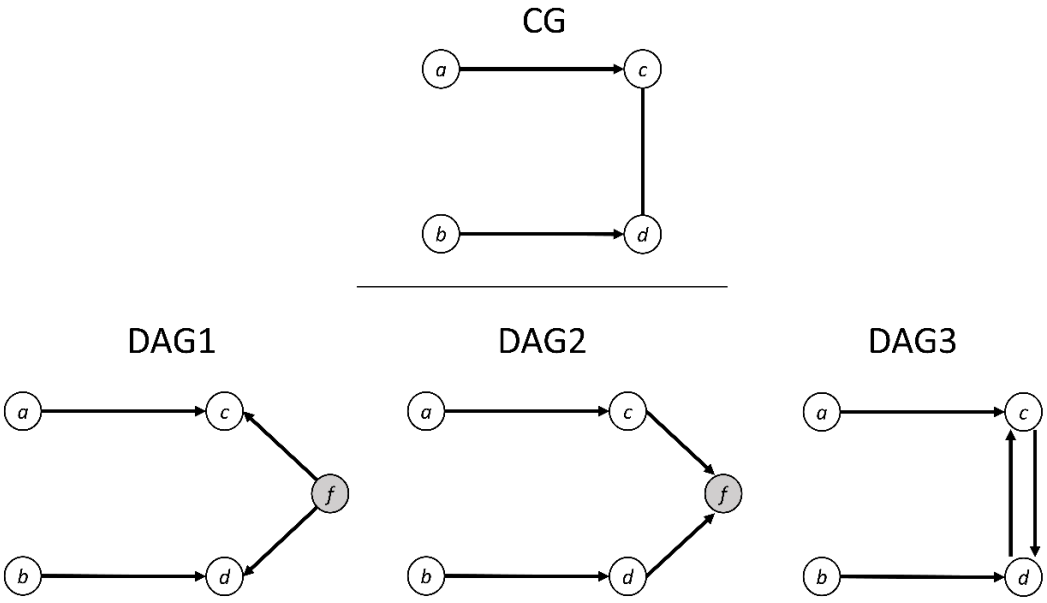


Figure 2. A CG (top) and three different DAGs that have different factorizations.

The DAGs on the bottom, despite being usually used to interpret the undirected edge on the CG, will not attain the same conditional distribution, being the joint density of DAG1, DAG2, and DAG3 represented by, respectively

$$a \perp \{b, d\}, \quad b \perp | \{a, c\}, \quad a \perp d | \{b, c\}, \quad b \perp c | \{a, d\}, \tag{2}$$

$$a \perp d | \{b, c\}, \quad b \perp c | \{a, d\}, \quad a \perp b, \tag{3}$$

$$a \perp b, \quad a \perp b|\{c,d\}, \quad a \perp d|\{b,c\}, \quad b \perp c|\{a,d\}. \tag{4}$$

This shows that CGs are used when correlations are better thought just as correlations, instead of data with common latent causes (DAG1) or common latent effects (DAG2). These interpretations are very commonly used in psychology, usually named as reflexive and formative models of measurement (Howell, Breivik, & Wilcox, 2007).

When structure is high dimensional, CGs will be divided in a path of blocks of variables, known as dependence chains (Wermuth & Lauritzen, 1990). Dependence chains are used to illustrate and properly condition causal effects in CGs because, within blocks, variables have edges, but between blocks there are arrows. Usually, if groups of variables are all predictors of other group of variables, they are considered to be part of the same block, even if the block is a disconnected graph—when there is not a path between every pair of vertices. If instead we define the blocks by clusters of variables, and if it is possible to test or assume the disjointness condition (i.e., to find arrows between blocks), then we have a special type of CG, which we call a PCG, as illustrated in Figure 3. Despite the fact that the factorization of both the CG and the PCG in Figure 3 are identical, the PCG representation allows further inference about the clustering process of the predictor variables and the causal chains as a whole.

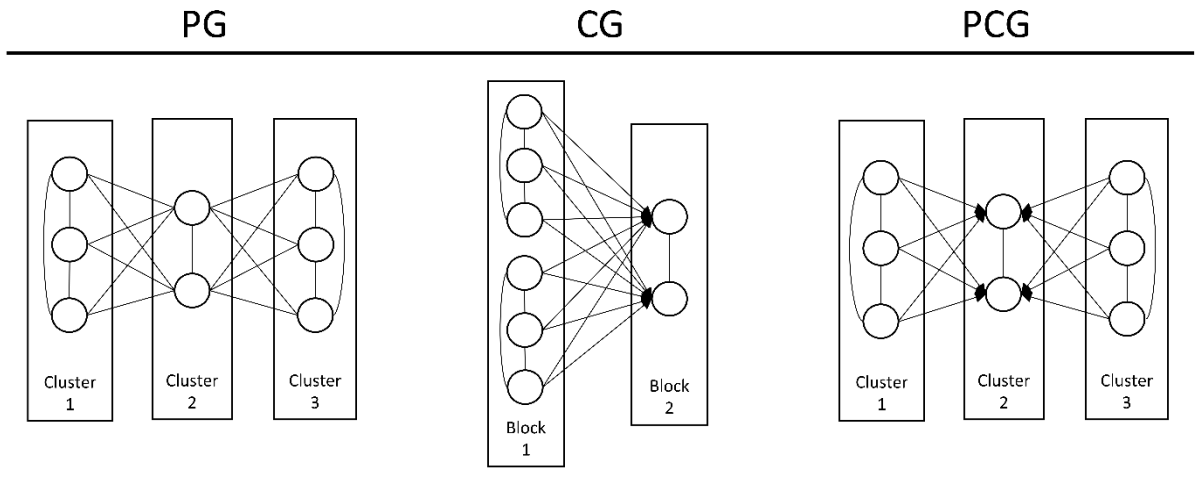


Figure 3. Comparison between dependencies represented with PG, CG and PCG.

Power chain graphs (PCGs)

Combining PGs and CGs, we define power chain graphs (PCGs) as *P*(*V*', *A*'), which is a DAG mapping to *G*(*V*, *B*). V' represents a set of sets which elements are nodes of similar nodes, or power vertexes, and A' represents an arrow set which elements are arrows between V's, or power arrows. Similarity between two nodes v_i and v_j is defined as a high correlation $\rho_{i,j}$ between these nodes and that $\rho_{i,k} \approx \rho_{j,k}$, meaning that the correlations of v_i and v_j with a third node v_k are approximately equal. Therefore, *V*' is but a cluster of variables. In Figure 4 we present an example of a PCG and the CG implied by it.

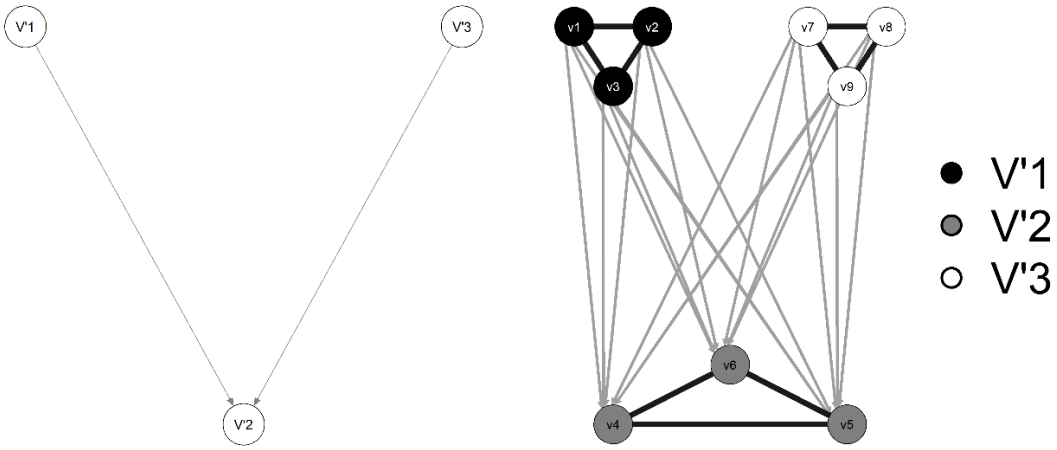


Figure 4. An example of PCG (to the left) and the CG (to the right) implied by it, with different colors for nodes representing different clusters.

It is possible to see that PCG is a DAG, where nodes are clusters of similar variables and the arrows represent a set of arrows. This means that PCGs assume strong connectivity, meaning that every vertex in a "cause" power vertex (clusters 1 and 3; *V*'1 and *V*'3) have an arrow directed to an "effect" power vertex (cluster 2; *V*'2). This characteristic implies that, for learning the causal structure of a PCG, we do not account for interference effects (Peña, 2018). Interference effects happen when a cause has effects on units other than those to which it has an arrow to. For this "default" PCG, which has no missing arrows between causes and effects, the distinction between the proper causal properties is of a secondary

matter. Therefore, we propose that learning PCGs as a two-step approach: first, the clustering of variables is applied; then the directions of the arrows between the clusters are learned.

The clustering procedure we propose departs both from our definition of similar nodes and from the fact proved by Rao (1979) that correlations estimated by maximum likelihood from the same sample are asymptotically distributed as multivariate normal with appropriate mean and dispersion matrix. Using model-based clustering parameterized with finite Gaussian mixture models we can properly find the number of variables' clusters of similar nodes departing from the correlation matrix of the data. In our application of model-based clustering, or correlation Gaussian mixture models (CGMMs), the estimated correlations $p = (p_1, \dots, p_k)$ are assumed to be generated by a mixture model density with *G* clusters:

$$f(p) = \prod_{i=1}^{k} \sum_{l=1}^{G} \tau_l f_l(p_i|\theta_l), \tag{5}$$

where $f_l(p_i|\theta_l)$ is the normal probability distribution with parameters θ_l, and τ_l is the probability of belonging to the l^{th} cluster. The parameters of the model are usually estimated by maximum likelihood using the Expectation-Maximization algorithm (Dempster, Laird, & Rubin, 1977), finding the value of *G* that better fits the data (i.e., the estimated correlation matrix). In multivariate settings, the volume, shape, and orientation of the covariances can be constrained to be equal or variable across clusters, resulting in 14 possible models with different geometric characteristics. A modified version of the Bayesian information criterion can be used to choose the best solution (Fraley & Raftery, 2007).

Polychoric correlation (Olsson, 1979) is the preferred procedure for estimating the correlation matrix used in the CGMM due to the discrete nature of psychological data. Polychoric correlations assume that the data is a logit discretized version of originally normally distributed variables and thus makes it always true that the correlation estimates are originated from normal data (Jöreskog, 1994). This procedure has also been shown to

improve results from other psychometrical approaches, as traditional Pearson correlations underestimate the true correlations between ordinal data (Holgado–Tello, Chacón–Moscoso, Barbero–García, & Vila–Abad, 2010). Nevertheless, caution should be taken as sometimes polychoric correlations may fail to generate proper positive semidefinite matrixes (Holgado–Tello et al, 2010). In this case, it is possible to use one of two alternatives. The first, applied in our simulation study, is to use Higham's (2002) algorithm to find the closest positive definite correlation matrix. The second, applied in our empirical example, is to use Spearman rank correlation, which has a homeomorphism with polychoric correlation (Ekström, 2011). This means that Spearman rank correlation is asymptotically equivalent to a proper positive definite correlation matrix estimated by polychoric correlation.

The clusters identified with this procedure will give which variables have mutualistic relations (i.e., no causal relations). By our definition of a PCG, variables from different clusters can only be connected by arrows and, if one variable of a cluster is connected to a variable in another cluster, all variables in the first cluster are connected to all variables in the second cluster, with the arrows directed the same way. Departing from our procedure for clustering, where correlations are assumed to follow normal distributions, weighted power edges E'_{lhgj} can be calculated by

$$E'_{lhgj} = r\left(\frac{1}{(N_l + N_h)}\sum_{g=1}^{N_l}\sum_{j=1}^{N_h} z_{gj}\right), \tag{6}$$

where N_l and N_h represents the number of vertexes included in the communities *l* and *h*, respectively, z_{gj} is the Fisher-transformed correlation of the *g*th variable of a community *l* with the *j*th variable of a community *h*, and *r*() is equal to

$$r(z') = \frac{\exp(2z') - 1}{\exp(2z') + 1}. \tag{7}$$

This procedure will generate an $G \times G$ averaged correlation matrix that can be used to learn the direction of the power arrows of a PCG, using the PC-stable algorithm (Colombo & Maathuis, 2014). This procedure can return, for instance, the PCG represented in Figure 4. The CG represented in Figure 4 is only implied by the PCG, not tested for its adequacy of the Markov properties of a CG (Lauritzen, 1996). Therefore, for learning a PCG, the procedures shown so far suffice. Nevertheless, for further evaluating the CG implied by the PCG, other causal discovery algorithms should be applied.

Benchmarks for the clustering procedure

From a psychometrical perspective, the most common approach for finding clusters of variables is through exploratory factor analysis (EFA; Fabrigar, Wegener, MacCallum, & Strahan, 1999). In this approach, *p* latent common causes (also known as factors) are assumed to exist for a set of *m* observed variables

$$\mathbf{X} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\varepsilon}, \tag{8}$$

where $\boldsymbol{\Lambda}$ is a $m \times p$ matrix of factor loadings, and $\mathbf{X}$ and $\boldsymbol{\xi}$ are random vectors of length *m* containing indicators and the factors. In this form, the model cannot be estimated in a simple way, thus it is usually reformulated using the observed $m \times m$ correlation matrix $\mathbf{C}$

$$\mathbf{C} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\psi}, \tag{9}$$

where ψ is an $m \times m$ diagonal matrix containing the unique factor variances and $\boldsymbol{\Phi}$ is the correlation matrix of the factors. Factor loadings can then be used to infer which factors are strongly related to which observed variables.

For proper estimation, one must choose beforehand what the value of *p* is; i.e., how many factors there are in the data. Several criteria are used, including the Kaiser criterion and Cattell's scree plot which are probably the most famous (Hayton, Allen, & Scarpello, 2004). Nevertheless, simulation studies have shown that these procedures are biased (Timmerman &

Lorenzo-Seva, 2011). From a factor analysis perspective, parallel analysis (PA; Horn, 1965) is probably one of the most reliable procedures, largely improving over the Kaiser criterion and Cattell's scree plot. PA is carried out by computing the eigenvalues for **C** and drawing a set of random ordered eigenvalues. If the eigenvalue of real data is larger than the random eigenvalue, then the factor is included in the model. The λ eigenvalues can also be adjusted for the sample error-induced inflation (Horn, 1965) by

$$\boldsymbol{\lambda_{Of} - \lambda_{Sf}}, \tag{10}$$

where is $\boldsymbol{\lambda_{Of}}$ the f^{th} eigenvalue of the observed data and is $\boldsymbol{\lambda_{Sf}}$ the corresponding mean eigenvalue of the random eigenvalues. For comparison with our clustering procedure, we will use PA and EFA (PA-EFA) with adjusted eigenvalues.

From a network psychometrics perspective, the exploratory graph analysis (EGA; Golino & Epskamp, 2017) procedure has been show to outperform PA-EFA in a number of conditions (Golino et al, 2018). EGA is a two-step procedure. First, it fits a least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996) regularized network to the data, guarding against overfitting in traditional estimation of the partial correlation matrix Θ by allowing some partial correlations to be exactly zero. Following Friedman, Hastie and Tibshirani (2008), using both Θ and *S*, the empirical covariance matrix, this alleviation of overfitting is achieved by maximizing the log-likelihood

$$\log\det\Theta - \mathrm{tr}(S\Theta) - \rho\|\Theta\|_1, \tag{11}$$

where det denotes the matrix determinant, tr denotes the trace, $\|\Theta\|_1$ is the sum of the absolute values of the elements of Θ and ρ is a tuning parameter. Previous studies showed that choosing the value of the tuning parameter according to the smallest extended Bayesian information criterion (EBIC; Chen & Chen, 2008) of at least 100 models will result in a robust estimation of the true graph (Foygel & Drton, 2010).

The second step of the EGA procedure is to identify the clusters of nodes in the graph estimated with the LASSO procedure. Golino and Epskamp (2017) achieved this by using the walktrap algorithm proposed by Pons and Latapy (2006). The basic idea behind this algorithm is to use Euclidean distances between variables (or nodes) and to generalize these distances to distances between clusters. The algorithm will then try to find the solution that minimizes the distance between variables within the same cluster and maximizes the distance between clusters. The walktrap algorithm is similar to the hierarchical clustering algorithm used by Royer et al (2008) for identifying clusters in PGs. Nevertheless, both procedures are limited by the fact that they depend on a previously estimated sparse UG.

Notwithstanding the fact that the CGMM procedure is more adequate for our definition of similarity, both PA-EFA and EGA can be used in the clustering step of PCGs. The decision between any of these three procedures relies upon how accurate they recover the clusters of variables, as well as the interpretability of the recovered cluster. Previous studies have compared EGA to PA-EFA and showed that, overall, EGA outperforms PA-EFA (e.g., Golino & Epskamp, 2017). Therefore, it is necessary for our procedure, if it is to be used in learning PCGs, to, at least, perform as well as EGA. Another point is that, usually, PA-EFA and EGA are compared in accuracy on data with associational underlying structure, while our procedure is directed for evaluating clustering in mixed data, with both associational and causal relations.

Causal discovery: Theory and CG tuning

One can use structural causal models to find causal relations in correlational data (Pearl, 2009). This idea is based on the fact that different causal paths (i.e., connections) result in different conditional distributions on a set of variables. This can be shown by the factorization of the three fundamental connections, represented in Figure 5: serial connection;

divergent connection; and convergent connection (also known as collider or v-structure). The factorization of the probability distributions of these connections can be expressed, respectively, as serial connection

$$\Pr(X_i)\,\Pr(X_j|X_i)\,\Pr(X_k|X_j)\,; \tag{12}$$

divergent connection

$$\Pr(X_i|X_j)\,\Pr(X_j)\,\Pr(X_k|X_j); \tag{13}$$

and v-structure

$$\Pr(X_i)\,\Pr(X_j|X_i,X_k)\,\Pr(X_k)\,; \tag{14}$$

being it easy to see that Eq 12 is equivalent to Eq 13.

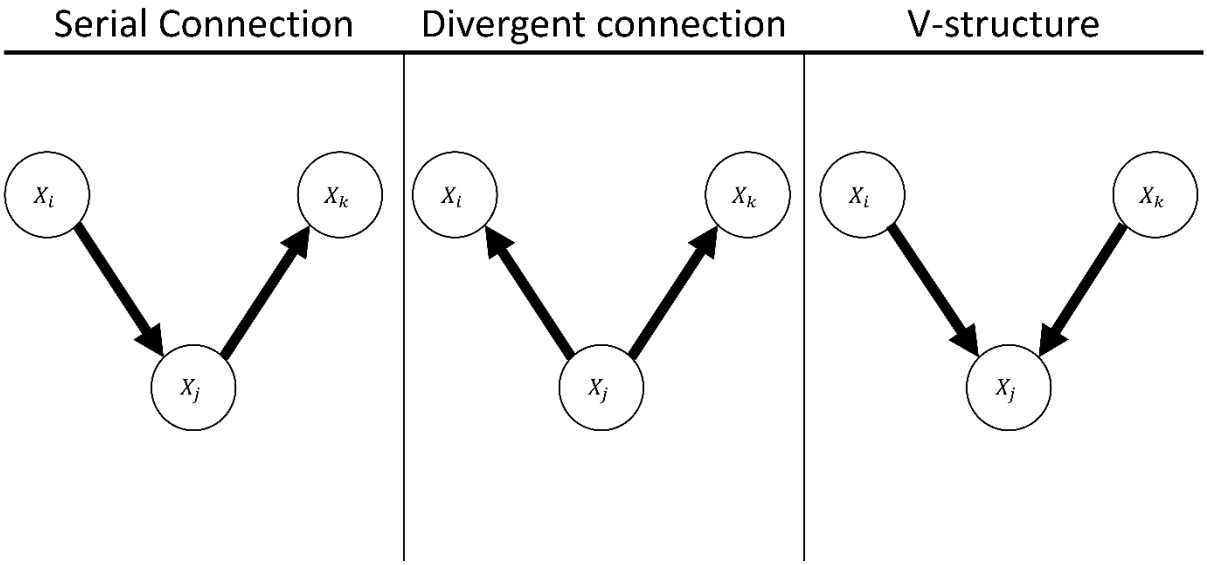


Figure 5. Three fundamental connections between three variables.

The fact that v-structures have a different factorization implies that—assuming non-confounders and no cycles (i.e., mutual causation)—it is possible to identify causal effects even if there are only correlational data, as long as v-structures, as expressed in Figure 5, can be found (Rohrer, 2018). This led to the development of three classes of approaches for learning causal paths (i.e., structure learning) from data (Scutari & Denis, 2015): constraint-based; score-based; and hybrid. Constraint-based algorithms are based on conditional tests of

triads of variables, aiming at finding v-structures and leaving the unidentified edges undirected (Verma & Pearl, 1990). This means that this type of algorithm will return complete partially DAGs instead of DAGs, given that a complete partially DAG can have edges that could be directed to any direction (i.e., an equivalence class consisting of the same v-structures and edges due to model equivalence between divergent and serial connections).

One of the first constraint-based algorithms to be implemented was the PC algorithm proposed by Spirtes, Glymour and Scheines (2000), which we use to learn the arrows in the PCG. The output of the original PC algorithm depends on the order in which the possible v-structures are tested. A simple modification proposed by Colombo and Maathuis (2014), called PC-stable, yields order-independent adjacencies in the skeleton, which is the partial correlations' UG generalized from a DAG. This means that PC-stable finds a UG for the data before trying to direct the edges, reducing the computational expense of testing the possible v-structures due to the size of the graph.

Score-based algorithms, on the other hand, apply heuristic optimization techniques to the problem of structure learning (Russell & Norvig, 2009). This class of algorithms assigns network scores (i.e., a goodness-of-fit index) to possible structures, which are then selected based on how well they fit the data. The greedy equivalence search (GES; Chickering, 2002) and the hill climbing (HC; Daly & Shen, 2007) algorithms use locally optimal choices at several iterations, until a solution is found. They have shown good performance when compared to the PC algorithm or others score-based algorithms. Nevertheless, they do not evaluate the existence of v-structures and, therefore, may be less conservative then constraint-based algorithms.

Hybrid learning algorithms, as the name may suggest, combine both constraint-based and score-based algorithms to trying to overcome the limitations of single class algorithms (Friedman, Peér, & Nachman, 1999; Tsamardinos, Brown, & Aliferis, 2006). One simple

approach for a hybrid learning procedure would be, for example, to learn the skeleton of a DAG by means of any constraint-based algorithm (what is called a restriction phase), and then to direct the edges accordingly to a score-based algorithm (called maximization phase). The Max-Min Hill Climbing (MMHC; Tsamardinos, Aliferis, & Statnikov, 2003) algorithm performs this exact procedure by combining both the PC-stable and the HC algorithms.

Despite the fact that the PC-stable is the best algorithm to learn the PCG, as it can be applied using only the correlation matrix, the other algorithms can be used for "fine tuning" the CG implied by the PCG. This means that other algorithms, or even the PC-stable, can be used for, after learning the structure of PCG, with the implied CG, remove arrows that may not hold true causal relations between nodes from different power nodes. Therefore, this step of "fine tuning", despite not essential for a PCG, can shed a light on how to improve structure learning of CGs (Peña, 2018).

Simulation study

Method

Data generating process (DGP). The DGP demanded four characteristics to properly address our aim: variables should be distributed in groups of similar nodes; different groups should have causal relations; the simulated PCG should be the only element on its equivalence class; and our simulated data should resemble empirical data from psychological research. The first two characteristics are simply characteristics of PCGs, exposed in the introduction, and somewhat unique to the present study. The third characteristic guarantees that, in the absence of unmeasured confounders and measurement error, there is only one best causal model to explain the dependencies in the data. Finally, the fourth characteristic is used to guarantee typical levels of measurement error are present in the data and, therefore, the results from the simulations can be more representative of true contexts of application.

The DGPs were based on three different multidimensional item response models (Liu, Magnus, O'Connor, & Thissen, 2018). Each model reflected a different PCG, as represented on Figure 6. The independent variables, represented by the power vertexes $V'1$, $V'2$, $V'4$, and $V'6$, were drawn from a multivariate normal distribution with correlations and means fixed to zero. The dependent variables were then generated by simply summing the independent variables that directly caused them. An ordinal Rasch model was used to transform the normally distributed variables to an ordinal level, emulating a 5-points Likert scale. The difficulties of items related to different power vertexes were also drawn from a multivariate normal distribution with correlations and means fixed to zero.

DGP1 was chosen because it represents a v-structure, the simplest possible model to be unique in its equivalence class. Both DGP2 and DGP3 were built upon DGP1, increasing the number of nodes, but keeping the same nodes and arrows, as well as the characteristic of being unique in its equivalence class. For instance, *V*'2 causes *V*'3 in all the DGPs. On the other hand, *V*'2 only causes *V*'5 in DGP2 and DGP3. This strategy reflects an important aspect of DAGs (Pearl, 2009): as long as confounders do not change the dependency relations in the graph, they can be unmeasured without loss of efficiency of the causal discovery algorithms. Therefore, even in the absence of some nodes, the algorithm applied to DGP1 should perform at least as well as when compared to its applications to DGP2 and DGP3.

Two other conditions were used: the number of variables per power vertex (or cluster size; 5 or 10) and the sample size (100, 250 or 500). Each of these two conditions was repeated 100 times and they were chosen based on other simulation studies in network psychometrics (e.g., Golino & Epskamp, 2017) and characteristics often found in psychological studies (e.g., Fraley & Vazire, 2014). These conditions were also used to reflect the fourth necessary characteristics of generalization of the findings to real empirical data in psychological research.

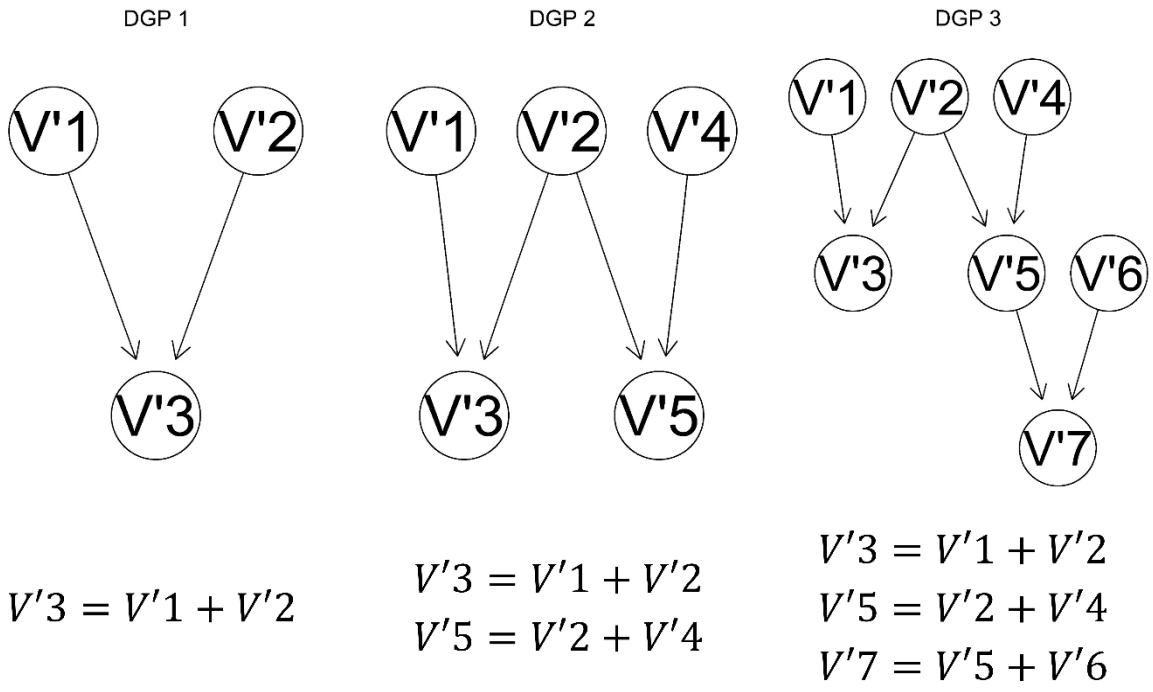


Figure 6. DGPs of the PCGs in the present study.

Performance indexes. For assessing the quality of the clustering procedures, we used both accuracy related indexes (Glaros & Kline, 1988) and cluster comparison indexes (Hennig & Liao, 2013). For comparing the structure learning algorithms we only used accuracy related indexes. The accuracy related indexes were accuracy (Acc), positive predictive value (PPV), and true positive and negative rates (TPR and TNR, respectively). The cluster comparison indexes were the graph adjusted Rand index (GARI; Zhang, Wong, & Shen, 2012), variation of information (VI; Meilă, 2007), normalized mutual information (NMI; Haghighat, Aghagolzadeh, & Seyedarabi, 2011), and whether the clustering procedure had identified the correct number of dimensions (HitND). GARI, VI and NMI are measures of similarity between two data clusterings, but each based on different theories. GARI is an especial type of accuracy measure, specific for clustering solutions. VI is a true metric calculated from distances between two clustering solutions. NMI is a dependence measure calculated from the Kullback–Leibler divergence of two clustering solutions.

Acc, PPV, TPR, TNR and NMI vary between 0 and 1, being values closer to 1 preferred. When we take the average of HitND for the whole simulation, it can be interpreted as the power of the clustering procedures, also varying between 0 and 1, being values closer to 1 preferred as they mean that the procedure has always found the correct number of clusters. GARI can vary from negative values to 1, where negative values meaning that the clusters being compared have completely different structures and 1 that they have the exact same structure. Finally, VI values closer to 0 are preferred as they mean all variables that should be clustered together were properly clustered together.

Implementation. All simulations and data analyses were done in R (R Core Team, 2019). For the community detection part of the simulation, the PA used was the one implemented in the `paran` package (Dinno, 2018) and the EFA implemented in the `psych` package (Revelle, 2019). EGA was estimated using the implementation in the `EGAnet` package (Golino & Christensen, 2019). To be able to start the CGMM procedure, the polychoric correlations were estimated by the implementation in the `psych` package and the Gaussian mixture model for clustering analysis was the one implemented in the `mclust` R package (Scrucca, Fop, Murphy, & Raftery, 2016). VI and NMI were calculated by the implementation in the `fpc` package (Hennig, 2018) and GARI by the implementation in the `loe` package (Terada & von Luxburg, 2016). The R codes with the full simulation are included in the a Supplemental File.

Results 1: Comparison between clustering procedures

Table 1 presents the results for the comparison between the clustering procedures averaged over the three different DGPs' conditions represented on Figure 7. We used bold numbers to emphasize the better performing procedure at each condition. It is possible to see that CGMM performed better independent of the DGP over all indexes but when the data generating

process is the DGP1 and the performance is evaluated with HitND. This means that, when the true causal structure is a PCG in form of a v-structure, EGA will find the correct number of clusters more often than both CGMM and PA-EFA. It is also possible to see that all procedures perform about the same over different DGPs. PA-EFA will always perform worse than the other procedures, showing a negative GARI on DGP3 and HitND equals to 0 over all conditions. This negative GARI means that PA-EGA give clusters that have no correct variable connect with each other and the HitND equals to 0 means that PA-EFA never founds the correct number of clusters with our DGPs.

Table 1
Performance comparison between different DGPs.

Fit	DGP1			DGP2			DGP3		
	PA-EFA	**EGA**	**CGMM**	**PA-EFA**	**EGA**	**CGMM**	**PA-EFA**	**EGA**	**CGMM**
Acc	.713	.914	**.938**	.745	.932	**.979**	.707	.945	**.980**
PPV	.675	.939	**.994**	.714	.916	**.993**	.694	.928	**.988**
TPR	.794	.881	**.882**	.776	.957	**.964**	.676	.969	**.971**
TNR	.632	.947	**.995**	.714	.906	**.994**	.739	.921	**.988**
GARI	.298	.841	**.915**	.159	.740	**.965**	-.082	.707	**.944**
VI	.937	.292	**.167**	1.215	.346	**.082**	1.663	.377	**.104**
NMI	.472	.870	**.934**	.515	.876	**.975**	.447	.890	**.973**
HitND	.000	**.802**	.625	.000	.462	**.845**	.000	.327	**.763**

Note. DGP = data generating process; PA-EFA = parallel analysis with exploratory factor analysis; EGA = exploratory graph analysis; CGMM = correlation Gaussian mixture model; Acc = accuracy; PPV = positive predictive value; TPR = true positive rate; TNR = true negative rate; GARI = graph adjusted Rand index; VI = variation of information; NMI = normalized mutual information; HitND = simulation rate of hits of number of clusters.

Table 2 focus on the effect of the sample size on the performance of the clustering procedures. Again, CGMM had the best performance over all conditions and over all indexes, but when EGA was evaluated by the TPR in both the 250 and 500 cases' conditions. PA-EFA is again the worst performer over all conditions and indexes, and gives a negative GARI in the 100 cases' condition. There also seems to be an influence of the sample size on the performance of the procedures: increasing from 100 cases to 250 cases will increase the performance

notably. Still, increasing from 250 to 500 cases seems to increase the performance of PA-EFA considerably, but has a smaller influence on EGA and especially smaller influence on CGMM.

Table 2
Performance comparison between different sample sizes.

Fit	n = 100			n = 250			n = 500		
	PA-EFA	**EGA**	**CGMM**	**PA-EFA**	**EGA**	**CGMM**	**PA-EFA**	**EGA**	**CGMM**
Acc	.621	.851	**.931**	.734	.966	**.983**	.811	.975	**.983**
PPV	.610	.866	**.977**	.708	.956	**.998**	.766	.962	**1.000**
TPR	.592	.834	**.883**	.772	**.981**	.967	.883	**.992**	.966
TNR	.651	.867	**.979**	.696	.950	**.998**	.738	.957	**1.000**
GARI	-.075	.563	**.872**	.135	.855	**.974**	.315	.870	**.978**
VI	1.732	.670	**.261**	1.211	.186	**.049**	.872	.159	**.043**
NMI	.305	.755	**.915**	.498	.934	**.982**	.632	.947	**.984**
HitND	.000	.282	**.540**	.000	.638	**.842**	.000	.670	**.852**

Notes: PA-EFA = parallel analysis with exploratory factor analysis; EGA = exploratory graph analysis; CGMM = correlation Gaussian mixture model; Acc = accuracy; PPV = positive predictive value; TPR = true positive rate; TNR = true negative rate; GARI = graph adjusted Rand index; VI = variation of information; NMI = normalized mutual information; HitND = simulation rate of hits of number of clusters.

For evaluating the influence of the clusters' sizes, results in Table 3 shows that CGMM outperforms all the other procedures in all conditions but when there are five variables per cluster. In this case, EGA will have the largest TPR. PA-EFA is, once more, the worst performing procedure in all the cases. Finally, increasing the cluster size will increase the performance of EGA and CGMM, but decrease the performance of PA-EFA.

Table 3
Performance comparison between different cluster sizes.

	v = 5			v = 10		
Fit	**PA-EFA**	**EGA**	**CGMM**	**PA-EFA**	**EGA**	**CGMM**
Acc	.762	.908	**.958**	.681	.953	**.973**
PPV	.727	.893	**.988**	.662	.962	**.995**
TPR	.813	**.930**	.928	.684	.941	**.949**
TNR	.712	.885	**.989**	.679	.964	**.996**
GARI	.203	.651	**.925**	.046	.874	**.958**
VI	1.051	.454	**.137**	1.493	.222	**.098**
NMI	.562	.837	**.954**	.394	.920	**.967**

HitND	.000	.328	**.689**	.000	.732	**.800**

Note. PA-EFA = parallel analysis with exploratory factor analysis; EGA = exploratory graph analysis; CGMM = correlation Gaussian mixture model; Acc = accuracy; PPV = positive predictive value; TPR = true positive rate; TNR = true negative rate; GARI = graph adjusted Rand index; VI = variation of information; NMI = normalized mutual information; HitND = simulation rate of hits of number of clusters.

Results 2: Comparison between structure learning algorithms

Using the PC-stable algorithm with the averaged correlation matrix of the clusters and, assuming the true clusters are known, we generated the results shown in Table 4. The overall results, which average over all conditions, show that the PC-stable algorithm will recover most of the arrows correctly. This means that all edges that should be identified as directed was identified as such, as well as their correct directions were also identified. We see that the DGP affects the efficiency of the algorithm, as the results are almost perfect when there are only three clusters of variables with a simple v-structure between them; i.e., DGP 1.

Regarding the effects of the sample size, the results are somewhat ambiguous: both the conditions with the largest sample and the one with the smallest sample outperform the other in two indexes. Finally, we found that increasing the cluster size will improve the average performance of the algorithm. Nevertheless, this improvement is quite small; for most of the indexes it is equal to .003.

Table 4
Performance of the PC-stable algorithm applied to learning the power arrows of the PCG.

Standard PCG									
Fit	**Overall**	**Data Generating Process**			**Sample Size**			**Cluster size**	
		DGP1	**DGP2**	**DGP3**	**100**	**250**	**500**	**5**	**10**
Acc	.963	**.999**	.943	.948	.945	.972	**.973**	.961	**.966**
PPV	.961	**1.00**	.920	.964	**.974**	.955	.955	.960	**.963**
TPR	.969	**.998**	.975	.933	.915	.993	**.997**	.966	**.971**
TNR	.958	**.000**	.911	.963	**.975**	.950	.950	.957	**.960**

Note. PCG = power chain graph; DGP = data generating process; Acc = accuracy; PPV = positive predictive value; TPR = true positive rate; TNR = true negative rate.

The second analysis was used to evaluate how applying causal discovery algorithms would change the total number of arrows in the CG implied by the PCG. Table 5 shows our findings in terms of the sparsity of the CGs discovered by the algorithms, where sparsity is simply the false negative rates. The first algorithm column, named as CG, is simply the CG implied by the PCG. Because of how our simulations were conducted, this column should always be the one with the lowest sparsity and, therefore, it is used as a benchmark for the other procedures. It is possible to see that the HC algorithm is always the one that keeps most of the arrows, followed by the PC-stable algorithm and, finally, the MMHC algorithm. No best–worst comparison is adequate in this case. Nevertheless, these results show that further tuning of the CG implied by the PCG will be considerably sparser and, therefore, causal analysis of CGs estimated by these procedures will need to be sensible to interference effects.

Table 5.
Sparsity comparison between tuning algorithms.

		Algorithms		
Condition	**CG**	**PC**	**HC**	**MMHC**
Overall	0.031	0.782	**0.554**	0.921
DGP				
DGP1	0.002	0.752	**0.540**	0.912
DGP2	0.025	0.783	**0.557**	0.919
DGP3	0.067	0.810	**0.565**	0.931
Sample size				
100	0.085	0.897	**0.719**	0.954
250	0.007	0.788	**0.548**	0.922
500	0.003	0.661	**0.394**	0.886
Cluster size				
5	0.034	0.706	**0.446**	0.883
10	0.029	0.858	**0.662**	0.959

Note. CG = chain graph; PC = PC-stable algorithm; HC = hill-climbing algorithm; MMHC = min-max hill-climbing algorithm; DGP = data generating process.

Empirical example

For illustration purposes, the present empirical example will evaluate the concept of empathy. Davis (1980) proposed a multidimensional measurement of empathy, a questionnaire composed by four seven-item subscales: perspective-taking; fantasy; empathic concern and personal distress. Fantasy is the tendency to get involved in the actions of fictional characters from diverse media. Perspective-taking is the tendency to comprehend others' point of view. Empathic concern is the tendency of feeling concern and sympathy for people in distress. Personal distress is the tendency of feeling unease in difficult, tense or emotional situations.

As usual in psychometrical literature, this measure was evaluated by Davis (1980) by means of exploratory factor analysis, assuming no causal relations between the factors. This analysis resulted in what is usually considered as good evidence of validity (items theorized to be together were indeed clustered by the same latent factor) and reliability (Cronbach's alpha above 0.70). Briganti, Kempenaers, Braun, Fried and Linkowski (2018) applied the EGA procedure and found the same structure as the original study by Davis (1980). In the present example, we will apply our PCG procedure to the open dataset (Braun, Rosseel, Kempenaers, Loas, & Linkowski, 2015) provided by Briganti et al. (2018).

Method

Participants. The dataset was composed of 1,973 French-speaking students in several colleges for higher education, from a diverse set of courses. The age ranged between 17 and 25 years (M = 19.6 years, SD = 1.6 years), with 57% females and 43% males. From the original sample of 1,973 students, only 1,270 answered the full questionnaire. Missing data were imputed by Briganti et al. (2018) according to the Gaussian graphical model they fitted by means of the EGA procedure. The items of the empathy measurement are displayed in

Table 6, as well as the original factorial solution which was validated and theoretically substantiated.

Analysis. We applied both the EGA and the CGMM procedure to identify the number of clusters in the data. The similarity of the solutions was evaluated by means of the graph adjusted Rand (GARI) index. After identifying the clusters, the average correlation matrix was calculated for both solutions (EGA's and CGMM's) and then used in the PC-stable algorithm to discover the arrows in the PCG. Finally, the tuning algorithms were applied to reduce the CG implied by the PCG, with the results displayed both graphically and with the percentage of removed arrows textually described. The R codes for these analyses are also included in the Supplemental File.

Table 6.
Description of the items of an instrument on empathy and the original assignment of items to factors (Davis, 1980).

Item	Factor	Item description
1	Fantasy	I daydream and fantasize, with some regularity, about things that might happen to me.
2	Empathic concern	I often have tender, concerned feelings for people less fortunate than me.
3R	Perspective-taking	I sometimes find it difficult to see things from the "other guy's" point of view.
4R	Empathic concern	Sometimes I don't feel very sorry for other people when they are having problems.
5	Fantasy	I really get involved with the feelings of the characters in a novel.
6	Personal distress	In emergency situations, I feel apprehensive and ill-at-ease.
7R	Fantasy	I am usually objective when I watch a movie, and I don't often get completely caught up in it.
8	Perspective-taking	I try to look at everybody's side of a disagreement before I make a decision.
9	Empathic concern	When I see someone being taken advantage of, I feel kind of protective towards them.
10	Personal distress	I sometimes feel helpless when I am in the middle of a very emotional situation.
11	Perspective-taking	I try to understand my friends better by imagining how things look from their perspective.
12R	Fantasy	Becoming extremely involved in a good book or movie is somewhat rare for me.
13R	Personal distress	When I see someone get hurt, I tend to remain calm.
14R	Empathic concern	Other people's misfortunes do not usually disturb me a great deal.
15R	Perspective-taking	If I'm sure I'm right, I don't waste much time listening to other people's arguments.
16	Fantasy	After seeing a play or movie, I have felt as though I were one of the characters.
17	Personal distress	Being in a tense emotional situation scares me.
18R	Empathic concern	When I see someone being treated unfairly, I sometimes don't feel very much pity for them.
19R	Personal distress	I am usually pretty effective in dealing with emergencies.
20	Fantasy	I am often quite touched by things that I see happen.
21	Perspective-taking	I believe that there are two sides to every question and try to look at them both.

22	Empathic concern	I would describe myself as a pretty soft-hearted person.
23	Fantasy	When watching a good movie, I can easily put myself in the place of a leading character.
24	Personal distress	I tend to lose control during emergencies.
25	Perspective-taking	When I'm upset at someone, I usually try to "put myself in his shoes" for a while.
26	Fantasy	When reading an interesting story, I imagine how I would feel if it was happening to me.
27	Personal distress	When I see someone who badly needs help in an emergency, I go to pieces.
28	Perspective-taking	Before criticizing somebody, I try to imagine how I would feel if I were in their place.

Note. An R after the item indicates it is a reversed item.

Results from the empirical example

We started the analysis by applying both the EGA and the CGMM procedure to the dataset. The solutions are displayed in Figure 7. EGA resulted exactly in the theoretical structure displayed in Table 6. CGMM, on the other hand, resulted in a six cluster solution. Comparing both, it is possible to see that six items (3, 6, 9, 10 15, and 17) were clustered by CGMM differently from what was expected, resulting in a value of .719 for the accuracy measure GARI. Cluster 1 in Figure 7 corresponds to the original Personal distress factor. Cluster 2, to the original Fantasy factor. Cluster 3, to the original Perspective-taking factor (except from item 9 in the CGMM solution). Cluster 4, to the original Empathic concern factor. Clusters 5 and 6 have no specific interpretation.

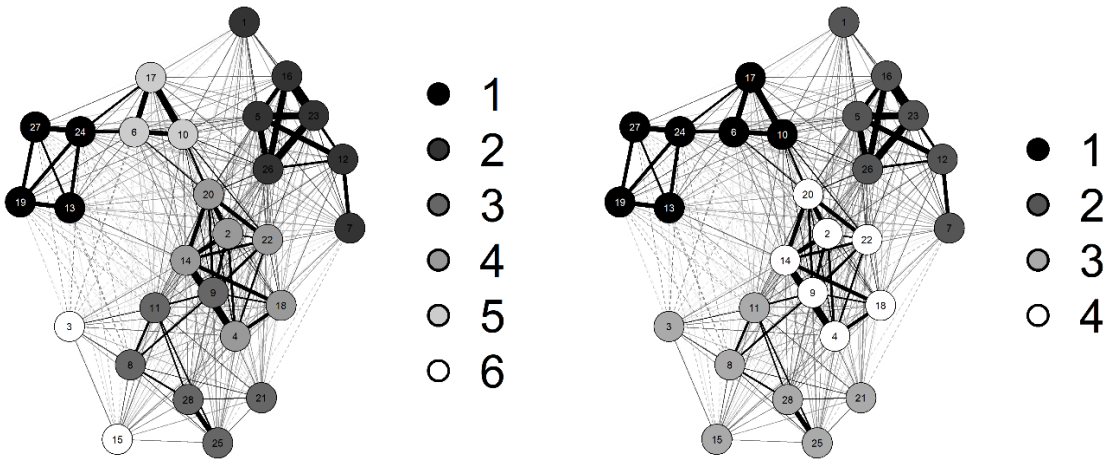


Figure 7. CGMM's (on the left) and EGA's (on the right) clustering solution, with different numbers and associated colors representing different clusters.

In the next step, we calculated the average correlation matrix, based on the clusters identified by both procedures. For assuring interpretability, from the cluster solution identified by the CGMM procedure, we removed both clusters 5 and 6, as well as item 9. Therefore, we applied the PC-stable algorithm to the averaged correlation matrix calculated from the solution to the right in Figure 7 and for a similar averaged correlation matrix, after removing items 3, 6, 9, 10, 15 and 17. The PCGs displayed in Figure 8 were estimated by the PC-stable algorithm. It is possible to see that, by removing the previous six items, the graph on the left (estimated from CGMM's averaged correlation matrix) didn't identify three arrows identified when using the other averaged correlation matrix (calculated from EGA): from power node *V*'1 to *V*'2; from *V*'1 to *V*'4; and from *V*'4 to *V*'2.

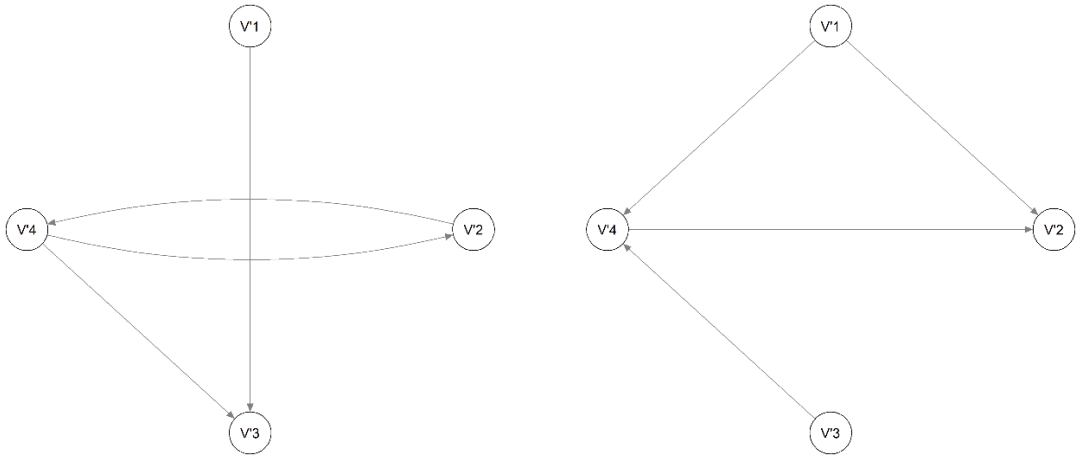


Figure 8. PCGs estimated with the PC-stable algorithm using CGMM's (on the left) and EGA's (on the right) averaged correlation matrix.

From the graph on the left, it is possible to infer that both Personal distress (*V*'1) and Perspective-taking (*V*'4) are the causes for Empathic concerns (*V*'3). This graph does not allow one to claims with certainty if Perspective-taking causes Fantasy (*V*'2) or if it is the contrary. From the graph on the right, on the other hand, it is possible to infer that Personal distress is the cause for both Perspective-taking and Fantasy. It is possible to infer as well that Empathic concerns cause Perspective-taking which in turn causes Fantasy.

Given that there are some arrows not properly identified on the graph to the left of Figure 8, the final analyses were conducted based only on the PCG to the right of Figure 8. For the final analysis, we proceeded as in the simulation study. First, the absent arrows and the direction of the arrows of the CG implied by the PCG were fixed. After that, the PC-stable, the HC and the MMHC algorithms were applied so unnecessary arrows could be removed. Tuning with the PC-stable algorithm resulted in 74.48% of the arrows removed. With the HC algorithm, 65.31% of the arrows were removed. With the MMHC algorithm, 86.22% of the arrows were removed. Each corresponding CG is displayed in Figure 9.

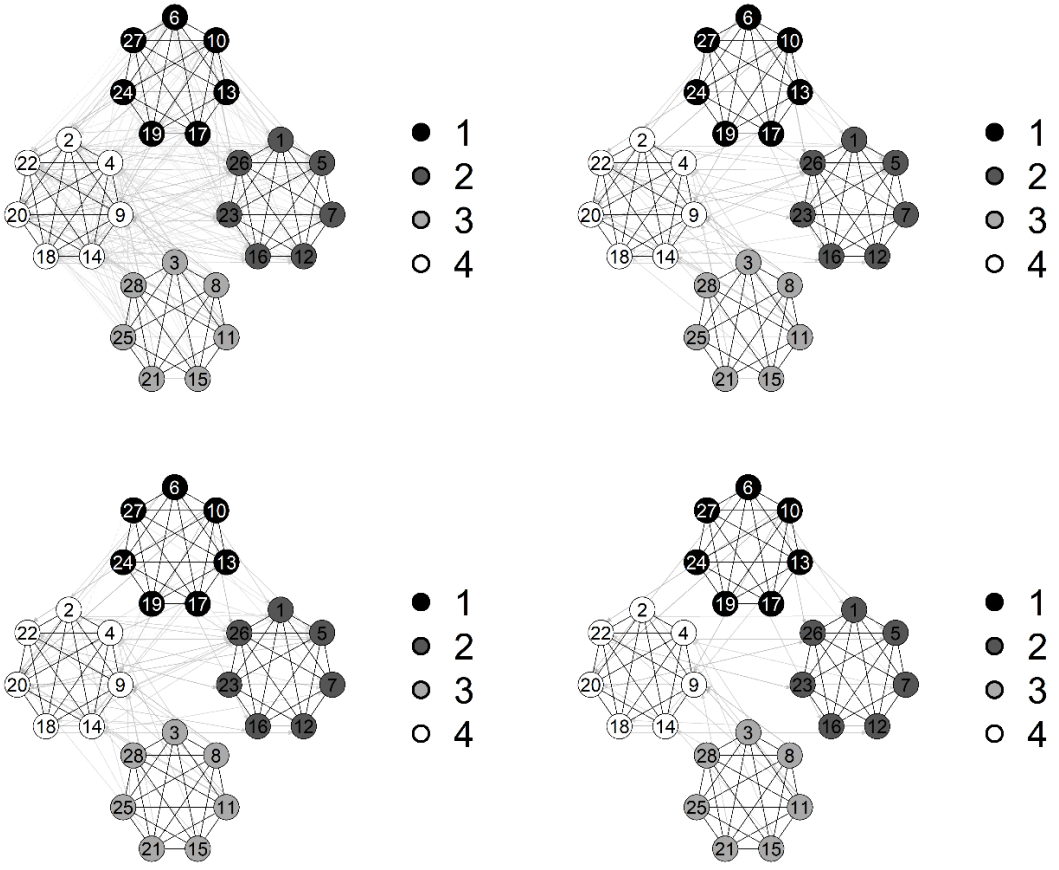


Figure 9. The original CG implied by the estimated PCG (top-left), a CG tuned by the PC-stable algorithm (top-right), a CG tuned by the HC algorithm (bottom-left) and a CG tuned by the MMHC algorithm (bottom-right).

Discussion

The main aim of the present study was to develop a new procedure of structure learning of power chain graphs (PCGs). The secondary aim of the study was to compare clustering algorithms and causal discovering algorithms that can be used for a learning process of the structure of the PCGs. To this end, we used simulated data. Empirical data were used for illustrative purposes of a potential application of the procedure. We showed that PCG can properly recover both the clusters of variables, by means of the correlation Gaussian mixture models (CGMM) procedure, and the correct direction of the underlying causal relations, by means of averaged correlation matrix and the PC-stable algorithm. Our simulations have also illustrated how power chain graphs (PCGs) can be used to further investigate the structure of chain graphs (CGs), through tuning of the CG implied by the PCG using three different classes of causal discovery algorithms.

To assess the performance of the CGMM in finding the correct number of clusters, we used a number of accuracy related indexes and cluster comparison indexes. We also compared the performance of the CGMM with that of two other more traditional procedures in psychometric analyses: parallel analysis used in exploratory factor analysis (PA-EFA) and exploratory graph analysis (EGA). Our results show that CGMM outperformed both procedures in a number of different conditions of sample sizes, number of variables per cluster and the total number of clusters. Nevertheless, our results are somewhat unexpected, mainly because of the performance of the PA-EFA procedure. Previous studies (Golino et al, 2018; Timmerman & Lorenzo-Seva, 2011) have shown high accuracy of the PA-EFA procedure, which was not replicated in the present study. One of the main differences between our study and these previous ones is the fact that our simulation has an underlying causal structure. Previous studies used only correlational underlying structures. This

difference causes data to be factorized in a different manner (Lauritzen, 1996), probably resulting in the differences of performance found in our study.

The PC-stable algorithm performed well in a number of different conditions. Changing the sample sizes, number of variables per cluster and the total number of clusters had little to no clear effect on the accuracy related indexes. The second part of the causal discovery analysis, which compared different algorithms, has some implications for structure learning procedures of CGs (e.g., Drton & Perlman, 2008; Ma, Xie, & Geng, 2008; Peña, Sonntag, & Nielsen, 2014). Most of these procedures try to find the best fitting CG by brute force (Peña, 2018): fitting a large number of different models and choosing the best fitting one. Learning the structure using the full PCG procedure and then using its implied CG as an initial step to obtain the estimated true CG can improve the learning procedures of the structure of CGs.

There are two main limitations of the present study. The first is the CGMM procedure. Notwithstanding its good performance in the simulation study, the results may be limited to the particular setting of the data generating process. Theoretical studies are necessary to better understand in what conditions CGMM may perform the best, so that the present studies regarding CGMM can be properly generalized. Nevertheless, this has little impact on the PCG procedure as a whole, as any clustering procedure can be used in place of the CGMM, as long as it respects the “similarity” condition for clustering the variables. The second main limitation is related to the averaging of the correlation matrix. Despite its relation with the definitions of a power edge and similarity between nodes, there is neither logical implication nor model constraint that requires for the power edges to be estimated like this. Therefore, it is also necessary that future studies evaluate which procedures to estimate or calculate power edges can improve the performance of the PCG procedure.

Regarding the empirical example, although it was used more as an illustration than as a central test of our study, it is robust in the sense that any empathy researcher could have used the PCG procedure and would have found the same results we have. It is interesting to see that removing some items cause completely different graphical models with different theoretical implications. These implications are direct consequences of the application of the PCG procedure. For researchers on empathy, our results may be of interest for replication, possible to be further investigated with longitudinal (e.g., Zhou et al., 2002) and neurological (e.g., Singer et al., 2004) studies on empathy. Researchers in other areas can also infer from these results that maybe constructs first hypothesized to be have only associational relations, can actually present some underlying causal structure.

Future studies on PCG can focus on both theoretical aspects of CGMM and how causal discovering algorithms can be extended for learning the causal structure from the averaged correlation matrix. For instance, what would happen if instead of using polychoric or Spearman correlations, measures of dependence (Paul & Shill, 2017; Zhao, Zhou, Zhang, & Chen, 2016), such as the maximal information coefficient or distance correlations, were used in CGMM? What would be the impact of using partial distance correlations' tests (Székely & Rizzo, 2014) instead of partial correlations' tests in the PC-stable algorithm? Is it possible to incorporate a clustering procedure such as CGMM directly as a step in the score-based algorithms? These questions help to set the environment for prolific research about PCGs and all its constituent elements: clustering; causal discovery; and CG structure learning.

References

Aalbers, G., McNally, R. J., Heeren, A., de Wit, S., & Fried, E. I. (2018). Social media and depression symptoms: A network perspective. *Journal of Experimental Psychology: General*. Advance online publication. http://dx.doi.org/10.1037/xge0000528

Bagozzi, R. P. (2007). On the meaning of formative measurement and how it differs from reflective measurement: Comment on Howell, Breivik, and Wilcox (2007). *Psychological Methods*, *12*(2), 229-237.

Braun, S., Rosseel, Y., Kempenaers, C., Loas, G., & Linkowski, P. (2015). Self-report of empathy: A shortened French adaptation of the interpersonal reactivity index (IRI) using two large Belgian samples. *Psychological Reports*, *117*(3), 735-753.

Briganti, G., Kempenaers, C., Braun, S., Fried, E. I., & Linkowski, P. (2018). Network analysis of empathy items from the Interpersonal Reactivity Index in 1973 young adults. *Psychiatry Research*, *265*, 87-92.

Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), 759-771.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, *3*, 507-554.

Colombo, D., & Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, *15*(1), 3741-3782.

Daly, R., & Shen, Q. (2007). Methods to accelerate the learning of Bayesian network structures. In *Proceedings of the 2007 UK Workshop on Computational Intelligence*.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, *10*, 85.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1-22.

Dinno, A. (2018). paran: Horn's test of principal components/factors. R Package retrieved from *https://cran.r-project.org/web/packages/paran/index.html*.

Drton, M., & Perlman, M. D. (2008). A SINful approach to Gaussian graphical model selection. *Journal of Statistical Planning and Inference*, *138*(4), 1179-1200.

Ekström, J. (2011). *On the relation between the polychoric correlation coefficient and Spearman's rank correlation coefficient*. Department of Statistics, UCLA.

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: a tutorial paper. *Behavior Research Methods*, *50*(1), 195-212.

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*, *23*(4), 617-634.

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika*, *82*(4), 904-927.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods, 4*(3), 272-299.

Foygel, R., & Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. *Advances in neural information processing systems, 2010*,604-612.

Fraley, C., & Raftery, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification, 24*(2), 155-181.

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS one, 9*(10), 1-12.

Fried, E. I., & Cramer, A. O. J. (2017). Moving forward: Challenges and directions for psychopathological network theory and methodology. *Perspectives on Psychological Science, 12*(6), 999–1020.

Friedman, N., Peér, D., & Nachman, I. (1999). Learning Bayesian network structure from massive datasets: the "sparse candidate" algorithm. In *Proceedings of the 15th conference on uncertainty in artificial intelligence* (pp. 206-215). Burlington: Morgan Kaufmann Publishers.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics, 9*(3), 432-441.

Glaros, A. G., & Kline, R. B. (1988). Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model. *Journal of Clinical Psychology, 44*(6), 1013-1023.

Golino, H.F., & Christensen, A. (2019). EGAnet: Exploratory graph analysis: A framework for estimating the number of dimensions in multivariate data using network psychometrics. R Package retrieved from *https://cran.r-project.org/web/packages/EGAnet/index.html*.

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PLOS ONE*, *12*(6), 1-26.

Golino, H.F., Shi, D., Garrido, L. E., Christensen, A. P., Nieto, M. D., Sadana, R., & Thiyagarajan, J. A. (2018). *Investigating the performance of Exploratory Graph Analysis and traditional techniques to identify the number of latent factors: A simulation and tutorial*. https://doi.org/10.31234/osf.io/gzcre

Haghighat, M. B. A., Aghagolzadeh, A., & Seyedarabi, H. (2011). A non-reference image fusion metric based on mutual information of image features. *Computers & Electrical Engineering, 37*(5), 744-756.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*(2), 191-205.

Hennig, C. (2018). fpc: Flexible procedures for clustering. R Package retrieved from *https://cran.r-project.org/web/packages/fpc/index.html*.

Hennig, C., & Liao, T. F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *62*(3), 309-369.

Hevey, D. (2018). Network analysis: A brief overview and tutorial. *Health Psychology and Behavioral Medicine*, *6*(1), 301-328.

Higham, N. J. (2002). Computing the nearest correlation matrix—a problem from finance. *IMA Journal of Numerical Analysis*, *22*(3), 329-343.

Holgado–Tello, F. P., Chacón–Moscoso, S., Barbero–García, I., & Vila–Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*, *44*(1), 153-166.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179-185.

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods*, *12*(2), 205-218.

Jain, P. M., & Shandliya, V. K. (2013). A survey paper on comparative study between Principal Component Analysis (PCA) and Exploratory Factor Analysis (EFA). *International Journal of Computer Science and Applications*, *6*(2), 373-375.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, *59*(3), 381-389.

Koller, D., Friedman, N., & Bach, F. (2009). *Probabilistic graphical models: Principles and techniques*. Cambridge: MIT press.

Lauritzen, S. L. (1996). *Graphical models*. Oxford: Clarendon Press.

Lauritzen, S. L., & Richardson, T. S. (2002). Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(3), 321-348.

Liu, Y., Magnus, B., O'Connor, H., & Thissen, D. (2018). Multidimensional item response theory. In P. Irwing, T. Booth & D. J. Hughes (Eds.), *The Wiley Handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 445-493). Hoboken, NJ: Wiley-Blackwell.

Ma, Z., Xie, X., & Geng, Z. (2008). Structural learning of chain graphs via decomposition. *Journal of Machine Learning Research*, *9*(Dec), 2847-2880.

Meilă, M. (2007). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, *98*(5), 873-895.

Mair, P. (2018). *Modern Psychometrics with R*. New York: Springer.

Nenov, M., & Nikolov, S. (2015). Employing power graph analysis to facilitate modeling molecular interaction networks. *International Journal Bioautomation*, *19*(1), 37-42.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443-460.

Paul, A. K., & Shill, P. C. (2017). Reconstruction of gene network through backward elimination based information-theoretic inference with maximal information coefficient. In *2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR)* (pp. 1-5). IEEE.

Pearl, J. (2009). *Causality*. Cambridge: Cambridge University Press.

Peña, J. M. (2018). Unifying Gaussian LWF and AMP Chain Graphs to model interference. *arXiv preprint arXiv:1811.04477*.

Peña, J., Sonntag, D., & Nielsen, J. (2014). An inclusion optimal algorithm for chain graph structure learning. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics* (pp. 778-786).

Pons, P., & Latapy, M. (2006). Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, *10*(2), 191-218.

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rao, D. C. (1979). Joint distribution of z transformations estimated from the same sample. *Human Heredity*, *29*(6), 334-336.

Revelle, W. (2019). psych: procedures for psychological, psychometric, and personality research. R Package retrieved from *https://cran.r-project.org/web/packages/psych/index.html*.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27-42.

Roverato, A., & Rocca, L. L. (2006). On block ordering of variables in graphical modelling. *Scandinavian Journal of Statistics*, *33*(1), 65-81.

Royer, L., Reimann, M., Andreopoulos, B., & Schroeder, M. (2008). Unraveling protein networks with power graph analysis. *PLoS Computational Biology*, *4*(7), 1-17.

Russell, S. J., & Norvig, P. (2009). *Artificial intelligence: A modern approach*. New York: Prentice Hall.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, *8*(1), 205-233.

Scutari, M., & Denis, J. B. (2015). *Bayesian networks: With examples in R*. Boca Raton, FL: CRC Press.

Singer, T., Seymour, B., O'doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, *303*(5661), 1157-1162.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. Cambridge: MIT Press.

Székely, G. J., & Rizzo, M. L. (2014). Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*, *42*(6), 2382-2412.

Terada, Y., & von Luxburg, U. (2016). Loe: Local ordinal embedding. R Package retrieved from *https://cran.r-project.org/web/packages/loe/index.html*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(2), 209-220.

Tsamardinos, I., Aliferis, C. F., & Statnikov, A. (2003). Time and sample efficient discovery of Markov blankets and direct causal relations. In *Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 673-678). ACM.

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, *65*(1), 31-78.

Verma, T. S., & Pearl, J. (1990). Equivalence and synthesis of causal models. *Proceedings of the Conference on Uncertainty in Artificial Intelligence* (pp. 220-227). Cambridge, MA.

Wermuth, N., & Lauritzen, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *52*(1), 21-50.

Zhang, S., Wong, H. S., & Shen, Y. (2012). Generalized adjusted rand indices for cluster ensembles. *Pattern Recognition*, *45*(6), 2214-2226.

Zhao, J., Zhou, Y., Zhang, X., & Chen, L. (2016). Part mutual information for quantifying direct associations in networks. *Proceedings of the National Academy of Sciences*, *113*(18), 5130-5135.

Zhou, Q., Eisenberg, N., Losoya, S. H., Fabes, R. A., Reiser, M., Guthrie, I. K., ... & Shepard, S. A. (2002). The relations of parental warmth and positive expressiveness to children's empathy‐related responding and social functioning: A longitudinal study. *Child Development*, *73*(3), 893-915.

FINAL CONSIDERATIONS

This dissertation started by asking “What does it mean to measure something?” Our first study made the point that this question can only be properly answered in a psychological research context if certain assumptions are made explicit. The following assumptions: (1) the structural validity assumption; (2) the process assumption; and (3) the construct assumption were already partially acknowledged in previous studies (e.g., Michell, 2019; Trendler, 2009). Nevertheless, these assumptions were named and identified explicitly for the first time in the present dissertation as cornerstones for psychometric research. The purpose of this strong statement is to make it very clear that measurement in psychology needs to be the consequence of thoughtful models (Sijtsma, 2012). “Thoughtful” in the present case means that psychometric models should avoid being too general, or at least be explicit in the assumptions that are made to affirm that a numerical representation generated by that particular model is a good measurement.

From the three identified assumptions, research agendas can be elaborated for psychometric researchers. For those researchers who believe process is not the main aim of their research, currently available models and theories, such as item response models, can be used in their current existing form. But, similarly to our second study, computational or statistical improvement of more traditional models can be investigated by psychometric researchers, seeking to generate more accurate estimates with the same type of data. This is to say that improvement can be attained by simply changing the structural validity assumptions of the models. Particularly for the more traditional psychometric literature, we believe that nonparametric models could be the default at least for initial development of psychometric scales (e.g., Straat, Van der Ark, & Sijtsma, 2013). In our second study we found that the nonparametric model used can better recover the true probability density of the true scores, but at the cost of increasing the bias of the estimated scores. Future studies might focus on

the possibility of decreasing bias and keeping the estimated density as close as possible to the true one.

If the underlying process is the main aim of the research, then cognitive modeling approaches can be used to identify or test concurrent hypotheses and theories. Cognitive modeling is the name given to developing psychometric-like models for measuring the magnitude of latent variables (Farrell & Lewandowsky, 2018). Nevertheless, this link between cognitive modeling and psychometrics is rarely identified by researchers in psychology (Embretson, 2010). It was even argued in the present dissertation that psychometrics and cognitive modeling should not necessarily be considered different things. On the contrary, if psychometrics is taken as the study of measurement in psychology, cognitive modeling is but a specific type of measurement, where context and process are of utmost importance. Another important take home message, inferred from our third study and its operationalization of Lewin's equation (Lewin, 1936), is that not only cognitive psychology can influence our modeling techniques. In fact, we used a traditional social psychology theory as the basis for the approach we proposed. Any theory or hypothesis on process can be used to create models of measurement.

Finally, if it is considered of paramount importance to exclude latent variables (e.g., Trendler, 2009), measurement can still be realized using multivariate statistical or physical approaches. In our fourth study we presented how this can be accomplished using power chain graphs (PCGs), as an alternative to structural equation modeling. PCGs can further be studied also using the physical approach to measurement, as some authors have done with the Rasch model (e.g., Perline, Wright, & Wainer, 1979). We believe the present dissertation has helped to set not only a clear direction for researchers who want to perform better measurements, but also more validity to the definition of psychometrics as the field of study concerned with the theory and technique of psychological measurement (Maul, 2017).

References

Embretson, S. E. (2010). *Measuring psychological constructs: Advances in model-based approaches*. New York: American Psychological Association.

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge: Cambridge University Press.

Lewin, K. (1936). *Principles of topological psychology*. New York: McGraw-Hill.

Maul, A. (2017). Moving beyond traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, *15*(2), 103-109.

Michell, J. (2019). The fashionable scientific fraud: Collingwood's critique of psychometrics. *History of the Human Sciences*. https://doi.org/10.1177/0952695119872638

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*(2), 237-255.

Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory & Psychology*, *22*(6), 786-809.

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification*, *30*(1), 75-99.

Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory & Psychology*, *19*(5), 579-599.