UvA-DARE (Digital Academic Repository)

Interpreting psychometric models

van Bork, R.

DOI

[10.31237/OSF.IO/X6A7S](https://doi.org/10.31237/OSF.IO/X6A7S)

Publication date

2019

Document Version

Final published version

License

CC BY

[Link to publication](#)

Citation for published version (APA):

van Bork, R. (2019). *Interpreting psychometric models*. [Thesis, fully internal, Universiteit van Amsterdam]. <https://doi.org/10.31237/OSF.IO/X6A7S>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

The background of the cover features several hand-drawn geometric shapes in a watercolor style. There is a large triangle in the upper left, a square in the center-right, and a circle in the lower left. These shapes are drawn with thick, textured lines in shades of pink, orange, and blue, which overlap and intersect each other. The overall composition is abstract and artistic.

INTERPRETING PSYCHOMETRIC MODELS

RIET VAN BORK

Interpreting Psychometric Models

RIET VAN BORK

© 2019 RIET VAN BORK
ALL RIGHTS RESERVED.

THE RESEARCH IN THIS THESIS WAS SUPPORTED BY A MARIE CURIE CAREER INTEGRATION GRANT, GRANT NUMBER 631145.

DOI: [10.31237/OSF.IO/X6A7S](https://doi.org/10.31237/osf.io/x6a7s)
COVER DESIGN: [NIKKI RITMEIJER, NIKKIRITMEIJER.COM](http://nikkiritmeijer.com)
LAYOUT: [ADAPTED FROM DISSERTATE.IO](http://dissertate.io)
TYPESET WITH: $\text{\LaTeX}$
PRINTED BY: IPSKAMP DRUKKERS, ENSCHEDE
DIGITAL COPY: [DARE.UVA.NL](http://dare.uva.nl) & [THESISCOMMONS.ORG](http://thesiscommons.org)

Interpreting Psychometric Models

ACADEMISCH PROEFSCHRIFT

TER VERKRIJGING VAN DE GRAAD VAN DOCTOR
AAN DE UNIVERSITEIT VAN AMSTERDAM
OP GEZAG VAN DE RECTOR MAGNIFICUS
PROF. DR. IR. K.I.J. MAEX
TEN OVERSTAAN VAN EEN DOOR HET COLLEGE VOOR
PROMOTIES INGESTELDE COMMISSIE,
IN HET OPENBAAR TE VERDEDIGEN IN DE AGNIETENKAPEL
OP VRIJDAG 14 JUNI 2019, TE 14:00UUR
DOOR RIET VAN BORK
GEBOREN TE AMSTERDAM

PROMOTIECOMMISSIE

PROMOTOR:	PROF. DR. D. BORSBOOM	UNIVERSITEIT VAN AMSTERDAM
CO-PROMOTOR:	DR. M. T. RHEMTULLA	UNIVERSITY OF CALIFORNIA DAVIS
OVERIGE LEDEN:	PROF. DR. H. L. J. VAN DER MAAS	UNIVERSITEIT VAN AMSTERDAM
	PROF. DR. E. M. WAGENMAKERS	UNIVERSITEIT VAN AMSTERDAM
	DR. M. MARSMAN	UNIVERSITEIT VAN AMSTERDAM
	PROF. DR. J. W. ROMEIJN	RIJKSUNIVERSITEIT GRONINGEN
	PROF. DR. K. SIJTSMA	TILBURG UNIVERSITY
	DR. N. E. WEINBERGER	TILBURG UNIVERSITY

FACULTEIT DER MAATSCHAPPIJ- EN GEDRAGSWETENSCHAPPEN

CONTENTS

I	INTRODUCTION	9
1.1	Psychometric models	9
1.2	Interpretation of psychometric models	14
1.3	Chapter outline	16
I	INTERPRETING THE LATENT VARIABLE MODEL	19
2	A CAUSAL INTERPRETATION OF THE COMMON FACTOR MODEL	21
2.1	Introduction	22
2.2	Statistical vs. causal models	26
2.3	Correlation does not entail causation	27
2.4	What is the common factor in the descriptivist approach?	28
2.5	Toward a causal interpretation	29
2.6	How to assess whether a common cause structure is correct	32
2.7	Conclusion	35
3	A CAUSAL THEORY OF ERROR SCORES	37
3.1	Introduction	38
3.2	Error in modern test theory	41
3.3	A causal theory of errors	52
3.4	Two rationales for the response variable as a random variable	61
3.5	Item Bias	65
3.6	Local homogeneity	68
3.7	Cronbach's alpha and test-retest correlation	69
3.8	Discussion	71
4	COMPOSITES CAN BE CAUSAL TOO	75
4.1	Commentary to Mõttus (2016)	75
II	EQUIVALENT EQUATIONS, DIFFERENT FRAMEWORKS	79
5	SHORT INTRODUCTION TO NETWORK MODELS	81
5.1	Introduction	81
5.2	Graphical models	82
5.3	Correlations to identify connections	85
5.4	Partial correlations to identify connections	86

5.5	Approximations of the covariance matrix to identify connections	89
6	LATENT VARIABLE MODELS AND NETWORKS: STATISTICAL EQUIV- ALENCE AND TESTABILITY	93
6.1	Introduction	94
6.2	Equivalence is not Identity	96
6.3	The meaning of model equivalence: general model classes versus specific model interpretations	98
6.4	Distinguishing between specific models	106
6.5	Reasoning with equivalence and testability: A practical example	112
6.6	Discussion	119
7	UNIDIMENSIONAL FACTOR MODELS IMPLY WEAKER PARTIAL COR- RELATIONS THAN ZERO-ORDER CORRELATIONS	123
7.1	Introduction	124
7.2	Unidimensional factor models	125
7.3	Weaker partial correlations than zero-order correlations	127
7.4	Application of the result	128
7.5	Empirical illustration	132
7.6	Discussion	133
8	COMMON EFFECT MODELS: POSITIVE OR NEGATIVE MANIFOLD?	137
8.1	Introduction	138
8.2	Common cause vs direct effects as explanations for the positive manifold . .	139
8.3	Equivalence of the Ising model and IRT model	141
8.4	The Kruis-Maris model	143
8.5	Kruis-Maris models vs causal indicator models	144
8.6	Causal indicator models imply a negative manifold when conditioning on the common effect	146
8.7	Positive versus negative manifold	148
8.8	Discussion	150
9	WHAT IS P? SOME RISKS OF GENERAL FACTOR MODELING	153
9.1	Introduction	153
9.2	The problem of nearly-equivalent models	155
9.3	The risk of affirming the consequent	159
9.4	Conclusions	164
10	DISCUSSION	167
10.1	The meaning of a causal interpretation	168
10.2	Implications of a causal interpretation	177
	APPENDIX A SUPPLEMENT TO CHAPTER 6	181
A.1	Items Irrational Procrastination Scale	181

APPENDIX B	SUPPLEMENT TO CHAPTER 7	183
B.1	Example I: UFM	183
B.2	Example II: random correlation matrix	186
B.3	Simulation study: Performance of the empirical bootstrap test	189
APPENDIX C	SUPPLEMENT TO CHAPTER 8	195
C.1	Proof of Theorem 1	195
C.2	Proof of Theorem 2	196
REFERENCES		199
NEDERLANDSE SAMENVATTING		221
SUMMARY		229
PUBLICATIONS		235
DANKWOORD		239

1

Introduction

The field of psychometrics – the field that is concerned with the theory of psychological measurement – heavily relies on the use of statistical models to connect observable behavior to psychological constructs. These models, in turn, rely on concepts such as *random variables*, *observed and latent variables*, *model equivalence*, and *measurement error*. While such concepts are the cornerstone of psychometrics, the question of what we mean when using these concepts is rarely investigated. Instead, psychometrics seems to rely on some common understanding of what such concepts mean. But, what is meant by *error*? What does it actually mean for a variable to be *random*? And, do we actually *observe* variables? Similarly, if models are mathematically equivalent, does that mean that these models are *identical*, that is, that they are *the same* model? This latter question is strongly related to the question of whether we think of models as mere descriptions of data, or as causal hypotheses about the external world. These, and similar questions, concern the interpretation of psychometric models, which is the topic of this dissertation. In the following I introduce the two main components of the title of this dissertation: *interpretation* and *psychometric models*. What are psychometric models, and what does it mean to interpret them?

1.1 PSYCHOMETRIC MODELS

Psychological constructs such as personality traits, psychopathological disorders, attitudes, abilities and skills are reflected in observable behavior. For example, extraversion is reflected in whether someone often hangs out with friends, depression is reflected in whether someone describes herself as unhappy, and math ability is reflected in whether someone knows how to solve numerical problems. Psychometric theories differ in how they link observed behavior to the

psychological constructs. For example, latent trait theory (also sometimes called modern test theory; Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1978) conceptualizes psychological constructs as unobservable human characteristics that underlie observed behavior. These human characteristics are called *latent traits* or *latent variables* and *account for* observed behavior (Lord & Novick, 1968, p. 358). In this framework, math ability is measured by observing people's performance on several numerical problems under the assumption that the differences between people in test performance are explained by differences in their math ability. Similarly, within this framework a psychological disorder is measured by administering a patient's symptoms under the assumption that the disorder accounts for these symptoms, just like a medical disease manifests itself in its symptoms by causing them.

In classical test theory, the most important concept, the true score, is defined as the expected score on a range of repeated administrations of a test (Lord & Novick, 1968). The psychological construct (or true score) is thus defined in terms of the test that is used, which can be seen as an operationalist conceptualization of psychological constructs. In operationalism, concepts are synonymous with the set of operations that are used to measure them (Bridgman, 1927). In this view, the link between the psychological construct and the observed behaviors is that the psychological construct is a label for the set of behaviors that are used to measure the construct. In this framework, math ability is the number of math items a person solves on average in a certain math test.

In the last decade, a new framework has been proposed in which psychological constructs are considered networks of inter-connected affective, cognitive and behavioral components that causally influence each other (Cramer et al., 2012, p. 415). The causal effects between these components explain why certain emotions, thoughts and behaviors are inter-connected or associated. This alternative view on psychological constructs has sprouted several methods to estimate network models from psychological data (Borsboom & Cramer, 2013; Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012; Van Borkulo et al., 2014), and these methods have been applied to a wide range of constructs (Boschloo et al., 2015; Costantini et al., 2015; Dalege et al., 2016; Dekker, Blanken, & van Someren, 2017; Deserno, Borsboom, Begeer, & Geurts, 2017a). Psychological network theory, which I use to refer to the theoretical framework that is used as a rationale for the use of network models in psychology, could arguably be considered a psychometric theory as well, and conceptualizes psychological constructs as clusters of behaviors that are positively associated because they mutually reinforce each other. So, according to psychological network theory, enjoying parties, being talkative and easily making new friends are associated because they reinforce each other, and this cluster of associated behaviors could

be called ‘extraversion’ (Cramer et al., 2012).

These different theories about how to conceptualize psychological constructs and how to relate these constructs to observable behavior can be formally defined in a set of equations and assumptions that make up a *psychometric model*. The chapters in this dissertation focus on two types of psychometric models: Latent variable models and network models. Part I of the dissertation focuses on the interpretation of the latent variable model. Part II of the dissertation makes a comparison between latent variable models and network models. In what follows, I will briefly introduce these two models, before outlining the individual chapters.

1.1.1 LATENT VARIABLE MODELS

Latent variable models were first introduced into psychometrics by Spearman (1904) and quickly gained prominence. The basic idea of latent variable models is that manifest or observed variables reflect a latent variable impossible to observe directly. For example, Spearman (1904) abstracted a latent general factor from a range of cognitive tasks; a technique that is now known as *factor analysis*. The general factor, which he named *general intelligence* or *g* (Spearman, 1927), explained the shared variance among the cognitive tasks, so that associations between the cognitive tasks (e.g., people that have better mathematical skills are also better in memory tasks and spatial visualization) disappear when conditioning on this factor. That is, *g* explains why these cognitive tasks correlate. Spearman (1927) proposed several hypotheses about what *g* represents, of which the most extensively described hypothesis is that *g* measures mental energy. So, while *mental energy* cannot be directly observed, it is reflected in the latent variable that explains the correlations between observed item scores on cognitive tasks. This idea is implemented in a range of different models, such as factor models (Bartholomew & Knott, 1999; Bollen, 1989a; Jöreskog, 1971; Sörbom, 1974; Spearman, 1904), Item Response Theory (IRT; Birnbaum, 1968; Lord, 1952; Mellenbergh, 1994a; Mokken, 1971; Rasch, 1960) models and latent class models (Goodman, 1974; Lazarsfeld & Henry, 1968), that all fall into the wider category of latent variable models.

The latent variable model is formalized as follows. Let $\mathbf{X} = \{X_1, X_2, \dots, X_j, \dots, X_J\}$ denote a vector of J observed variables (e.g., item scores of J items, such as whether someone enjoys going to parties, whether that person is talkative and whether that person easily makes new friends) and let Θ denote a latent variable (e.g., a psychological construct such as ‘Extraversion’). In the latent variable model, the observed variables are a function of the latent variable and an error variable that captures the influence of all other predictors of the observed variable.

When the observed variables are normally distributed continuous variables, the function that links the latent variable to the observed variables (i.e., the Item Response Function [IRF]) can be linear, as is the case in the unidimensional factor model,

$$\begin{aligned} X_1 &= v_1 + \lambda_1 \Theta + E_1 \\ X_2 &= v_2 + \lambda_2 \Theta + E_2 \\ &\vdots \\ X_J &= v_J + \lambda_J \Theta + E_J, \end{aligned} \tag{1.1}$$

where E_j denotes the error score variable, λ_j denotes the factor loading of response variable X_j on the latent variable, and v_j denotes the intercept (Jöreskog, 1971). Here X_j , Θ and E_j are random variables of which the sample space is defined by randomly selecting people from a population (Lord & Novick, 1968, pp. 32–34). With the additional assumption of local homogeneity (Ellis & Van den Wollenberg, 1993), the error score variable and response variable are also random variables within any specific¹ person i , and the model in Equation (1.1) also holds for any person i ,

$$X_{ij} = v_j + \lambda_j \theta_i + E_{ij}. \tag{1.2}$$

For the random variables X_{ij} and E_{ij} , the sample space is defined by sampling an occasion from all possible occasions in which item j could have been administered (i.e., the *replication space* in Lord & Novick, 1968, p. 47). This further specification of the model, resulting in the homogeneous latent variable model, is explained in Chapter 3. When the observed variables are binary variables that are Bernoulli distributed the IRF can be a logistic function as is the case in, for example, the Rasch model (Rasch, 1960), or can be cumulative normal, as is the case in the normal ogive model (Lord, 1952). Latent variable models assume that the joint distribution of $(\mathbf{X}, \Theta)$ satisfies restrictions, of which the most basic class of latent variable models satisfy the following three restrictions (Holland & Rosenbaum, 1986; Junker & Ellis, 1997):

1. Θ is unidimensional
2. $\mathbf{X}$ is conditionally independent given Θ
3. $P(X_j > x | \Theta)$ is monotonically nondecreasing in Θ

¹Lord and Novick (1968) use the subscript $*$ to denote random selection of a person from the population and subscript a to denote the selection of a specific person a , I leave the subscript $*$ out (i.e., I use X_j instead of X_{*j}) and use subscript i to denote the selection of a specific person.

These assumptions make it possible to identify the common factor Θ from the observed variables in $\mathbf{X}$ when $\mathbf{X}$ consists of at least three variables. The common factor can be identified as that dimension that renders the observed variables conditionally independent, which is given by assumption (2). This assumption means that all marginal associations between the observed variables can be explained by their shared reliance on a common factor, and is also known as the assumption of *local independence*. This assumption implies that, given someone's level of extraversion, the observed behaviors (being talkative, enjoying to go to parties and easily making new friends) are no longer associated; whatever these behaviors have in common is captured by the construct extraversion. The assumption that a single underlying dimension can explain away the associations between a set of observed behaviors nicely goes with the view that the psychological construct functions as a common cause of the observed behaviors. This is because *if* a common cause underlies a set variables, this implies that (1) the variables are marginally associated, and (2) the variables are conditionally independent when conditioning on this common cause. These implications of a common cause structure are entailed in the assumptions of a latent variable model given above.

1.1.2 NETWORK MODELS

The psychological construct as a common cause of the observed behaviors works well to explain positive associations between certain behaviors. For example, suppose that *because* someone is extraverted, this person likes to go to parties, is talkative and easily makes friends. In that case, the observation that someone enjoys parties is indicative of this person being extraverted. The observation that the person enjoys parties also indicates that this person is likely to be talkative and easily makes friends, by the logic that enjoying parties indicates extraversion and an extraverted person is likely to engage in these behaviors as well. Thus, enjoying parties, being talkative and easily making friends correlate because extraversion is a common cause of these behaviors. However, such correlations between the behaviors can be explained in other ways as well. For example, at parties, it is easier to meet people and make friends, who in turn might invite you to more parties, which then increases the joy of going to parties. That is, these behaviors mutually reinforce each other (Cramer et al., 2012).

In psychological network theory, psychological constructs refer to clusters of behaviors that directly influence each other. The network model in which this theory is formalized represents observed variables as nodes and the relations between these observed variables as edges that connect the nodes. Instead of a shared reliance on a latent variable, the observed variables are func-

tions of each other. For example, one way of fitting a network model is by regressing each of the variables in $\mathbf{X}$ on all other variables in $\mathbf{X}$,

$$\begin{aligned} X_1 &= \beta_{12}X_2 + \beta_{13}X_3 + \dots + \beta_{1J}X_J + E_1 \\ X_2 &= \beta_{21}X_1 + \beta_{23}X_3 + \dots + \beta_{2J}X_J + E_2 \\ &\vdots \\ X_J &= \beta_{J1}X_1 + \beta_{J2}X_2 + \dots + \beta_{J(J-1)}X_{(J-1)} + E_J. \end{aligned} \tag{1.3}$$

In these regression equations, β_{jk} denotes the regression weight of node j on node k . When β_{jk} equals zero², this means that X_j and X_k are conditionally independent given the other variables in $\mathbf{X}$. The network models used in psychology are typically Markov Random Fields (Koller & Friedman, 2009), which means that nodes are connected by an edge if and only if the variables are conditionally dependent. It is thus important to identify which variables are conditionally independent (these variables should not be connected with an edge) and which variables are conditionally dependent (these variables should be connected by an edge). That is, one wants to identify which β -coefficients in Equation (1.3) are different from zero in the population and connect the variables for which this is the case. One method for identifying edges is to estimate the above regression weights with maximum likelihood estimation and only connect X_j and X_k with an edge if the estimated regression weights $\hat{\beta}_{jk}$ and $\hat{\beta}_{kj}$ are significantly different from zero. Another method for identifying edges that are zero, is to use regularization techniques that penalize the number of edges that are used to describe the covariance structure of the data (Friedman, Hastie, & Tibshirani, 2008). There have been developed multiple different estimation techniques for network models of psychological data (see Epskamp et al., 2012; Epskamp & Fried, 2018; Van Borkulo et al., 2014), of which some are introduced in Chapter 4 of this dissertation.

1.2 INTERPRETATION OF PSYCHOMETRIC MODELS

Earlier, I described that psychometric theories can be formalized in statistical models. But, while a psychometric theory can form a rationale for a psychometric model, the model does not imply the rationale. For example, the belief that general intelligence is the common cause of perfor-

²When β_{jk} equals zero, β_{kj} also equals zero since in that case the covariation between X_j and X_k given all other variables in $\mathbf{X}$ is zero.

mance on a set of cognitive tasks can be a rationale for fitting a factor model to the test scores and interpreting the common factor as the common cause, that is, as *general intelligence*. But, one can also fit a factor model to the test scores on a set of cognitive tasks without any causal interpretation of the model, and interpret the common factor as *the statistical dimension that renders the test scores conditionally independent* or as *the variation in the scores on the different cognitive tasks that is shared among these tasks*. Similarly, one can hypothesize that a set of behaviors influence each other, and therefore try to estimate these direct effects by estimating a network model. The edge weights could then be interpreted as causal hypotheses. But one can also choose to refrain from any such causal interpretation and interpret the edge weights as *partial correlations* or as *conditional dependencies*. While a statistical model does not *imply* a causal interpretation of the model, I will argue in this dissertation that a causal interpretation of psychometric models has several benefits.

I here briefly mention two such benefits to give an idea of what such benefits can be. The most important benefit of a causal interpretation is that a purely statistical interpretation of psychometric models lacks substantive theory and therefore cannot be falsified. For example, consider the general factor model that is estimated from a set of test scores on cognitive tasks. When this model is interpreted as reflecting the hypothesis that there is a common cause that underlies the test scores (e.g., mental energy), which can be true or false, one can find support for this theory or falsify this theory. When g is interpreted as *the shared variance* among the test scores, and the model is interpreted as *a description of the data*, then the model does not bear any relation to a substantive theory that can be supported or falsified. Rather than providing evidence for or against a substantive theory, the fit of the model quantifies how well the statistical factor can summarize the information required to represent the data (Jonas & Markon, 2016).

A second benefit of a causal interpretation of psychometric models is that it enables generalization beyond the observed dataset. If g is merely the shared variance among those exact test scores, then a general factor that is estimated from a new dataset (later timepoint, or different cognitive tasks) does not refer to the same g . That is, because the factor lacks a referent outside the model, statistical analyses on different datasets do not build on each other, and cannot be used as separate tests to strengthen the same theory. Next, I outline the chapters of this dissertation and explain how these chapters are related to the interpretation of psychometric models.

1.3 CHAPTER OUTLINE

The chapters in this dissertation are grouped in two parts. Part one includes chapters that focus on a causal interpretation of the latent variable model. Chapter 2 compares a descriptive interpretation of the common factor model with a causal interpretation of the common factor model, and presents several benefits of the latter over the former. Chapter 3 presents a causal theory of error scores and uses this causal interpretation of the error to make sense of what it means for response variables to be random variables, and also connects this interpretation of the error to psychometric concepts such as reliability, item bias and local homogeneity. Chapter 4 is a commentary that argues that composite variables can also be interpreted causally. Part two includes chapters that focus on the comparison of models from different modeling frameworks and the tensions of *comparing* such models while classes of these models are mathematically equivalent. For example, consider a researcher who has test scores on a set of cognitive tasks, and can fit a network model or a latent variable model to these data. But, it has been demonstrated that the class of network models and the class of latent variable models are statistically equivalent (Epskamp, Maris, Waldorp, & Borsboom, 2018; Marsman et al., 2018; Marsman, Maris, Bechger, & Glas, 2015). Is it then sensible to compare these models, or should they be considered the same model? And if a comparison of these models is sensible, is it even possible to find support for one model over the other? Chapter 5 provides a brief introduction to network models and how they can be estimated from empirical data. Chapter 6 discusses the equivalence of network models and latent variable models, and in what cases a comparison of these models is possible based on diverging predictions for empirical data. The chapter derives two diverging predictions for two subclasses of latent variable models and network models: the unidimensional factor model (Jöreskog, 1971) and the sparse network model (Epskamp, Rhemtulla, & Borsboom, 2017). One of these diverging predictions is that the unidimensional factor model implies that partial correlations are always weaker than zero-order correlations, while this is not the case for network models. Chapter 7 provides a proof of this result and proposes a bootstrap test to identify partial correlations that are significantly stronger than the corresponding zero-order correlation. Chapter 8 focuses on the common effect model in which the observed variables are causes of a latent common effect. The interpretation of common effect models is discussed, and the model is compared to the common cause model, in which the observed variables are effects of the latent variable instead of causes. The chapter demonstrates that the reliance on a common cause and the conditioning on a common effect imply opposite covariance structures: whereas the covariance matrix implied by a common cause model is a diagonal matrix *plus* a rank one matrix

(i.e., observed variables can always be rescaled so that all covariances are positive), conditioning on a common effect implies that the covariance matrix is a diagonal matrix *minus* a rank one matrix (i.e., observed variables can always be rescaled so that all covariances are negative). Chapter 9 discusses the comparison of equivalent models in the context of Caspi et al. (2014) who abstracted the *p* factor of psychopathology from a broad range of psychopathological disorders. The chapter discusses some of the risks of comparing models solely based on model fit when the models are mathematically closely related. In addition, it is argued that support for the conclusion of a general *p* factor that underlies the disorders would benefit from not only comparing models within a single framework (in Caspi et al.'s paper the latent variable framework) but also considering alternative explanations for the data in different frameworks. Finally, Chapter 10 ends with a discussion in which I specify what I mean by *causal relations* in the context of psychometric models, and in which I discuss some of the implications of a causal interpretation of psychometric models.

Part I

Interpreting the latent variable model

2

A causal interpretation of the common factor model

Abstract

Psychological constructs such as personality dimensions or cognitive traits are typically unobserved and are therefore measured by observing so-called indicators of the latent construct (e.g., responses to questionnaire items or observed behavior). The Common Factor Model (CFM) models the relations between the observed indicators and the latent variable. In this article we argue in favor of interpreting the CFM as a causal model rather than merely a statistical model, in which common factors are only descriptions of the indicators. When there is sufficient reason to hypothesize that the underlying causal structure of the data is a common cause structure, a causal interpretation of the CFM has several benefits over a merely statistical interpretation of the model. We argue that (1) a causal interpretation conforms with most research questions in which the goal is to explain the correlations between indicators rather than merely summarizing them; (2) a causal interpretation of the factor model legitimizes the focus on shared, rather than unique variance of the indicators; and (3) a causal interpretation of the factor model legitimizes the assumption of local independence.

This chapter is adapted from Van Bork*, R. , Wijsen*, L. D., & Rhemtulla, M. (in press). Toward a causal interpretation of the common factor model. To appear in *Disputatio*.

*These authors contributed equally.

2.1 INTRODUCTION

One of the many pursuits of psychology is to establish causal relations between properties of the mind and human behavior. Psychologists are interested in the motivations, cognitive abilities and personality traits that explain why people behave in a certain way. This endeavor has proved to be anything but easy. Part of the problem comes down to the very nature of psychological constructs. In psychology, the attributes that are used to explain human behavior (e.g., personality characteristics that explain individual differences in behavior) are typically *latent variables*, that is, theoretical constructs that are unobserved (Borsboom, Mellenbergh, & Van Heerden, 2003). Because latent variables are by definition unobserved, psychological tests are constructed to measure them. The observed responses on such tests are believed to reflect the latent variable that underlies them. For example, *intelligence* cannot be directly observed. Yet, it is assumed that intelligence can be measured by administering a set of IQ items to which the responses can be observed. The measurement model on which this mechanism is based is the *reflective model* (Figure 2.1). The IQ test is supposed to measure intelligence because it is built on the assumption that the responses to the items are a direct *effect* of one unobserved entity, intelligence. Consequently, the variance shared among the test scores is assumed to reflect this theoretical entity. The variance that is unique to each item is assumed to reflect measurement error as well as unique causes of the responses to that item (represented by ε_1 to ε_4 in Figure 2.1). The reflective model is an example of a common cause model because the latent variable functions as a cause of the observed variables. In the remainder of the paper, we refer to the observed variables (such as the test scores in this example) as the *indicators* of the latent variable.

Latent variable modeling originated with the construction of the *common factor model* (CFM; Spearman, 1904). Spearman observed what is now known as the positive manifold: item responses on a variety of mental ability tests were all positively correlated with one another. Using the principle of the common cause in which a correlation between two variables is explained by a third variable, Spearman abstracted the general factor of intelligence (*g*-factor) from these test scores, arguing that one underlying entity explains the shared variance in all branches of intellectual activity: namely, general intelligence. In the CFM all indicators are a linear function of the common factor and the indicators are statistically independent conditional on this latent variable. After all, if a common cause explains the covariation between two indicators, the indicators no longer correlate when conditioning on this common cause. So, a CFM of general intelligence implies that all branches of intellectual activity are rendered independent, conditional on general intelligence. This principle is fundamental to Spearman's CFM and to

Figure 2.1: The reflective model. Squares represent indicators (e.g., X_1 to X_4 could be the responses on four different IQ items) and η represents the latent variable (e.g., intelligence). ε_1 to ε_4 refer to measurement error and unique causes of the indicators. Eliminating these unique factors would imply that the indicators are fully determined by the latent variable, which is typically not assumed for causal processes in psychology

reflective models in general, and is called the *principle of local independence*. The process of fitting a CFM to data is called *factor analysis*.

Factor analysis enabled psychologists to discover and test theories about plausible explanations for human behavior. It is therefore not surprising that CFMs have become very popular, not only in intelligence research, but also in the fields of psychopathology (e.g., Asmundson et al., 2000; Caspi et al., 2014) and personality (e.g., McCrae & Costa, 1987; Musek, 2007).

We reserve the name *reflective model* for the theoretical causal interpretation of the CFM. In contrast to the reflective model, the CFM is not defined as a causal model but is typically defined as the set of equations that equate each indicator X_i to a function of the latent variable η and a unique component ε_i that is typically called *the residual* in factor analysis. For example, consider four indicators X_1 to X_4 of the same latent variable η . Each indicator is a linear function of the same latent variable and a unique residual component:

$$X_1 = \lambda_1 \eta + \varepsilon_1 \quad (2.1)$$

$$X_2 = \lambda_2 \eta + \varepsilon_2$$

$$X_3 = \lambda_3 \eta + \varepsilon_3$$

$$X_4 = \lambda_4 \eta + \varepsilon_4$$

These equations are typically graphically represented in the same way as the reflective model (see Figure 2.1). The covariance matrix of the indicators, Σ , is a function of the vector of factor loadings, λ , the variance of the latent variable, ψ , and the covariance matrix of the residuals, Θ :

$$\Sigma = \lambda\psi\lambda + \Theta \quad (2.2)$$

Although the equations above are agnostic with respect to causality, we argue that when factor analysis is used to measure a psychological construct, the CFM benefits from being interpreted as a reflective model, that is, interpreting the latent variable as the common cause of the indicators. The reflective model in Figure 2.1 can be seen as a specific case of the CFM: it includes the CFM equations above, and adds to them a causal interpretation (Bollen & Lennox, 1991; Edwards & Bagozzi, 2000).

Even though latent variable modeling has had a profound influence on psychological research, methodologists and modelers have not reached a general consensus about how to understand the nature of latent variables (Bollen, 2002; Borsboom et al., 2003; Jonas & Markon, 2016). As latent variables are not directly observed, it is typically impossible to perform a manipulation on them to determine (a) whether they exist at all, and (b) whether they really cause the observed responses that they are supposed to cause. For example, it is impossible for us to manipulate a person's intelligence to test whether the answers on an IQ test change as a result of this intervention¹.

It is not only the unobservable character of psychological constructs that makes causal modeling in psychology a complicated endeavour. Factor analysis is most frequently performed on cross-sectional data, gathered at a single time point. Such data cannot be used to distinguish between models that have different causal structures but are statistically equivalent. Whenever a CFM is fit to a dataset, there are alternative causal structures that may equally well have generated the data (Van der Maas et al., 2006). Because of this ambiguity, many psychometricians advocate sticking to a strict statistical interpretation of the CFM, to avoid inferring causality without direct evidence for it. They argue instead for a descriptivist approach in which the latent variable merely represents the shared variance among a set of indicators. The descriptivist approach understands latent variables as a parsimonious summary of the data, rather than an underlying cause of the indicators (Jonas & Markon, 2016).

¹Note that it is possible to hypothesize an intervention on intelligence, e.g. drinking a lot of alcohol (Borsboom et al., 2003). However, it is problematic to test whether this intervention actually affects intelligence or whether it instead affects the indicators directly, because intelligence is only accessible via its reflection in the indicators.

Thus, although the CFM was developed as a model in which the common factor is hypothesized to represent an existing causal entity (i.e., general intelligence) that explains patterns in different branches of intellectual activity (Spearman, 1904), the descriptivist approach views the common factor that is obtained with factor analysis as “just a convenient way of summarizing patterns of observed relationships” (Jonas & Markon, 2016, p. 91). Jonas and Markon (2016, pp. 91-92) argue:

“Reflective latent variable models are agnostic with regard to the nature of the etiological process: this is the heart of the descriptivist paradigm. Reflective latent variable models can describe data generated from any number of etiologic processes; consequently, the form of the latent variable model cannot arbitrate questions of causality.”

For any arbitrary set of indicators that share variance, the shared variance will constitute a common factor, no matter what caused this shared variance. However, in order to measure a psychological construct of interest, one does not consider any randomly chosen set of variables, but rather a specific set of variables that are hypothesized to be affected by the construct. The hypothesis that the indicators reflect the psychological construct of interest enables the researcher to interpret the common factor as this construct, by the logic that, *if* the indicators share a common cause, the shared variance among these variables reflects this cause (Edwards & Bagozzi, 2000).

In this paper, we argue against the view that the common factor is merely a convenient summary of the data and argue for a causal interpretation of the CFM when used to measure constructs. When a causal interpretation is not justified, for example because it is unlikely that the construct of interest causes the indicators, other statistical models might be more appropriate (see for example, Bollen, 2011; Diamantopoulos, Riefler, & Roth, 2008; Edwards & Bagozzi, 2000). Although a causal interpretation of the CFM is not necessary, it offers several benefits over a purely statistical reading of the model. To make this argument we first distinguish between statistical models and causal models, and explain why researchers might be reluctant to interpret the CFM causally. Subsequently, we outline the descriptivist approach in more detail. Finally, we present three arguments that highlight the benefits of a causal interpretation over a purely statistical interpretation of the factor model.

2.2 STATISTICAL VS. CAUSAL MODELS

We distinguish between statistical models on the one hand, and statistical models with a causal interpretation on the other. For brevity, we use *causal model* to refer to the latter, though we acknowledge that causal models need not be statistical. A statistical model is a set of probability distributions on some sample space (McCullagh, 2002). This means that no conceptual interpretation is yet included; a statistical model only includes the statistical dependencies among indicators (Moneta & Russo, 2014). In contrast, a causal model has an additional causal interpretation in which one or more of the parameters in the model reflect causal relations. A causal model could be seen as a representation of a real-world phenomenon, rather than a model that only plots the relations in the data. We are aware that there exists a large literature on the definition of a model and the distinction between different types of models in philosophy of science (e.g., Frigg & Hartmann, 2006; Van Fraassen, 2008), but a discussion of this literature is beyond the scope of this paper.

Two models that represent different causal relations can be statistically equivalent. For example, many structural equation models are statistically equivalent to a model in which one of their structural relations is reversed (MacCallum, Wegener, Uchino, & Fabrigar, 1993). When two models are statistically equivalent, any observational dataset provides equal support for both of them; however, because they have different causal implications, it is possible to distinguish them via experimental intervention. Consider two statistically equivalent models: (I) $A \rightarrow B \rightarrow C$ and (II) $A \leftarrow B \rightarrow C$. Any possible observed correlation matrix between the variables A , B , and C will provide equal support for both models, though the arrow connecting A and B is reversed. The causal relation between A and B is therefore different for each model. Now suppose that A , B and C are measured at two points in time, and that between these time points, variable B is experimentally manipulated. Both the correlational structure at time point 1 and time point 2 will provide equal support for both models, but model (I) and (II) imply different expectations with respect to the increase or decrease in variable A as a result of the intervention on variable B . If intervening on B changes A to the extent predicted by model (II), that is evidence for model (II) over model (I).

Because causally different models can be statistically equivalent, some may argue that it is unjustified to make any causal interpretations of statistical models, and prefer the descriptivist approach. Against this view, we argue that it is often best to enrich a given statistical model with a causal interpretation and thus, we argue for the interpretation and use of the CFM as a causal model. In the next section, we take issue with one of the main arguments for interpreting the

CFM merely as a statistical model rather than a causal model. According to this view, a causal interpretation is not justified because factor models are typically estimated from the covariance structure of cross-sectional data without considering interventions on any of the variables over time.

2.3 CORRELATION DOES NOT ENTAIL CAUSATION

Factor analysis is typically applied to the correlation structure of cross-sectional data. However, a causal interpretation of the CFM implies that factor analysis infers causal relations between the latent variable and the indicators from observational data without having manipulated or intervened on anything. So how can factor analysis be used to hypothesize causal relations if it is based on correlational data?

Although correlations do not entail causality, causality does entail covariation. In addition, particular causal structures imply particular patterns of conditional independence among variables. For example, when $A \leftarrow B \rightarrow C$ is hypothesized, in which B is a common cause of A and C , then one expects covariation between A , B and C , and one expects that A and C are conditionally independent given B (Reichenbach, 1956). Also, one expects that the correlation between A and C is smaller than the correlations between A and B and between B and C . These are all implications of a causal structure that can be verified from a correlation matrix. Similar expectations can be laid out for the reflective model (Figure 2.1), in which the latent variable is a common cause of its indicators: one expects the indicators to covary in a certain way, e.g., two indicators that are strongly correlated with each other should also share much variance with the other indicators. After all, indicators that consist of relatively more shared variance and less unique variance are more reliable indicators than those that consist of relatively little shared variance and more unique variance. Reliable indicators will also correlate more strongly with each other. Unlike in the 3-variable example cited earlier, however, in latent variable modeling it is impossible to test conditional independence given the common cause, because the common cause is unobserved. All in all, causation does lay out some expectations for observational data, though one should always consider the possibility of alternative causal structures that imply similar expectations for the data (e.g., the expectations summarized for $A \leftarrow B \rightarrow C$ also correspond to the structure $A \rightarrow B \rightarrow C$).

These implied constraints of the common cause structure are tested in the fit of a CFM. Violation of local independence will result in poor model fit. As noted earlier, although the hypothesis of a common cause can be rejected because of the testable constraints it puts on the

data (e.g., tetrad rank constraints; see Bollen & Ting, 1993), the common cause structure cannot be verified from observational data alone because there will always exist alternative causal structures that can explain the observations in the data. But even though factor analysis cannot verify a common cause structure, hypothesizing such a structure justifies the use of factor analysis to measure the latent variable in the first place. After all, the constraints that are tested when fitting a CFM to the data are all implied by the common cause structure.

In the next section, we elaborate on the descriptivist approach and then we proceed to developing three arguments in favor of a causal interpretation of the CFM.

2.4 WHAT IS THE COMMON FACTOR IN THE DESCRIPTIVIST APPROACH?

As described by Jonas and Markon (2016), the descriptivist approach entails that a common factor should not be interpreted as a real-world entity, but rather as mere shared variance that is nothing but a parsimonious summary of the data. From this perspective, latent variables are not postulated as concrete entities that have direct causal effects on their indicators, but rather as summaries of the covariance among indicators. When the latent variable is defined as the shared variance of a set of indicators, it cannot also be a common cause of the indicators. After all, the shared variance among indicators does not cause the shared variance among indicators; something cannot have caused itself. Thus, when the latent variable is the shared variance rather than being a cause of the shared variance, the arrows in Figure 2.1 pointing from the latent variable to the indicators should be interpreted as “is part of”. The latent variable *is part of* the indicators, by the logic that the variance in the indicators consists of shared variance (arrow pointing from a common component to the indicator) and unique variance (arrow pointing from a unique component to the indicator). That is, when the latent variable is merely shared variance, the relation *intelligence* $\rightarrow$ *IQ* test should be interpreted in the same manner as *boys* $\rightarrow$ *people*: just as the set of boys is part of the set of people, intelligence is part of IQ test scores. Whether a variable is part of another variable or causes another variable, statistically speaking there is no difference between these two relations. More concretely, whether the latent variable is a part of the variance of the indicators (i.e., the variance that is shared among indicators) or is a cause of the indicators, both result in the same statistical model.

2.5 TOWARD A CAUSAL INTERPRETATION

Whereas descriptivists argue for a statistical reading of the factor model, we believe that a causal interpretation of the model is in many cases more appropriate. It is important to note here that our objections to the descriptivist approach are limited to cases in which the CFM is used as a measurement model rather than as a method for data reduction. When factor analysis is used as a data reduction method, the common factor can be used to predict other variables, but the common factor does not refer to anything outside the model itself. In data reduction, researchers are only interested in bringing a large set of variables down to a smaller set of dimensions. In other words, they are interested in a concise summary of the data, rather than in causal connections between latent variables and indicators.

In the next subsections we present three arguments for a causal interpretation of a CFM over a merely statistical interpretation when the goal is to measure psychological constructs. We argue that (1) establishing causal relations conforms with most research questions in which the goal is to explain the correlations between indicators rather than merely summarizing them; (2) a causal interpretation of the CFM legitimizes why we are interested in the shared variance rather than in the unique variance of the indicators; and (3) a causal interpretation of the CFM legitimizes the assumption of local independence.

2.5.1 “LOOK, WE FOUND SHARED VARIANCE!”

Social scientists are typically interested in the best possible explanation for their observations. They want to know how and why observations occur the way they do. Borsboom et al. (2003) argued that latent variable models require a realist ontology in order to use them for establishing causal connections between latent variables and indicators. The choice for latent variable models, in which a set of indicators is assumed to covary *because* of this latent variable, already implies that one assumes this latent variable exists. Choosing a latent variable model therefore naturally implies a realist ontology. In other words, when the aim is to detect and theorize about a putative causal relation, it does not make sense to avoid a realist, causal interpretation all together.

To illustrate this point, we will use a study performed by Caspi et al. (2014) as a special case that in our view highlights why a causal reading is sensible. In their study, Caspi et al. (2014, p. 120) “evaluate alternative hypotheses about the latent structure underlying 10 common mental disorders”. The authors conclude that a bifactor model, with three group factors and one general factor, best explains the structure of psychopathology, and they name this general psy-

chopathology factor *the p factor*. The descriptivist view would only allow the researchers to conclude that their study provides support for shared variance among a set of disorders. With a causal interpretation though, their findings would provide support for a meaningful hypothesis, namely that the p factor causes mental disorders to covary, and explains comorbidity. Here, a descriptivist approach is simply unsatisfying and uninteresting.

Of course it is legitimate to claim that a single dimension is able to account for most of the co-variation among disorders, resulting in a parsimonious description of the data. But if the goal is to explain the structure of psychopathology, this claim is greatly unsatisfying: shared variance as such has no explanatory value. In contrast, when the shared variance reflects a common cause, this common cause does explain the correlations between indicators, resulting in a better understanding of the observations and an opportunity to further the research programme by searching for the identity of the common cause. A common cause not only renders the indicators independent when accounted for, the indicators correlate because they share a common cause. Explanation therefore has additional value to mere description, and in the end, establishing causal relations and finding explanations is essential to science. Thus, the p factor model has much greater theoretical import as a causal model than as a merely statistical model. Researchers should, however, try to verify whether this causal model for the structure of psychopathology is justified in each case.

2.5.2 WHY SHARED VARIANCE AT ALL?

The CFM distinguishes between shared variance (that which is shared by all indicators) and unique variance (variance which is unique to each indicator). Shared variance is attributed to the latent variable, while unique variance is attributed to residual influences, such as measurement error. The distinction between shared and unique variance, of which only the former is of interest in latent variable modeling, makes sense when a common cause underlies the indicators but is not sensible under alternative underlying causal structures. Put differently, the belief that a common cause underlies the indicators legitimizes that only shared variance is of interest rather than unique variance of the indicators.

Consider an example in which the indicators reflect some variable that is not a common cause of the indicators. For example, *healthy eating* is a variable that can be indicated by responses to questions like “do you eat a lot of junk food?”, “do you often eat fruit?”, and “how much sugar do you eat?”. These three items are all indicators of *healthy eating*, however, healthy eating is probably not a cause of these indicators. In this case it is not sensible to only take into account

the variance that is shared by these items: the unique variance of the indicators is equally important to the construct *healthy eating*, even though these indicators are correlated and would likely result in a well-fitting reflective model.

The items *eating vegetables* and *avoiding sugar* are both relevant to the construct *healthy eating*, regardless of what causes them or how much of that variance is shared. That is, it does not matter *what* causes someone to eat a lot of vegetables (e.g., having a big vegetable garden) and whether that is the same cause as for not eating too much sugar (e.g., having strict parents). Both of these are indicators of healthy eating, including their unique variance. When a factor model is fit to these data, variance due to these unique causes is relegated to the residual component of the model. But arguably, the unique variance due to vegetable gardens and strict parents is no less relevant to *healthy eating* than the shared variance that may be due to causes such as a person's motivation to be healthy. As such, when the construct of interest is not a common cause, there is no theoretical reason to disentangle the shared and unique variance of the indicators. Doing so risks a biased representation of the construct of interest.

The above argument can just as well be applied to psychological constructs of which the exact nature is unclear. For example, *extraversion* is typically measured with a reflective model. Suppose that *extraversion* is just a summary of the scores on a set of extraversion items. In that case, why use the shared variance as a summary, rather than the unique variance or all variance? What is the justification for discarding unique variance if the shared variance holds no special status? There is none. The shared variance is just one way of reducing high dimensional data to fewer dimensions, and there are alternative techniques for data reduction that do not discard unique variance, e.g., Principal Component Analysis (PCA; James, Witten, Hastie, & Tibshirani, 2013). Thus, when the status of the construct is either known to be not a common cause (as in the *healthy eating* example) or is unknown (as in the *extraversion* example), the practice of interpreting only the shared variance among a set of items is not well supported.

In contrast, when one's theory explicitly states that the construct is a common cause of a set of items, it is immediately legitimate to interpret only the shared variance of the indicators because it is precisely this shared variance that must be due to the common cause. Returning to the Caspi et al. (2014) study, we can draw a similar conclusion. The reason why the authors are at all interested in the shared variance rather than the total variance, or the unique variance, is because they must have reasons to believe that the shared variance reflects something that has certain explanatory power. Thus, the difference between a latent variable that is interpreted as merely shared variance and a latent variable that is believed to be a common cause of the indicators is that the former *results* from the distinction between unique and shared variance in

a CFM, while the latter *legitimizes* this distinction.

2.5.3 LOCAL INDEPENDENCE

The principle of local independence is a fundamental assumption of the reflective model and comprises the idea that when a common cause underlies a set of indicators, conditioning on this common cause renders the indicators statistically independent (Borsboom et al., 2003). When factor analysis is performed on a dataset, it searches for a factor solution that meets this criterion. The assumption of local independence does not apply to all data reduction models. As stated before, PCA also results in a parsimonious summary of the data but in contrast to factor analysis, PCA does not use local independence as a criterion. PCA rather composes a variable that accounts for most of the variability in the indicators (James et al., 2013). In this process, PCA takes all variance into account, not only shared variance. Thus in the case of PCA, the indicators are not rendered independent, yet PCA provides a parsimonious summary of the data.

A purely statistical interpretation of the reflective model does not legitimize the principle of local independence and therefore the use of the reflective model. In contrast, a causal interpretation of the reflective model implies that local independence should hold, so a latent variable model should be preferred when theory holds that a common cause is responsible for covariation among items. Getting back to the example of the p factor, the reason why the p factor is constructed in a way that it renders the disorders independent is because it is believed to reflect a psychopathology factor that forms a common cause to these disorders. A mere summary of the data is not bound to such a constraint on the data. So not only does a causal theory justify a focus on the shared variance, it also justifies the assumption of local independence. Instead of local independence merely *resulting from* absorbing shared variance, a common cause *explains why* local independence is assumed.

2.6 HOW TO ASSESS WHETHER A COMMON CAUSE STRUCTURE IS CORRECT

Using a CFM in the descriptivist framework does not require that a specific causal structure is hypothesized. It simply does not matter exactly what causal structure underlies the data for the latent variable to be a useful summary of the data. For the causal interpretation, in contrast, it is of central importance that the model accurately represents the underlying causal structure. Attaching causal meaning to a CFM that does not accurately represent the actual causal system

will result in incorrect predictions.

Consider four indicators that all influence each other, resulting in shared variance among these variables. A CFM can be applied to the correlation matrix of these variables, and, depending on the strengths of the causal relations between the indicators, this may result in a well-fitting model. In such a situation, the factor model may be useful as a parsimonious prediction model, but as a causal model this factor model makes incorrect predictions with respect to intervention on any of the indicators. For example, a causal interpretation of the factor model implies that intervention on any of the indicators would not affect the other indicators, because all covariation is purported to arise due to a single latent common cause. In contrast, the true underlying causal structure implies that intervention on any of the indicators leads to changes in the other indicators, because the shared variance truly reflects direct causal relations among indicators. So, a causal interpretation of an estimated model can lead to false conclusions when the causal structure of the data-generating mechanism is not represented accurately.

This begs the question of how to assess whether the true data-generating mechanism has a common cause structure. How do we figure out whether the principle of the common cause applies in cases in which we do not have easy access to the real underlying causal structure? As stated before, ideally one would manipulate the common cause, and see what happens to the variation in the indicators (Borsboom et al., 2003). If this were possible, one could directly observe the effects of the manipulation and consequently give causal meaning to the model. But when the common cause is latent, such direct manipulation can be impossible. In some cases, however, an alternative causal model might offer different predictions for manipulations on the indicators. For example, if an alternative model states that the indicators cause each other, such that the shared variance is explained by causal influences between the variables rather than by a common cause, intervening on the indicators would differentiate the two models. Whereas a model with causal influences between indicators implies that certain indicators change as a result of interventions on other indicators, a common cause model implies that indicators are not affected by other indicators. These diverging predictions enable the researcher to differentiate between such alternative models that explain the shared variance among a set of variables.

When intervention is not possible, empirical tests of the goodness of fit of the reflective model may shed light on whether the common cause model is a likely generating model for a particular data set. Although a researcher cannot determine what causal structure underlies cross-sectional data, the common cause model does put testable constraints on the covariance structure of the data. Widely used test statistics and fit indices for confirmatory factor analysis test the hypothesis that the covariance structure of the data matches the covariance structure

implied by a common cause model. Tests based on the pattern of partial correlations in the data may also allow researchers to determine whether a common cause model is more likely to underlie the data than an alternative model that posits direct effects among indicators (Van Bork, Rhemtulla, & Borsboom, 2015; Van Bork, Rhemtulla, Waldorp, & Borsboom, 2016).

We believe that it is especially important to think about the data-generating process in the phase of test construction. As we mentioned before, researchers do not pick an arbitrary set of indicators but select those indicators that are hypothesized to be affected by the construct of interest. This is an important point that concerns the stage of test construction rather than test-analysis. Borsboom, Mellenbergh, and Van Heerden (2004, p. 1067) write:

“a century of experience with test construction and analysis clearly shows that it is very hard to find out where the scores are coming from if tests are not constructed on the basis of a theory of item response processes in the first place”

They continue with the conclusion:

“Thus, it is suggested here that the issue may not be first to measure and then to find out what it is that is being measured but rather that the process must run the other way. It does seem that if one knows exactly what one intends to measure, then one will probably know how to measure it, and little if any validation research will be necessary.”

Another possibility to give causal meaning to the factor model, is that a factor model can be posited as a plausible hypothesis for a certain phenomenon (Haig, 2005). When theoretical considerations imply that the principle of the common cause describes a certain mechanism accurately, the factor model can be used as a method for theory generation. Specifically, exploratory factor analysis can be used for the generation of plausible theories, and through confirmatory factor analysis, these theories can then be evaluated. Through methods such as cross-validation and replication, the theory can gain additional support. This of course does not result in absolute certainty that the causal reading is in fact the true reading, but by choosing the models that have more explanatory value over those that are less explanatory, the best model is left standing. This way, it is possible to gather evidence for the hypothesis that a common cause structure is indeed applicable to the phenomenon in question.

All in all, an inevitable conclusion of the causal interpretation of the factor model we defend, is that the data-generating process matters. Therefore, one should always try to assess whether it is plausible that the data-generating mechanism has a common cause structure.

2.7 CONCLUSION

If the goal is to measure psychological constructs and make meaningful claims about how such constructs relate to each other, we argue for a causal interpretation of the CFM rather than a merely statistical interpretation proposed by the descriptivist approach. First, a causal interpretation matches the scientific aim of explaining observed patterns in behavior rather than summarizing them. Second, a causal interpretation of the CFM legitimizes why the shared variance is of interest rather than the unique variance of the indicators. And lastly, a causal interpretation of the CFM legitimizes the assumption of local independence. A statistical reading of the model does not offer these three things.

We are aware that a causal interpretation of reflective models brings a host of other problems. Reading a reflective model causally does not suddenly turn a non-causal mechanism into one that is in fact causal. Reflective models are only tools to posit plausible causal relations, and they need additional input in the form of causal knowledge or assumptions to establish what kind of causal relation it is. These models are only meaningful when users impose meaning onto them, and it is up to these users to then verify whether the hypotheses represented by the model are empirically supported. In other words, researchers construct plausible causal hypotheses, fit a model that imposes these hypotheses on the data, and then consider how the output supports their hypothesis.

We also acknowledge that there are situations in which a statistical non-causal interpretation of the CFM makes sense. When a common cause structure is unlikely to underlie the data, a purely statistical interpretation of the reflective model can still be useful for prediction. For example, although the existence of general intelligence is often disputed, the g-factor can be used to merely predict school success, without having to establish its objective existence. But even for data reduction, other methods are available (e.g., PCA) that do not rely on latent variables, local independence and a distinction between shared and unique variance in the indicators. It is not entirely clear why factor analysis should be used rather than other data reduction methods such as PCA, when the first model implies constraints on the data that can not be justified from a descriptivist approach whereas the latter does not.

Ultimately, we plead for an understanding of the CFM as an explicit causal hypothesis that can be falsified. In situations where the data-generating mechanism matters for the theory about a psychological construct (e.g., should depression be understood as a cause of its symptoms or as the name for a system of interacting symptoms?), treating the CFM as a summary rather than a hypothesis about this causal structure takes away the need to test the data-generating mecha-

nism. Rather than concealing the urge to infer causal relations by restricting the interpretations of these models to descriptions of the data without referent outside of the data, we argue that one should dare to hypothesize. By making hypotheses explicit (e.g., “we hypothesize that construct A is the common cause of the indicators X_1 to X_j ”), they are open to falsification. Additionally, interpreting the CFM causally stresses the need for theories about how differences in the latent variable result in differences in the responses; theories that we think are crucial for the measurement of psychological constructs.

3

A causal theory of error scores

Abstract

Modern test theory uses probabilistic models to relate the latent variable to response variables (i.e., an item score or a test score). In these models, the response variables are random variables. For the model to be probabilistic it is essential to include errors; without errors the model is deterministic. Despite this important implication of including errors in a model, a substantive interpretation of error, that is, an interpretation that goes beyond ‘a residual that is unexplained by the rest of the model’ is often left implicit. We present a causal interpretation of errors, in which the error score results from all unique determinants of the response variable. In this framework, we distinguish unreliable determinants, which we call “circumstance variables” and reliable determinants, which we call “characteristic variables”. We show how different assumptions about error (1) imply different psychometric models, (2) have different implications for the chance experiment that makes the response variables random variables, and (3) have implications for the reliability of item scores, item bias and local homogeneity.

This chapter is in preparation for publication as Van Bork, R., Rhemtulla, M., Sijtsma, K., & Borsboom, D. (in preparation). A causal theory of error scores.

3.1 INTRODUCTION

A student who takes an IQ test the day after a long night out might claim that her score is not reflective of her intelligence because her fatigue affected her score. Likewise, for a student who finds it hard to concentrate, a particularly good night of sleep or a cup of strong coffee can positively affect his IQ score. Or, when a person takes a math test but is unfamiliar with some of the words used to introduce the math problem, and as a result the person does not understand the introduction well, it can be argued that the score resulting from this test is not a good measure of this person's math ability. These claims stem from the intuitive ideas that "if I had done this same test on a day after a good night's sleep I would have scored higher", "if I had not had a cup of coffee before the IQ test I would have scored lower" or "if I had understood these words in the introductory story, I would have been able to solve the math problem".

Such arguments rest on the idea that the test score is not perfectly determined by the attribute that is measured by the test, but that factors that are unrelated to the measured construct, such as fatigue or familiarity with the vocabulary, also influence the test performance. These factors can cause two important psychometric phenomena: (1) test scores are not perfectly reliable (i.e., even when a person's attribute level is fixed, the test scores vary over test occasions, such as temporary changes in energy level as a result of the amount of sleep or a cup of coffee; Lord & Novick, 1968; Mellenbergh, 1996) and (2) test scores are biased with respect to the attribute being measured (e.g., when test results are influenced by a group factor unrelated to the measured attribute, such as mother tongue affecting people's familiarity with words used to introduce a math problem in a math test; Mellenbergh, 1989; Meredith, 1993; Millsap, 1997, 2007). Whereas (1) concerns variation of an individual's score across testing occasions, given a fixed attribute level, (2) concerns variation between people or groups of people, given the same attribute level. Modern Test Theory (MTT) is a paradigm in which latent variable models (Junker & Ellis, 1997; Mellenbergh, 1994a; Moustaki & Knott, 2000) are used to analyze and construct tests that measure mental attributes. In MTT, both types of factors make up the error component of the model and are typically referred to as (1) "random measurement error" and as (2) "systematic error" or "unique factors".

The examples illustrate an intuition about what is considered error. For many measurement procedures one can come up with examples of factors that can be classified as error, based on an intuition of which factors one wants to have an influence on the test score (the measured attribute) and for which factors one wants to reduce the influence on the test scores as much possible (the factors that are considered error). Despite this intuitive understanding of which

factors qualify as error, a substantive interpretation of the error scores in a statistical model is typically left implicit. That is, while most variables in the model are interpreted as reflecting meaningful constructs, errors are commonly described as residuals, the remaining variability that is not accounted for by the rest of the model (Kline, 2016, p. 13; Newsom, 2015, p. 2; Raykov and Marcoulides, 2000, p. 13). We turn this around and start with a causal theory of error, and show that many concepts in MTT such as reliability, item bias and the underlying chance experiments that make the response variables random variables, follow from different assumptions about error.

Latent variable models such as the Item Response Theory (IRT) model (e.g., Birnbaum, 1968; Goldstein & Wood, 1989; Guttman, 1950; Lord, 1952, 1980; Mellenbergh, 1994a; Mokken, 1971; Moustaki & Knott, 2000; Rasch, 1960; Thissen & Steinberg, 1986; Van der Linden & Hambleton, 2013) and Confirmatory Factor Analysis (CFA) models (e.g., Bartholomew & Knott, 1999; Bollen, 1989a; Jöreskog, 1971; Sörbom, 1974; Spearman, 1904) are *statistical models*. A statistical model is defined as a set of probability distributions on the sample space (McCullagh, 2002), where the sample space is the set of all possible outcomes of a chance experiment (Knight, 2000). For example, one can randomly sample a person from some population and establish this person's age. The sample space consists of all possible ages that could have been the result of the chance experiment. A random variable is a function that maps the sample space to the real line (Knight, 2000), so that each realization is somewhere on this line and has a probability associated with it. For example, it is more probable to sample a person aged 40 than aged 100, so that 'age' can be a random variable for which the realization '40' has a higher probability than the realization '100'.

The use of statistical models to draw inferences from item responses thus relies on the assumption that the response variables are *random variables*, and thus that some chance experiment, such as random sampling, underlies their realizations. More specifically, in MTT the response variables have a probability distribution conditional on the latent variable. For example, responses to a math item may result in a binary response variable that has two realizations: 0 (incorrect/negative response) and 1 (correct/positive response), which both have a probability given a fixed math ability level. Higher levels of math ability are associated with higher probabilities of a correct response, as defined by the Item Response Function (IRF) used to describe the relation between attribute levels and item response probabilities (Borsboom & Molenaar, 2015; Irwing, Booth, & Hughes, 2018; Mellenbergh, 1994a).

Holland (1990) distinguished two rationales for interpreting the response variables in an IRT model as random variables conditional on the latent variable. They are the *stochastic subject*

rationale, in which an individual person's item response is random, and the *random sampling* rationale, in which the individual's item response is fixed, but randomness arises because the individual is sampled from a population. These two rationales give a different interpretation to the phrase: "the conditional probability of a positive response to the item '2 + 7' given a fixed math ability level θ (i.e., the targeted attribute) is .7". In the stochastic subject rationale, this phrase is interpreted as "each person with math ability level θ has a probability .7 of giving a correct response to the item '2 + 7'". In the random sampling rationale, this phrase is interpreted as "the probability is .7 to sample a person who gives a correct response to the item '2+7' from the population of people with math ability level θ ".

In the current article, we propose a causal theory of errors in psychometric models, in which the errors refer to all unique factors that influence the item responses other than the target attribute. This framework reveals that the two chance experiments described by Holland (1990) refer to two competing empirical hypotheses about how item responses originate, with diverging empirical predictions. Thus, while it may seem that the interpretation of the conditional probabilities in MTT is a theoretical choice (Klauer, 1995, p. 97; Molenaar, 1995, p. 5), we argue that the question of which chance experiment is a more accurate description of reality is an empirical question, not a theoretical one.

The causal theory presented in this article rests on the assumption that the variables in the psychometric models that we discuss influence each other in a *deterministic* system. That is, while one can argue that the act of responding to an item itself is a stochastic event (de Boeck & Wilson, 2004, pp. 49-50), for example, because the cognitive process resulting in the response is a stochastic process (Van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011), we assume that responding is not intrinsically stochastic but that stochasticity results from either sampling a person from a population, or sampling a measurement occasion from all possible occasions in which the item could have been administered. Such measurement occasions are characterized by uncontrolled factors in the measurement procedure that cause variability in the item responses. We elaborate on such factors and what we mean by occasions in the main argument of the article. More specifically, in the famous hypothetical example of Mr. Brown who is repeatedly asked a question while being brainwashed in between questions (Lazarsfeld, 1959, in Lord and Novick, 1968, p. 29), we assume that if Mr. Brown is brainwashed in a way that his psychological state is the same when attempting a question, and the circumstances in which he answers questions is also held exactly constant – that is, there is no variability in measurement occasions – he will give the exact same response to each repeated question.

In the following section, we introduce IRT and CFA models. Next, we explain the notion

of reliability in the Classical Test Theory (CTT) framework, because we will use the notion of reliability to disentangle the empirical implications of the stochastic subject rationale and the random sampling rationale. We then propose the causal framework of error scores. In this framework, we distinguish two interpretations of error: (1) error is a composite of *circumstance variables*; these are perfectly *unreliable* variables that are independent of the measured attribute but influence a person's response to that specific item, and (2) error is a composite of *characteristic variables*; these are perfectly *reliable* variables independent of the measured attribute that influence the response to that specific item. We argue that these different assumptions about the reliability of the factors that make up the error correspond to the two rationales proposed by Holland (1990). In the causal framework, the stochastic subject rationale implies that the error score variable is a composite of only circumstance variables, while the random sampling rationale implies that the error score variable is a composite of only characteristic variables. Finally, we discuss how the causal theory of errors relates to the notion of item bias, local homogeneity and reliability.

3.2 ERROR IN MODERN TEST THEORY

Some problems can be described by means of a simple mathematical model of the general form (Lord & Novick, 1968)

$$X = \phi(\Theta), \quad (3.1)$$

where X is an observed dependent variable, Θ is an independent variable, and ϕ is some known function that relates X to Θ . In this function, X is completely determined by Θ and the model is therefore considered to be a deterministic model. Fixing Θ to a specific value θ determines a realization of X ,

$$x = \phi(\theta). \quad (3.2)$$

In practice, one often finds that given a fixed value $\Theta = \theta$, the dependent variable still varies a little bit over repeated measurements, due to uncontrolled determinants that the model does not explain. When the remaining variation in X given $\Theta = \theta$ (henceforth abbreviated to ' X given θ ') is small enough to be neglected, Equation (3.1) can be useful even though it ignores this variation.

In psychology, the residual variation in X given θ is often quite large compared to the total variation in X , which means that X can only be predicted inexactly from Θ and the remaining variance cannot be ignored. For this reason, deterministic models are rarely or never used in

psychology, and instead an error term is introduced, resulting in a probabilistic model

$$X = \phi(\theta) + E. \quad (3.3)$$

We use $\mathbf{X} = [X_1, \dots, X_j, \dots, X_J]$ to denote the vector of random variables, in which X_j denotes a random variable with scores on item j , of which the sample space is defined by the random selection of people from a population (Lord & Novick, 1968, pp. 32–34)¹. Let $\mathbf{x} = [x_1, \dots, x_j, \dots, x_J]$ denote the realizations of $\mathbf{X}$. In MTT, $\mathbf{X}$ is regressed on Θ , in which Θ refers to a psychological attribute (e.g., math ability) on which people show systematic variation. We focus on the situation where Θ is a unidimensional latent variable. In this case, θ_i exhaustively characterizes the attribute level of person i .

Most of the important MTT models are specializations of Equation (3.3) by (a) substituting appropriate distribution functions for $\mathbf{X}$ and Θ , and (b) choosing the appropriate IRF to relate $\mathbf{X}$ to Θ . For example, in CFA, $\mathbf{X}$ consists of continuous variables that are normally distributed, and the IRF that relates these variables to Θ is linear. In the IRT models we discuss in this article, $\mathbf{X}$ consists of binary variables that are Bernoulli distributed and the IRF that relates these variables to Θ is logistic or cumulative normal. These models are probabilistic models of the general form in Equation (3.3) in which E accounts for residual variation in X given θ .

In the following, we propose a causal interpretation of error scores in CFA models and IRT models, and show that error can be a composite of both reliable and unreliable determinants of the response variable. To make this argument, we first use the next section to introduce the notion of reliability in the CTT framework.

3.2.1 ERROR AND RELIABILITY IN CTT

In CTT, response variable X_j is a function of two variables. The variables are the true-score variable T_j and a random measurement-error variable E_j , which are both random variables of which the sample space is defined by the random selection of people from a population (Lord & Novick, 1968, pp. 32–34),

$$X_j = T_j + E_j. \quad (3.4)$$

¹Lord and Novick(1968, pp. 32–34) use the subscript $*$ to denote the random selection of people from a population: X_{j*} . This differs from the subscript i , which denotes the selection of a specific person i . We leave out the subscript $*$.

The true-score of a person i is operationally defined as the expected value of the response variable of person i over repeated measures (Lord & Novick, 1968, p. 30),

$$\tau_{ij} = E(X_{ij}), \quad (3.5)$$

where X_{ij} is a random variable for a specific person i defined over measurement occasions. That is, for a fixed person i , X_{ij} has a probability distribution that includes the probabilities for all realizations of X_{ij} in repeated measurements. The subscript i that distinguishes X_j from X_{ij} is important, because it distinguishes a variable that is defined over persons (X_j), so that the mean μ_{X_j} is the population mean, and a variable for person i that is defined over measurement occasions (X_{ij}), so that the mean $\mu_{X_{ij}}$ is person i 's mean over repeated measures. The probability distribution of X_{ij} is what Lord and Novick (1968, p. 47) call a *propensity distribution*. The true score of person i on item j , τ_{ij} , is a parameter, hence we use τ rather than T , which is a random variable defined over persons. The measurement error is operationally defined as the discrepancy between the observed item score and the true score,

$$E_{ij} = X_{ij} - \tau_{ij}, \quad (3.6)$$

where E_{ij} just like the response variable X_{ij} is a random variable for a fixed person i .

The main reason for setting up true-score theory is to define the concept of reliability (Cronbach, 1947; Lord & Novick, 1968; Miller, 1995; Sijtsma & van der Ark, 2015). The reliability of the response variable X_j , is defined as the proportion of observed-score variance of X_j that is true-score variance,

$$\rho_{X_j T_j}^2 = \frac{\sigma_{T_j}^2}{\sigma_{X_j}^2}. \quad (3.7)$$

The square root of the reliability is the correlation between the true-score variable, T_j , and the observed score variable, X_j . When the reliability of a measure equals 1, the measure gives the same results over repeated measurement occasions and thus all variance in X_j reflects variability in true scores between people. When the reliability of a measure is 0, people have the same true score and all variance in X_j reflects variability over repeated administrations. The error score E_j in Equation (3.4) thus has reliability 0 by definition.

For the purpose of this article, it is useful to distinguish perfectly reliable variables (which have unit reliability in the CTT sense) and perfectly unreliable variables (which have zero reliability). When X_j is perfectly *reliable*, all variability in X_j is variability between persons and given any fixed person it no longer is a variable but a constant across replications. The probabil-

ity of a realization of X_j , $P(X_j = x_j)$, is the probability of sampling a random person with this realization from the population (i.e., sampling a person i for whom $x_{ij} = x_j$). When X_j is perfectly *unreliable*, X_{ij} has a probability distribution (i.e., the propensity distribution in Lord & Novick, 1968, p. 47) for a fixed person i , and person i 's score on X_j is a draw from person i 's probability distribution of X_{ij} . Furthermore, all persons have the same expected value for X_{ij} .

3.2.2 ERROR IN MTT MODELS CAN BE RELIABLE

Because the error term in CTT is operationally defined as the discrepancy of the item score and the expected value of the item score, of all predictors that together determine the response variable, the error component consists of only unreliable predictors, and all reliable predictors of the item score together form the true-score. In the current section, we discuss the CFA model and IRT model in more detail. Error in MTT can be perfectly unreliable but, in contrast to error in CTT, this is not necessarily so.

CONFIRMATORY FACTOR MODELS

In unidimensional CFA models (also known as Jöreskog's model of congeneric item responses, factor analytic models and unidimensional latent trait model with continuous item responses, Jöreskog, 1971; Mellenbergh, 1994b), the response variable X_j is a linear function of the latent variable Θ and an error variable E_j ,

$$X_j = v_j + \lambda_j \Theta + E_j, \quad (3.8)$$

where λ_j denotes the factor loading of item j on the latent variable, and v_j denotes the intercept (Jöreskog, 1971). In contrast to the true score, latent variable Θ is not just all reliable predictors of the item scores but reflects a dimension that renders the set of response variables conditionally independent (Ellis & Junker, 1997). When the model in Equation (3.8) is correct and $\mathbf{X}$, the set of response variables that load on the factor, consists of three or more variables, Θ can be identified and is reflected in the shared variance among the variables in $\mathbf{X}$. Variables that uniquely predict a single item are absorbed by the error of that item. Variables that predict a subset of the variables in $\mathbf{X}$ result in correlated errors of these items, which results in bad fit of the model in Equation (3.8). The covariation between these errors can be modeled by including a specific factor (i.e., a factor that explains the shared variance among a group of items but is more specific

than the common factor) to the model.

In the factor modeling literature (e.g., Bollen, 1989a), the error score E_j is also sometimes called the *uniqueness* of an item. The error score E_j in Equation (3.8) reflects both reliable predictors of the item (systematic predictors of that item that are unique to that item and are unrelated to Θ , such as mother tongue influencing familiarity with the language that is used in a specific item) and measurement error, which is unreliable (such as fluctuations in energy level or distractions in the testing environment during the math ability test). This differs from CTT, in which E_j is random (i.e., perfectly unreliable) error by definition, and in which unique reliable influences of an item are part of the true score of that item.

If E_j in Equation (3.8) is perfectly reliable (i.e., no random error), then, for person i with attribute level θ_i , the error score, ε_{ij} , is a constant, and so is the item response, x_{ij} ,

$$x_{ij} = v_j + \lambda_j \theta_i + \varepsilon_{ij}. \quad (3.9)$$

If the error score is not perfectly reliable, then for person i with attribute level θ_i , the error score E_{ij} is not fixed over repeated measurements, but has a probability distribution over all possible realizations that is defined over repeated measurements (Rozeboom, 1966). That is, E_{ij} is a random variable and so is the item score X_{ij} ,

$$X_{ij} = v_j + \lambda_j \theta_i + E_{ij}. \quad (3.10)$$

The error score, E_j , is usually assumed to be at least partly unreliable, but in contrast to the error score in CTT, which is perfectly unreliable by definition, in MTT the extent to which E_j is reliable depends on the extent to which the unique determinants that influence the item score are reliable. Next, we introduce the IRT model in which a logistic or cumulative normal function connects binary or polytomous response variables to the latent variable. For simplicity we only consider binary response variables. Similar to the CFA model, error in the IRT model can be partly reliable.

ITEM RESPONSE THEORY MODELS

We make a distinction between IRT models in which the item responses are binary and CFA models in which the item responses are continuous, but the modeling frameworks are closely related. The CFA model can be described as an IRT model for continuous item responses (Mellenbergh, 1994a, 1994b) and the IRT model is equivalent to factor analysis with binary items

Table 3.1: Summary of how models differ based on diverging assumptions about errors.

Assumptions about error	Latent response formulation for X_j	Conditional probability formulation for X_j	Latent response formulation for X_{ij}	Conditional probability formulation for X_{ij}	Model
1) $E_j \sim N(0, \sigma^2)$ 2) E_j is perfectly reliable	$X_j = 1$ iff $\Theta - b_j + E_j > 0$	$\pi_j = \Phi(\theta - b_j)/\sigma$	$x_{ij} = 1$ iff $\theta_i - b_j + \varepsilon_{ij} > 0$	The random variable X_{ij} does not exist, only x_{ij} .	1-parameter Normal ogive model
1) $E_j \sim N(0, \sigma_j^2)$ 2) E_j is perfectly reliable	$X_j = 1$ iff $\Theta - b_j + E_j > 0$	$\pi_j = \Phi(\theta - b_j)/\sigma_j$	$x_{ij} = 1$ iff $\theta_i - b_j + \varepsilon_{ij} > 0$	The random variable X_{ij} does not exist, only x_{ij} .	2-parameter Normal ogive model
1) $E_j \sim \text{Logistic}(0, \sigma)$ 2) E_j is perfectly reliable	$X_j = 1$ iff $\Theta - b_j + E_j > 0$	$\pi_j = \frac{\exp(\mathbf{a}_j(\theta - b_j))}{\exp(1 + \mathbf{a}(\theta - b_j))}$	$x_{ij} = 1$ iff $\theta_i - b_j + \varepsilon_{ij} > 0$	The random variable X_{ij} does not exist, only x_{ij} .	1-PL model
1) $E_j \sim \text{Logistic}(0, \sigma_j)$ 2) E_j is perfectly reliable	$X_j = 1$ iff $\Theta - b_j + E_j > 0$	$\pi_j = \frac{\exp(\mathbf{a}_j(\theta - b_j))}{\exp(1 + \mathbf{a}_j(\theta - b_j))}$	$x_{ij} = 1$ iff $\theta_i - b_j + \varepsilon_{ij} > 0$	The random variable X_{ij} does not exist, only x_{ij} .	2-PL model
1) $E_j \sim N(0, \sigma^2)$ 2) E_j is perfectly unreliable 3) $\sigma_{ij}^2 = \sigma^2$ given θ	$X_j = 1$ iff $\Theta - b_j + E_j > 0$	$\pi_j = \Phi(\theta - b_j)/\sigma$	$X_{ij} = 1$ iff $\theta_i - b_j + E_{ij} > 0$	The random variable X_{ij} does not exist, only x_{ij} .	homogeneous 1-parameter Normal ogive model
1) $E_j \sim N(0, \sigma_j^2)$ 2) E_j is perfectly unreliable 3) $\sigma_{ij}^2 = \sigma_j^2$ given θ	$X_j = 1$ iff $\Theta - b_j + E_j > 0$	$\pi_j = \Phi(\theta - b_j)/\sigma_j$	$X_{ij} = 1$ iff $\theta_i - b_j + E_{ij} > 0$	$\pi_{ij} = \Phi(\theta_i - b_j)/\sigma_j$	homogeneous 2-parameter Normal ogive model
1) $E_j \sim \text{Logistic}(0, \sigma)$ 2) E_j is perfectly unreliable 3) $s_{ij} = \sigma$ given θ	$X_j = 1$ iff $\Theta - b_j + E_j > 0$	$\pi_{ij} = \frac{\exp(\mathbf{a}(\theta - b_j))}{\exp(1 + \mathbf{a}(\theta - b_j))}$	$X_{ij} = 1$ iff $\theta_i - b_j + E_{ij} > 0$	$\pi_{ij} = \frac{\exp(\mathbf{a}(\theta_i - b_j))}{\exp(1 + \mathbf{a}(\theta_i - b_j))}$	homogeneous 1-PL model
1) $E_j \sim \text{Logistic}(0, \sigma_j)$ 2) E_j is perfectly unreliable 3) $s_{ij} = \sigma_j$ given θ	$X_j = 1$ iff $\Theta - b_j + E_j > 0$	$\pi_j = \frac{\exp(\mathbf{a}_j(\theta - b_j))}{\exp(1 + \mathbf{a}_j(\theta - b_j))}$	$X_{ij} = 1$ iff $\theta_i - b_j + E_{ij} > 0$	$\pi_{ij} = \frac{\exp(\mathbf{a}_j(\theta_i - b_j))}{\exp(1 + \mathbf{a}_j(\theta_i - b_j))}$	homogeneous 2-PL model

Bold symbols denote differences between the 1-parameter models and 2-parameter models.

Note that the assumptions in the first column are not sufficient for the models in the sixth column, but only describe assumptions related to error. In addition to the assumptions in the first column, the models in the sixth column (all examples of monotone unidimensional IRT models), also assume (1) unidimensionality, (2) local independence and (3) monotonicity (Ellis & Van den Wollenberg, 1993; Junker & Ellis, 1997).

(Takane & de Leeuw, 1987). Although the models are similar, a distinction between IRT models and CFA models that is important for this article is that the typical formulation of the IRT model does not make the error score explicit, whereas the error score is explicit in the typical formulation of the CFA model; see Equation (3.8). In IRT, it is common to model the *probability of a positive response* as a function of a latent variable rather than the *observed item scores* as a function of the latent variable. This section provides an extensive overview of different IRT models and uses the latent response formulation (Muthén & Asparouhov, 2002) of these models to show for each model which assumptions about error result in a particular IRT model. We summarize this overview in Table 3.1.

In IRT, the item scores are Bernoulli distributed given θ with expected value π_j , where π_j denotes the probability of a positive response given θ (Mellenbergh, 1994a). These probabilities are functionally related to the attribute Θ using one or more item parameters. The 2-parameter logistic (2-PL) model (Birnbau, 1968), distinguishes a difficulty parameter of item j , b_j , and a discrimination parameter of item j , a_j . In Birnbau's 2-PL model, the probability of a positive response to item j given θ , π_j , is specified by means of the following nonlinear function,

$$\pi_j = P(X_j = 1 \mid \theta) = \frac{\exp(a_j(\theta - b_j))}{1 + \exp(a_j(\theta - b_j))}, \quad (3.11)$$

in which π_j ranges between 0 and 1, and π_j approaches 1 as $\theta - b_j$ grows large positive and π_j approaches 0 when $\theta - b_j$ grows large negative. Note that $\pi_j = 0.5$ when $\theta = b_j$.

If local homogeneity is assumed, which is the assumption that persons with the same level θ have identical response probabilities (Ellis & Van den Wollenberg, 1993), the following formulation of the homogeneous 2-PL model is obtained,

$$\pi_{ij} = P(X_{ij} = 1 \mid \theta_i) = \frac{\exp(a_j(\theta_i - b_j))}{1 + \exp(a_j(\theta_i - b_j))}, \quad (3.12)$$

where π_{ij} denotes the expected value of person i 's distribution of item scores over repeated administrations.

When all items are constrained to have identical discrimination parameters, the model in Equation (3.11) simplifies to the one-parameter logistic model (1-PL model),

$$\pi_j = P(X_j = 1 \mid \theta) = \frac{\exp(\theta - b_j)}{1 + \exp(\theta - b_j)}, \quad (3.13)$$

and with the assumption of local homogeneity we obtain the homogeneous 1-PL model,

$$\pi_{ij} = P(X_{ij} = 1 \mid \theta_i) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)}. \quad (3.14)$$

This form of the model is also known as the Rasch model (Fischer, 1995; Rasch, 1960).

A transformation function can be used to transform probabilities to a scale from negative to positive infinity. Two possible transformation functions for the binomial distribution are the logit function, which is the inverse of the logistic function, and the probit function, which is the inverse of the cumulative normal function. When a logit transformation function is used on π_j as specified in Equation (3.11), we obtain $g(\pi_j)$ which ranges from negative to positive infinity and is a linear function of Θ ,

$$g(\pi_j) = \ln \left(\frac{\pi_j}{1 - \pi_j} \right) = a_j(\theta - b_j). \quad (3.15)$$

In Equation (3.11) to (3.14), the error score is not an explicit variable like it is in the CFA model; see Equation (3.8). The reason is that π_j is not the observed item score but its conditional expected value, given θ .

To explicate the error score in the IRT model, we derive an alternative formulation for the IRT model from the latent response variable formulation for regular logistic regression in an econometric approach (Liao, 1994). The latent response variable formulation for logistic regression and probit regression is common in the theory of discrete-choice models. Let $\mathbf{X}$ denote the set of predictor variables and Y the binary response variable. Let $Y^* = \mathbf{x}'\boldsymbol{\beta} + E$, with E distributed according to a standard logistic distribution. In the latent response variable formulation of logistic regression, the latent variable Y^* is thresholded to obtain the binary outcome variable Y ,

$$Y = \begin{cases} 1 & \text{if } \mathbf{x}'\boldsymbol{\beta} + E > 0 \\ 0 & \text{if } \mathbf{x}'\boldsymbol{\beta} + E \leq 0. \end{cases} \quad (3.16)$$

When E is normally distributed, Equation (3.16) represents probit regression (Agresti, 2007; Harrell, 2015; Wooldridge, 2002, p. 387). It can be shown that Equation (3.16) is mathematically equivalent to the generalized linear model formulation of logistic regression in which a transformation (e.g., logit or probit) of the expected value $p = P(Y = 1 \mid \mathbf{X} = \mathbf{x})$ (similar to π_j in IRT models) is regressed on a set of predictor variables (Wooldridge, 2002). The equivalence between these two formulations can be shown using the fact that E is distributed according to

a standard logistic distribution. Equation (3.16) implies

$$P(Y = 1 \mid \mathbf{x}) = P(Y^* > 0 \mid \mathbf{x}) \quad (3.17)$$

$$= P(\mathbf{x}'\boldsymbol{\beta} + E > 0) \quad (3.18)$$

$$= P(E > -\mathbf{x}'\boldsymbol{\beta}). \quad (3.19)$$

Because $E \sim \text{Logistic}(0, s)$

$$P(E < x) = \text{logit}^{-1}(x). \quad (3.20)$$

The logistic distribution is symmetric, which implies

$$P(E > -\mathbf{x}'\boldsymbol{\beta}) = P(E < \mathbf{x}'\boldsymbol{\beta}) \quad (3.21)$$

$$= \text{logit}^{-1}(\mathbf{x}'\boldsymbol{\beta}) \quad (3.22)$$

$$= p. \quad (3.23)$$

The same route can be followed to derive the probit model from the above formulation by assuming that E is standard normally distributed (Wooldridge, 2002, p. 457), in which case the resulting MT'T model is the normal ogive model (Lord, 1952). Both the logistic and normal distribution are symmetric, but because the logistic distribution has heavier tails, the two models differ from each other in the tails.

In IRT models, the relation between the item score and the latent variable is also a logistic regression function, in which the response variable X_j is regressed on the latent variable Θ , hence, this model can also be reformulated with a latent variable $X_j^* = \Theta - b_j + E_j$ (Muthén, 1978; Takane & de Leeuw, 1987). For example, Equation (3.13) represents the conditional probability formulation of the 1-PL model, while the latent response variable formulation (Muthén & Asparouhov, 2002) of the 1-PL model is

$$X_j = \begin{cases} 1 & \text{if } \Theta - b_j + E_j > 0 \\ 0 & \text{if } \Theta - b_j + E_j \leq 0, \end{cases} \quad (3.24)$$

where $\Theta - b_j + E_j$ is a latent variable underlying the binary response variable X_j , and E_j follows a standard logistic distribution. Similar to the CFA model in Equation (3.8), if E_j is perfectly reliable, then, for person i with attribute level θ_i , the error score, ε_{ij} , is fixed, and so is

the item response, x_{ij} ,

$$x_{ij} = \begin{cases} 1 & \text{if } \theta_i - b_j + \varepsilon_{ij} > 0 \\ 0 & \text{if } \theta_i - b_j + \varepsilon_{ij} \leq 0. \end{cases} \quad (3.25)$$

If E_j is not perfectly reliable, then for person i with ability level θ_i , the error score is not fixed over repeated measurements, but has a probability distribution over all possible realizations that is defined over repeated measurements. That is, the error score, E_{ij} , and item score, X_{ij} , are random variables,

$$X_{ij} = \begin{cases} 1 & \text{if } \theta_i - b_j + E_{ij} > 0 \\ 0 & \text{if } \theta_i - b_j + E_{ij} \leq 0. \end{cases} \quad (3.26)$$

As we will explain in more detail in the section “Local Homogeneity”, if E_j is perfectly unreliable, Equation (3.26) with the additional assumption that each person with the same level θ has the same variance $\sigma_{E_{ij}}^2$ implies local homogeneity (Ellis & Van den Wollenberg, 1993). The fact that local homogeneity can be expressed in assumptions about the error score becomes important when we in later sections relate the causal theory of errors that we set out in this article, to the notion of item bias and local homogeneity.

The latent response variable formulation is mathematically equivalent to the conditional probability formulation. One could argue that while the formulations are mathematically equivalent, the latent response formulation implies determinism in the sense that all causes of the item response together fully determine a latent response which is thresholded to obtain categorical responses, while the conditional probability formulation does not imply this determinism. Since in this article we make the additional assumption of a deterministic causal framework, the conditional probability formulation is interpreted like the latent response formulation. That is, in the deterministic causal framework, the conditional probabilities imply a continuous latent response that is thresholded.

In the latent response variable formulation of the IRT model, the discriminating power of item j , a_j , is a function of the variance of E_j . The error score E_j is distributed according to a logistic distribution with mean zero and a scale parameter s_{E_j} , which is proportional to the standard deviation of E_j , σ_{E_j} . A smaller variance of E_j corresponds to a stronger discriminating power of item j , which is reflected in a steeper item characteristic curve at the inflexion point of the IRF (i.e., the point where $\theta = b_j$). In the extreme case, when E_j has no variance and thus is a constant, the IRF is a step function representing the Guttman model, which is deterministic.

When the error score E_j in the latent response formulation is normally distributed, the resulting model is a normal ogive model (Lord, 1952; Lord & Novick, 1968, p. 365) in which the transformation function that maps π_j to a variable $g(\pi_j)$, which ranges from negative infinity to positive infinity, is a probit function instead of a logit function. In the normal ogive model, the standard deviation of E_j , σ_{E_j} , functions as the discrimination parameter (Tucker, 1946). The function for the two-parameter normal ogive model is (Takane & de Leeuw, 1987)

$$\pi_j = \Phi(\theta - b_j)/\sigma_{E_j}, \quad (3.27)$$

where Φ is the cumulative normal function. The normal-ogive model has a similar form to the 2-PL model. In the normal ogive model σ_{E_j} is comparable to $1/a_j$.

The latent response variable formulation of a normal-ogive IRT model is equivalent to the formulation that is used in factor analysis with dichotomous items (Christofferson, 1975; Muthén, 1978; Takane & de Leeuw, 1987)

$$X_j = \begin{cases} 1 & \text{if } \lambda_j\theta - t_j + E_j > 0 \\ 0 & \text{if } \lambda_j\theta - t_j + E_j \leq 0. \end{cases} \quad (3.28)$$

In this formulation, t_j denotes the threshold, and E_j is assumed to be normally distributed. This model is used in the software package Mplus for CFA with categorical outcomes (Muthén & Asparouhov, 2002). The difficulty parameter, b_j , and discrimination parameter, a_j , in the IRT parameterization relate to this formulation of factor analysis for dichotomous variables in the following way,

$$b_j = \frac{t_j}{\lambda_j}, \quad (3.29)$$

$$a_j = \frac{\lambda_j}{\sigma_{E_j}}. \quad (3.30)$$

The latent response variable formulation explicates the error that determines the item score in addition to the latent trait. The probability of a positive response, π_j , in the conditional probability formulation is related to the latent response variable formulation in the following way,

$$\pi_j = P(X_j = 1 \mid \theta) \quad (3.31)$$

$$= P(\theta - b_j + E_j > 0) \quad (3.32)$$

$$= P(\theta + E_j > b_j). \quad (3.33)$$

Figure 3.1: The error score is a composite of all unique influences (denoted “UI” in panel A) on the response variable. These unique influences can be circumstance variables (denoted “CI” in panel B) or characteristic variables (denoted “CH” in panel C). In all panels, each of the errors E_1 to E_3 is a composite of causal effects, but the figure only shows these causal effects for E_1 .

Altogether, Equation (3.24) represents the 2-PL model if E_j is distributed according to a standard logistic distribution. Similarly, Equation (3.26) represents the homogeneous 2-PL model if for any fixed person i the error score E_{ij} is distributed according to a standard logistic distribution and persons with the same level θ have the same error variance. Equation (3.24) represents the 1-PL model if, in addition, E_j is constrained to have the same variance for all items, and this is equal to constraining all items to have the same discrimination parameter. If E_j in Equation (3.24) is normally distributed, the function in Equation (3.24) represents the normal ogive function. These results are summarized in Table 3.1.

Similar to the CFA model, the error component in the IRT model accounts for the unique variation in the item scores, and can reflect both reliable and unreliable predictors of the item scores. As a result, like the CFA model, the extent to which E_j is reliable is an empirical question about the extent to which unique determinants of the item are reliable.

3.3 A CAUSAL THEORY OF ERRORS

In the previous section we showed that, in contrast to the error in CTT, the error in MTT comprises all unique predictors of the item so that the error can be perfectly reliable, perfectly

unreliable or anything in between depending on whether the unique predictors are reliable or not. In the current section, we build on the distinction between reliable and unreliable unique predictors of the item score and present a causal framework for the error. In this framework, we consider the error in MTT to reflect all unique causes of the item score that, together with Θ , fully determine the response variable. We interpret the latent variable as the common cause of the response variables (see e.g., Borsboom et al., 2003). In the following we define what we mean by E_j as a composite of all unique causal effects on the response variable (see Figure 3.1A).

Let us call any variable that is a determinant of the item responses of only item j , a *unique influence* of X_j . Let UI_f denote the f^{th} unique influence on X_j . UI_f is a random variable of which the sample space is defined by randomly selecting people from the population. Let C_j denote a weighted sum of all unique influences on the response variable. The unique influence UI_f is weighted by γ_f , which represents the importance of UI_f for predicting the unique variability in the response variable. For example, in Figure 3.1A, C_1 is a weighted sum of F causal effects on X_1 ; that is, $C_1 = \gamma_1 UI_1 + \dots + \gamma_f UI_f + \dots + \gamma_F UI_F$. The unique influences are not uniquely identified and the weights are unknown. The response variable X_j is a function of Θ and these unique influences,

$$X_j = \lambda\Theta + \gamma_{j1}UI_{j1} + \dots + \gamma_{jf}UI_{jf} + \dots + \gamma_{jF}UI_{jF}, \quad (3.34)$$

and the error is a composite of the causal effects UI_1 to UI_F ,

$$C_j = X_j - \lambda\Theta = \gamma_{j1}UI_{j1} + \dots + \gamma_{jf}UI_{jf} + \dots + \gamma_{jF}UI_{jF}, \quad (3.35)$$

of which the mean is centered at zero by subtracting the expected value of C_j from C_j ,

$$E_j = C_j - E(C_j). \quad (3.36)$$

The expected value of C_j is reflected by the intercept v_j in the factor model

$$v_j = E(C_j). \quad (3.37)$$

Defining the intercept as the expected value of C_j is a new result that follows from the causal theory of the error employed here. This definition of the intercept is important for the interpretation of item bias that follows from the causal theory of errors, which we discuss in a later

section. Equations (3.34) to (3.37) imply

$$X_j = \lambda\Theta + \gamma_{j1}UI_{j1} + \cdots + \gamma_{jf}UI_{jf} + \cdots + \gamma_{jF}UI_{jF} = v_j + \lambda\Theta + E_j. \quad (3.38)$$

An interpretation of the error as a sum of causal effects on the response variable is not the only possible causal interpretation of the error, but we stick to this definition for simplicity. The response variable is completely determined by the F unique determinants and Θ .

A causal interpretation of the error requires an additional interpretational step that is not captured in the CFA models and IRT models we discussed in the previous section. That is, the interpretation of error as a composite of unique causes is not given by the equations we used to define CFA and IRT models. We have established that (1) in CTT the error has zero reliability by definition, and (2) in the IRT model and the CFA model from MTT, the error does not necessarily have zero reliability; the error can be reliable or unreliable (choosing either one would entail an additional assumption to the model, see Ellis & Van den Wollenberg, 1993). While the reliability of the error being zero or not necessarily being zero, is implied by how the model is defined, the interpretation of the error as an aggregate of causal effects on the response variable is not implied by the models but requires an interpretational step. By making this interpretational step, we follow the work of Chen and Pearl (2013, p. 1) who describe the error term in a structural equation such as $Y = \mathbf{x}'\beta + E$ as representing “all other direct causes of Y , accounting for the difference between $\mathbf{x}'\beta$ and the actual values of Y ”. We interpret errors in the same way to build the conceptual framework that is proposed in this article.

In the next section, we explain the difference between characteristic variables that are perfectly reliable determinants of the item score, and circumstance variables that are perfectly unreliable determinants of the item score, to give a conception of what reliable and unreliable determinants could be. Both types of variables are causally relevant to the item score, and both types of variables are *unique* determinants of the item scores.

3.3.1 CIRCUMSTANCE VARIABLES

Unreliable determinants of the item score are variables that are not controlled in the measurement procedure so that they vary over replications of the measurement procedure. As a result of such uncontrolled factors in the measurement procedure, one expects to find variability in the item scores when person i repeatedly responds to the same item j . Rozeboom (1966, p. 380) introduces the notion of ‘circumstances’ to refer to variables in the measurement procedure that are causally relevant to the item score:

“the score actually assigned to a subject on a given test occasion is determined not merely by the subject’s own attributes, but by a great many details of the surrounding circumstances as well.”

All circumstances together form a detailed description of the test circumstance. The circumstance description is defined by the set of values on all circumstance variables. Rozeboom gives the example of a body weighing procedure in which some variables are controlled (i.e., standardized over repeated measurements, such as subjects’ bodily activity and clothing worn at the moment of testing) and some variables are uncontrolled (i.e., not standardized over repeated measures, such as air temperature and the subject’s foot position). The test circumstance description specifies the values all of these variables have, but whereas the subject’s bodily activity and clothing worn are fixed over different circumstance descriptions, the air temperature and subject’s foot position vary.

The circumstances Rozeboom (1966) described are similar to what Ellis and Van den Wollenberg (1993) call *occasions* on which the test can be administered. Occasions include all uncontrolled factors that determine the item-score pattern of a subject, such as uncontrolled environmental variables but also temporary changes in the person, such as unpredictable mood fluctuations and concentration lapses. Ellis and Van den Wollenberg base their interpretation of the variability over occasions on Lord and Novick (1968, p. 38) who describe the *error-score random variable* as “a disturbance that is due to a composite of a multitude of factors not controlled in the measurement procedure”. This error-score random variable consists of components associated with the physical measurement process, the measurement environment and temporal changes in the person (Lord & Novick, 1968, pp. 38–39).

Let CI_{jg} denote the g^{th} circumstance variable that influences the item scores of item j . The realizations of CI_{jg} are scores of different people on that circumstance variable. Each person’s score on CI_{jg} is a single draw from this person’s distribution over measurement occasions for that circumstance variable (Figure 3.1B). That is, each person has a propensity distribution for each of the circumstance variables. Circumstance variables cause randomness in the sense that any time subject i responds to an item, a random draw of scores on the total set of circumstance variables together compose ε_{ij} . Over different circumstances, person i has a probability distribution over all possible realizations of the error variable for item j and thus E_{ij} is a random variable, and therefore X_{ij} also is a random variable. The randomness in X_{ij} thus results from sampling a circumstance from all possible circumstances in which person i could have responded to item j . Examples of circumstance variables may be temporary changes in the person such as tiredness or distraction in the context of an ability test, or mood fluctuations that influ-

ence how someone responds to a questionnaire about life satisfaction. Another example is that if someone just had a nice experience going out and meet a lot of new people the night before being administered a personality questionnaire, this person might be more inclined to give responses that indicate a higher level of extraversion than if this person had filled out the same questionnaire after a stressful week at the office and going to bed early.

3.3.2 CHARACTERISTIC VARIABLES

We define characteristic variables as variables that, just like Θ , vary *over* persons and are fixed *within* a person over repeated measures, but in contrast to Θ are causally relevant to the score on a *unique* item. These are what Ellis and Van den Wollenberg (1993) call *item specific traits*, what Shealy and Stout (1993a, 1993b) call *nuisance determinants* and what Rozeboom (1966, p. 380) calls *the subject's own attributes*, which he contrasts to the *details of the surrounding circumstances*. Let CH_{jh} denote the h^{th} characteristic variable that influences the item scores of item j . The realizations of CH_{jh} are scores of different people on characteristic variable g . Each person's score on CH_{jh} is fixed over occasions (Figure 3.1C).

An example of a characteristic variable is 'familiarity with the English language' that was considered in the introduction of this article as a potential influence on the item responses of a math ability item that uses difficult words to introduce the math problem. Similarly, a culture-specific word in a math item can have as a result that 'knowledge of the meaning of that word' causes some subjects to give a correct response to this item, while others not familiar with the meaning of that word give an incorrect response to the item, even though they all have the same math ability level θ . This characteristic of knowing the word or not could be something that varies over people (some people will know the meaning of the word while others will not) but that hardly varies over measurement occasions. That is, by repeatedly administering the math test, it will be the same people that know the word and the same people that do not know the word. In that case, all variability in the variable 'knowing the word or not' is true-score variance, i.e., the variable is reliable. Another example of a characteristic variable is whether a person taking an IQ test has recently seen a certain movie that gave information relevant to one of the IQ items. The variable 'having seen movie x ' may not be related to people's intelligence but can help a person that saw the movie to solve an IQ item that someone with the same intelligence but who has not seen that movie cannot solve. This other person might know the correct answer to an item that draws on knowledge about biology, because her mother works in that discipline. These two persons with the same level of intelligence may roughly have the same total score, while having

A

B

Figure 3.2: In panel A, the error E_j is perfectly reliable, which is consistent with the random sampling rationale; given a fixed person i , ε_{ij} is a parameter. In panel B, the error E_j is perfectly unreliable, which is consistent with the stochastic subject rationale; given a fixed person i , E_{ij} is a random variable defined over occasions. Note that when E_j is perfectly unreliable, all persons have the same expected value, $E(E_{1j}) = \dots = E(E_{4j})$, and thus the distribution of E_j in panel B does not reflect true score variance. This is in contrast to the variance of E_j in panel A, in which all variance is true score variance.

different items that they would consistently answer correctly, when repeatedly presented with the same item.

The point here is not whether it is plausible whether unique determinants of an item can have unit reliability, but the distinction between unreliable circumstance variables and reliable characteristic variables serves to outline two extremes in a continuous spectrum. Next, we discuss this continuum.

3.3.3 RELIABLE VS UNRELIABLE ERROR

If the error is a composite of only characteristic variables, E_j is perfectly reliable implying that the error score of item j for a person i , ε_{ij} , is fixed (panel A in Figure 3.2). As a result, a person has probability 1 or 0 of giving a positive response to item j . Out of the group of all people

with the same level θ , some have probability 1 and others have probability 0 of giving a positive response depending on their fixed error score. One might wonder how it is possible that two people have the same level θ if one has some characteristics that provide her a probability 1 of giving a positive response to item j while the other has characteristics that provide him a probability 0. Does this not imply that the first person has a higher value of Θ ?

The answer is no, because these characteristics are unique to item j and thus for some other item k with the same difficulty as item j , the second person may have other characteristics (and thus error score) that give him probability 1 of giving a positive response. More generally, for an infinite set of items, these two people would have the same proportion of items correct (Ellis & Junker, 1997; McDonald, 2003). Thus, in the long run, considering a sufficiently large sample of items, the characteristics that advantage and disadvantage the two people will counterbalance.

If the error is a composite of only circumstance variables, E_j is perfectly unreliable and thus the error score of item j for a person i , E_{ij} , is a random variable; the realization ε_{ij} is a single draw from person i 's propensity distribution over possible circumstances for item j (panel B in Figure 3.2).

3.3.4 DETERMINANTS ARE NEITHER PERFECTLY RELIABLE NOR PERFECTLY UNRELIABLE!

Let person 1 and person 2 denote two persons with the same attribute level θ . The first column in Table 3.2 presents different possible hypotheses about C_j . The implications of these hypotheses about the weighted sum of all unique causal effects on the response variable, C_j , for the chance experiment and the empirical predictions are discussed in the next three sections. So far, we have described two extremes, which are: (1) The hypothesis that all unique determinants of an item are characteristic variables (Figure 3.1B, and first and second row of Table 3.2), which implies that E_j is perfectly reliable, and (2) the hypothesis that all unique determinants of an item are circumstance variables (Figure 3.1C, and the fifth row of Table 3.2), which implies that E_j is perfectly unreliable. Figure 3.3 presents a common factor model (e.g., unidimensional IRT model or unidimensional CFA model) with three response variables, X_1 , X_2 and X_3 . Although all three response variables have unique influences on the response variable, Figure 3.3 only shows these influences for X_2 , for which the set of influences consists of G circumstance variables and H characteristic variables. When the error is perfectly reliable, the variance of the circumstance variables is zero: $\sigma_{C_{I_1}}^2 = \dots = \sigma_{C_{I_G}}^2 = \dots = \sigma_{C_{I_G}}^2 = 0$. When the error is perfectly unreliable, the variance of the characteristic variables is zero: $\sigma_{CH_1}^2 = \dots = \sigma_{CH_H}^2 =$

Table 3.2: Implications of different hypotheses about C_j that result in the responses on item j for two persons, $i = 1$ and $i = 2$, who have the same attribute level θ .

Hypotheses about C_j	Implications for conditional distributions and probabilities	Chance experiment	Empirical predictions
C_j is composite of perfectly unreliable determinants AND $\sigma_{C_{1j}}^2 = \sigma_{C_{2j}}^2$	$E(C_{1j}) = E(C_{2j})$; <i>true</i> $\sigma_{E_{1j}}^2 = \sigma_{E_{2j}}^2$; <i>true</i> $P(X_{1j} = 1) = P(X_{2j} = 1)$; <i>true</i> (= local homogeneity holds)	Stochastic subject hypothesis: <i>true</i> Random sampling hypothesis: <i>not true</i>	Absence of item bias: <i>true</i> Cronbach's α equals test-retest correlation: <i>true</i> Test-retest correlation is 1: <i>not true</i>
C_j is composite of perfectly unreliable determinants AND $\sigma_{C_{1j}}^2 \neq \sigma_{C_{2j}}^2$	$E(C_{1j}) = E(C_{2j})$; <i>true</i> $\sigma_{E_{1j}}^2 = \sigma_{E_{2j}}^2$; <i>not true</i> $P(X_{1j} = 1) = P(X_{2j} = 1)$; <i>not true</i>	Stochastic subject hypothesis: <i>not true</i> Random sampling hypothesis: <i>not true</i>	Absence of item bias: <i>not true</i> Cronbach's α equals test-retest correlation: <i>true</i> Test-retest correlation is 1: <i>not true</i>
C_j is composite of both reliable and unreliable determinants	$E(C_{1j}) = E(C_{2j})$; <i>not true</i> $\sigma_{E_{1j}}^2 = \sigma_{E_{2j}}^2$; <i>true</i> iff $\sigma_{C_{1j}}^2 = \sigma_{C_{2j}}^2$ $P(X_{1j} = 1) = P(X_{2j} = 1)$; <i>not true</i>	Stochastic subject hypothesis: <i>not true</i> Random sampling hypothesis: <i>not true</i>	Absence of item bias: <i>not true</i> Cronbach's α equals test-retest correlation: <i>not true</i> Test-retest correlation is 1: <i>not true</i>
C_j is composite of partly reliable determinants	$E(C_{1j}) = E(C_{2j})$; <i>not true</i> $\sigma_{E_{1j}}^2 = \sigma_{E_{2j}}^2$; <i>true</i> iff $\sigma_{C_{1j}}^2 = \sigma_{C_{2j}}^2$ $P(X_{1j} = 1) = P(X_{2j} = 1)$; <i>not true</i>	Stochastic subject hypothesis: <i>not true</i> Random sampling hypothesis: <i>not true</i>	Absence of item bias: <i>not true</i> Cronbach's α equals test-retest correlation: <i>not true</i> Test-retest correlation is 1: <i>not true</i>
C_j is composite of perfectly reliable determinants	$E(C_{1j}) = E(C_{2j})$; <i>not true</i> $\sigma_{E_{1j}}^2 = \sigma_{E_{2j}}^2$; <i>true</i> (for both individuals variance is 0) $P(X_{1j} = 1) = P(X_{2j} = 1)$; <i>not true</i>	Stochastic subject hypothesis: <i>not true</i> Random sampling hypothesis: <i>true</i>	Absence of item bias: <i>not true</i> Cronbach's α equals test-retest correlation: <i>not true</i> Test-retest correlation is 1: <i>true</i>

The stochastic subject hypothesis means that each person with the same attribute level θ has the same probability of giving a positive response.
The random sampling hypothesis means that each person has either probability 1 or probability 0 of giving a positive response.

Figure 3.3: The error is a composite of both circumstance variables (denoted “CI”) and characteristic variables (denoted “CH”). Each of the errors E_1 , E_2 and E_3 is a composite of unique causal effects on the response variables, but for simplicity, the figure only shows these causal effects for X_2 .

$$\dots = \sigma_{CH_H}^2 = 0.$$

These extreme positions are a simplification in at least two important respects. First, in practice, the error will likely be a composite of both reliable and unreliable determinants (both circumstance variables and characteristic variables explain parts of the variance in E_j). The third row of Table 3.2 represents a situation in which the error is a composite of both characteristic variables and circumstance variables. The error likely reflects very different determinants in different tests, so that in some tests error is more reliable than in some other tests. For example, memory tasks may be sensitive to circumstance variables such as whether the person slept well, or whether the environment in which the person takes the test is quiet and nobody suddenly walks in the room. On the other hand, general knowledge questions may be more sensitive to characteristic variables such as whether a person has read a book or heard his/her parents talk about a topic related to some item of the test.

Second, in practice, many determinants of the response variable cannot be classified as either reliable or unreliable but rather somewhere in between. For example, ‘temporary changes within the person’, which Ellis and Van den Wollenberg (1993) present as an example of a determinant that is part of the occasion, causes variability in the item scores over replications. Some people may experience stronger changes over time than other people, so that the variance in this

variable differs over persons. For example, the energy levels of both person 1 and person 2 fluctuate over repeated measurements of a math ability test, but for person 1 these fluctuations are stronger than for person 2. When the variance of such a determinant differs over persons (e.g., $\sigma_{E_{1j}}^2 \neq \sigma_{E_{2j}}^2$ in Figure 3.2), but the expected value is the same over persons ($E(E_{1j}) = E(E_{2j})$ in Figure 3.2), the error is still perfectly unreliable. This situation is represented in the second row of Table 3.2. But, when a unique determinant of the response variable both varies for a fixed person over replications and has different expected values over these replications for different people ($E(E_{1j}) \neq E(E_{2j})$ in Figure 3.2), the determinant is partly reliable. This situation is represented in the fourth row of Table 3.2. For example, both person 1 and person 2 fluctuate in their energy levels but person 1 has a higher energy level on average which helps to be more concentrated on average so that for a single item for which concentration is important, person 1 has a higher expected value of the error and thus also of the item score, even though both persons have the same math ability level.

The hypothesis that E_j is a composite of only circumstance variables can be seen as one extreme of a continuous spectrum, in which E_j has reliability 0, and the hypothesis that E_j is a composite of only characteristic variables can be seen as the other extreme of that spectrum, in which E_j has reliability 1. Cases in which both characteristic variables and circumstance variables influence the response variable, or in which partly reliable determinants influence the response variable, are somewhere in between these extremes on the spectrum.

In the next section, we connect the causal framework to two rationales explaining why the response variables are random variables, which are the stochastic subject rationale and random sampling rationale (Holland, 1990). We argue that the choice of rationale for X_j as a random variable depends on whether E_j is composed of characteristic variables or circumstance variables.

3.4 TWO RATIONALES FOR THE RESPONSE VARIABLE AS A RANDOM VARIABLE

In the introduction, we argued that it is key to CFA and IRT models that the response variables are random variables, and more specifically, that the response variables have a probability distribution conditional on θ . That is, given θ , the realizations of the response variable are the result of some chance experiment such as random sampling. Holland (1990) compared the stochastic subject and the random sampling interpretations of the response variable X_j as a random variable in IRT models. The current section connects the causal interpretation of error proposed

in the previous section to these chance experiments to show that both rationales imply an empirical hypothesis that is at one extreme of the continuum formulated in the previous section.

In the stochastic subject rationale, each person with attribute level θ has the same probability of giving a positive response to an item j . So, $P(X_j = 1 \mid \theta) = .7$ means that for each person with attribute level θ the probability of giving a positive response to item j is .7. This implies that the probability of sampling a person who gives a positive response to item j from the population with attribute level θ , $P(X_j = 1 \mid \theta)$, equals the probability that any person i with attribute level θ gives a positive response, $P(X_{ij} = 1 \mid \theta)$. X_{ij} is a random variable for person i defined over repeated measures. Holland (1990) used a quote from Rasch (1960, p. 73) to explain this rationale:

“Even if we know a person to be very capable, we cannot be sure that he will solve a certain difficult problem, nor even a much easier one. There is always a possibility that he fails – he may be tired or his attention is led astray, or some other excuse may be given. And a person of slight ability may hit upon the correct solution for a difficult problem. Furthermore, if the problem is neither “too easy” nor “too difficult” for a certain person, the outcome is quite unpredictable. But we may in any case attempt to describe the situation by ascribing to every person a probability of solving each problem correctly, and this probability will be our indicator of how easily the problem is solved. The probability that a very able person solves a very easy problem is near unity, but not necessarily equal to 1, and the probability that a person of small ability solves a difficult problem is very near to 0.”

Lord and Novick (1968, p. 29) quote Lazarsfeld (1959) to provide an explanation of what the probability distribution of item scores for a single individual over repeated measurements could mean. Lazarsfeld gives the hypothetical example of some individual, Mr. Brown, who is questioned repeatedly whether he is in favor of the United Nations. After each question, Mr. Brown is brainwashed and questioned again. Since Mr. Brown is not certain about whether he is in favor of the United Nations, he sometimes gives a positive response and sometimes a negative response. $P(X_j = 1)$ is the proportion of positive responses in this hypothetical brainwashing experiment.

In the random sampling rationale, there is no stochasticity within the person. The conditional probability distribution of the response variable given the latent attribute, reflects the probability of drawing a person from the population of people with attribute level θ , who gives a positive response to the item. So, $P(X_j = 1 \mid \theta) = .7$ means that in the population of people with attribute level θ , 70% of the people give a positive response and 30% give a nega-

tive response. Because there is no stochasticity within persons, a person who gives a positive response to an item has probability 1 of giving a positive response to this item.² That is, if the person were brainwashed, the person would consistently give a positive response over repeated administrations of the item. Similarly, a person that gives a negative response has probability 0 of giving a positive response. The conditional probability $P(X_j = 1 \mid \theta) = .7$ thus means that 30% of the people from the population with attribute level θ have probability 0, and 70% of the people from this population with attribute level θ have probability 1 of giving a positive response to item j .

In both the stochastic subject rationale and the random sampling rationale, the conditional probability of the item response given a fixed attribute level is only a probability if there is an error term in the model. If the response variable were determined entirely by the attribute Θ (i.e., there were no uncontrolled determinants), persons with the same attribute level would either all have probability 1 or probability 0 since these persons would give the same response over repeated measurement occasions. Without an error component, neither of the rationales above are sensible, because there is no variation in item scores between people with the same attribute level (random sampling rationale) and there is also no variation in the item scores of a single person over repeated administrations of the item (stochastic subject rationale). Error thus is essential in explaining the different types of variation on which the two rationales are based.

In the random sampling rationale, an interpretation of the error is tantamount to explaining why people with the same attribute level θ nevertheless vary with respect to their item responses (i.e., some people with attribute level θ have probability 1 of giving a positive response, while others have probability 0). The error thus explains the difference in item responses between people with the same attribute level θ . In contrast, in the stochastic subject rationale, an interpretation of the error is tantamount to explaining why one person with a fixed attribute level θ

²By making the step that because there is no stochasticity within persons, people have either probability 1 or 0 of giving a positive response, we follow the work of e.g., Borsboom et al. (2003, p. 205) and Molenaar (1995, p. 6). Holland (1990), however, discussed an interpretation that hinges on defining a person as only existing at the moment that the item is administered. That is, in a repeated administration of the item, person i changed into person i^* , and the population P changed into a new population P^* . As a result, individuals do not have a probability 1 or 0 of giving an item response, as there only exists one unique item administration for that person. The problem with this solution is that the definition of a person or population becomes very narrow, because inferences about a person i or a population P only refer to that person or population at that exact moment of the item administration, and cannot generalize to later or earlier timepoints. In the following, we therefore choose to use the reading of the random sampling rationale in which persons have either a probability 1 or probability 0 of giving a positive response over repeated measures.

varies over measurement occasions. In this rationale, the error explains the variety in responses for a single person on the same item over different occasions.

In the random sampling rationale, individuals have probability 1 or 0 of giving a positive response to an item, and this implies that the error is perfectly reliable. As such, the random sampling rationale implies the empirical hypothesis that all other determinants of the response variables are characteristic variables. These characteristic variables explain why people with the same attribute level θ vary in their item responses. By an empirical hypothesis we mean that the claim that ‘all other determinants of the response variables are characteristic variables’ is a statement about how the item responses originate, that can be true or false, and this hypothesis has empirical consequences. This hypothesis implies that the response variable is perfectly reliable, which can be tested by assessing the test-retest correlation. We come back to this implication in the section ‘Cronbach’s alpha and test-retest correlation’ and in the discussion of this article. The random sampling hypothesis corresponds with the assumptions about the error in the first four rows of Table 3.1. In the stochastic subject rationale, individuals with the same attribute level have the same probability of giving a positive response to an item, which implies a perfectly unreliable error. The stochastic subject rationale thus implies the empirical hypothesis that all unique determinants of the response variables are circumstance variables. This hypothesis implies that the error is random and thus that the response variable is not perfectly reliable. More specifically, this hypothesis implies that Cronbach’s alpha equals the test-retest correlation, given that the response variables follow the monotone unidimensional latent variable model. The stochastic subject hypothesis corresponds with the assumptions about the error in the last four rows of Table 3.1.

Perfectly reliable error thus is a necessary condition for the random sampling hypothesis and perfectly unreliable error is a necessary condition for the stochastic subject hypothesis. However, although perfectly reliable error is also a sufficient condition for the random sampling hypothesis (fifth row of Table 3.2), perfectly unreliable error is not a sufficient condition for the stochastic subject hypothesis. In addition to perfectly unreliable error (C_j is perfectly reliable and thus so is E_j), any pair of persons with the same attribute level also needs to have the same error variance over occasions ($\sigma_{C_{1j}}^2 = \sigma_{C_{2j}}^2$ and thus $\sigma_{E_{1j}}^2 = \sigma_{E_{2j}}^2$). Consider two persons 1 and 2, and let C_j be perfectly unreliable so that persons 1 and 2 have the same expected value for the error. Now suppose that person 1 has a smaller variance over repeated administrations of item j than person 2 has ($\sigma_{C_{1j}}^2 < \sigma_{C_{2j}}^2$), then if $b_j < \theta$, person 2 has a higher probability of giving a positive response to item j than person 1, and if $b_j > \theta$, person 1 has a higher probability of giving a positive response to item j than person 2. For this reason, the stochastic subject

hypothesis is only implied by the hypothesis that the error is perfectly unreliable, and that the variance of the error is equal given θ (first row of Table 3.2), and not just by the hypothesis that the error is perfectly unreliable (the second row of Table 3.2).

Using the causal interpretation of the error to make sense of the chance experiment underlying the item responses, we conclude that the stochastic subject rationale and random sampling rationale are not just two interpretations, but reflect empirical hypotheses that are on opposite extremes of a continuum. The opposites are the hypothesis that the error has reliability 0 (stochastic subject rationale) and the hypothesis that the error has reliability 1 (random sampling rationale).

We note that in both these approaches the conditional probability refers to a frequency over repeated sampling from a population, which is not the only way to conceptualize probability. Different definitions of what a probability is (e.g., Bayesian probabilities; Jeffreys, 1961), are beyond the scope of this article. In the next section, we relate the causal interpretation of the error to the notion of measurement invariance (Meredith, 1993) or item bias (Mellenbergh, 1989).

3.5 ITEM BIAS

So far, we have employed a causal interpretation of error and related this interpretation to the different chance experiments that result in a conditional probability distribution for the response variable given latent variable Θ . In this section, we show that the causal interpretation of error is also relevant for the interpretation of item bias. We show that while circumstance variables result in unreliable measures, characteristic variables imply item bias.

An item is unbiased when the conditional distribution of the response variable given θ is independent of the subpopulation G . Formally, item j is unbiased with respect to a nominal variable G indicating group membership if and only if

$$F(X_j \mid \theta, G = g) = F(X_j \mid \theta), \quad (3.39)$$

where $F(\cdot)$ denotes the (cumulative) distribution function (Mellenbergh, 1989). This definition of item bias is identical to the definition of measurement invariance (Meredith, 1993) and of subpopulation invariance (Ellis & Van den Wollenberg, 1993). It also implies Lord's (1980) definition of lack of bias (Meredith, 1993). In the special case of a dichotomous response variable X_j , Equation (3.39) implies that an item is unbiased with respect to G (Mellenbergh, 1989),

if and only if

$$P(X_j = 1 \mid \theta, g) = P(X_j = 1 \mid \theta), \quad (3.40)$$

which means that the IRF of item j does not depend on group membership. While this is the definition adopted in IRT, the definition in Equation (3.39) is more general and also holds for continuous response variables. In the following, we relate the causal interpretation of error to item bias for both binary response variables in the IRT framework and continuous response variables in the CFA framework.

3.5.1 ITEM BIAS FOR CONTINUOUS RESPONSE VARIABLES

To see how the causal interpretation of the error relates to the notion of item bias, consider Equations (3.34) to (3.37). For continuous response variables, Meredith (1993) showed that for lack of bias it is required that strict factorial invariance holds. A set of continuous response variables $\mathbf{X}$ is said to be strictly factorial invariant with respect to selection based on a grouping variable G if (1) the vector of factor loadings is equal over subpopulations based on G , (2) the vector of intercepts is equal over subpopulations based on G , and (3) the error variances are equal over subpopulations based on G (Meredith, 1993). In Equation (3.34), this means that λ has to be equal over groups, $\sigma_{E_j}^2 = \sigma_{C_j}^2$ is equal over groups, and that $E(C_j) = v_j$ is equal over groups.

We defined a characteristic variable as a unique reliable causal effect on the item response. The definition of a characteristic variable as a reliable variable implies that this variable can be used to select subpopulations. The presence of a characteristic variable in the weighted sum C_j (i.e., at least one of the unique influences is a characteristic variable) implies that the population can be grouped in subpopulations so that the expected value of the composite of causal effects varies across subpopulations; that is,

$$\text{presence of CH}_{jh} \quad (\text{i.e., } \rho_{TC_j, C_j}^2 > 0) \implies E(C_j) \neq E(C_j \mid g). \quad (3.41)$$

The presence of a characteristic variable thus implies that the intercepts differ across groups based on G , and this violates strict invariance.

The absence of characteristic variables (i.e., C_j is a weighted sum of only circumstance variables) is a necessary condition for lack of item bias. However, the absence of characteristic variables is not a sufficient condition for lack of item bias; also when C_j is perfectly unreliable, there may be item bias because subpopulations differ with respect to the variance of C_j or with re-

spect to the strength of the effect of Θ on the response variable, (i.e., differences in λ).

3.5.2 ITEM BIAS FOR BINARY RESPONSE VARIABLES

For binary response variables, we can use the latent variable formulation in which an underlying continuous latent response variable X_j^* is thresholded

$$X_j^* = \lambda\Theta + C_j, \quad (3.42)$$

$$X_j = \begin{cases} 1 & \text{if } X_j^* > 0 \\ 0 & \text{if } X_j^* \leq 0. \end{cases} \quad (3.43)$$

Since the intercept v_j is defined as $E(C_j)$, we can reformulate Equation (3.43) to

$$X_j = \begin{cases} 1 & \text{if } v_j + \lambda_j\Theta + E_j > 0 \\ 0 & \text{if } v_j + \lambda_j\Theta + E_j \leq 0. \end{cases} \quad (3.44)$$

The intercept is the negative of a threshold, so a formulation equivalent to Equation (3.44) is

$$X_j = \begin{cases} 1 & \text{if } \lambda_j\Theta + E_j > t_j \\ 0 & \text{if } \lambda_j\Theta + E_j \leq t_j. \end{cases} \quad (3.45)$$

where $t_j = -v_j$. Equation (3.44) represents the equation for factor analysis with dichotomous variables (Christofferson, 1975; Muthén, 1978; Takane & de Leeuw, 1987), which equals

$$X_j = \begin{cases} 1 & \text{if } (\lambda_j\Theta - t_j) + E_j > 0 \\ 0 & \text{if } (\lambda_j\Theta - t_j) + E_j \leq 0. \end{cases} \quad (3.46)$$

We showed that the presence of a characteristic variable implies that the population can be grouped in subpopulations so that the expected value of the composite of causal effects varies across subpopulations (i.e., $E(C_j) \neq E(C_j \mid g)$; see Equation (3.41)). As a result, it follows from Equation (3.42) that $E(X_j^* \mid \theta) \neq E(X_j^* \mid \theta, g)$, hence, $P(X_j = 1 \mid \theta) \neq P(X_j = 1 \mid \theta, g)$, which means that X_j is biased with respect to grouping variable G . This type of bias implies that intercept v_j in Equation (3.44) (and equivalently the threshold t_j in Equation (3.45)) differs over subpopulations based on G .

For binary response variables, an item is unbiased if and only if the IRFs of X_j in all subpopulations coincide. That is, finding that either the discrimination parameter or the difficulty parameter differs over subpopulations indicates item bias. Since the difficulty parameter is a function of t_j and λ_j ($b_j = t_j/\lambda_j$; see Equation (3.29)), the bias that results from a characteristic variable manifests itself in different difficulty parameters b_j between subpopulations that are matched on the latent variable.

Uniform item bias exists when the IRFs for different groups do not coincide but when there is no interaction between Θ and group membership (Mellenbergh, 1982; Swaminathan & Rogers, 1990). That is, the difference between subgroups in the probabilities of a positive response is equal at all levels of Θ . For the 2-PL model, this can be expressed in the hypothesis that the IRFs of the item are equally discriminating over groups and only differ with respect to the difficulty parameter. When the variance of C_j is equal over subpopulations (i.e., $\sigma_{C_j}^2 = \sigma_{C_j|g}^2$), characteristic variables result in uniform item bias. After all, the discrimination parameter is a function of the variance of E_j ($a_j = \lambda_j/\sigma_{E_j}$, see Equation (3.30)), and $\sigma_{E_j}^2 = \sigma_{C_j}^2$. Different variances of the error over subpopulations (i.e., $\sigma_{C_j}^2 \neq \sigma_{C_j|g}^2$) implies different discrimination parameters in these subpopulations and thus implies nonuniform bias.

We conclude that in the causal framework, if $\sigma_{C_j}^2 \neq \sigma_{C_j|g}^2$, then item j is nonuniformly biased; the discrimination parameter is a function of the variance of C_j . This is independent of whether the unique influences that are summed in C_j , are characteristic variables or circumstance variables. If $\sigma_{C_j}^2 = \sigma_{C_j|g}^2$, then there are two options: (1) if all unique influences are circumstance variables (i.e., error is perfectly unreliable), then there is no item bias; and (2) if at least one of the unique influences is a characteristic variable (i.e., error is partly reliable), then there is uniform item bias. These conditions for absence of item bias are represented in Table 3.2 (first row), in which only the hypothesis of perfectly unreliable error plus the hypothesis of equal error variances given θ implies absence of item bias.

3.6 LOCAL HOMOGENEITY

The notion of item bias is closely related to the notion of local homogeneity. Local homogeneity holds when people with the same attribute level have identical response probabilities (Ellis & Van den Wollenberg, 1993). In contrast to the concept of item bias in which probabilities are defined for groups of people with the same attribute level θ , the assumption of local homogeneity takes the stochasticity explicitly to the subject level; each person with the same attribute level θ has the same probability of a positive response. Ellis and Van den Wollenberg (1993) show that

the stochastic subject interpretation holds if and only if the assumption of local homogeneity is met. Local homogeneity implies that there are no effective specific traits (Ellis & Van den Wollenberg, 1993, p.423), which in the causal framework discussed in this study refer to characteristic variables, or even any unique causal effect on the response variable that is not perfectly unreliable. More generally, perfectly unreliable error is a necessary condition for local homogeneity. Since we showed that unreliable error is a necessary condition for the stochastic subject hypothesis but not a sufficient condition, it may come as no surprise that perfectly unreliable error is not a sufficient condition for local homogeneity either. Local homogeneity only holds if, in addition to perfectly unreliable error, given the same attribute level, each person has the same variance of C_{ij} over repeated administrations. Thus, for any pair of individuals with the same attribute level, given that C_j is perfectly unreliable and $\sigma_{C_{1j}}^2 = \sigma_{C_{2j}}^2$, these people have the same response probabilities; that is, local homogeneity holds. This situation is represented in the first row of Table 3.2.

3.7 CRONBACH'S ALPHA AND TEST-RETEST CORRELATION

Whereas the presence of characteristic variables implies item bias and violates the assumption of local homogeneity, the presence of circumstance variables has implications for the reliability of the item scores. The following illustrates some of the implications of circumstance variables and characteristic variables for two measures of test-score reliability; Cronbach's α and the test-retest correlation. Equation (3.7) provided the reliability of an item in CTT. The reliability of a test can be defined in a similar way. Let X_+ denote the total score or test score

$$X_+ = \sum_{j=1}^J X_j. \quad (3.47)$$

The reliability of a test is defined as (Lord & Novick, 1968)

$$\rho_{X_+X'_+} = \frac{\sigma_T^2}{\sigma_{X_+}^2}, \quad (3.48)$$

where σ_T^2 is true-score variance of test X_+ , $\sigma_{X_+}^2$ is the total variance of test X_+ and X'_+ is a test parallel to X_+ . That is, the correlation between parallel tests X_+ and X'_+ is equivalent to the proportion of variance in X_+ that is true-score variance; both represent the reliability of X_+ . However, this information is not available when one only has data from a single test administration (and in the absence of parallel tests), and thus alternative methods have been

proposed to estimate the reliability from data of a single test administration (Sijtsma, 2009).

One such coefficient that is often used as an estimate of the reliability of a test is Cronbach's α (Cronbach, 1951)

$$\alpha = \frac{J}{J-1} \left(1 - \frac{\sum_{j=1}^J \sigma_{X_j}^2}{\sigma_{X_+}^2} \right). \quad (3.49)$$

Coefficient α is shown to be equal to the reliability $\rho_{X_+X'_+}$ if the items in the test are essentially tau-equivalent (Novick & Lewis, 1967). Two variables are essentially tau-equivalent if and only if for two items j and k and a constant c_{jk} (Sijtsma, 2009)

$$T_j = T_k + c_{jk} \quad \text{for all items } j \neq k. \quad (3.50)$$

To understand the implications of characteristic variables and circumstance variables for test reliability, consider the model in Figure 3.3 in which three response variables X_1 , X_2 and X_3 , are caused by the same latent variable Θ , and each variable is also caused by characteristic variables CH_1 to CH_H and by circumstance variables CI_1 to CI_G . If the errors of the response variables in $\mathbf{X}$ are all composites of only circumstance variables (i.e., the variance of CH_1 to CH_H is 0), then all variance in X_j that is true-score variance is explained by Θ . That is, for $X_j = v_j + \lambda_j \Theta + E_j$ (see Equation (3.8)), the true-score variable, T_j , equals $v_j + \lambda_j \Theta$. Thus when λ_j (or equivalently the discrimination parameter in IRT) is the same over items, that is $\lambda_j = \lambda$, the items are essentially tau-equivalent and α is $\rho_{X_+X'_+}$. In this case, the reliability $\rho_{X_+X'_+}$, and α both equal the proportion of variance in the test scores explained by Θ . When λ_j differs over items, α is smaller than $\rho_{X_+X'_+}$ (α is a lower bound of the reliability; Guttman, 1945), and only the latter equals the proportion of variance in the test scores explained by Θ .

If the errors of the response variables in $\mathbf{X}$ are composites of both circumstance variables and characteristic variables, the items are not essentially tau-equivalent, in which case $\alpha < \rho_{X_+X'_+}$. In the extreme, if the errors are composites of only characteristic variables (i.e., the variance of CI_1 to CI_G is 0), $\rho_{X_+X'_+} = 1$, whereas α is smaller as a larger part of the variance of the test scores is due to characteristic variables rather than Θ . When the variance of Θ is 0, that is, when the response variables are only caused by characteristic variables, $\alpha = 0$, while $\rho_{X_+X'_+} = 1$. Note that when the test scores are to some extent caused by characteristic variables, $\rho_{X_+X'_+}$ no longer equals the proportion of variance in the test scores explained by Θ , but refers to the proportion of variance that is explained by Θ and the characteristic variables.

Another estimate of the reliability of a test is obtained by administering the test repeatedly and calculating the correlation between test scores. While circumstance variables vary over dif-

ferent measurement occasions, characteristic variables are consistent over repeated measures. The test-retest correlation coefficient (Silk, 1977), which is the correlation between test scores on the same test administered at two different measurement occasions, t_1 and t_2

$$\rho_{\text{test-retest}} = \frac{\text{cov}(X_{+t_1}, X_{+t_2})}{\sqrt{\sigma_{X_{+t_1}}^2 \sigma_{X_{+t_2}}^2}}. \quad (3.51)$$

quantifies the *relative consistency* of test scores, that is, “the consistency of the position or rank of individuals in the group relative to others” (Weir, 2005, p. 231). While α as a measure of reliability ignores true-score variance that is due to characteristic variables, the test-retest correlation coefficient does not. If the errors in the model in Figure 3.3 are a composite of only characteristic variables, $\rho_{\text{test-retest}} = 1$ (ignoring learning effects or development of the characteristic variables over time). If there is no common cause and the response variables are only caused by characteristic variables, then $\alpha = 0$, while the test-retest correlation coefficient $\rho_{\text{test-retest}} = 1$. If the errors are a composite of only circumstance variables, and the items are essentially tau-equivalent, then α and $\rho_{\text{test-retest}}$ are the same, and as more of the error variance is due to characteristic variables, the two indicators of reliability diverge. This implication is included in the last column of Table 3.2.

3.8 DISCUSSION

The error scores in psychometric models are often called residuals (Kline, 2016, p. 13; Newsom, 2015, p. 2; Raykov and Marcoulides, 2000, p. 13), which literally means ‘an amount still remaining after the main part is subtracted or accounted for; a remainder’ (Oxford English Dictionary, n.d.). Residual variance, which is the variance of the error, is the remaining variance in the response variables after attributing variance of the response variables to latent variables that are included in the model. While a lot of thinking goes into the interpretation of latent variables (e.g., this is the subject of construct validity research; Cook & Beckman, 2006; Cronbach & Meehl, 1955), a substantive interpretation of the error, that is, an interpretation that goes beyond ‘variability that is unexplained by the rest of the model’, is typically left implicit. We turned this around by starting with a causal interpretation of the error and showed how different assumptions about error (1) imply different models (see Table 3.1), (2) have different implications for the chance experiment that makes the response variables random variables, and (3) have implications for the reliability of item scores, item bias and local homogeneity. Table 3.2 provides a summary of (2) and (3).

In the causal theory of error scores, we defined error as a composite of all unique determinants of the response variable, and we made a distinction between reliable determinants, which we named ‘characteristic variables’ and unreliable determinants, which we named ‘circumstance variables’. Greater variance in the circumstance variables of an item results in less reliable item scores, and more variance in the characteristic variables of an item results in larger item bias.

Although we showed that the presence of characteristic variables entails the presence of item bias, this does not automatically mean that the bias is harmful. Even when error is perfectly reliable and thus consists of only characteristic variables, it is possible that each characteristic variable only has a negligible influence on the item responses. In such a case, to define groups for which the probability of a positive response given θ is significantly different, one needs a composite of many of these characteristic variables to define the groups. For example, it is possible that to define groups that have significantly different probabilities given θ , one needs to specify groups based on whether a person has read book x , and whether the person has seen movie y and whether the person has had experience z , etc., that is, a composite of a whole set of experiences that characterizes a person. The resulting group factor may not even correlate with any known subpopulations (e.g., subpopulations based on gender, culture, language, etc.). Such group factors that do not map onto any social categories, are not typically what one is after in research about item bias.

For the predictions about Cronbach’s α and the test-retest correlation, we assumed for simplicity that the attribute is constant over repeated measures. This assumption might hardly ever hold in practice. For example, for some attributes the researcher might expect that the subjects develop over time. Furthermore, in practice when administering a test a second time, one often expects learning effects. Changes in the true scores over time deflate the test-retest correlation as an indication of the reliability of the test. Nonetheless, we find it of crucial importance to show that given the assumption of a constant attribute certain observations can be expected. Also, a deflated test-retest correlation implies a smaller divergence between Cronbach’s α and the test-retest correlation in the presence of characteristic variables, since the latter is expected to be higher than the former in the presence of characteristic variables. Thus, when one finds that α is lower than the test-retest correlation, this can be interpreted as an indication of the presence of characteristic variables.

A next step is to apply the theoretical framework presented in this study to empirical research. Not only is this framework useful in thinking through consequences of the implicit hypotheses one may have about errors, also formulating explicit substantive theories about error can be a research topic in itself. Similar to studies on the validation of a measurement instrument, one

could study the empirical predictions derived from assuming a certain theory about error for a measurement instrument, to get a deeper insight into the types of factors that influence the item responses, and derive strategies to improve reliability or reduce item bias. As a follow-up on the theories that result from such studies that focus on the error of different measurement instruments, these theories can be compared between different substantive fields. For example, how do the circumstance variables and characteristic variables that influence the test scores differ between personality questionnaires and ability tests?

Substantive theories about error can also be used to validate the ideas presented in this study. For example, it might be hypothesized that in a memory task, mostly circumstance variables influence the test scores while in a knowledge test more characteristic variables can be identified. From such hypotheses the prediction can be formulated that, for the test scores on the memory task, Cronbach's α will be closer to the test-retest correlation than for the test scores on a knowledge test. The same sort of predictions can be formulated on the item level. For example, based on the content of the items in a test and theories about potential characteristic variables or circumstance variables that may influence the responses on these items, one can predict which items will be more reliable. The items can then be grouped based on whether they are expected to be more influenced by characteristic variables or more by circumstance variables, after which one can assess the test-retest correlation and Cronbach's α for these groups of items.

To summarize, the present article has approached fundamental concepts of modern test theory from the perspective of the error. The causal theory of errors employed in this study has resulted in a framework that is useful in thinking through concepts such as stochastic response variables, reliability, item bias and local homogeneity. In addition, some well known models in modern test theory were shown to follow from assumptions about error. We conclude that rather than approaching errors as residuals, the framework presented in this study shows the importance of a deep understanding of error and calls for more substantive theories of the factors composing error.

4

Composites can be causal too

Abstract

“In his article ‘Towards more rigorous personality trait–outcome research’, Mõttus (2016) gives the impression that composites, as well as other models in which traits are a result rather than a cause of their indicators, require “emergent properties” to have causal power. We argue that this is not necessary; composites can be considered causally relevant by themselves when they mediate the relation between their constituents and the outcome variable.”

4.1 COMMENTARY TO MÕTTUS (2016)

Mõttus (2016) describes a number of alternatives to reflective measurement models in personality. Examples are (a) the model proposed by McCrae (2015) in which traits are unions of their semi-autonomous constituents, (b) network models as proposed by Cramer et al. (2012), and (c) Wood, Gardner, and Harms (2015)’s model, in which traits are formed by behaviors that covary due to shared functional values. Mõttus (2016, p. 21) suggests that for such models in which semi-autonomous behaviors constitute a trait (rather than reflecting it), causal power is more accurately ascribed to the constituents rather than to the trait:

“If [traits are artificial constructions], attributing causality to traits as such seems equally questionable regardless of whether their constituents have similar or different associations with the outcome at hand. Even if the associations generalize across

This chapter is published as Van Bork, R. Rhemtulla, M., & Borsboom, D. (2016). Composites can be causal too. *European Journal of Personality*, 30, 307 – 308. doi:10.1002/per.2060

trait constituents, causal interpretations may be more fruitfully based on these constituents rather than the summary-level traits.”

As an example of an “artificial construction”, Möttus gives socioeconomic status (SES): a composite of education level, income, occupational status and the quality of one’s residence. A composite is a function of its constituents, which completely determine it; an example is a sum score based on questionnaire items, which is completely determined by the item scores. Möttus (2016, p. 22) concludes that we should not interpret SES as a cause of its associated outcomes “because SES itself is then the *result* rather than the cause of its constituents and, unless it takes on emergent properties, it thereby owes its outcome correlations to these constituents”. We disagree. We argue that composites can in fact have causal relevance over and above their indicators, and that this is a realistic possibility in the context of personality.

Introducing a composite as a cause of a particular outcome involves a conjunctive hypothesis; namely, it implies that scores on one constituent can make up for scores on any other. Consider as a small example two constituents: (1) the number of males on a train and (2) the number of females on the train, which together entirely determine the composite ‘the number of people on the train’. It is entirely reasonable to conclude that the composite itself (rather than its constituents) causes the outcome variable ‘the time it takes to find an empty seat on the train’. In this example, the composite itself is causal because it fully mediates the relation between the constituents and the outcome variable: If one knows the number of people on the train, the number of women on the train does not predict any additional variance in the outcome. But does that mean that the composite variable ‘number of people in the train’ has emergent properties with respect to its constituents? That seems implausible. A composite can have causal force without having emergent properties in any interesting sense of the word.

In the example, the constituents show causal unity; they are linked with the outcome in a similar way. However, if the outcome were different (say, ‘the number of high heels on the train’), the composite may no longer mediate the relation between ‘the number of females on the train’ and the outcome variable. Thus, a composite may screen off the relation between indicators and outcomes, but does not necessarily do so. With respect to the variable, ‘the number of high heels’, a constituent variable (i.e., the number of females on the train) may have a unique causal relation with the outcome.

Could personality traits function like people on a train? We think they could. All that is required is for the constituents to play compensatory roles with respect to the outcomes of interest. Consider the impulsiveness items, “I have trouble resisting my cravings”, “When I am having my favorite foods, I tend to eat too much”, and “I sometimes eat myself sick”: It is not

Figure 4.1: SEM models for testing item-outcome associations for impulsiveness and BMI. Upper panel: reflective latent variable model. Lower panel: composite model. Indicators i7 and i8 refer to eating-related behaviors. A significant β_1 coefficient would mean that the trait as a whole is related to BMI. The β_2 and β_3 coefficients reflect unique effects of i7 and i8 on BMI.

implausible that obesity could be caused by a high sum of these constituents, whether that sum is due entirely to any one, two, or a combination of all three items.

Möttus argues convincingly that researchers must test whether trait-outcome relations are due to the unique influence of specific items and facets of the trait. We strongly agree and note that it is possible to perform such a test whether the trait is conceptualized as a common-cause latent variable or as a composite. In the former case, a structural equation model can be used to model the item-trait-outcome relations explicitly and examine unique effects. In the latter case, the composite variable must be defined independently of the outcome (e.g., by weighting all constituents equally; Howell, Breivik, & Wilcox, 2007). Figure 4.1 depicts what these two test models might look like for the Impulsiveness $\rightarrow$ BMI example from Terracciano et al. (2009).

Rather than dismissing composites as lacking causal power at the level of the trait, we think it is important to take seriously the possibility that personality traits may be the result of a set of behaviors. Composites can have causal relevance without being emergent, but constituents can also have unique causal force. As Möttus argues persuasively, just like we should not assume that causality is at the level of the trait rather than the item or facet, but demonstrate this, we should do the same when considering the causal power of composites.

Part II

Equivalent equations, different frameworks

5

Short introduction to network models

5.1 INTRODUCTION

The first part of this dissertation has focused on a causal interpretation of the latent variable model. The second part of this dissertation will focus on the implication of a causal interpretation of psychometric models for comparing different modeling frameworks. The main implication is that even though models from a different modeling framework may be statistically equivalent, a causal interpretation implies that the models differ in the data-generating mechanism they represent. For example, a causal interpretation of the latent variable model means that the model represents the hypothesis that the observed variables are effects of a latent common cause. A second framework that we will compare to the latent variable framework is the network framework, in which the observed variables directly influence each other in a causal network. Before comparing these frameworks, the current chapter provides a more elaborate introduction in network models than was already provided in the introduction chapter of this dissertation. This chapter starts with an introduction in graphical models and then briefly discusses some methods for estimating networks from continuous data.

This chapter is adapted from Van Bork*, R., van Borkulo*, C. D., Waldorp, L. J., Cramer, A. O. J., & Borsboom, D. (2018). *Network models for clinical psychology*. In E. J., Wagenmakers (Ed.) *Stevens' Handbook of Experimental psychology and Cognitive Neuroscience* (Fourth Edition, Volume 5), doi: 10.1002/9781119170174

*These authors contributed equally

5.2 GRAPHICAL MODELS

Networks consist of nodes and edges. Nodes can represent anything, for example entities such as train stations or variables such as psychological test scores. The edges represent relations between the nodes, for example whether the train stations are connected by a railway line or, when nodes represent test scores, the extent to which these scores correlate. Edges can be directed (e.g., variable x causes variable y , indicated by an arrow pointing from x to y) or undirected (e.g., correlations or partial correlations, indicated by a line between nodes). In recent decades, the conception of complex systems of interacting entities as networks has led to the development of a set of powerful empirical research methods, known as network analysis (Borsboom & Cramer, 2013).

In the network approach, psychological constructs can be understood as clusters of closely related variables that have direct (i.e., pairwise) causal relations with each other. But how do we model such a network of direct relations between variables? One way to model direct relations is by estimating dependencies between variables while conditioning on other variables. Consider a set of variables of which you believe a causal network underlies the observed associations between these variables. Many of the associations will be induced by relations via other variables in the network. For example, when ‘sleep problems’ lead to ‘fatigue’ and ‘fatigue’ leads to ‘concentration problems’, then ‘sleep problems’ and ‘concentration problems’ will have an association as well. However, part of this association cannot be explained by a direct relation but is due to the mediation of ‘fatigue’. Therefore, the direct relation between ‘sleep problems’ and ‘concentration problems’ is more accurately approximated by the association between ‘sleep problems’ and ‘concentration problems’ while *conditioning on* ‘fatigue’ than by the simple association between ‘sleep problems’ and ‘concentration problems’.

In disciplines such as physics, probability theory, and computer science, probabilistic graphical models are used to model the conditional dependencies between a set of random variables. Two examples of probabilistic graphical models are Markov random fields (or Markov networks; see Figure 5.1a) and Bayesian networks (see Figure 5.1b). Both graphs consist of a set of nodes that represent random variables and a set of edges that represent conditional dependencies between the nodes they connect. A Markov random field is an undirected graph (i.e., the edges have no direction) while a Bayesian network is a directed acyclic graph (DAG), which means that edges are directed but without forming cycles. Let $\perp\!\!\!\perp$ denote independence, let $|$ denote a conditional event and let *iff* denote ‘if and only if’. Missing edges in a Markov random field correspond to all pairs of variables for which the pairwise Markov property holds: $X_i \perp\!\!\!\perp X_j | X_{V \setminus \{i,j\}}$ *iff*

(a) Markov random field: Undirected graph. A and D are conditionally independent given B and C. Also D and B are conditionally independent given A and C.

(b) Bayes network: Directed acyclic graph. Note that A, B and C do not form a cycle, and all edges are directed.

Figure 5.1: Two examples of a probabilistic graphical model to represent the conditional dependencies between the variables A, B, C and D.

$\{i, j\} \notin E$, with $\mathbf{X}$ being a set of variables, E the set of all edges, V the set of all nodes, and $V \setminus \{i, j\}$ the set of nodes except nodes i and j . This means that for every two nodes in the Markov random field that are not connected, the variables represented by these nodes are conditionally independent given all other variables in the network, while for every two nodes that are connected by an edge, the variables represented by these nodes are conditionally dependent given all other variables in the network.

A Bayesian network is a DAG that satisfies the local Markov property

$$X_v \perp\!\!\!\perp X_{V \setminus de(v)} | X_{pa(v)} \quad (5.1)$$

for all v in V (Koller & Friedman, 2009). This means that given its parents $X_{pa(v)}$ every node in the network is conditionally independent of its non-descendants $V \setminus de(v)$. For every two nodes that are connected, the parent is the node connected to the tail of the arrow (i.e., the cause) while the descendent is the node connected to the head of the arrow (i.e., the effect). A node can have none, one or multiple parents and none, one or multiple descendants. In Figure 5.1b node A is conditionally independent of D (a non-descendent of A as there is no arrow pointing from A directly to D) given its parents (B and C). Node B has only one parent (C) but is also conditionally independent of D given its parent C.

These probabilistic graphical models play an important role in the development of network-construction methods that are used to model psychological constructs and relations between psychological variables. The Ising model, one of the earliest types of Markov random fields, forms the basis for constructing networks of binary variables and partial correlation networks are a Gaussian version of a Markov random field. The correspondence between a Markov random field and a partial correlation network will be explained more thoroughly in section 5.4: Partial correlations to identify connections.

Note that dependencies in a Markov random field do not necessarily indicate direct relations. A dependence between two nodes could also be induced by a common cause of these nodes that is not included as a node in the network and therefore is not conditioned on. For example, when ‘concentration problems’ and ‘sleep problems’ are both caused by a noisy environment, but this noise is not included as a node in the network and thus is not conditioned on, this induces a dependency between ‘concentration problems’ and ‘sleep problems’ in the network. This edge between ‘concentration problems’ and ‘sleep problems’ however does not reflect a causal relationship; the edge reflects their dependence on the same common cause. Another reason why two nodes may show a dependency in a Markov random field that does not reflect a direct causal relation is when these nodes share a common effect (i.e., when two variables have a causal effect on the same variable as is the case with nodes B and C that both cause A in Figure 1). Because of these alternative explanations for conditional dependencies, one should always be careful with interpreting such edges in a network.

In the next section we discuss methods that are currently used to identify the connections in a network. Since most of the later chapters in this book build on network models for Gaussian data, we here only discuss methods for the estimation of networks from Gaussian data. Not all of the networks discussed in this chapter rest on conditional independencies. For example, edges in a correlation network (see section 5.3), reflect marginal dependencies. Such marginal dependencies (e.g., zero-order correlations) may often reflect spurious relationships that disappear when the other nodes in the network are conditioned on. For this reason, to obtain an accurate estimate of the underlying direct relations between nodes, conditional independencies are preferred over marginal dependencies. Nevertheless, correlation networks can provide a quick insight in the structure of the data.

5.3 CORRELATIONS TO IDENTIFY CONNECTIONS

A fairly straightforward way of constructing a network of mental disorders is to use the correlation matrix observed in clinical test scores. For example, a set of n MDD or Generalized Anxiety Disorder (GAD) items will result in a symmetric matrix of $n \times n$ correlations (or polychoric correlations when symptoms are measured with ordinal items). Each of these correlations refers to the linear association across individuals between the scores on the two items corresponding to the row and column in that matrix (with the diagonal consisting of ones). A correlation matrix consists of $n(n - 1)/2$ unique elements; the lower or upper triangle of the matrix. This number of unique elements corresponds to the maximum number of edges in a correlation network. In a correlation network, every two nodes are connected by an edge when their correlation differs from zero. Therefore, the number of unique non zero elements of the correlation matrix corresponds to the set of edges in the correlation network. Note that freely estimated correlations always differ somewhat from zero, resulting in a fully connected network. For this reason, one might prefer to set a minimum value for the correlations that are included as edges in the network, or only include edges that correspond to significant correlations. Every edge in the correlation network represents the correlation between the two nodes it connects. Edges can differ in thickness, corresponding to the strength of the correlation, and in color, corresponding to the sign of the correlation. The upper right panel of Figure 5.2 is a hypothetical example of a correlation network based on simulated data of symptoms of MDD and GAD, in which green edges represent positive correlations and red edges represent negative correlations. The position of the nodes in this network is based on the Fruchterman-Reingold algorithm which places nodes that strongly correlate more closely together. This also causes that nodes that weakly connect with other nodes are positioned in the periphery of the network while clusters of strongly connected nodes form the center of the network (Borsboom & Cramer, 2013).

Correlation networks provide information on which nodes cluster together. However, as mentioned before, correlations do not reveal which of these associations reflect direct relations. After all, a correlation between two variables could be explained by a direct causal relation but also by a third mediating variable or a common cause of the two variables. For this reason the direct relation between two variables is better captured by the correlation between these variables while conditioning on all other variables in the network. The correlation between two variables while conditioning on a set of other variables, is called a partial correlation. In the next section the partial correlation network will be discussed.

Figure 5.2: Hypothetical example of a network based on simulated data of symptoms of MDD and GAD. The upper left network represents a hypothetical data generating network of direct relations between symptoms. The upper right network represents a correlation network based on simulated data from that data generating network. The lower left network represents the partial correlation network of these simulated data. The lower right network represents the network that is estimated from the simulated data using EBIC glasso. The R package `qgraph` was used to make all four networks in this Figure (Epskamp et al., 2012).

5.4 PARTIAL CORRELATIONS TO IDENTIFY CONNECTIONS

A partial correlation is the correlation between two variables while a set of other variables is controlled for (or partialled out). To deduce conditional independencies from partial correlations,

multivariate normality is assumed.¹ For example, if one is interested in the direct relation between ‘sleep problems’ and ‘concentration problems’ one could control for all other symptoms. Conditioning on these variables results in the removal of the part of the simple correlation between ‘sleep problems’ and ‘concentration problems’ that is explained by other symptoms such as ‘fatigue’; leaving the partial correlation between ‘sleep problems’ and ‘concentration problems’. In a partial correlation network every edge corresponds to the partial correlation between the variables represented by the nodes that are connected by that edge, controlling for all other variables in the network. Consider a network in which V is the set of nodes, i and j are two nodes in this network and X_i is the variable that corresponds to node i and X_j is the variable that corresponds to node j . To obtain the partial correlation between X_i and X_j , the other variables that are partialled out, $X_{V \setminus \{i,j\}}$, are used to form the best linear approximation of X_i and X_j (denoted as resp. $\hat{X}_{i;V \setminus \{i,j\}}$ and $\hat{X}_{j;V \setminus \{i,j\}}$) (Cramér, 1999). $\hat{X}_{i;V \setminus \{i,j\}}$ represents the part of the variation in X_i that is explained by the other variables in the network (i.e., the variance of X_i that is explained by $X_{V \setminus \{i,j\}}$). The residual of X_i is $X_i - \hat{X}_{i;V \setminus \{i,j\}}$ (denoted as $\hat{X}_{i \cdot V \setminus \{i,j\}}$ ²) and corresponds to the part in X_i that is not accounted for by $X_{V \setminus \{i,j\}}$. The partial correlation between X_i and X_j (denoted as $\hat{\rho}_{ij \cdot V \setminus \{i,j\}}$) is the simple correlation between $\hat{X}_{i \cdot V \setminus \{i,j\}}$ and $\hat{X}_{j \cdot V \setminus \{i,j\}}$ (i.e., between the residuals of X_i and X_j). In this way one obtains the correlation between X_i and X_j that is not explained by other variables in the network (e.g., the relation between ‘sleep problems’ and ‘concentration problems’ that is not explained by the other symptoms).

Just as for the correlation matrix, the partial correlation matrix consists of $n(n-1)/2$ unique elements, and every non-zero element of these unique elements corresponds to an edge in the partial correlation matrix. The lower left panel of Figure 5.2 is an hypothetical example of a partial correlation network based on simulated data of MDD and GAD symptoms. Compared to the correlation network in the upper right panel of this figure, several strong edges between nodes have vanished, because these correlations can be explained by other nodes in the network. As could be derived from the explanation of partial correlations above, a partial correlation of zero (or the lack of an edge) corresponds to a conditional independence. As such, partial corre-

¹Other types of relations or types of variables are possible for computing partial correlations but are beyond the scope of this introductory chapter.

²Here, ‘ $\cdot$ ’ can be understood as ‘in terms of’, so variable X_i *in terms of* the other variables in the network is the linear combination of the other nodes that best approximates X_i . The symbol ‘ $\cdot$ ’ can be understood as ‘controlled for’ so $X_{i \cdot V \setminus \{i,j\}}$ means the variable X_i while controlling for the other variables in the network, which is obtained by subtracting the linear combination of these other nodes from X_i .

lation networks are the multivariate Gaussian version of a Markov random field.

To understand how the partial correlation matrix corresponds to a Markov random field, it is important to understand how a partial correlation network relates to the covariance matrix, and how the covariance matrix relates to a Markov random field. The off-diagonal of the partial correlation matrix equals -1 times the off-diagonal of the standardized inverse of the covariance matrix, Σ^{-1} (called the precision matrix, denoted P ; see Koller & Friedman, 2009; Lauritzen, 1996). More precisely, the following relation holds between correlations and elements of the covariance matrix, Σ ,

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}, \quad (5.2)$$

and this relation is similar to the relation between partial correlations and elements of the precision matrix, P ,

$$\rho_{ij \cdot V \setminus \{i,j\}} = -\frac{p_{ij}}{\sqrt{p_{ii}p_{jj}}}. \quad (5.3)$$

Note that this relation implies that elements on the off-diagonal of P that equal zero, result in corresponding partial correlation of zero. In addition to the relation between the partial correlation matrix and the covariance matrix, another important is the relation between the covariance matrix and a Markov random field. With Gaussian multivariate data, zeros in the precision matrix correspond to conditional independencies (Rue & Held, 2005)

$$X_i \perp\!\!\!\perp X_j | X_{V \setminus \{i,j\}} \text{ iff } p_{ij} = 0. \quad (5.4)$$

Thus, a multivariate normal distribution forms a Markov random field iff missing edges correspond to zeros in the precision matrix. The following example explains why zeros in the precision matrix correspond to conditional independencies. To understand this example, two statistical facts should be explicated. Let $\mathbf{X} = [X_1, \dots, X_k]^\top$ be a vector of dimension k where $\top$ denotes the transpose of a matrix. Let $f_{\mathbf{x}}$ denote the density function of the variables in $\mathbf{X}$. First, the following proportional relationship holds for the multivariate Gaussian distribution when the covariance matrix, Σ , is positive definite (Koller & Friedman, 2009)

$$f_{\mathbf{x}}(X_i, X_j, \dots, X_k) \propto \exp\left(-\frac{1}{2}\mathbf{X}^\top \Sigma^{-1} \mathbf{X}\right) \quad (5.5)$$

Second, two variables, X_1 and X_2 are independent iff the following equivalence holds

$$f_{\mathbf{x}}(X_i, X_j) = f_{\mathbf{x}}(X_i) \times f_{\mathbf{x}}(X_j) \quad (5.6)$$

Consider two variables X_i and X_j ($\mathbf{X} = [X_i, X_j]^\top$) for which we define two different precision matrices to illustrate the independence principle for Gaussian data. In Equation (5.7) the element $p_{ij} = p_{ji} = 0.3$ (non-zero), and in Equation (5.8) the element $p_{ij} = p_{ji} = 0$.

$$P = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix} \quad (5.7) \quad P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (5.8)$$

These two matrices can be plugged in for Σ^{-1} in Equation (5.4):

$f_{\mathbf{x}}(X_i, X_j)$ $\propto \exp\left(-\frac{1}{2} \begin{bmatrix} X_i & X_j \end{bmatrix} \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix} \begin{bmatrix} X_i \\ X_j \end{bmatrix}\right)$ $\propto \exp\left(-\frac{1}{2} (X_i^2 + 0.3X_iX_j + 0.3X_iX_j + X_j^2)\right)$ $\propto \exp\left(-\frac{1}{2} (X_i^2) - \frac{1}{2} (X_j^2) - \frac{1}{2} (0.6X_iX_j)\right)$ $\propto \exp\left(-\frac{1}{2} (X_i^2)\right) \times \exp\left(-\frac{1}{2} (X_j^2)\right) \times \exp\left(-\frac{1}{2} (0.6X_iX_j)\right)$ $f_{\mathbf{x}}(X_i, X_j) \neq f_{\mathbf{x}}(X_i) \times f_{\mathbf{x}}(X_j)$	$f_{\mathbf{x}}(X_i, X_j)$ $\propto \exp\left(-\frac{1}{2} \begin{bmatrix} X_i & X_j \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_i \\ X_j \end{bmatrix}\right)$ $\propto \exp\left(-\frac{1}{2} (X_i^2 + X_j^2)\right)$ $\propto \exp\left(-\frac{1}{2} (X_i^2) - \frac{1}{2} (X_j^2)\right)$ $\propto \exp\left(-\frac{1}{2} (X_i^2)\right) \times \exp\left(-\frac{1}{2} (X_j^2)\right)$ $f_{\mathbf{x}}(X_i, X_j) = f_{\mathbf{x}}(X_i) \times f_{\mathbf{x}}(X_j)$
---	---

This example shows that a zero in the precision matrix results in an independency for the multivariate distribution. This example extends to more than two variables and implies that if p_{ij} equals zero, X_i and X_j are conditionally independent given all other variables $\in \mathbf{X}$.

5.5 APPROXIMATIONS OF THE COVARIANCE MATRIX TO IDENTIFY CONNECTIONS

From the previous section it is clear that for multivariate Gaussian data the covariance matrix is sufficient to determine conditional independencies (i.e., partial correlations). The inverse of the covariance matrix, the precision matrix P , then holds all information on the unique (linear) relation between pairs of variables without the influence of other variables.

When sufficient observations are available, i.e., $k < n$, then it is common to estimate the covariance matrix using maximum likelihood (ML; Bickel & Doksum, 2015; Bilodeau & Bren-

ner, 1999). The ML estimate is obtained by maximising the likelihood for a particular P given the observed data $X = x$. We can then maximize the function $\log f_\theta(X)$ where θ contains all relevant parameters. Suppose as before the mean is zero and we only require Σ , with its inverse P , and let $S = \frac{1}{n} \sum_{i=1}^n X^\top X$ be the estimate of Σ . Then $\theta = \Sigma$ and the ML estimate can be obtained by maximising the log-likelihood

$$L_\Sigma(X) = \log f_\Sigma(X) = -\log |\Sigma| - \text{tr}(\Sigma^{-1}S). \quad (5.9)$$

The maximum of L_Σ over all positive definite matrices gives the ML estimate $\hat{\Sigma} = S$. An unbiased version is $n/(n-1)S$. The ML estimate is consistent, meaning that $S \rightarrow \Sigma$ in probability as $n \rightarrow \infty$ (Ferguson, 1996; Van der Vaart, 1998).

In many cases of network analysis there is an insufficient number of observations such that S is not positive definite. That means that the ML estimate $\hat{\Sigma}$ cannot be used because it cannot be inverted to get $\hat{P}$; $\hat{\Sigma}$ is singular. In this (high-dimensional) situation one of the most popular ways to obtain an estimate of the precision matrix P is the lasso (least absolute shrinkage and selection operator), introduced by Friedman et al. (2008). The general idea is to introduce a penalty to Equation (5.9) such that many parameters will be set to zero exactly. Let $P = \Sigma^{-1}$, the precision matrix. The lasso (or ℓ_1) penalty is $\|P\|_1 = \sum_{i=1}^k \sum_{j < i}^k |p_{ij}|$, such that only the sum of absolute values of the lower triangular part of the matrix P is used in the penalty with $k(k-1)/2$ unique covariance parameters. The lasso algorithm to obtain the estimate $\hat{P}$ minimizes

$$\log g_\Sigma(X) = \log |P| - \text{tr}(PS) + \lambda \sum_{i=1}^k \sum_{j < i}^k |p_{ij}|. \quad (5.10)$$

A high value of λ puts more p_{ij} to zero than a low value. This algorithm is often referred to as the glasso for graphical lasso (Friedman et al., 2008).

In order to obtain $\hat{P}$, the penalty (regularization) parameter λ needs to be determined. One of the promising methods to do so is the extended Bayesian Information Criterion (EBIC; Foygel & Drton, 2010). Let S be a subset of $1, 2, \dots, k$ and $s = |S|$ the cardinality of this subset. In the EBIC the parameter λ is obtained by minimizing (Foygel & Drton, 2011)

$$\text{EBIC}_\gamma(S) = -2L_\Sigma(X) + s \log(n) + 2\gamma \log(p). \quad (5.11)$$

Foygel and Drton (2011) show that the EBIC leads to consistent networks as long as γ is rela-

tively high. The lower right panel of Figure 5.2 is an example of a network that is estimated on simulated data of GAD and MDD symptoms, using EBIC glasso which is implemented in the package `qgraph` in R (Epskamp et al., 2012). The upper left panel of Figure 5.2 represents the data generating network from which data is simulated. The other three networks (the correlation network, partial correlation network and glasso network) are based on these data. Comparing these three networks to the data generating network shows that the network that is estimated with glasso comes closest to the data generating network. The correlation network includes many spurious relations that can be explained by other nodes in the network. This is illustrated by the fact that many edges in the correlation network are no longer present in the partial correlation network in which the other variables are controlled for. But, just like the correlation network, the partial correlation network still includes many spurious relations that result from random noise in the data. In the glasso network most of the spurious relations are put to zero, rendering the network that is closest to the data generating network.

6

Latent variable models and networks: Statistical equivalence and testability

Abstract

Networks are gaining popularity as an alternative to latent variable models for representing psychological constructs. Whereas latent variable approaches introduce unobserved common causes to explain the relations among observed variables, network approaches posit direct causal relations between observed variables. While these approaches lead to radically different understandings of the psychological constructs of interest, recent papers have established mathematical equivalences that hold between network models and latent variable models. In the current paper we make this general equivalence concrete by discussing some examples and argue that the fact that for any model from one class there is an equivalent model from the other class does not mean that both models are equally plausible alternative candidate theories on the data-generating mechanism. In addition, network theory and latent variable theory support different types of constraints that do not automatically translate into interpretable constraints in the equivalent model. Finally, we discuss a number of diverging predictions for the relation between zero-order correlations and partial correlations implied by two subsets of network models and latent variable models: sparse network models and unidimensional factor models. The diverging implications of these models are illustrated in an empirical data example with items that measure procrastination.

This chapter is submitted for publication as Van Bork, R., Rhemtulla, M., Waldorp, L. J., Kruis, J., Rezvanifar, S., & Borsboom, D. (under review). Latent variable models and networks: Statistical equivalence and testability. Submitted to *Multivariate Behavioral Research*

6.1 INTRODUCTION

Psychological constructs have traditionally been approached using a latent variable perspective, in which a set of observed variables (e.g., psychopathology symptoms, self-reports on questionnaire items, or performance on cognitive tests) is assumed to reflect an underlying psychological construct (e.g., depression, extraversion, or general intelligence; Cronbach & Meehl, 1955). In such approaches, researchers typically assume that the construct being measured (e.g., depression) creates the shared variance between these observed variables (e.g., symptoms of depression). If this is the case, the shared variance of the observed variables *reflects* the latent construct, and hence appropriate functions of the test scores (e.g., total scores or factor score estimates) can be interpreted as *measures* of the construct (Bollen & Lennox, 1991; Borsboom et al., 2003, 2004; Edwards & Bagozzi, 2000).

Recently, however, an alternative approach has been proposed, based on the idea that many observed variables in psychology reflect basic psychological states or processes that engage in mutual (causal) interactions (Cramer, Waldorp, van der Maas, & Borsboom, 2010; Van der Maas et al., 2006). As such, correlations between observed variables are not attributable to their dependence on a common latent variable, but rather may result from local interactions between the relevant processes. For instance, in intelligence research, increasing short-term memory capacity leads to the availability of new cognitive strategies (Siegler & Alibali, 2005), the application of which results in improvements in short-term memory; in depression research, a symptom like insomnia may lead to concentration problems, which may lead to worry, which in turn may lead to insomnia (Cramer et al., 2010); and in personality, a person who likes to go to parties may develop greater social skills, which may lead him or her to like parties even more (Cramer et al., 2012). In all of these cases, mutualistic interactions between observable variables plausibly generate at least some of the common variance among indicators.

In the past decade, researchers have therefore started to reconsider psychological constructs that have traditionally been viewed from a latent variable perspective, such as intelligence, personality traits and psychopathology (Caspi et al., 2014; McCrae & Costa, 1987; Spearman, 1904), from a network perspective (e.g., Cramer et al., 2012; McNally et al., 2014; Van der Maas et al., 2006). These studies have resulted in a novel perspective on the nature and etiology of psychological constructs, and the ensuing research program has led to alternative interpretations of phenomena such as comorbidity (Cramer et al., 2010) and spontaneous recovery (Cramer et al., 2016), as well as to new substantive findings. Examples include the observation that depressed individuals have more strongly connected networks of negative emotional states (Pe et al., 2015),

and the finding that network structure is related to recovery from depression (Van Borkulo et al., 2015). As a result of these developments, network approaches are quickly gaining popularity.

At first glance, the latent variable perspective and the network perspective propose radically different views on how to understand psychological constructs and the relations between observed variables: In the latent variable approach, shared variance of observed variables is assumed to reflect a latent construct, whereas in the network approach, it is assumed to reflect a causal network. These approaches propose contrasting data-generating mechanisms, which lead to different substantive interpretations of the statistical models. However, these divergent hypothesized causal processes do *not* necessarily translate into different statistical data structures. For example, Van der Maas et al. (2006) showed that simulating data according to a model with positive direct relations between observed variables can result in a well-fitting factor model (see also Epskamp, Rhemtulla, & Borsboom, 2017; Gignac, 2016; Marsman et al., 2015; Van der Maas & Kan, 2016), so that observing good fit for a factor model may just as well be taken as consistent with a network model. Similarly, if data are simulated according to a single factor model with positive factor loadings, a network analysis will result in a fully connected network model with positive edge weights. In this case, good fit for a (low-rank) network model is consistent with a latent variable hypothesis. Several papers in the past couple of years have formalized the statistical equivalence between network models and latent variable models for binary variables, that is, between Ising models (Ising, 1925) and Multidimensional Item Response Theory (MIRT) models (Epskamp, Maris, et al., 2018; Kruis & Maris, 2016; Marsman et al., 2018, 2015). This equivalence means that for any Ising model that fits the data, there exists an MIRT model that explains the data equally well, and for any MIRT model that fits the data, there exists an Ising model that explains the data equally well, suggesting that it is impossible to distinguish network models and latent variable models from empirical data alone. Some might use this statistical equivalence to argue that network models and latent variable models are two representations of the same model, suggesting that it is not sensible to compare network models and latent variable models at all.

In this paper, we argue that network models and latent variable models, while mathematically equivalent, have different substantive implications. As such, the question of how to empirically distinguish these two classes of model is relevant to scientific practice. We begin by explicating the equivalence of network models and latent variable models more precisely than has been done previously in the literature. We argue that, while each network model has an equivalent latent variable model, and vice versa, the network models that follow from substantively interesting hypotheses are not equivalent to substantively interesting latent variable models. Follow-

ing that analysis, we pinpoint possible avenues for distinguishing between the corresponding data-generating mechanisms based on statistical differences between the substantively relevant models.

In the next sections, we first briefly discuss the distinct substantive implications of network models and latent variable models. Second, we offer a systematic investigation of the meaning and implications of existing equivalence proofs as they obtain for the general model classes of latent variable and network models. We argue that statistical constraints that have a sensible interpretation in the one framework can translate to uninterpretable constraints in the other framework. As a result, the equivalent model may explain the data equally well but does not offer a plausible alternative generating mechanism for the data. Third, we show that, despite these equivalences, specific subsets of these models may still imply divergent statistical predictions, as we will explicate through a comparison of two such subsets: the unidimensional factor model and the sparse network model. We discuss three implications of the unidimensional factor model for the relation between partial correlations and zero-order correlations that do not hold for sparse network models. These implications suggest that a focus on the relation between marginal and conditional associations between observed variables in the data can help distinguish these two models empirically. Fourth, we illustrate the meaning of the equivalence as well as the use of the diverging implications of the unidimensional factor model and sparse network model in an empirical data example.

The arguments in this paper about latent variable models, apply to both Confirmatory Factor Analysis (CFA) models and Item Response Theory (IRT) models. In the following we use ‘response variables’ to mean binary or continuous observed variables and ‘latent variables’ to refer to common factors in CFA models or underlying traits in IRT models. For consistency we denote the latent variable in both the IRT framework and CFA framework with the Greek letter η . We use ‘common factor model’ to refer to both IRT models and CFA models.

6.2 EQUIVALENCE IS NOT IDENTITY

Statistical equivalence is a thorny concept. For instance, at first glance it may seem that statistically equivalent models are in fact identical, so that the choice between, say, a network or latent variable model is merely a choice between two alternative ways of representing the data. However, as Markus (2002) has pointed out, statistical equivalence of two models implicitly assumes that these models are not the same model, that is, they are not identical. The very fact that researchers find a case of statistical equivalence important enough to publish suggests that the

models in question should in fact not be taken to be identical; these equivalences are important precisely because they show that *different* models yield *the same* statistical distributions. Markus argues that models that are statistically equivalent are not semantically equivalent when models have different substantive implications. These distinct substantive implications refer to differences in the theories about the studied phenomenon that is reflected in the models. This is in line with other literature on model equivalence in Structural Equation Modeling (SEM), in which equivalent models have different path diagrams corresponding to different hypotheses about how the phenomenon operates (Bollen & Pearl, 2013; MacCallum et al., 1993; Raykov & Marcoulides, 2001). Bollen and Pearl (2013) argue that statistically equivalent models are different causal models to the extent that they imply different causal assumptions. These causal assumptions are reflected in the directed paths that go from one variable to another and can be understood as predictions about how the variable connected to the head of the arrow responds if one were to intervene on the variable connected to the tail of the arrow. When network models and latent variable models are understood as reflecting theories about the data-generating process, then also network models and latent variable models differ in their substantive implications. For example, a network model with direct interactions between variables may be equivalent to a single latent variable model, but the absence of direct paths between the indicators in the path diagram of the latent variable model implies the strong causal assumption that an intervention on any of the indicator variables does not have an effect on any of the other indicators.

Consider a more concrete example in which the associations between five depression symptoms can be modeled either as a common factor model or as a network model. Suppose that one of these symptoms is insomnia. The network model implies that an intervention on insomnia (e.g., through sleeping pills or sleep hygiene therapy) should influence other symptoms via its influence on insomnia. The common factor model implies that such an intervention, however, would not influence the other symptoms. Even if the intervention were to unknowingly also influence the latent variable, and via the latent variable all other symptoms, the implication of the common factor model would still be that the influence on the symptoms is proportional to their factor loadings while the network model implies that the influence on the other symptoms depends on the strength of the paths that connect that symptom to insomnia. As such, the network model and common factor model reflect very different predictions about the effect of interventions on variables in the model and as such feature different substantive implications.

In the following section we consider the statistical equivalence of Ising models and MIRT models and discuss the meaning of this equivalence in the light of models reflecting a theory

about the data-generating mechanism.

6.3 THE MEANING OF MODEL EQUIVALENCE: GENERAL MODEL CLASSES VERSUS SPECIFIC MODEL INTERPRETATIONS

In the last decade the field of psychometrics (re)discovered equivalence relationships between statistical models that each propose radically different explanations for the observed associations between a set of binary observed variables (Epskamp, Maris, et al., 2018; Kruis & Maris, 2016; Marsman et al., 2018, 2015). Specifically, the equivalence of two groups of models was demonstrated, namely common factor models and network models. The papers cited above focus on binary variables, so that the class of common factor models refers to Item Response Theory (IRT; Mellenbergh, 1994a) models, and the class of network models refers to Ising (1925) models, which describe networks of binary data. However, similar equivalence relationships exist between network models and latent variable models for other types of data (Epskamp, Maris, et al., 2018). In the literature cited above, the equivalence between Ising models and MIRT models is demonstrated in two steps. First, it is shown that the Rasch (1960) model, which is a special case of the class of IRT models, is equivalent to the Curie Weiss model (Kac, 1968), which is a special case of the Ising model. Second, it is shown that the broader class of MIRT models, is equivalent to the broader class of Ising models.

The finding that these equivalent representations exist shows that whenever a researcher compares different models within one such framework to adopt the best fitting model, there are necessarily alternatives within the other class that fit equally well to the data. Alternative models that are statistically equivalent can imply different data-generating mechanisms, but their fit to any dataset will be identical. It is important to specify the nature of this equivalence further. In particular, the equivalence proofs relate *model classes* to each other. This means that the equivalence can be understood as stating ‘for any model in the class of MIRT models, there is an equivalent representation as an Ising model, and vice versa’. But these classes are very broad: the class of Ising models refers to all semi-positive definite connectivity matrices, and the class of MIRT models refers to all eigenvalue decompositions with nonnegative eigenvalues.

In any particular case, it is not only important to note that for any network or latent variable model, there exists *a* model within the other class of models that is equivalent, but it is equally important to consider *which* model from the other class is equivalent to the given model, since the general model classes also allow for complex uninterpretable models that do not form a plausible candidate for the data-generating mechanism. While statistically equivalent repre-

sentations of these models thus always exist, the theoretical implications that flow from these alternative representations need not be equally plausible. By plausible we do not refer to the probability of the observed data under these models, as this is exactly equal for equivalent models, but refer to the interpretability of the theories that these models reflect. While there are probably many different aspects on which these implied theories can differ in their plausibility, we here focus on the extent to which statistical constraints of the model in one representation are interpretable when translated to another representation. That is, we regard a model implausible when it is not clear how to interpret the constraints of the model in the first place.

6.3.1 EQUIVALENCE OF ISING MODELS AND MIRT MODELS

In the following we briefly introduce the MIRT model and Ising model before turning to the equivalence between the two. The response variables in both the Ising model and MIRT model are binary variables. In psychology, binary response variables can for example represent correct/incorrect responses to educational testing items that are used to model math ability, or the presence/absence of symptoms that are used to model depression.

In the Ising model the binary response variables are typically coded $+/-1$. Let $\mathbf{X}^T = [X_1, X_2, \dots, X_p]$ denote a vector of p binary response variables that can take on the values $+1$ or -1 , and let $\mathbf{x}^T = [x_1, x_2, \dots, x_p]$ denote the realization of this vector. The Ising model is defined as follows:

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \exp\left(\sum_i \tau_i x_i + \sum_{\langle ij \rangle} \omega_{ij} x_i x_j\right) \quad (6.1)$$

$$Z = \sum_{\mathbf{x}} \exp\left(\sum_i \tau_i x_i + \sum_{\langle ij \rangle} \omega_{ij} x_i x_j\right), \quad (6.2)$$

in which $\sum_i$ takes the sum over all response variables, $i = 1, 2, \dots, p$, and $\sum_{\langle ij \rangle}$ takes the sum over all distinct pairs of variables i and j . The τ_i parameter can be interpreted as the preference of the variable to be in the $+1$ state, and is referred to as the threshold parameter. τ_i ranges from $-\infty$ to ∞ so that $\tau_i = 0$ means that the i^{th} variable does not prefer one state ($-1/+1$) over the other. The ω_{ij} parameter is referred to as the edge weight of the edge connecting X_i and X_j . When $\omega_{ij} = 0$, there is no edge that connects X_i and X_j . $\sum_{\mathbf{x}}$ in Equation (6.2) takes the sum over all 2^p realizations of $\mathbf{X}$, so that Z in Equation (6.1) is a normalizing constant and the probability function sums to one over all possible realizations of $\mathbf{X}$. The edge weights are elements of a symmetric connectivity matrix $\mathbf{\Omega}$, which is a $p \times p$ matrix with an

arbitrary diagonal. Note that the Ising models that are considered in this paper are used to model psychological constructs, and therefore these Ising models typically consist of relatively few variables compared to some other fields in which the Ising model is used, and as such also have relatively few parameters.

In the IRT model the binary response variables are typically coded 0/1. Let $\mathbf{Y}^T = [Y_1, Y_2, \dots, Y_p]$ denote a vector of p binary response variables that can take on the values 0 or 1, and let $\mathbf{y}^T = [y_1, y_2, \dots, y_p]$ denote the realization of this vector. In the unidimensional two parameter logistic (2PL) model the probability of a correct response is a function of the latent ability η , the discrimination of the i^{th} response variable a_i and the difficulty of the i^{th} response variable b_i (Birnbaum, 1968):

$$P(Y_i = 1 | \eta, a_i, b_i) = \frac{\exp(a_i(\eta - b_i))}{1 + \exp(a_i(\eta - b_i))} \quad (6.3)$$

The a_i parameter represents the maximum slope of the item characteristic function of the i^{th} response variable and is referred to as the discrimination parameter, such that a steeper slope is associated with a more discriminating item. The b_i parameter is on the same scale as η , ranging from $-\infty$ to ∞ , and is referred to as the difficulty parameter. The difficulty of an item can be interpreted as the η level for which the probability of a correct response is 0.5, so that higher values correspond to more difficult items. For the multidimensional extension of this model, the M2PL model (Reckase, 2009), the above formulation is rewritten in slope/intercept form: Let d denote $-ab$, it follows that $a(\eta - b) = a\eta - ab = a\eta + d$. Now we can replace the scalar a_i in Equation (6.3) with $\mathbf{a}_i$, the $K \times 1$ vector of discrimination parameters of the i^{th} response variable for all K latent dimensions, and replace η with a $K \times 1$ vector $\boldsymbol{\eta}$:

$$P(Y_i = 1 | \boldsymbol{\eta}, \mathbf{a}_i, b_i) = \frac{\exp(\mathbf{a}_i' \boldsymbol{\eta} + d_i)}{1 + \exp(\mathbf{a}_i' \boldsymbol{\eta} + d_i)}, \quad (6.4)$$

in which $\mathbf{a}_i' \boldsymbol{\eta} + d_i = a_{i1}\eta_1 + a_{i2}\eta_2 + \dots + a_{iK}\eta_K + d_i$. In this formulation a_{ik} denotes the slope of the i^{th} response variable on the k^{th} dimension, and the scalar d_i denotes the intercept of the i^{th} response variable.

For consistency, the M2PL model in Equation (6.4) can be rewritten such that it incorporates $+/-1$ coded variables:

$$P(X_i = x_i | \boldsymbol{\eta}, \mathbf{a}_i, b_i) = \frac{\exp(x_i(\mathbf{a}_i' \boldsymbol{\eta} + d_i))}{\exp(-\mathbf{a}_i' \boldsymbol{\eta} - d_i) + \exp(\mathbf{a}_i' \boldsymbol{\eta} + d_i)}. \quad (6.5)$$

In the M2PL model the local independence assumption ensures that the latent variables $\boldsymbol{\eta}$ lead to a product measure for the response variables $\mathbf{X}$, and so we can describe the joint distribution as a product over the response variables:

$$P(\mathbf{X} = \mathbf{x} | \boldsymbol{\eta}, \mathbf{a}_i, b_i) = \prod_i \frac{\exp(x_i(\mathbf{a}'_i \boldsymbol{\eta} + d_i))}{\exp(-\mathbf{a}'_i \boldsymbol{\eta} - d_i) + \exp(\mathbf{a}'_i \boldsymbol{\eta} + d_i)}. \quad (6.6)$$

Now that we have defined both the MIRT model and Ising model as functions of a set binary $+/-1$ coded response variables, we can turn to the equivalence of these models. We will not prove the equivalence of the models defined above, as this has been done extensively in the papers cited earlier (see Epskamp, Maris, et al., 2018; Kruis & Maris, 2016; Marsman et al., 2018, 2015). Rather, we will make this equivalence concrete by zooming in on specific instances of these models.

Equivalent models have the same number of freely estimated parameters, and therefore to understand the equivalence between network and common factor models it is insightful to consider what parameters are estimated in each of these models. In an Ising model, each item has a threshold and each pair of two items can be connected with an edge. When all thresholds and edge weights are freely estimated, this would result in $p(p-1)/2$ edge weights and p thresholds resulting in $p + p(p-1)/2$ freely estimated parameters. This is the maximum number of parameters that can be estimated from the available information. Degrees of freedom can be gained by constraining some parameters. A possible constraint could be to set an edge weight to a specific value, for example zero. Another possible constraint could be to set multiple edge weights to the same value, or impose a proportionality constraint in which all edge weights form a function of a smaller number of parameters.

As previously demonstrated (Epskamp, Maris, et al., 2018; Kruis & Maris, 2016; Marsman et al., 2015), the equivalent MIRT model for an Ising model is obtained by choosing an arbitrary diagonal for $\boldsymbol{\Omega}$ such that $\boldsymbol{\Omega}$ becomes a positive-semidefinite matrix. The eigenvalue decomposition of a positive-semidefinite matrix results in p orthogonal eigenvectors and p nonnegative eigenvalues. Any positive semi-definite matrix can be represented as an MIRT model with as many latent variables as positive eigenvalues. In fact, the diagonal of $\boldsymbol{\Omega}$ can always be chosen such that the lowest eigenvalue equals zero. An eigenvalue decomposition with p nonnegative eigenvalues of which one equals zero, can be interpreted as an MIRT model with $p-1$ orthogonal latent variables.

A model with $p-1$ positive eigenvalues results in $p(p-1)$ discrimination parameters, which represent the reliability of each of the p response variables with respect to each of the $p-1$ latent

variables. Each discrimination parameter a_{ik} is a function of the i^{th} element of the eigenvector associated with the k^{th} latent dimension, and the k^{th} eigenvalue (Epskamp, Maris, et al., 2018). Additionally each of the p response variables has an intercept d_i . However, not all $p + p(p - 1)$ parameters are freely estimated as such a model would be under-identified from the information that is available in p observed variables. Rather, the constraint that the eigenvectors are mutually orthogonal implies constraints on the discrimination parameters such that the total number of freely estimated parameters is $p + p(p - 1)/2$, the same number of freely estimated parameters as in a freely estimated Ising model.

Now what does this mean for the equivalence of the Ising model and MIRT model? It means that any freely estimated Ising model with no constraints on the edge weights is equivalent to an MIRT model with $p - 1$ orthogonal latent variables, in which all items are related to all $p - 1$ latent variables. This does not entail that any MIRT model that is equivalent to an Ising model necessarily contains $p - 1$ latent variables. Ising models with particular constraints imposed may be equivalent to MIRT models with fewer than $p - 1$ latent variables. An example is the Curie-Weiss model (Kac, 1968), in which all edge weights are constrained to be equal. The Curie-Weiss model is equivalent to a Rasch model, which is an IRT model with only one latent variable and in which all discrimination parameters are constrained to be equal to each other. The Rasch model uses p parameters to estimate p item difficulties. In addition, the Rasch model uses one degree of freedom (df) to estimate the discrimination a of the items or, equivalently, uses one df to estimate the distribution of the latent variable¹. The discrimination parameter can thus be fixed to an arbitrary value without affecting the likelihood if the scale of the latent variable is allowed to be free (Johnson, 2007). When the distribution of the latent variable is fixed, the discrimination parameter cannot be fixed to one without loss of information. Instead, the Rasch model then includes one parameter for the discrimination of all items. The Rasch model, thus, estimates $p + 1$ parameters. The Curie Weiss model estimates p thresholds and a single parameter for all edge weights, therefore just like the Rasch model, the Curie Weiss model estimates $p + 1$ parameters.

The equivalence of the number of parameters is important because neglecting it may lead to misunderstandings. For example, because any unidimensional latent variable model implies a fully connected clique of variables, it may be tempting to argue that any fully connected clique in a network model represents a single latent variable. However, this is misleading because in

¹This number of parameters for the Rasch model is based on the marginal maximum likelihood in which the distribution of the latent variable is conditionally Gaussian given a fixed observation $\mathbf{X} = \mathbf{x}$. This distribution of the latent variable results in a closed form solution that is equivalent to the Ising model (Epskamp, Maris, et al., 2018).

order for the fully connected clique to represent a unidimensional latent variable model, the edge weights have to meet certain constraints. In order for the models to be *equivalent*, proportionality constraints should hold for the edge weights, such that the matrix of edge weights Ω plus some diagonal matrix is of rank one (Epskamp, Maris, et al., 2018). Put differently, any single common cause implies a fully connected cluster with constraints on the edge weights such that the edge weights are a function of the discrimination parameters, but not any fully connected cluster can be explained by a single latent variable. Consider a fully connected cluster of p variables. This cluster is equivalent to a unidimensional 2PL model only if the edge weights are constrained such that the edge weight between each pair of variables is a function of the discrimination parameters of the variables. As such, the $p(p - 1)/2$ edge weights are a function of only p discrimination parameters.

So, what MIRT model *is* equivalent to a fully connected network? That depends on the constraints that are put on the edge weights between these nodes. Different MIRT models correspond to different constraints in the Ising model. If there are no constraints the equivalent model is a $p - 1$ latent variable model. If all edge weights are constrained to be equal, the equivalent model is a Rasch model. Often though, a regularization technique is used to obtain sparsity (i.e., multiple edge weights are set to exactly zero; see Boschloo et al., 2015; Dalege et al., 2016; Van Borkulo et al., 2014, for examples of sparse Ising models). To what kind of MIRT model would an Ising model with edge weights constrained to zero be equivalent? Constraining an edge weight to zero in the Ising model implies a complex proportionality constraint in the equivalent MIRT model. So, although for any Ising model there is an equivalent MIRT model, this MIRT model is often either complex in the sense that the number of orthogonal latent variables is almost as large as the number of observed variables, or complex in the sense that the constraints are no longer interpretable (it is not trivial how one edge weight set to zero translates to a constraint in an MIRT model and what this latter constraint would mean when interpreting the model). This latter complexity arises because the type of constraints that make sense in a MIRT framework (e.g., eliminating latent variables) results in constraints that do not often make sense in the network framework (proportionality constraints on the edge weights) while constraints that make sense in the network framework (e.g., constraining edges to zero) result in complex proportionality constraints in the MIRT model that are hard to defend from the latent variable perspective.

Figure 6.1 presents an example of two equivalent models. The model in Figure 6.1 (a) represents an Ising model with four different edges (ω_{31} and ω_{41} are constrained to zero). This model thus has eight freely estimated parameters: four thresholds and four edge weights. The

a

$$\Omega = \begin{bmatrix} 0 & \omega_{12} & 0 & 0 \\ \omega_{21} & 0 & \omega_{23} & \omega_{24} \\ 0 & \omega_{32} & 0 & \omega_{34} \\ 0 & \omega_{42} & \omega_{43} & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0.30 & 0 & 0 \\ 0.30 & 0 & 0.20 & 0.25 \\ 0 & 0.20 & 0 & 0.35 \\ 0 & 0.25 & 0.35 & 0 \end{bmatrix}$$

b

$$\mathbf{A} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & 0 \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & 0 \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & 0 \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0.407 & -0.780 & 0.113 & 0 \\ 0.796 & -0.372 & -0.126 & 0 \\ 0.739 & 0.423 & 0.249 & 0 \\ 0.782 & 0.386 & -0.166 & 0 \end{bmatrix}$$

Figure 6.1: The model in panel (a) represents the Ising model that is associated with the connectivity matrix Ω . The model in panel (b) represents the MIRT model that is equivalent to the Ising model in panel (a). The discrimination parameters of this MIRT model are presented in the matrix $\mathbf{A}$.

model in Figure 6.1 (b) represents the equivalent MIRT model, that results from the eigenvalue decomposition of Ω in which the diagonal is set equal to -1 times the lowest eigenvalue of Ω so that the lowest eigenvalue equals zero. Let $\mathbf{Q}$ represent the $p \times p$ eigenvector matrix in which each row is associated with an observed variable X_i and each column is associated with a latent variable η_k . Let $\mathbf{r}$ represent the vector of eigenvalues of the $p \times p$ covariance matrix, of which the last element equals zero such that the total number of latent variables equals $p - 1$. a_{ik} , the discrimination parameter for variable X_i on the latent variable η_k , is a function of q_{ik} , the eigenvector value of that variable that is associated with η_k and the eigenvalue r_k associated with η_k : $a_{ik} = -2\sqrt{r_k/2} \times q_{ik}$. For a more extensive explanation of how to obtain $\mathbf{A}$, the matrix of discrimination parameters, from Ω , we refer the reader to Epskamp, Maris, et al. (2018) and Marsman et al. (2015). Note that although the MIRT model seems to be more complex in the number of parameters, the number of freely estimated parameters of the two models is exactly equal due to proportionality constraints on the discrimination parameters. That is, all discriminations parameters are a function of the smaller number of ω parameters.

To summarize, low rank MIRT models are equivalent to low rank Ising models, freely estimated networks with no constraints on the edge weights are equivalent to MIRT models with $p - 1$ latent variables, and Ising models in which edge weights are constrained to zero are equivalent to MIRT models with complex constraints on the discrimination parameters.

In contexts where psychometric models are intended to represent a theory about the data-generating process, constraints reflect hypotheses that can be rejected and as such make the theory implied by the model falsifiable (Popper, 1959). In this perspective on models, the constraints are interpreted in the light of the theory reflected by the model. For example, the constraint that residual variables in a one-factor model are independent makes sense when the model reflects the theory that a common cause underlies the data. After all, local independence follows logically from the theory that all response variables are caused by a single common cause. These independence constraints make the theory falsifiable, since when the residual variables do correlate, the one-factor model with residual correlations constrained to zero does not fit.

The different modeling frameworks are characterized by different theories about the data-generating process and therefore different constraints follow naturally from these frameworks. In a latent variable model a rank constraint corresponds with the hypothesis that a smaller set of latent variables can explain the covariation among a larger set of observed variables. Constraining a network model to be sparse by setting edge weights to zero, corresponds to the hypothesis that some variables are not directly related and that all covariance between these variables is due to other variables in the network. The interpretation of a rank constraint in a network model is unclear, just like the proportionality constraints on the discrimination parameters in an MIRT model that are implied by the sparsity constraints in the equivalent network, are unclear. The lack of a clear interpretation of these constraints decreases the plausibility of these models.

All in all, it seems that although the Rasch model and Curie Weiss model form two elegant models that are equivalent, most of the considerably more realistic models in which not all edge weights or discrimination parameters are assumed to be equal are equivalent to models that are hard to defend from the other framework and as such do not form plausible explanations for the data. Thus, while the model equivalences established in the literature are important from a theoretical psychometric perspective, the fact that for any model from one class there is an equivalent model from the other should not be taken to imply that these models are equally plausible. Also, the fact that ‘natural’ model restrictions in the two frameworks are so different serves to illustrate that network models and latent variable models while statistically equivalent, are not semantically equivalent (Markus, 2002). In addition, it turns out that in many important special cases, plausible candidate models taken from the latent variable and network realm do in fact have different statistical implications that can be used to distinguish them, a topic we turn to next.

6.4 DISTINGUISHING BETWEEN SPECIFIC MODELS

The previous section has established that model equivalences, while important, should not be interpreted to mean that network models and latent variable models are basically the same. In this section, we drive this point further home by identifying divergent statistical predictions that can be derived from two important submodels that have often been proposed as candidate models in the psychological literature. First, the unidimensional latent variable model (Jöreskog, 1971); second, the sparse network model (SNM; Epskamp, Rhemtulla, & Borsboom, 2017). Because distributional representations become rather unwieldy for categorical variables, in this section we consider Gaussian variables, which means that the unidimensional latent variable model translates to the unidimensional factor model (UFM; Mellenbergh, 1994a), and the SNM translates to a Gaussian Graphical Model (GGM; Epskamp, Rhemtulla, & Borsboom, 2017). Before we examine divergent implications that follow from these models, we briefly introduce them to the reader.

6.4.1 UNIDIMENSIONAL FACTOR MODELS AND SPARSE NETWORK MODELS

In factor models, just like for MIRT models, the observed variables are considered fallible indicators of a latent variable, which typically represents the construct of interest. Let Σ denote the covariance matrix of $\mathbf{y}$, in which $\mathbf{y}$ denotes the vector of continuous observed variables. Let λ denote the vector of factor loadings and Θ the residual covariance matrix. In a UFM all variables in $\mathbf{y}$ are a linear function of a single factor, η , and independent errors, ε ,

$$y_i = \lambda_i \eta + \varepsilon_i. \quad (6.7)$$

The UFM implies that the covariance matrix of the indicators admits the following representation

$$\Sigma = \lambda \lambda' + \Theta. \quad (6.8)$$

Here, we assume that Θ is diagonal, that is, that the observed variables do not covary over and above what is explained by the latent factor. In other words, the principle of local independence holds (Bollen, 1989b). We furthermore let η have unit variance. Equation (6.8) implies that the covariance among observed variables is a function of their factor loadings. More precisely, because Θ is a diagonal matrix, the covariance between two variables y_i and y_j equals $\lambda_i \lambda_j$, in

which λ_i and λ_j are elements of the vector $\boldsymbol{\lambda}$ and denote the factor loadings of y_i and y_j on the factor η . Figure 6.2b depicts a factor model with three indicators. All variables in $\mathbf{y}$ are standardized so that $\boldsymbol{\Sigma}$ simplifies to a correlation matrix. The reason for doing so is that standardized factor loadings are more easily interpreted because they range from -1 to 1 . Note that the results in this paper extend to unstandardized variables because Equation (6.8) also holds for an unstandardized $\boldsymbol{\Sigma}$ and $\boldsymbol{\lambda}$. Furthermore, we assume that standardized factor loadings are never exactly -1 or 1 , as this would imply that the common factor is observed. We also assume that factor loadings are not exactly zero, as this would imply that the corresponding observed variable is not an indicator of the common factor.

For Gaussian variables, as we consider here, the Ising network model discussed above translates into a partial correlation network. In a partial correlation network for Gaussian variables, edges have a particularly simple representation, namely as the partial correlation between two variables that results when all other variables in the network are partialled out. These partial correlations are computed by standardizing elements of the precision matrix, $\mathbf{P} = \boldsymbol{\Sigma}^{-1}$,

$$\rho_{ij \cdot V \setminus \{i, j\}} = -\frac{p_{ij}}{\sqrt{p_{ii}p_{jj}}}, \quad (6.9)$$

where $V \setminus \{i, j\}$ denotes the set of all nodes except for i and j . This equation implies that off-diagonal elements of $\mathbf{P}$ that equal zero correspond to partial correlations of zero.

Whereas in the factor modeling approach the covariance matrix of $\mathbf{y}$ is a function of factor loadings and the residual covariance matrix, in the network approach the covariance matrix, $\boldsymbol{\Sigma}$, is defined as

$$\boldsymbol{\Sigma} = \boldsymbol{\Delta}(\mathbf{I} - \boldsymbol{\Omega})^{-1}\boldsymbol{\Delta}, \quad (6.10)$$

in which $\boldsymbol{\Omega}$ is a weight matrix with zeros on the diagonal and off-diagonal elements ω_{ij} that reflect the edge weights between nodes, and $\boldsymbol{\Delta}$ is a diagonal matrix in which the elements are functions of the diagonal elements of the precision matrix, $\delta_{jj} = p_{jj}^{-\frac{1}{2}}$ (Epskamp, Rhemtulla, & Borsboom, 2017).

When $\boldsymbol{\Omega}$ is obtained by standardizing the inverse sample covariance matrix and multiplying the off-diagonal elements by -1 , one obtains a saturated model, which perfectly reproduces the sample partial correlation matrix (Epskamp, Rhemtulla, & Borsboom, 2017). In this case, all nodes will be connected to all other nodes (with the rare exception of sample partial correlations that equal exactly zero). In practice, modelers gain degrees of freedom by assuming that the true network is sparse, that is, that relatively few direct interactions are necessary to fully explain the covariance between observed variables (Costantini et al., 2015; Deserno, Bors-

boom, Begeer, & Geurts, 2017b; Epskamp, Rhemtulla, & Borsboom, 2017; Epskamp, Waldorp, Möttus, & Borsboom, 2018; Isvoranu et al., 2016). Sparsity is modeled by constraining some elements of Ω to equal zero; these zeroes correspond to pairs of variables that are conditionally independent given all other variables in the network. In the following we refer to this model as a sparse network model (SNM). As such, the SNM is the multivariate normal version of a pairwise Markov Random Field (MRF; Koller & Friedman, 2009). A pairwise MRF is an undirected network in which nodes represent observed random variables and any edge between two nodes represents the conditional dependence between these nodes given all other nodes in the network. Correspondingly, any missing edge represents a conditional independence between these nodes given all other nodes and as such a MRF encodes the conditional independence structure of a set of nodes.

6.4.2 DIVERGENT IMPLICATIONS

When considering two specific models, such as the UFM and SNM, rather than two general model classes (e.g., all factor models versus all network models), we can derive targeted comparisons of particular phenomena in the data for which the relevant models have divergent statistical predictions. We explicate one such implication in this section.

PARTIAL CORRELATIONS AND ZERO-ORDER CORRELATIONS

For unidimensional factor models, each partial correlation lies between zero and the zero-order correlation (Van Bork, Grasman, & Waldorp, 2018). This result includes three more specific implications of UFM: (1) the UFM implies that partial correlations cannot equal zero, (2) the UFM implies that partial correlations cannot switch sign compared to the zero-order correlation and (3) the UFM implies that partial correlations cannot be stronger than the zero-order correlation. In the following we will explain these three implications.

PARTIAL CORRELATIONS OF ZERO

Let y_i , y_j and y_z be elements of $\mathbf{y}$, the vector of observed variables. The partial correlation between y_i and y_j given y_z is defined as follows (Chen & Pearl, 2014):

$$\rho_{ij.z} = \frac{\rho_{ij} - \rho_{iz}\rho_{jz}}{\sqrt{(1 - \rho_{iz}^2)(1 - \rho_{jz}^2)}} \quad (6.11)$$

Figure 6.2 shows two possible causal mechanisms that result in nonzero correlations between y_i , y_j and y_z . In Figure 6.2a, two direct relations (ω_{zi} and ω_{jz}) cause all three variables to be correlated. By plugging in the correlations that correspond to Figure 6.2a ($\rho_{iz} = \omega_{zi}$, $\rho_{jz} = \omega_{jz}$ and $\rho_{ij} = \omega_{zi}\omega_{jz}$) in Equation (6.11), the numerator of Equation (6.11) becomes zero. If a common factor underlies the covariance of y_i , y_j , and y_z (depicted in Figure 6.2b), however, $\rho_{ij.z}$ cannot equal zero. It follows from Equation (6.8) that, if y_i , y_j and y_z are influenced by a single common factor like in Figure 6.2b, their correlations can be expressed as follows:

$$\begin{aligned}\rho_{ij} &= \lambda_i \lambda_j \\ \rho_{iz} &= \lambda_i \lambda_z \\ \rho_{jz} &= \lambda_j \lambda_z\end{aligned}\tag{6.12}$$

The partial correlation $\rho_{ij.z}$ equals zero when $\rho_{ij} = \rho_{iz}\rho_{jz}$. Using expression 6.12, in the case of a UFM, this equivalence holds if and only if $\lambda_i \lambda_j = (\lambda_i \lambda_z)(\lambda_j \lambda_z)$, which only holds if either λ_i or λ_j equals 0, or when $\lambda_z = 1$ or $\lambda_z = -1$. Put differently, partial correlations of 0 in a UFM coincide with standardized factor loadings of 0, 1 or -1 . A standardized factor loading of zero implies that the corresponding variable does not load on the common factor and thus none of its correlations with other variables can be explained by the common factor. A variable with a standardized factor loading of 1 or -1 is perfectly correlated with the latent factor, which, as a result, is no longer latent. In this case, the common factor is observed, and conditioning on the variable with factor loading 1 or -1 would render all other variables statistically independent, due to the assumption of local independence (Bollen, 1989b).

Holland and Rosenbaum (1986) proved that the simple example above with three observed variables applies to all UFM; that is, UFM imply that partial correlations between the observed variables can never be exactly zero. However, do note that, given a unidimensional factor model, as the set of observed variables increases (i.e., as $\mathbf{y}$ contains more elements), partial correlations may become very small, as the set of observed variables partialled out yields an increasingly better approximation to the latent factor (see also, Guttman, 1940, 1953). This suggests that, unless one has access to a very large sample so that one can distinguish between zero partial correlations and very small partial correlations, this divergent implication may be most efficiently utilized for relatively small sets of variables.

(a) Direct relations underlying the correlations between y_i, y_j and y_z . The causal structure on the left implies the conditional independence structure on the right.

(b) Single common cause underlying the correlations between y_i, y_j and y_z .

Figure 6.2: Two different models that explain the correlational structure of three observed variables.

SIGN SWITCH IN PARTIAL CORRELATIONS

Holland and Rosenbaum (1986) also showed that UFM's imply that the partial correlation has to have the same sign as the zero-order correlation. Consider the earlier example with three variables y_i, y_j and y_z again. The denominator in Equation (6.11) is positive. This means that if ρ_{ij} is positive, the partial correlation $\rho_{ij.z}$ only switches sign (i.e., is negative) when the numerator $\rho_{ij} - \rho_{iz}\rho_{jz}$ is negative. Suppose now that y_i, y_j and y_z are indicators in a UFM. If ρ_{ij} is positive, then $\lambda_i\lambda_j$ is positive. This means that the numerator in Equation (6.11) cannot be negative since we established that for a UFM the numerator equals $\lambda_i\lambda_j - (\lambda_i\lambda_z)(\lambda_j\lambda_z)$ which equals $\lambda_i\lambda_j - \lambda_i\lambda_j\lambda_z^2$. For this to be negative λ_z^2 has to be larger than one. In contrast, a network model does not imply that the partial correlation should have the same sign as the zero-order correlation. Collider structures in which two variables have a common effect, are a well known example in which the correlation between the causes can be (slightly) positive but the partial correlation between the causes conditioning on the common effect is negative (Lauritzen, 1996; Pearl, 2000a). For example, when y_i and y_j have a weak positive correlation but both have a strong positive influence on y_z , then the partial correlation between y_i and y_j given y_z will be negative. This result suggests that the observation of partial correlations that have a different sign than the zero-order correlation indicate evidence in favor of an SNM over a UFM.

INCREASE IN ABSOLUTE PARTIAL CORRELATIONS

The third implication of a UFM that diverges from an SNM, is that partial correlations between two indicators in a UFM cannot be stronger than the corresponding zero-order correlation between these indicators. That is, the *absolute* partial correlation between two variables can never be larger than the corresponding *absolute* zero-order correlation.

To get an intuitive sense for why this implication follows, consider again Equation (6.11). Suppose that ρ_{ij} is positive, then $\rho_{ij.z}$ is stronger than ρ_{ij} only if $\rho_{iz}\rho_{jz}$ is negative. For $\rho_{iz}\rho_{jz}$ to be negative, either ρ_{iz} or ρ_{jz} has to be negative, but not both. That is, exactly one of the three correlations must be negative. Suppose that ρ_{ij} is negative, then $\rho_{ij.z}$ is stronger than ρ_{ij} only if $\rho_{iz}\rho_{jz}$ is positive. For $\rho_{iz}\rho_{jz}$ to be positive both ρ_{iz} and ρ_{jz} have to be negative or both ρ_{iz} and ρ_{jz} have to be positive. As a result, for any three variables there are two correlational structures that result in partial correlations that are stronger than the corresponding zero-order correlations: (I) one negative correlation and two positive correlations or (II) three negative correlations.

Neither of these two correlational structures, however, can be explained by a UFM, as be-

comes clear from examining expression (6.12). Given any three variables, to arrive at a correlational structure with either one or three negative correlations at least one factor loading ought to be negative. However, both one and two negative factor loadings result in two negative correlations, while three negative factor loadings results in no negative correlations. In sum, there is no way to choose factor loadings such that the model results in a correlation structure that complies with partial correlations that are stronger than their corresponding zero-order correlations.

The above example extends to more than three variables. Let $\hat{y}_{i;V \setminus \{i,j\}}$ and $\hat{y}_{j;V \setminus \{i,j\}}$ denote the best linear approximation of respectively y_i and y_j in terms of the variables that are partialled out, $V \setminus \{i, j\}$ (i.e., the variation in y_i that is explained by $V \setminus \{i, j\}$). The residual of y_i is $y_i - \hat{y}_{i;V \setminus \{i,j\}}$ and is denoted as $\hat{y}_{i \cdot V \setminus \{i,j\}}$ (because it represents the part of y_i that is not predicted by other variables in $\mathbf{y}$). The partial correlation between y_i and y_j (denoted $\hat{\rho}_{ij \cdot V \setminus \{i,j\}}$) is the zero-order correlation between $\hat{y}_{i \cdot V \setminus \{i,j\}}$ and $\hat{y}_{j \cdot V \setminus \{i,j\}}$ (i.e., between the residuals² of y_i and y_j ; Cramér, 1999). In any UFM, each observed variable (y_i) is a function of the factor (η) and some random error (ε_i). The principle of local independence implies that the observed variables only share variance due to η . Thus, by subtracting a linear combination of the other observed variables from both y_i and y_j , variance related to η is pulled out, while none of the random error (ε_i or ε_j) is pulled out as this is not shared with the other observed variables. Because η is precisely what y_i and y_j share, the correlation between $\hat{y}_{i \cdot V \setminus \{i,j\}}$ and $\hat{y}_{j \cdot V \setminus \{i,j\}}$ (i.e., the partial correlation between y_i and y_j) must be smaller in absolute size than the zero-order correlation. Note that the partial correlation does not become zero because the observed variables that are partialled out also contain random error. However, if any of the partialled-out variables contained no measurement error, all of η would be subtracted from y_i and y_j , resulting in a partial correlation of zero. This is in correspondence with the previous section, in which we showed that partial correlations can only equal zero if at least one standardized factor loading equals one (i.e., there is no measurement error). For a complete proof of the general case with p variables see Van Bork et al. (2018).

6.5 REASONING WITH EQUIVALENCE AND TESTABILITY: A PRACTICAL EXAMPLE

The above discussion of equivalence and testability suggests that, in practical research situations, model equivalences can sharpen and inform the way researchers approach their data. For

²Note here that the term ‘residual’ refers to $\hat{y}_{i \cdot V \setminus \{i,j\}}$ rather than to ε_i (i.e., $\hat{y}_{i \cdot \eta}$ in a UFM).

Figure 6.3: The UFM that was estimated from the item responses in the procrastination dataset. The estimated factor loadings are standardized.

instance, the equivalence proofs can be used to construct alternative hypotheses on the data-generating mechanism, by considering which latent variable model would be able to mimic data from a network or vice versa (see Epskamp, Kruis, & Marsman, 2017, for some examples). In addition, the divergent statistical predictions identified in this paper can be used to gain insight into the relative likelihoods of particular models under investigation. To illustrate this, we examine a dataset containing items designed to assess procrastination. We consider the alternative hypotheses that the item responses arise from a UFM or from an SNM.

The dataset consists of 414 observations on items of the Irrational Procrastination Scale (IPS; Steel, 2010), and is part of an ongoing data collection programme of the Allameh Tabataba'i University in Tehran. The IPS is a self-report measure of procrastination consisting of nine items that are scored on a 5-point Likert scale. The items and the polychoric correlation matrix of the data are included in Appendix A, included at the end of this dissertation. The internal consistency of the IPS is $\alpha = 0.91$ in the validation sample (Steel, 2010), and $\alpha = 0.81$ in the sample we use for our analyses. We used the polychoric correlation matrix of these 9 variables to estimate a UFM and SNM. Steel (2010) found that a single factor was sufficient to explain the data. These findings have been replicated in other samples which resulted in the same conclusion of

Figure 6.4: The network that is equivalent to the estimated UFM in Figure 6.3. All weights are a function of the nine factor loadings, so that the weights conform to a rank one matrix. Note that the network is not an SNM; it is a fully connected network.

a single dimension underlying the procrastination items (Rozental et al., 2014; Svartdal, 2017). Rozental et al. (2014) actually found a second factor for the reversed items but concluded that this is an artifact reflecting that participants missed the negation in these items. The best fitting UFM is represented in Figure 6.3. The equivalent network model is represented in Figure 6.4. Note here that the number of parameters is not equal to the number of non-zero edges: the network is fully connected but all edge weights are a function of the nine factor loadings. The best fitting SNM is represented in Figure 6.5. While the exploratory nature of the lasso regularization used to estimate sparse networks involves that the zeroes are not fixed in location, for

Figure 6.5: The SNM that was estimated from the item responses in the procrastination dataset.

explanatory purposes we consider the following SNM:

$$\Omega = \begin{bmatrix} 0 & 0 & \omega_{13} & \omega_{14} & 0 & \omega_{16} & \omega_{17} & \omega_{18} & \omega_{19} \\ 0 & 0 & \omega_{23} & \omega_{24} & \omega_{25} & \omega_{26} & 0 & \omega_{28} & \omega_{29} \\ \omega_{13} & \omega_{23} & 0 & \omega_{34} & \omega_{35} & 0 & 0 & \omega_{38} & 0 \\ \omega_{14} & \omega_{24} & \omega_{34} & 0 & \omega_{45} & 0 & \omega_{47} & \omega_{48} & \omega_{49} \\ 0 & \omega_{25} & \omega_{35} & \omega_{45} & 0 & \omega_{56} & \omega_{57} & \omega_{58} & 0 \\ \omega_{16} & \omega_{26} & 0 & 0 & \omega_{56} & 0 & \omega_{67} & \omega_{68} & \omega_{69} \\ \omega_{17} & 0 & 0 & \omega_{47} & \omega_{57} & \omega_{67} & 0 & \omega_{78} & 0 \\ \omega_{18} & \omega_{28} & \omega_{38} & \omega_{48} & \omega_{58} & \omega_{68} & \omega_{78} & 0 & \omega_{89} \\ \omega_{19} & \omega_{29} & 0 & \omega_{49} & 0 & \omega_{69} & 0 & \omega_{89} & 0 \end{bmatrix} \quad (6.13)$$

in which $\omega_{12} = \omega_{15} = \omega_{27} = \omega_{36} = \omega_{37} = \omega_{39} = \omega_{46} = \omega_{59} = \omega_{79} = 0$. Let us here assume that the remaining edge weights are freely estimated so that there are no additional proportionality constraints on the edge weights. We can understand the above SNM model as the set of all fully parameterized models in which the above constraints hold (i.e., the edges

set to zero). The SNM in Figure 6.5 is an element of this set. For each element in this set we can construct the equivalent factor model by following the same logic as is used to prove the equivalence of the Ising model and MIRT. For example, for the estimated network model in Figure 6.5 we can take the correlation matrix that is implied by that SNM and interpret the eigenvalue decomposition of this matrix as a representation of a factor model. Note again here that we, for simplicity, ignore the fact that this model is estimated with lasso regularization and instead treat this estimated SNM as if it were estimated as an element of the model in Table 6.13, that is, with nine specific edge weights constrained to zero. The covariance matrix of the estimated SNM is of full rank and thus the eigenvalue decomposition results in as many factors as observed variables, i.e., all nine positive eigenvalues can be seen to represent a factor. This factor solution is represented in Figure 6.6. Note that in the continuous case the diagonal of the connectivity matrix is not arbitrary and can therefore not be set to a value that results in one eigenvalue of zero. As a result the model has nine instead of eight factors. The factor loadings in the factor model equivalent to the SNM in Table 6.13, are not freely estimated; the edge weights that are constrained to zero in the SNM translate to constraints in the factor model that are not trivial to define but result in the same number of parameters being estimated in the SNM and equivalent 9-factor model.

The interesting comparison here is not between the SNM and its equivalent latent variable model (the models in Figure 6.5 and 6.6) or between the UFM and its equivalent network model (the models in Figure 6.3 and 6.4), but rather between the best fitting SNM and best fitting UFM, and these two classes of models (UFMs and SNMs) diverge in their predictions for empirical data, making them distinguishable. To account for sampling variability we compare the likelihoods of the UFM and SNM for a specific statistic based on two of the three diverging implications of UFMs and SNMs described earlier.

In the previous section, we discussed three diverging predictions for empirical data: (1) UFMs imply that the population partial correlations cannot equal zero (2) UFMs imply that population partial correlations cannot have a different sign than the corresponding zero-order correlations and (3) UFMs imply that the population partial correlations can not be stronger than the corresponding zero-order correlations. None of these three implications hold for SNMs. Considering these alternative hypotheses of SNMs versus UFMs, we can assess whether the proportion of partial correlations that have a different sign than the corresponding zero-order correlations in the data is more consistent with a UFM or with an SNM. We can also assess whether the proportion of partial correlations that are stronger than the corresponding zero-order correlations in the data is more consistent with a UFM or with an SNM. We can not assess the number

Figure 6.6: The eigenvalue decomposition of the correlation matrix implied by the SNM in Figure 6.3. Q is a matrix of the eigenvectors and $\mathbf{r}$ is a vector of the eigenvalues.

of partial correlations in the data that are exactly zero, since sample partial correlations will rarely be exactly zero even when the population partial correlation is zero. We could circumvent this problem by assessing the proportion of partial correlations that are significantly different from zero, but since the UFM implies very small partial correlations (especially with more indicators or stronger factor loadings) it is difficult to determine when the power is sufficient to detect zero partial correlations. We therefore use (2) and (3) in the illustration below.

First, we can interpret the proportion of partial correlations that have the opposite sign as the zero-order correlation as a test statistic, and compare the value of this test statistic to the sampling distribution that would arise (a) under the hypothesis that the best fitting UFM were true, and (b) under the hypothesis that the best fitting SNM were true. The relevant sampling distributions, while analytically intractable, are not difficult to simulate. Appendix B provides code to get sampling distributions for the proportion partial correlation with a different sign than the zero-order correlation, under the best fitting UFM and SNM. Figure 6.7A shows the sampling distributions of the test statistic under the best fitting UFM and the best fitting SNM for the

Figure 6.7: The probability mass functions of the best fitting SNM and UFM for (A) the proportion of partial correlations that have the opposite sign as their corresponding zero-order correlations and (B) the proportion of partial correlations that are stronger than their corresponding zero-order correlations. The black vertical line represents the observed proportion in the data.

nine procrastination items. The vertical black line represents the observed proportion of partial correlations in the procrastination data that have a different sign than their corresponding zero-order correlation, t_{sign} . The sampling distributions function as a probability mass function and thus the probability mass of the best fitting UFM for the observed proportion t_{sign} equals the likelihood of this UFM given t_{sign} : $P(t_{sign}|\text{ML}(\text{UFM}))$, in which $\text{ML}(\text{UFM})$ refers to the maximum likelihood estimate of the UFM given the observed data. Analogously the likelihood of the best fitting SNM equals the probability mass of the SNM for the observed t_{sign} .

As we can see from Figure 6.7A, the best fitting SNM has a higher likelihood given the observed proportion partial correlations that switched sign, than the best fitting UFM. The best fitting SNM assigns a probability of .23 to the observed proportion of partial correlations in the data that switched sign compared to the zero-order correlation. The best fitting UFM assigns a probability of only .054 to the observed proportion in the data. So, the observed proportion of partial correlations with a switched sign is more probable under the estimated SNM than under the estimated UFM.

Second, we can assess the number of partial correlations that are stronger than the corresponding zero-order correlations. This proceeds roughly in the same way as the evaluation of partial correlations that switch sign: we compute the number of partial correlations that have a higher absolute value than the corresponding correlations, and evaluate that against the im-

plied sampling distributions as derived from the best fitting UFM and SNM. These are given in Figure 6.7B.

As we can see from the figure, the observed proportion of partial correlations that are stronger than the corresponding zero-order correlations is more probable under an SNM than under a UFM. The best fitting SNM assigns a probability of .47 to the observed proportion of partial correlations in the data that are stronger than their corresponding zero-order correlation. The best fitting UFM only assigns a probability of .02 to the observed proportion in the data. For this statistic (proportion of partial correlations stronger than zero-order correlations) the probability mass functions are distributed over less possible outcome values. Under the estimated UFM the only three proportions that were observed were 0, $1/45$ and $2/45$, corresponding to respectively zero, one and two partial correlations that are stronger than the zero-order correlations. Under the UFM it is most probable to draw a sample with a proportion of 0 (i.e., no partial correlations stronger than zero-order correlations); the UFM assigns a probability of .77 to the proportion 0.

The results of these two comparisons suggest that a focus on the relation between the partial correlations and zero-order correlations in the data provides evidence for the estimated SNM over the estimated UFM. Given the observed proportion of partial correlations that have a different sign than the zero-order correlation and given the observed proportion of partial correlations that are stronger than the zero-order correlation, the SNM has a higher likelihood than the UFM.

6.6 DISCUSSION

Psychology has a history of modeling theoretical constructs as latent factors that act as common determinants for a set of observed variables (e.g., Caspi et al., 2014; McCrae & Costa, 1987; Spearman, 1904). In the past decade, however, multiple studies have shown that networks offer reasonable alternative ways of understanding such constructs (e.g., Cramer et al., 2012; McNally et al., 2014; Van der Maas et al., 2006). At this point in time, it is possible to model the correlation structure in a dataset using either a factor model (explaining the correlations by a common cause) or a network model (explaining the correlations by relatively few direct interactions). In the past couple of years multiple papers have proved an equivalence between latent variable models and network models (Epskamp, Maris, et al., 2018; Kruis & Maris, 2016; Marsman et al., 2015) which might lead some to conclude that the latent variable model and network model are just two different representations of the same model, suggesting it is senseless to compare these

models in their plausibility to have generated the data. We argued against this view: the difference in substantive implications of these modeling frameworks makes the distinction between network models and latent variable models crucial. But, even if it is important to distinguish these models, one may take the statistical equivalence between these models to mean that it is impossible to make comparisons based on empirical data alone. We argued that while statistical equivalence holds for the broad classes of models, this is typically not the level on which comparisons between models are being made.

Our argument throughout this paper was twofold. First, we showed how the models are equivalent only in the broadest sense; any semi-positive definite matrix can be represented as a network model and as a factor model. We argued that although the general group of latent variable models is equivalent to the general group of network models, it is important to consider what models exactly are equivalent because in many cases one of the equivalent models will not form a plausible candidate for the data-generating mechanism. The equivalence between network models and latent variable models relies on the fact that any positive semi-definite covariance matrix results in an eigenvalue decomposition with nonnegative eigenvalues, which can be interpreted as a latent variable model with as many latent variables as positive eigenvalues (Epskamp, Maris, et al., 2018; Kruis & Maris, 2016). Thus, any network model will also imply a positive definite matrix that can be explained with as many latent variables as positive eigenvalues. The resulting factor models, however, are in many cases not plausible as a data-generating mechanism. As we have shown in this paper, many models will consist of as many orthogonal latent variables as observed variables in the data with uninterpretable constraints on the factor loadings or discrimination parameters. The other way around, the constraints that form a sensible hypothesis in the latent variable framework (e.g., omitting latent variables) translate to constraints that are difficult to interpret substantively in the network framework (rank constraints).

Second, we shifted the focus from the broad classes of latent variable models and network models to more specific subsets of models within these broader classes that are theoretically interesting to compare. We believe that in most cases researchers do not want to compare just any multi-dimensional factor model to any network model but rather focus on the comparison of subsets of these models. Rather than comparing the full group of network models to the full group of factor models, researchers might want to compare two models that fit the data and have theoretical support. We showed that for example SNMs and UFM diverge in their predictions for empirical data, and as such observations in the data can lend support for the one model over the other model. We intend this to be a first step towards a comprehensive research program

that could be directed at the question of which sets of models from the different frameworks are empirically distinguishable.

In order to identify more of such subsets of models that diverge in their empirical predictions it is important to note that the group of SNMs and UFM models that are considered in this paper are mutually exclusive; there are no UFM models that are statistically equivalent to a SNM and vice versa. The group of SNMs is a subset of the set of all possible network models, namely those that have missing edges, and the set of unidimensional factor models is a subset of the set of all factor models. Our results are important because they show that, although it might not be possible to distinguish a given network model from all possible factor models, there are groups of models within the class of factor models and network models that can be statistically compared, and further research might identify more of such groups. However, it should be clearly kept in mind that such comparisons only pit specific models against each other. It is always possible that some other model has generated the data, so the conclusion that one specific model explains certain features of the data better than another does not license the conclusion that the model is correct. As is generally the case in model comparisons, the best fitting model may still be entirely wrong.

We believe it is important to explore to what kinds of factor models other than a UFM the implications discussed in this paper may extend. The three implications of UFM models discussed in this article do not generalize to all factor models, as cross-loadings as well as orthogonal factors make it possible to arrive at, for example, partial correlations of zero. However, the implications of UFM models discussed in this paper likely do generalize to some types of factor models other than unidimensional ones. Possible candidates are higher-order factor models with one factor at the apex and thus also the subset of correlated factor models that are equivalent to such higher-order factor models. Future research may be directed to evaluate alternative scenarios, and investigate which properties of the data are best suited to distinguish between networks and latent variable models in these scenarios.

7

Unidimensional factor models imply weaker partial correlations than zero-order correlations

Abstract

In this paper we present a new implication of the unidimensional factor model. We prove that the partial correlation between two observed variables that load on one factor given any subset of other observed variables that load on this factor, lies between zero and the zero-order correlation between these two observed variables. We implement this result in an empirical bootstrap test that rejects the unidimensional factor model when partial correlations are identified that are either stronger than the zero-order correlation or have a different sign than the zero-order correlation. We demonstrate the use of the test in an empirical data example with data consisting of fourteen items that measure extraversion.

This chapter is published as Van Bork, R., Grasman, R.P.P.P., & Waldorp, L. J. (2018). Unidimensional factor models imply weaker partial correlations than zero-order correlations. *Psychometrika*, 83, 443–452.

7.1 INTRODUCTION

Unidimensional factor models (UFMs) are widely used throughout the social and behavioral sciences, and many of its implications have already been revealed. In this paper, we reveal yet another implication, namely that the partial correlation between two observed variables given any subset of other observed variables is always closer to zero than the zero-order correlation between these observed variables. As we will show, this implication adds to a series of implications generated by the work of, among others, Holland and Rosenbaum (1986), Ellis (2014) and Guttman (1940) and can be used to detect misfit of the UFM for the observed data.

In the remainder of this section we describe some of the implications of the UFM that have been revealed previously in order to situate our contribution. We then turn to the actual proof of our contribution, before discussing possible ways in which it can be used in future research. To illustrate the potential use of the result we consider its use in a bootstrap test that can detect misfit of the UFM and can also indicate for which observed variables this misfit is obtained.

The basic assumption of UFMs is that the observed variables are conditionally independent given the latent factor. This condition is sometimes called *latent conditional independence* (Holland & Rosenbaum, 1986) but is more widely known as *local independence* in Item Response Theory (IRT; Lord, 1980). The model implied covariance matrix of a UFM with local independence equals the sum of a diagonal matrix representing the residual covariance matrix, and a matrix of rank one. This characteristic of the UFM has resulted in multiple implications for the covariance structure of the observed variables.

For example, building on Rosenbaum (1984)'s work on IRT models, Holland and Rosenbaum (1986) showed that nonnegative associations between observed variables in a UFM imply nonnegative conditional associations. When the observed variables are multivariate Gaussian, the result of Holland and Rosenbaum (1986) implies that nonnegative correlations imply nonnegative partial correlations. That is, for any set of multivariate Gaussian observed variables that load on a single factor, partial correlations cannot have the opposite sign of their corresponding zero-order correlation.

Ellis (2014) shows that in a unidimensional latent variable model for any pair of two binary observed variables given any subset of the other binary observed variables and their products, the partial correlation is nonnegative. This condition implies that the unconditional correlation matrix must show a certain ordinal pattern (Ellis, 2014). Yet another implication of the UFM is that the partial correlation between two observed variables given the other observed variables cannot equal zero (de Fátima Salgueiro, Smith, & McDonald, 2008). In addition, de

Fátima Salgueiro et al. (2008) show that certain patterns of signs in the partial correlation matrix are incompatible with a UFM.

A final implication of the factor model that is of importance here is that Guttman claimed that as the number of observed variables increases to infinity the square of the multiple correlation coefficient of an observed variable on the other observed variables tends to the communality and all partial correlations between this observed variable and the other observed variables tend to zero (Guttman, 1940, 1953).

In this paper we present a related implication of the UFM: the partial correlation between two observed variables given any subset of the other observed variables is always closer to zero than the zero-order correlation between these observed variables. This, together with the implications mentioned previously, that partial correlations cannot equal zero and cannot switch sign compared to the zero-order correlations, means that the UFM implies that the partial correlation between any two observed variables given any subset of the other observed variables lies between zero and the zero-order correlation between these two observed variables. This result implies that as the number of observed variables increases to infinity the partial correlations not only tend to zero but tend *monotonically* to zero.

7.2 UNIDIMENSIONAL FACTOR MODELS

We first define UFM's before presenting a proof. Let Σ denote the covariance matrix of $\mathbf{y}$, in which $\mathbf{y}$ denotes the vector of observed variables. We assume that Σ is nondegenerate and so is positive definite. Let λ denote the vector of factor loadings and Θ the residual covariance matrix. In a UFM all observed variables in $\mathbf{y}$ are a linear function of the same factor, η , and independent residuals, ε :

$$y_i = \lambda_i \eta + \varepsilon_i. \quad (7.1)$$

We assume $\text{var}(\eta) = 1$. We also assume mutually uncorrelated residuals, that is, Θ is diagonal. In the following, we take Σ to be standardized and assume $\forall i, |\lambda_i| \in (0, 1)$, so that the observed variables are correlated to the factor but not perfectly correlated to the factor. The model implied covariance matrix of the observed variables is a function of the factor loadings and the residual covariance matrix:

$$\Sigma = \lambda \lambda' + \Theta. \quad (7.2)$$

Equation (7.2) implies that the covariance among observed variables is a function of their factor loadings. More precisely, because Θ is a diagonal matrix, the covariance between two variables

y_i and y_j equals $\lambda_i \lambda_j$.

Consider three variables y_1, y_2 and z . The partial correlation between y_1 and y_2 given z can be expressed in terms of their zero-order correlations by (e.g., Chen & Pearl, 2014):

$$\rho_{y_1 y_2 \cdot z} = \frac{\rho_{y_1 y_2} - \rho_{y_1 z} \rho_{y_2 z}}{\sqrt{(1 - \rho_{y_1 z}^2)(1 - \rho_{y_2 z}^2)}} \quad (7.3)$$

Some correlation structures for three variables imply that the partial correlation is stronger than the correlation. For example, a negative $\rho_{y_2 z}$ in combination with a positive $\rho_{y_1 y_2}$ and $\rho_{y_1 z}$ result in a partial correlation $\rho_{y_1 y_2 \cdot z}$ that is stronger than the zero-order correlation $\rho_{y_1 y_2}$. Langford, Schwertman, and Owens (2001) point out that the property of being positively correlated is not transitive: for three variables it is possible to have a correlation structure with one negative and two positive correlations. However, a structure with one negative and two positive correlations is not possible under a UFM as it is impossible to choose factor loadings such that they result in two positive correlations and one negative correlation. In fact all possible correlation structures that result in some partial correlations that are stronger than their corresponding zero-order correlation appear to be impossible under a UFM. For example, for three variables y_1, y_2 and z that are all elements of $\mathbf{y}$, we can substitute each correlation in equation (7.3) with factor loadings:

$$\rho_{y_1 y_2 \cdot z} = \frac{\lambda_{y_1} \lambda_{y_2} (1 - \lambda_z^2)}{\sqrt{1 - \lambda_{y_1}^2 \lambda_z^2} \sqrt{1 - \lambda_{y_2}^2 \lambda_z^2}} \quad (7.4)$$

Note that both the denominator $\sqrt{1 - \lambda_{y_1}^2 \lambda_z^2} \sqrt{1 - \lambda_{y_2}^2 \lambda_z^2}$ and $(1 - \lambda_z^2)$ are positive and that $\sqrt{1 - \lambda_{y_1}^2 \lambda_z^2} \sqrt{1 - \lambda_{y_2}^2 \lambda_z^2} > (1 - \lambda_z^2)$. It follows that $0 < \frac{1 - \lambda_z^2}{\sqrt{1 - \lambda_{y_1}^2 \lambda_z^2} \sqrt{1 - \lambda_{y_2}^2 \lambda_z^2}} < 1$, and thus that the partial correlation $\rho_{y_1 y_2 \cdot z}$ is weaker than the zero-order correlation $\lambda_{y_1} \lambda_{y_2}$. While this example only serves as a proof for $p = 3$ variables, in the following section it is proved that for any p the UFM implies that the partial correlations are necessarily weaker than the zero-order correlations.

7.3 WEAKER PARTIAL CORRELATIONS THAN ZERO-ORDER CORRELATIONS

We start with providing the assumptions defined in the the previous section.

Assumptions.

1. $\Sigma = \lambda\lambda' + \Theta$, and $\Sigma_{ii} = 1$ for all i
2. Θ is diagonal and positive definite,
3. $|\lambda_i| \in (0, 1)$ for all i .

These assumptions imply the following lemma.

Lemma. Assume 1 - 3 above. For a set of p continuous random variables that load on one common factor, the partial correlation between any two of them, conditional on any non-empty subset of the other $p - 2$ variables, is weaker than their marginal correlation.

Proof. The proof is for the partial correlation conditional on all $p - 2$ remaining variables. That the result holds for any subset follows from restricting Σ to the subset of variables.

Let $\mathbf{y} = [y_1, y_2, \dots, y_p]'$. Partition $\mathbf{y}' = (\mathbf{Y}', \mathbf{Z}')$, where $\mathbf{Y} = (y_1, y_2)'$, and $\mathbf{Z} = (y_3, \dots, y_p)'$. Partition $\lambda' = (\lambda_Y', \lambda_Z')$ accordingly. The partitioned covariance matrix of $\mathbf{y}$ is

$$\Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{pmatrix} = \begin{pmatrix} \lambda_Y \lambda_Y' + \Theta_{YY} & \lambda_Y \lambda_Z' \\ \lambda_Z \lambda_Y' & \lambda_Z \lambda_Z' + \Theta_{ZZ} \end{pmatrix},$$

where $\Theta_{YY} = \text{diag}(1 - \lambda_1^2, 1 - \lambda_2^2)$, and $\Theta_{ZZ} = \text{diag}(1 - \lambda_3^2, \dots, 1 - \lambda_p^2)$. The partial correlation between y_1 and y_2 , conditioned on all the other variables, can be computed from the partial covariance matrix

$$\begin{aligned} \Sigma_{YY \cdot Z} &= \Sigma_{YY} - \Sigma_{YZ} \Sigma_{ZZ}^{-1} \Sigma_{ZY} \\ &= (\lambda_Y \lambda_Y' + \Theta_{YY}) - \lambda_Y \lambda_Z' (\lambda_Z \lambda_Z' + \Theta_{ZZ})^{-1} \lambda_Z \lambda_Y' \\ &= \lambda_Y \lambda_Y' [1 - \lambda_Z' (\lambda_Z \lambda_Z' + \Theta_{ZZ})^{-1} \lambda_Z] + \Theta_{YY}. \end{aligned}$$

Reading off the partial variances $(\Sigma_{YY \cdot Z})_{11}$, $(\Sigma_{YY \cdot Z})_{22}$ and covariance $(\Sigma_{YY \cdot Z})_{12}$, the partial correlation can be expressed as

$$\begin{aligned} \rho_{y_1 y_2 \cdot Z} &= \frac{\lambda_1 \lambda_2 (1 - \gamma^2)}{\sqrt{\lambda_1^2 (1 - \gamma^2) + (1 - \lambda_1^2)} \sqrt{\lambda_2^2 (1 - \gamma^2) + (1 - \lambda_2^2)}} \\ &= \frac{\lambda_1 \lambda_2}{\sqrt{\lambda_1^2 + (1 - \lambda_1^2)/(1 - \gamma^2)} \sqrt{\lambda_2^2 + (1 - \lambda_2^2)/(1 - \gamma^2)}}, \end{aligned} \quad (7.5)$$

where $\gamma^2 = \lambda'_Z(\lambda_Z\lambda'_Z + \Theta_{ZZ})^{-1}\lambda_Z > 0$ by the assumption that $\lambda_i \neq 0$. Note that $\rho_{y_1 y_2 \cdot Z} = \rho_{y_1 y_2} = \lambda_1 \lambda_2$ is the marginal correlation if and only if $\gamma^2 \equiv 0$. The Sherman-Morrison formula (Sherman & Morrison, 1950) shows that $\gamma^2 < 1$:

$$\begin{aligned}\lambda'_Z(\lambda_Z\lambda'_Z + \Theta_{ZZ})^{-1}\lambda_Z &= \lambda'_Z(\Theta_{ZZ}^{-1} - \Theta_{ZZ}^{-1}\lambda_Z[1 + \lambda'_Z\Theta_{ZZ}^{-1}\lambda_Z]^{-1}\lambda'_Z\Theta_{ZZ}^{-1})\lambda_Z \\ &= \frac{\lambda'_Z\Theta_{ZZ}^{-1}\lambda_Z}{1 + \lambda'_Z\Theta_{ZZ}^{-1}\lambda_Z} < 1.\end{aligned}$$

As a consequence, $0 < 1 - \gamma^2 < 1$, and hence, the denominator in (7.5) is greater than 1, and $\rho_{y_1 y_2 \cdot Z}$ shrinks to 0 relative to $\rho_{y_1 y_2}$. $\square$

The factor γ^2 in the proof can be interpreted as the multiple correlation coefficient (coefficient of determination) between the latent variable η and the remaining $p - 2$ variables. The smaller this multiple correlation is, the closer the partial correlation is to the marginal correlation; the larger this multiple correlation, the closer to zero the partial correlation is. This is because a larger multiple correlation implies that the variables that are partialled out have strong factor loadings and thus conditioning on these variables pulls out much of the shared variance between y_1 and y_2 .

In our proof, we defined $\mathbf{Z}$ as consisting of all $p - 2$ variables other than $\mathbf{Y}$. However, the vector of variables that are partialled out does not need to include *all* variables loading on the factor other than $\mathbf{Y}$ in order for the partial correlation to be necessarily weaker than the zero-order correlation. In fact, the proof also holds for any subset of $\mathbf{Z}$. Given that the UFM also implies that the partial correlation cannot equal zero (de Fátima Salgueiro et al., 2008) and also cannot have a different sign as the zero-order correlation (Holland & Rosenbaum, 1986), we can conclude that the partial correlation implied by a UFM is bounded on two sides: the partial correlation between any two observed variables given a set of the other observed variables lies between zero and the zero-order correlation between these observed variables.

The proof implies that when a UFM is hypothesized to underlie some dataset, the identification of partial correlations stronger than their corresponding zero-order correlations is an indication of model misfit.

7.4 APPLICATION OF THE RESULT

In the previous section we proved that the UFM implies that the partial correlation between any two variables is the product of the zero-order correlation between these two variables and a

constant between zero and one. However, while this should hold for the correlations in the population, sampling variability can result in partial correlations that are not between zero and the zero-order correlation in the sample, even though a UFM is the true data generating model. For example, the upper triangle of the matrix in Table 7.1 represents a sample correlation matrix of which the true data-generating model is a UFM. The lower triangle represents the corresponding partial correlations. In this sample correlation matrix there are multiple partial correlations that are not between zero and the zero-order correlations (i.e., r_{16} , r_{23} , r_{24} , r_{25} , r_{26} , r_{34} , r_{35} , r_{36} , r_{46}). To decide whether the UFM should be rejected based on these observations, a test for statistical significance is needed. In the following we propose one possible version of such a test, an empirical bootstrap test, that uses the result developed in this paper to identify misfit of the UFM. There are four observations we use for the test that provide evidence against a UFM:

1. The correlation and corresponding partial correlation are both positive but the partial correlation is more strongly positive than the correlation.
2. The correlation and corresponding partial correlation are both negative but the partial correlation is more strongly negative than the correlation.
3. The correlation is negative while the corresponding partial correlation is positive.
4. The correlation is positive while the corresponding partial correlation is negative.

In the bootstrap test that we lay out in the following, the goal is to identify correlations for which any of the four observations above holds, accounting for sampling variability.

These four points can be summarized in one implication of the UFM, that is, the UFM implies $0 < \rho_{y_i y_j \cdot Z} / \rho_{y_i y_j} < 1$. If and only if observation (1) or (2) is the case $\rho_{y_i y_j \cdot Z} / \rho_{y_i y_j} > 1$. If and only if observation (3) or (4) is the case $\rho_{y_i y_j \cdot Z} / \rho_{y_i y_j} < 0$. Note that $0 < \rho_{y_i y_j \cdot Z} / \rho_{y_i y_j} < 1$ if and only if $|2\rho_{y_i y_j \cdot Z} / \rho_{y_i y_j} - 1| < 1$, and so the latter statistic can be used to test for all four observations above¹. The distribution of this ratio is not easily obtained. We therefore constructed a bootstrap procedure to test whether $|2\rho_{y_i y_j \cdot Z} / \rho_{y_i y_j} - 1| > 1$. Finding that $|2\rho_{y_i y_j \cdot Z} / \rho_{y_i y_j} - 1| > 1$ provides evidence against the UFM.

¹Let $x = \rho_{y_i y_j \cdot Z} / \rho_{y_i y_j}$. It follows that $0 < x < 1 \iff 0 < (x-0.5)+0.5 < 1 \iff -0.5 < x - 0.5 < 0.5 \iff |x - 0.5| < 0.5 \iff 2|x - 0.5| < 1 \iff |2x - 1| < 1$.

7.4.1 BOOTSTRAP TEST

We propose the following empirical bootstrap procedure to test the hypothesis that the data come from a UFM:

1. Sample n observations with replacement from the original data set
2. From the m -th bootstrap sample, for all $i, j = 1, \dots, p, i \neq j$,
 - (a) compute the correlations $r_{y_i y_j}$ and partial correlations $r_{y_i y_j \cdot Z_{ij}}$, where $Z_{ij} = \{y_k\}_{k \neq i, j}$, $|Z_{ij}| = p - 2$
 - (b) Compute the statistics $s_{ij}^m = |2r_{y_i y_j \cdot Z_{ij}} / r_{y_i y_j} - 1|$
3. Repeat steps 1 through 2b M times (e.g., $M = 1000$)
4. Use the bootstrapped $\{s_{ij}^m\}_{m=1}^M$ to construct a one-sided 95% confidence interval (CI). Any pair of variables y_i, y_j for which the 95% CI only includes values greater than one indicates misfit of the UFM.

The online supplementary materials provide an R function named `onefactor.test()` that implements the above algorithm (see <https://link.springer.com/article/10.1007/s11336-018-9607-z#SupplementaryMaterial>). The output of the test provides a 95% CI of the parameter $|2\rho_{y_i y_j \cdot Z} / \rho_{y_i y_j} - 1|$ for each of the $p(p-1)/2$ unique correlations and their corresponding partial correlation. The default significance level α is set to 0.05 corresponding to a 95% CI, but one can adjust α to account for multiple testing (the test constructs a CI for each pair of variables in the data and thus performs $p(p-1)/2$ tests).

Earlier we stressed the need for a test by arguing that sampling variability can result in partial correlations that are not between zero and the zero-order correlation even when a UFM underlies the data. The correlation matrix in Table 7.1, for example, corresponds to data that were simulated from a UFM with factor loadings close to zero. Sampling data from a UFM with small factor loadings results in a partial correlation matrix that is close to the marginal correlation matrix (when factor loadings are exactly zero, the population correlation matrix and the population partial correlation matrix are both diagonal matrices). For an illustration of the test it is most interesting to sample data from a population in which the partial correlation matrix is close to the marginal correlation matrix, because in that situation sampling variability can easily

	V ₁	V ₂	V ₃	V ₄	V ₅	V ₆
V ₁	1	0.212	0.139	0.231	0.193	0.079
V ₂	0.129	1	0.296	0.170	0.010	0.042
V ₃	0.128	0.319	1	-0.157	0.065	-0.060
V ₄	0.207	0.210	-0.266	1	0.193	-0.101
V ₅	0.158	-0.074	0.084	0.166	1	-0.088
V ₆	0.122	0.071	-0.110	-0.137	-0.076	1

Table 7.1: Upper triangle of matrix represents sample correlation matrix of dataset with 60 observations that is simulated from UFM with both positive and negative factor loadings. The absolute factor loadings are sampled from a uniform distribution over [0.05, 0.2]. The lower triangle of the matrix represents the corresponding partial correlations.

	V ₁	V ₂	V ₃	V ₄	V ₅	V ₆
V ₁	-	0.028	0.08	0.076	0.045	0.117
V ₂	-	-	0.216	0.129	0.065	0.084
V ₃	-	-	-	0.633	0.063	0.080
V ₄	-	-	-	-	0.068	0.103
V ₅	-	-	-	-	-	0.049
V ₆	-	-	-	-	-	-

Table 7.2: Lower bound of the CI obtained with the bootstrap test for each zero-order correlation and corresponding partial correlation in Table 7.1.

lead to partial correlations that are not between zero and the zero-order correlation. As mentioned earlier, the sample correlation and sample partial correlation matrices included in Table 7.1 include many partial correlations that are stronger than the zero-order correlation. Applying the bootstrap test to the data corresponding to this correlation matrix did not result in any significant results (all CIs for $|2\rho_{y_i y_j \cdot Z} / \rho_{y_i y_j} - 1|$ included the value 1, see Table 7.2) and thus the UFM was not rejected. We included a more detailed description of the results of this analysis, as well as R code to replicate the example, in the supplementary materials to this chapter that are found in Appendix B. Furthermore, these supplementary materials include (a) a similar illustration of the test on data simulated from a random correlation matrix, and (b) a simulation study that shows that the performance of the test in terms of controlling type I errors is adequate.

7.5 EMPIRICAL ILLUSTRATION

In this section, we illustrate the use of the statistical procedure by applying it to an empirical example concerning extraversion. Extraversion is an interesting application because the appropriateness of the latent variable model for personality has been contested. For example, Eysenck (1983) claims that extraversion and neuroticism are strong candidates for referring to a real trait underlying the item responses and according to Eysenck “there seems to be little doubt that personality traits have a firm basis in the individual’s biological structure and functioning”, while alternative theories of personality deny such an interpretation of the latent variables in psychometric models for personality (Cramer et al., 2012). Thus, it is interesting to evaluate to what extent the data are compatible with the unidimensional factor model, as Eysenck’s theory would predict.

We used data from the ‘Vijf PersoonlijkheidsFactoren Test’ (Five Personality Factors Test) developed by Elshout en Akkerman (Elshout & Akkerman, 1975). The data consists of 8954 observations on 14 extraversion items on a seven-point Likert scale. For a more detailed description of the dataset we refer to the paper of Smits, Dolan, Vorst, Wicherts, and Timmerman (2013) who made the data publicly available. We note that the exact wording of the items is not available due to copyright issues (Smits et al., 2013). We here focus on the Extraversion factor but because the complete data for all five factors is publicly available anyone who is interested can perform the test on any of the five personality factors.

The upper triangle of Table 7.3 includes the zero-order correlations between the 14 extraversion items. The lower triangle of Table 7.3 includes the corresponding partial correlations. As can be seen from Table 7.3 all zero-order correlations between the items are positive. We used the bootstrap function ‘`onefactor.test()`’ that is included in the online supplementary materials, and set the number of bootstraps to 100.000. We used a Bonferroni correction to correct for multiple testing. The results of the bootstrap test on the extraversion data are summarized in Table 7.4 which includes the lower bounds of the CI’s for each pair of a zero-order correlation and partial correlation between two variables. For example, consider the zero-order correlation and partial correlation between V_3 and V_{12} (Table 7.3 shows that these are 0.13 and -0.01 respectively). The lower bound of the CI associated with this pair is 0.62, which is not larger than 1 and therefore we conclude that although in the sample the sign for the partial correlation between V_3 and V_{12} is different from the sign of the zero-order correlation, this is not a significant result. Cells with a gray background refer to pairs of variables for which the partial correlation is either significantly stronger than the corresponding zero-order correlation or has a different

sign (i.e., the lower bound of the CI is larger than 1). There are 10 pairs of variables for which the partial correlation had a different sign than the corresponding zero-order correlation. This means that the bootstrap test rejects the UFM based on the implication of UFM that all partial correlations should be between zero and the zero-order correlation.

Because the bootstrap test is a local test of model fit, we can use the significant results in Table 7.4 to eliminate items that cause misfit. To select variables for elimination we used the selection rule that the variable that is part of the most pairs of two variables for which the correlation and partial correlation are significant in the bootstrap test is removed. For example, Table 7.4 shows that V12 is part of four pairs of variables for which the correlation has a different sign than the partial correlation ($r_{2,12}$, $r_{6,12}$, $r_{9,12}$ and $r_{11,12}$). We thus started with eliminating the variables V9 and V12 because these both were part of four pairs of variables for which the bootstrap test was significant. As a result, the number of pairs that were significant decreased from 10 to 4 pairs. V2 and V5 were both part of two pairs of variables for which the bootstrap test was significant, and thus we eliminated V2 and V5 as well. This resulted in a set of 10 variables (all variables except for V2, V5, V9 and V12) for which the bootstrap test does not reject the UFM.

7.6 DISCUSSION

In this paper we proved that the UFM implies that partial correlations between observed variables given any subset of the other observed variables are always closer to zero than the zero-order correlation between these observed variables. To the best of our knowledge this is a new result, which implies that the identification of partial correlations in the data that are significantly stronger than the zero-order correlation may cast doubt on whether the UFM is the data generating model. To facilitate the use of this result in data analysis, we presented a bootstrap test to evaluate the tenability of the implication and illustrated the use of the test in an empirical example.

Most of the fit indices for the UFM available in the literature rely on distributional assumptions or asymptotic normality (e.g., the gamut of fit measures based on normal theory LRT or approximate asymptotic distributions; see Bollen and Ting (1993) for an overview). The main result of this paper only relies on the structure of the covariance matrix, and not on properties of the statistical distribution such as the normal distribution. The proposed bootstrap method is essentially distribution free. Furthermore, as argued in Browne and Cudeck (1992), the factor model in itself is all but certain to be incorrect. Therefore, instead of focusing on the absolute fit of the factor model, it is more sensible to assess where the lack of fit occurs (Browne & Cud-

eck, 1992). Saris, Satorra, and van der Veld (2009) show that global fit indices (FIs) vary in their sensitivity to different types of misspecifications so that a model with a small substantively irrelevant misspecification may be rejected because the FI is sensitive to such misspecifications, while this FI may support the acceptance of a model with an important misspecification due to its low sensitivity to such misspecifications. Claeskens and Hjort (2008) even argue that the ‘best model’ depends on the parameter of interest and propose the use of focused information criteria tailored to parameters of interest. The inequality of the main result of this paper allows for a local fit evaluation that indicates where misfit occurs. In addition, the procedure can be used to guide the selection of items in a test: Items that are consistently involved in violations of the proposition should be considered for removal (see Clark and Watson (1995) for a discussion of item selection).

We believe that the result presented in this paper has the potential to help decide whether a UFM is a plausible candidate to have generated the observed data. However, we also think that the result is not useful to test the dimensionality of the factor model when comparing different factor models. This is because we think it is likely that the implication presented in this paper generalizes to, for example, correlated factor models or hierarchical factor models. That is, partial correlations between zero and the zero-order correlation are a necessary condition for the UFM but not a sufficient condition as this property likely holds for other factor models as well. In a correlated factor model, for example, the correlation between two observed variables that load on different factors equals the product of their factor loadings and the correlation between the factors these observed variables load on. As such, the correlation matrix and partial correlation matrix implied by such models will not only be a function of the factor loadings but also of Ψ , the correlation matrix of the latent variables. Future research might reveal how Ψ in the function for the correlation and partial correlation influences the relative size of the partial correlation with respect to the correlation. Another interesting extension would be to investigate whether the result presented in this paper generalizes to other types of latent variable models such as IRT models, in which the test scores are not continuous. Finding such extensions of the result to other latent variable models implies that this result should not be used to compare different factor models with each other within the realm of factor modeling, but rather provides a first test of whether one should enter the gateway to the factor modeling world at all. We hope that future research explores how general the class of latent variable models is for which this property holds. For now we can conclude that the identification of partial correlations that are not between zero and the zero-order correlation provide evidence against a UFM.

	V ₁	V ₂	V ₃	V ₄	V ₅	V ₆	V ₇	V ₈	V ₉	V ₁₀	V ₁₁	V ₁₂	V ₁₃	V ₁₄
V ₁	I	0.34	0.36	0.37	0.41	0.45	0.42	0.45	0.22	0.56	0.60	0.25	0.37	0.45
V ₂	0.08	I	0.16	0.3	0.13	0.33	0.16	0.27	0.24	0.26	0.28	0.04	0.40	0.35
V ₃	0.07	-0.02	I	0.26	0.26	0.40	0.22	0.21	0.14	0.38	0.30	0.13	0.17	0.28
V ₄	0.04	0.11	0.04	I	0.17	0.49	0.32	0.23	0.28	0.31	0.36	0.14	0.28	0.24
V ₅	0.07	-0.08	0.09	-0.02	I	0.24	0.23	0.57	0.06	0.43	0.33	0.27	0.20	0.31
V ₆	0.11	0.08	0.22	0.28	-0.03	I	0.31	0.35	0.37	0.37	0.41	0.10	0.31	0.31
V ₇	0.10	-0.08	0.01	0.10	-0.03	0.03	I	0.28	0.24	0.39	0.40	0.31	0.31	0.27
V ₈	0.11	0.10	-0.06	-0.02	0.46	0.1	0.05	I	0.19	0.38	0.38	0.17	0.28	0.33
V ₉	-0.04	0.06	-0.01	0.07	-0.07	0.21	0.1	0.06	I	0.15	0.28	0.01	0.24	0.18
V ₁₀	0.19	0.03	0.14	0.03	0.16	0.05	0.11	0.01	-0.04	I	0.46	0.32	0.29	0.47
V ₁₁	0.32	0.02	0.03	0.08	0.05	0.05	0.13	0.05	0.11	0.11	I	0.16	0.33	0.36
V ₁₂	0.05	-0.05	-0.01	0.04	0.13	-0.06	0.21	-0.03	-0.06	0.14	-0.06	I	0.14	0.21
V ₁₃	0.05	0.23	-0.03	0.05	0.00	0.03	0.13	0.03	0.07	-0.02	0.05	0.01	I	0.41
V ₁₄	0.11	0.14	0.07	-0.03	0.05	0.01	-0.02	0.03	0.03	0.21	0.01	0.05	0.21	I

Table 7.3: The upper triangle of this matrix represents the zero-order correlations of the Extraversion item responses. The lower triangle of this matrix represents the corresponding partial correlations. Each zero-order correlation with a gray background corresponds to a partial correlation with a gray background and the combination refers to a pair for which the bootstrap test was significant (the CI obtained with the bootstrap test did not include one).

	V ₁	V ₂	V ₃	V ₄	V ₅	V ₆	V ₇	V ₈	V ₉	V ₁₀	V ₁₁	V ₁₂	V ₁₃	V ₁₄
V ₁	-	0.31	0.39	0.57	0.49	0.38	0.33	0.34	1.03	0.20	0.00	0.35	0.54	0.36
V ₂	-	-	0.85	0.08	1.56	0.31	1.50	0.03	0.2	0.54	0.65	1.26	0.01	0.02
V ₃	-	-	-	0.44	0.11	0.00	0.61	1.18	0.68	0.09	0.57	0.62	0.91	0.29
V ₄	-	-	-	-	0.84	0.05	0.17	0.82	0.28	0.57	0.38	0.01	0.43	0.92
V ₅	-	-	-	-	-	0.88	0.99	0.53	1.76	0.14	0.51	0.00	0.69	0.48
V ₆	-	-	-	-	-	-	0.57	0.23	0.00	0.54	0.57	1.28	0.61	0.71
V ₇	-	-	-	-	-	-	-	0.44	0.00	0.29	0.21	0.21	0.00	0.86
V ₈	-	-	-	-	-	-	-	-	0.10	0.72	0.53	0.91	0.50	0.59
V ₉	-	-	-	-	-	-	-	-	-	1.01	0.04	1.67	0.18	0.38
V ₁₀	-	-	-	-	-	-	-	-	-	-	0.37	0.00	0.87	0.00
V ₁₁	-	-	-	-	-	-	-	-	-	-	-	1.22	0.47	0.70
V ₁₂	-	-	-	-	-	-	-	-	-	-	-	-	0.38	0.20
V ₁₃	-	-	-	-	-	-	-	-	-	-	-	-	-	0.00
V ₁₄	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 74: Lower bound of the CI obtained with the bootstrap test for each zero-order correlation and corresponding partial correlation between the extraversion items. The values that have a gray background refer to a combination of a zero-order correlation and partial correlation for which the CI obtained with the bootstrap test does not include one.

8

Common effect models: Positive or negative manifold?

Abstract

Observing exclusively positive associations among a set of variables (i.e., a positive manifold) is a robust finding in many areas in psychology. These positive associations can be explained by positing an underlying common cause or, alternatively, through positive direct effects among the variables. Recently, the Kruis-Maris model has been proposed as a third alternative mechanism to explain a positive manifold. In the Kruis-Maris model, observed variables are causes of a latent variable, such that conditioning on the latent variable induces associations between the observed variables. In this paper, we analyze the difference between the Kruis-Maris model and the more familiar causal indicator model. Both are examples of common effect models, but they imply different relations between the response variables and the effect variable. Whereas conditioning on the common effect in the Kruis-Maris model implies a positive manifold among the response variables, we show that conditioning on the common effect in the causal indicator model implies a *negative* manifold. We demonstrate this result for two specific cases: (1) in which both the response variables and the effect variable are normally distributed variables, and (2) in which the effect variable is normally distributed but the response variables are dichotomous variables.

This chapter is submitted for publication as Van Bork, R., Marsman, M., Rhemtulla, M., Epskamp, S., Kruis, J., & Borsboom, D. (under review). Common effect models: Positive or negative manifold? Submitted to *British Journal of Mathematical and Statistical Psychology*

8.1 INTRODUCTION

When variables are all positively associated (Holland & Rosenbaum, 1986) these variables are said to feature a positive manifold. A positive manifold is a very common observation in variables that are intended to measure the same thing. For example, the test scores on a set of items that are all intended to measure math ability typically all correlate positively with each other. Other examples of data where one can observe a positive manifold include personality data (Musek, 2007), test scores on cognitive tasks (Jensen, 1998; Spearman, 1904) and psychopathology data (Caspi et al., 2014; Lahey et al., 2012).

The positive manifold is often explained through the hypothesis that observed variables are causal effects of one or a few unobserved common causes. This is naturally compatible with factor modeling (Spearman, 1904). Another explanation that can be shown to be equivalent to the most dominant factor modelling techniques is that observed variables correlate positively because they interact with one another (Van der Maas et al., 2006). This alternative explanation has resulted in network modelling techniques in which the observed variables are nodes in a network that are connected by an edge when they are conditionally associated (Costantini et al., 2015; Epskamp, Borsboom, & Fried, 2018; Van Borkulo et al., 2014). Kruis and Maris (2016) recently introduced a third explanation: observed variables correlate positively with one another because one may (unknowingly) have selected only cases for which a common effect of the observed variables is present. We refer to this model as the Kruis-Maris model, which is a specific type of common effect model. The Kruis-Maris model as an alternative explanation for the positive manifold is remarkable because it is directly opposite in conceptualization to the common factor model in which the observed variables are functions of the latent factor instead of the other way around. Yet Kruis and Maris (2016) showed their model to be equivalent to the Item Response Theory (IRT) model, which can itself be described as a factor modeling technique for dichotomous variables (Takane & de Leeuw, 1987). In fact, the Kruis-Maris model, IRT model and Ising model (a network model for dichotomous variables) are all mutually equivalent. This equivalence is interesting because the models reflect very different explanations for the observation of solely positive correlations (i.e., dependence on a common cause, conditioning on a common effect and positive direct effects; Kruis & Maris, 2016).

In the current paper we zoom in on two types of common effect models: we compare the Kruis-Maris model to the more familiar causal indicator model (Bollen & Lennox, 1991; Edwards & Bagozzi, 2000), also known as the formative measurement model. The causal indicator model is also a common effect model, but is not equivalent to the IRT model. The goal of the

paper is to demonstrate that while the Kruis-Maris model, in which the effect is a quadratic function of the response variables, implies a *positive* manifold, the causal indicator model, in which the effect is a linear function of the response variables, implies a *negative* manifold among the response variables when conditioning on the common effect. To be precise, a causal indicator model with a single common effect implies a conditional covariance matrix given the effect that is a diagonal matrix *minus* a rank one matrix. This is the opposite structure of what is implied by a common cause model: a unidimensional factor model implies that the covariance structure is a diagonal matrix *plus* a rank one matrix.

We start by explaining how common cause models and network models provide alternative explanations for the positive manifold and explain the equivalence between the IRT model (a type of factor model) and Ising model (a type of network model). We then introduce the Kruis-Maris model as a third explanation for the positive manifold and explain how this model is equivalent to the Ising model and IRT model. Thereafter, we introduce the causal indicator model, which is a linear common effect model, and compare this model to the Kruis-Maris model, which is a quadratic common effect model. We prove that while the Kruis-Maris model, just like a common cause model, implies a positive manifold, the causal indicator model implies a negative manifold when conditioning on the common effect. We finish the paper by discussing some implications of this result.

8.2 COMMON CAUSE VS DIRECT EFFECTS AS EXPLANATIONS FOR THE POSITIVE MANIFOLD

One explanation for a positive manifold among a set of response variables is that the response variables are all effects of an underlying common cause. This explanation is consistent with reflective measurement in which a latent variable is identified by observing multiple indicators that are assumed to be effects of the latent variable, such that the shared variance among these indicators reflects their common cause (Bollen & Lennox, 1991; Edwards & Bagozzi, 2000). For example, if differences in math ability lead to differences in test scores on several math items, the shared dependence on math ability induces positive correlations among the test scores on these items. The test scores on math items can be used as *measures* of math ability because it is assumed that a higher math ability results in a higher probability of giving a correct response to each item. Since people with a better math ability have a higher probability correct for all math items, the test scores on these items will correlate positively over persons. This common cause explanation can be represented in a statistical model in which the observed variables are a

function of a latent common factor.

In this article we consider Item Response Theory (IRT) models in which the observed variables are dichotomous variables and the probability correct for each item is a logistic function of the latent variable. In factor models the observed variables are continuous variables that are a linear function of the latent variable. Any unidimensional factor model or IRT model implies that the observed variables can always be recoded such that all correlations between the observed variables become positive. The explanation for a positive manifold implied by a unidimensional factor model or IRT model, is that the common factor can be interpreted as a common cause of the observed variables. The left panel in Figure 8.1 shows an example of a common factor model with four response variables (X_1 to X_4).

Another way to explain a positive manifold among a set of variables, is to assume mutual direct effects among these variables (Van der Maas et al., 2006). For example, for a set of symptoms of major depression it may well be that the symptoms directly influence each other in such a way that having some problems (e.g., suffering from insomnia and worrying a lot), leads to more problems (e.g., concentration problems and negative thoughts; Cramer & Borsboom, 2015). To the extent that the presence of symptoms leads to the presence (rather than absence) of more symptoms, symptoms will be positively associated. This explanation for the associations between observed variables can be expressed in a network model.

Pairwise Markov random fields are undirected networks in which the edges represent conditional dependencies and the absence of an edge represents conditional independence between two nodes given all other nodes in the network (Lauritzen, 1996). The Ising model (Ising, 1925) is an undirected network with dichotomous variables and the Gaussian Graphical Model (GGM; Lauritzen, 1996) is an undirected network model with Gaussian variables. Both the Ising model and GGM satisfy the pairwise Markov property, such that nodes are connected with an edge when the corresponding variables are conditionally dependent given all other variables in the network. The right panel in Figure 8.1 shows an example of a Markov random field with four response variables (X_1 to X_4).

While providing substantively different explanations for the positive manifold, the Ising model and Multidimensional IRT model have been shown to be mathematically equivalent (Epskamp, Maris, et al., 2018; Kruis & Maris, 2016; Marsman et al., 2018, 2015). That is, for each Ising model there is some (multidimensional) IRT model that is equivalent, and what IRT model this is depends on the constraints in the Ising model. In the next section we elaborate on this equivalence.

Figure 8.1: Common factor model (a) and Markov random field (b). Squares represent observed variables and the circles represent latent variables. Model A represents a common factor model. The latent variables ε_1 to ε_4 represent the errors, so that ε_i accounts for the unique variance of X_i that is not explained by η . Model B represents a Markov random field; edges represent conditional dependencies, and missing edges represent conditional independencies given the other two variables in the network. More specifically, when the response variables are dichotomous, Model A represents a unidimensional IRT model and Model B an Ising model. When the response variables are multivariate normal, Model A represents a unidimensional factor model and Model B represents a GGM.

8.3 EQUIVALENCE OF THE ISING MODEL AND IRT MODEL

We here briefly introduce the Ising model and the unidimensional IRT model to give an intuition of why they both imply a positive manifold and how they are equivalent. Let $\mathbf{X}$ denote a vector of p dichotomous response variables of which the i th element takes on the value $X_i = 1$ or $X_i = -1$, and let $\mathbf{x}$ denote its realization. In the Ising model, the variables in $\mathbf{X}$ are a function of their direct neighbors. That is, a variable X_i is a function of all variables that are directly connected to this variable by an edge,

$$P(\mathbf{X} = \mathbf{x}) = \frac{\exp(\sum_{i=1}^p \mu_i x_i + \sum_{\langle ij \rangle} \omega_{ij} x_i x_j)}{\sum_{\mathbf{x}} \exp(\sum_{i=1}^p \mu_i x_i + \sum_{\langle ij \rangle} \omega_{ij} x_i x_j)}, \quad (8.1)$$

where ω_{ij} denotes the edge weight connecting X_i and X_j , μ_i denotes the intercept of X_i and $\sum_{\langle ij \rangle}$ takes the sum over all distinct pairs of variables X_i and X_j ($i > j$). $\sum_{\mathbf{x}}$ takes the sum over all 2^p realizations of $\mathbf{X}$ so that the denominator functions as a normalizing constant. The connectivity matrix $\mathbf{\Omega}$ is a symmetric $p \times p$ matrix consisting of all edge weights between the variables. The diagonal of $\mathbf{\Omega}$ is undefined. Any connectivity matrix with positive edge weights

implies a positive manifold between the variables in $\mathbf{X}$. An example is the Curie Weiss model (Kac, 1968), which is an Ising model with a single (positive) parameter for all edge weights (i.e., all variables are connected by an edge and all edge weights are constrained to be equal),

$$P(\mathbf{X} = \mathbf{x}) = \frac{\exp(\sum_{i=1}^p \mu_i x_i + \sum_{\langle ij \rangle} \omega x_i x_j)}{\sum_{\mathbf{x}} \exp(\sum_{i=1}^p \mu_i x_i + \sum_{\langle ij \rangle} \omega x_i x_j)}. \quad (8.2)$$

In the unidimensional IRT model, all variables in $\mathbf{X}$ are a function of a single latent variable η ,

$$P(\mathbf{X} = \mathbf{x} \mid \eta) = \prod_{i=1}^p \frac{\exp(x_i(a_i \eta + d_i))}{\sum_{x_i} \exp(a_i \eta + d_i)}, \quad (8.3)$$

where $\sum_{x_i}$ takes the sum over the two realizations of X_i . The item parameter d_i is an intercept and the item parameter a_i denotes the maximum slope of the item characteristic function of X_i and is referred to as the discrimination parameter, where more discriminating items are characterized by a steeper slope. One can assume a distribution for the latent variable such that integrating over the latent variable in Equation (8.3) to obtain the marginal probability distribution of $\mathbf{X}$ results in an Ising model with a connectivity matrix that is rank one (Marsman et al., 2015). In this case, all edge weights are a function of the p discrimination parameters. For any set of two variables for which the discrimination parameters have the same sign, the implied marginal association is then positive. This means that for the unidimensional IRT model (just like for its equivalent rank one Ising model) the response variables can always be recoded so that all marginal associations are positive (i.e., to obtain a positive manifold; Holland & Rosenbaum, 1986).

When all discrimination parameters in Equation (8.3) are constrained to be equal, the so-called Rasch model (Rasch, 1960) is obtained. The Curie Weiss model is equivalent to a marginal Rasch model (Epskamp, Maris, et al., 2018; Kruis & Maris, 2016; Marsman et al., 2018). Both the Rasch model and Curie Weiss model use $p + 1$ parameters: p intercept parameters and one parameter for the slopes (Rasch model) or for the edge weights (Curie Weiss model).

The general equivalence between Ising models and multidimensional IRT models relies on two steps. Since (1) the diagonal of the connectivity matrix of an Ising model is undefined and thus can always be chosen so that a semi-positive definite matrix is obtained and (2) any semi-positive definite matrix can be expressed as an IRT model with as many latent variables as positive eigenvalues, there is an equivalent IRT model for any Ising model and vice versa. For a more extensive explanation of this equivalence we refer the reader to earlier papers (Epskamp, Maris, et al., 2018; Marsman et al., 2015).

8.4 THE KRUIS-MARIS MODEL

In a recent paper, Kruis and Maris (2016) propose a third model that is equivalent to the Ising model and IRT model. The Kruis-Maris model reflects another alternative mechanism to explain the positive manifold, which is by conditioning on a common effect of the observed variables. Kruis and Maris (2016) show that if data are generated by an independence model (for which the state of one variable does not influence the state of another variable), but only cases are investigated for which a specific effect (a function of the observed outcomes) is present, a positive manifold is obtained. That is, conditioning on the common effect induces positive associations that can be described by an Ising model with positive edge weights.

In this section, we provide an introduction to the Kruis-Maris model. Kruis and Maris present a formula for the joint distribution of $\mathbf{X}$, and a binary effect variable, E : $p(\mathbf{X} = \mathbf{x}, E = e)$. Here e denotes the realization of E , $e = 1$ indicating that the effect is present and $e = 0$ indicating that the effect is absent. The chain rule for probability gives us that the joint probability $p(\mathbf{X} = \mathbf{x}, E = 1)$ equals $p(\mathbf{X} = \mathbf{x}) \times p(E = 1 | \mathbf{X} = \mathbf{x})$. The Kruis-Maris model consists of these two probability distributions. We discuss each factor separately.

The first factor in the formula,

$$p(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^p \frac{\exp(\mu_i x_i)}{\sum_{x_i} \exp(\mu_i x_i)}, \quad (8.4)$$

represents the marginal probability distribution of $\mathbf{X}$, in which μ_i denotes the intercept ranging from $-\infty$ to $+\infty$. This factor only includes intercepts and no interactions between the variables, so $\mathbf{X}$ is marginally independent. When μ_i is a large positive number then $x_i = 1$ results in a probability close to one and $x_i = -1$ results in a probability close to zero. When μ_i is a large negative number then $x_i = -1$ results in a probability close to one and $x_i = +1$ results in a probability close to zero. When μ_i equals zero, the probability for both possible values of x_i result in a probability of 0.5. The marginal probability distribution described above forms an empty Ising model in which all edge weights equal zero.

The second factor in the formula,

$$p(E = 1 | \mathbf{X} = \mathbf{x}) = \frac{\exp(\frac{1}{2}\omega[\sum_{i=1}^p x_i]^2)}{\exp(\frac{1}{2}\omega p^2)}, \quad (8.5)$$

implies that the probability of the effect being present depends on whether the variables in $\mathbf{X}$ are in the same state or in opposite states. The denominator functions as a normalizing constant so that when all variables are in the same state, the numerator equals the denominator and the

probability for the effect being present equals 1. Variables in the same state increase the probability of the effect being present while variables in opposite states decrease the probability of the effect being present. In the extreme, both response patterns in which all responses equal +1 as well as response patterns in which all responses equal -1 result in a probability 1 for the effect being present, whereas a response pattern in which half of the responses equal -1 and half of the responses equal +1, results in a probability of 1 for the effect being absent. Conditioning on the effect being present implies that data patterns in which the response variables are in the same state have a high probability, thus inducing positive connections between the response variables. The conditional probability distribution of the response variables given that the effect is present is therefore similar to the probability distribution of the response variables implied by a Curie Weiss model in which the response variables also have thresholds and also tend to be in the same state.

The fact that the effect variable here depends on whether variables are in the same state or in opposite states is the result of the square in the formula which results in a quadratic relation between the effect and the response variables. If the function in Equation (8.5) would be a linear function, the probability of the effect would increase as more responses equalled +1, and decrease as more responses equalled -1. Kruis and Maris (2016) show that the unidimensional Kruis-Maris model as presented above is equivalent to a Curie Weiss model, and the Ising model is equivalent to a Kruis-Maris model with multiple orthogonal effect variables.

8.5 KRUIS-MARIS MODELS VS CAUSAL INDICATOR MODELS

In the structural equation modeling literature, psychometric models in which the latent variable is a common effect of the observed variables are not new (Bollen & Bauldry, 2011; Bollen & Lennox, 1991; Edwards & Bagozzi, 2000; MacCallum & Browne, 1993). In the models described in this literature, the common effect variable η is a linear function of the response variables:

$$\eta = \mathbf{x}^T \boldsymbol{\gamma} + \zeta = \gamma_1 x_1 + \gamma_2 x_2 + \dots + \gamma_p x_p + \zeta. \quad (8.6)$$

In this equation ζ denotes the disturbance on the latent variable η , that is unexplained by the response variables $\mathbf{X}$. This model is typically referred to as a causal indicator model, since the response variables are indicators that cause the latent variable, although it is also known as the formative model, reflecting the idea that the latent variable is “formed” by the observed variables (Edwards & Bagozzi, 2000). An example with four observed variables is given in Figure

8.2. While the notion of common effect models is thus well established within psychology, it is interesting to notice that the models from this tradition are different from the Kruis-Maris model. We use this section to compare these two common effect model frameworks.

Although both the Kruis-Maris model and the causal indicator model include a latent variable that is a function of a set of observed variables, the relation between the effect variable and the response variables in the Kruis-Maris model described earlier is very different from the relation between the causal indicators and the latent common effect in the causal indicator model. Here we explain how these models differ and show that the models imply different conditional associations between the response variables. In the literature on causal indicator models both the latent effect and the observed causes are typically assumed to be Gaussian variables, although sometimes also categorical causal indicators are included (Edwards & Bagozzi, 2000; MacCallum & Browne, 1993). The latent effect is a linear function of the causal indicators, which is different from the Kruis-Maris model in which the presence of the effect is primarily dependent on whether the causes are in the same state or in opposite states from each other.

Conditioning on the effect in the Kruis-Maris model induces positive associations between the response variables because the effect being present implies that the variables are likely to be in the same state. Conditioning on the common effect in a causal indicator model, however, induces negative associations. If two variables X_1 and X_2 are marginally independent but both positively affect a third variable X_3 , conditioning on X_3 induces a negative dependency between X_1 and X_2 (Elwert & Winship, 2014; Pearl, 2000a).

As an example of the linear causal indicator model, suppose that ‘complexity of the topic’ (X_1) as well as ‘perfectionism of the student’ (X_2) both positively predict the time it takes for a student to finish an assignment (X_3). Further, suppose that the perfectionism of the student is not related to what topic is assigned to this student so that X_1 and X_2 are marginally independent. The observation that a student handed in the assignment quickly induces a dependency between the complexity of the topic and the perfectionism of this student: observing that this student actually had a complex topic makes it less likely that this student is perfectionistic, and observing that the student is perfectionistic leads to the inference that the topic must have been rather simple, given the small amount of time that the student put in the assignment.

That conditioning on a common effect induces bias is well-known. This bias is relevant in several commonly occurring situations that involve conditioning on a common effect. For example, when one studies relations between cognitive tasks in a sample in which only people with an IQ score higher than some cut-off value are selected, or when one studies the relations between symptoms in a sample in which only people with a diagnosis are selected. In these cases

one conditions on a common effect, and because in these examples the variables that are being considered all positively affect the common effect, this results in spurious negative relations between the observed variables that bias the existing correlations between these variables.

The above examples comply with the causal indicator model rather than with the Kruis-Maris model. This is the case because in these examples the effect variable is a linear function of the variables in $\mathbf{X}$, instead of being dependent on whether the variables are in similar or opposite states as is the case in the Kruis-Maris model. This difference between the common effect models becomes even more clear from taking a closer look at the implications of the Kruis-Maris model for one of these examples. Let the dichotomous causes X_1 to X_p represent the presence/absence of symptoms and let the dichotomous effect variable represent whether someone is diagnosed. A linear common effect model would imply that the probability of a diagnosis increases as the number of symptoms increases.

The Kruis-Maris model, in contrast, implies that individuals for whom all symptoms are present have the same probability of being diagnosed as individuals for whom none of the symptoms are present, because for both these groups all symptoms are in the same state (all present or all absent). There may be causes and effects that follow the implications of the Kruis-Maris model, but for many examples in psychology a linear relation between the causes and effect seems more plausible. Because many of the examples of selection bias encountered in psychological research concern a linear relation between the causes and the effect (e.g., selecting people with a high sum score), we take the next section to formulate more formally what the effect is of conditioning on a common effect in the causal indicator model.

8.6 CAUSAL INDICATOR MODELS IMPLY A NEGATIVE MANIFOLD WHEN CONDITIONING ON THE COMMON EFFECT

In the previous section we explained that conditioning on a common effect that linearly depends on two marginally independent variables induces a negative association between the two causes (Pearl, 1988). We show here that the causal indicator model with a single normally distributed common effect variable that is a linear function of its causes, implies that all conditional covariances among the causes given the effect are negative. Just like Kruis and Maris (2016), we assume marginally independent causes. We conjecture that this result holds for all models in which the effect is a linear function of the predictor variables, but we here show it for two specific instances. We show that the conditional associations between the predictor variables given the effect variable are all negative for (1) the case in which both the predictor variables and the effect variable

Figure 8.2: Causal indicator model. The latent common effect η is a linear function of the four observed variables $X_1 - X_4$. The double headed arrows that connect the observed variables represent the correlations between these variables.

are normally distributed, and (2) the case in which the predictor variables are dichotomous variables and the effect variable is a normally distributed variable. We call the observation of solely negative associations a *negative manifold*.

Let the common effect variable η be a linear function of a vector $\mathbf{X}$ of p random variables X_i , as in Equation (8.6):

$$\eta = \mathbf{x}^T \boldsymbol{\gamma} + \zeta,$$

where $\mathbf{x}$ is a vector of realizations of $\mathbf{X}$, $\boldsymbol{\gamma} \in \mathbb{R}^p$ is a vector of regression coefficients, and $\zeta \sim \mathcal{N}(0, \sigma^2)$. That is, the conditional distribution $p(\eta \mid \mathbf{x})$ is normal with mean $\mathbf{x}^T \boldsymbol{\gamma}$ and variance σ^2 .

Assuming that the predictor variables are independent of each other but all predict a single common effect, we conjecture that whenever the γ parameters are either all negative or are all positive, negative associations arise between the predictor variables $\mathbf{X}$ when conditioning on η . More specifically, the matrix of associations between the variables in $\mathbf{X}$ when conditioning on η is a diagonal matrix *minus* a rank one matrix in the case of normally distributed response variables or a zero matrix *minus* a rank one matrix with an undefined diagonal in the case of dichotomous response variables. Because the matrix that is subtracted is of rank one, the predictor variables can always be recoded such that a negative manifold arises, regardless of the sign of the γ parameters. We prove this for two particular cases, which we state as Theorems 1 and 2. In Theorem 1 we consider the case where the predictors X_i are independent normal random

variables.

Theorem 1. *If the variable η is defined as in Equation (8.6) using p independent normal predictor variables*

$$X_i \sim \mathcal{N}(\mu_i, \tau_i^2),$$

and the γ parameters all have the same sign, the conditional distribution $p(\mathbf{X} \mid \eta)$ implies a negative manifold.

The proof of Theorem 1 is given in Appendix C. Theorem 2 mirrors the results for Theorem 1 for the particular case of dichotomous predictor variables.

Theorem 2. *If the variable η is defined as in Equation (8.6) using p independent dichotomous predictor variables that have the distribution as in Equation (8.4), and the γ parameters all have the same sign, the conditional distribution $p(\mathbf{X} \mid \eta)$ implies a negative manifold.*

The proof of Theorem 2 is given in Appendix C, where it is shown that the conditional distribution $p(\mathbf{X} \mid \eta)$ is an Ising model with a fully negative connectivity matrix.

More precisely, the proof for Theorem 1 in Appendix C demonstrates that conditioning on the common effect in a causal indicator model implies a diagonal matrix *minus* a rank one matrix. The proof for Theorem 2 in Appendix C demonstrates that when the variables in $\mathbf{X}$ are dichotomous variables, conditioning on the common effect in a causal indicator model results in an Ising model with a connectivity matrix that is a zero matrix *minus* a rank one matrix. Note that the diagonal of this connectivity matrix is arbitrary as it drops out.

8.7 POSITIVE VERSUS NEGATIVE MANIFOLD

In the previous section we presented a theorem stating that conditioning on a common effect in a causal indicator model induces conditional associations between the marginally independent observed variables that comply to a diagonal matrix (or zero matrix for dichotomous response variables) minus a rank one matrix. Because the matrix that is subtracted from the diagonal matrix is of rank one, this implies that the observed variables can always be recoded so that this matrix is fully positive and hence subtracting this matrix from a diagonal matrix results in a matrix with solely negative off-diagonal elements. The associations implied by conditioning on the common effect thus feature a negative manifold. This is the opposite pattern of what is implied by a single common cause: the positive manifold.

A single common cause model implies that the covariance pattern among the observed variables is a diagonal matrix *plus* a rank one matrix. The common cause and common effect thus

Figure 8.3: Implied associations of a common cause model and causal indicator model. The associations between the response variables X_1 to X_4 (included in the matrix Σ) that result from the common cause model (model on the left) have the opposite sign of the conditional associations (included in the matrix $\Sigma | \eta$) that result from conditioning on the common effect in the causal indicator model (model on the right). Conditioning on a latent variable is visualized by drawing the circle inside a square.

have very different implications for empirical data. When a single common cause underlies the data one can recode the items such that all factor loadings are positive and as such that all covariances among the observed variables are positive. In contrast, when the associations are the result of conditioning on a single common effect in a causal indicator model, one can always recode the items such that all γ parameters are positive and as such that all conditional covariances among the observed variables are negative.

The possible covariance structures that can be explained by a common cause, and those that can be explained by conditioning on a common effect are symmetric in the sense that for any covariance matrix that can be explained by a single common cause, the off-diagonal elements can be multiplied by -1 to obtain a matrix that is consistent with a causal indicator model, and vice versa. Figure 8.3 depicts an example of this symmetry for four response variables.

Three mutually exclusive groups of correlation matrices can be distinguished: those for which the pattern of positive versus negative correlations is consistent with a common cause model, those for which it is consistent with a causal indicator model, and those which are inconsistent

with both models. An example of a pattern of positive and negative correlations that is inconsistent with both the common cause model and the causal indicator model is a covariance matrix for $p > 2$ variables with only one negative covariance. For such a covariance structure it is impossible to recode the variables such that either a positive or negative manifold is obtained.

8.8 DISCUSSION

We compared the Kruis-Maris model (Kruis & Maris, 2016) to the more familiar causal indicator model (Edwards & Bagozzi, 2000). The models imply very different relations between the effect variable and the response variables. In the Kruis-Maris model the relationship between the effect variable and response variables is a quadratic relationship; the effect variable has a higher probability of being present when more response variables are in the same state. In the causal indicator model the effect variable is a linear function of the response variables. The Kruis-Maris model and causal indicator model also differ in what they imply for the conditional associations among the response variables. Whereas the Kruis-Maris model implies positive associations between variables because the effect variable has a higher probability of being present when the response variables are in the same state, the causal indicator model implies negative associations between variables that positively affect the common effect.

In the last part of the paper we proved that a causal indicator model with a single latent common effect implies solely negative associations, which we dubbed *a negative manifold*, between the response variables conditioned on the common effect. This result reveals a symmetry in the association patterns that can result from a common cause and from conditioning on a common effect; when the response variables are effects of a single common cause the response variables can always be recoded such that a positive manifold is obtained. In contrast, when the associations between response variables are induced by conditioning on the common effect in a causal indicator model, the variables can always be recoded such that all associations are negative.

The result of a negative manifold was derived under the assumption of marginally independent response variables. This assumption is highly unlikely for psychological variables; it is a far more plausible scenario that the response variables are marginally associated (for example due to direct effects) and that conditioning on a common effect induces bias to these associations rather than turning the covariance matrix into a fully negative matrix. The phenomenon that conditioning on a common effect may induce spurious associations and as such bias the covariance structure is referred to in the literature as collider bias or endogenous selection bias (Elwert & Winship, 2014; Greenland, 2003; Pearce & Richiardi, 2014). The proof of Theorem

1 shows that the conditional distribution of $\mathbf{X}$ given η is a diagonal matrix minus a rank one matrix, in which the diagonal matrix represents the marginal covariance matrix of the variables in $\mathbf{X}$. Thus, when the response variables are not marginally independent, this matrix will no longer be a diagonal matrix, and thus the rank one matrix is subtracted from whatever marginal covariance structure exists between the response variables. For example, if the response variables feature strong positive marginal correlations, then conditioning on a common effect will shrink these correlations since a rank one matrix is subtracted from the fully positive matrix. The proof in Theorem 2 for dichotomous variables can be extended in a similar way. Future research may reveal whether the specific covariance structure that is induced by conditioning on a common effect as demonstrated in this paper (i.e., subtracting a rank one matrix from the marginal associations), can help identify selection bias and correct for it.

To conclude with an answer to the question raised in the title of this paper: a common effect model implies a positive manifold when the relation between the response variables and the common effect is quadratic, but implies a negative manifold when the relation between the common effect and response variables is linear.

9

What is p ?

Some risks of general factor modeling

Abstract

Recent research has suggested that a range of psychological disorders may stem from a single underlying common factor, which has been dubbed the p -factor. This finding may spur a line of research in psychopathology very similar to the history of factor modeling in intelligence and, more recently, personality research, in which similar general factors have been proposed. We point out some of the risks of modeling and interpreting general factors, derived from the fields of intelligence and personality research. We argue that: (a) factor-analytic resolution, i.e., convergence of the literature on a particular factor structure, should not be expected in the presence of multiple highly similar models; and (b) the true underlying model may not be a factor model at all, because alternative explanations can account for the correlational structure of psychopathology.

9.1 INTRODUCTION

In recent years, general factor models have been proposed to explain the correlational structure of psychopathological data. This line of research advocates the hypothesis that a wide range

This chapter is published as Van Bork, R., Epskamp, S., Rhemtulla, M., Borsboom, D., & van der Maas, H. L. J. (2017). What is the p -factor of psychopathology? Some risks of general factor modeling. *Theory & Psychology*, 27, 759–773

of symptoms that arise in psychopathology and personality disorders are determined by the pervasive influence of a single latent variable. In particular, Caspi et al. (2014; and replicated by Laceulle, Vollebergh, & Ormel, 2015) proposed that the correlational structure of psychopathology may be explained by a general factor that underlies a wide variety of disorders. Caspi et al. refer to this general factor as the *p*-factor or *p*. Sharp et al. (2015) found similar results in a wide set of personality disorders, and argued in a parallel fashion that a general factor may underlie personality pathology; in fact, Sharp et al. suggested that their general factor might be the same as the *p* identified by Caspi et al.

These research efforts are based on a methodology and line of reasoning that will sound familiar to researchers in the domains of intelligence and personality research, where models involving a general factor have long been proposed to explain the correlational structure of cognitive and personality tests, respectively (Musek, 2007; Rushton & Irwing, 2008; Spearman, 1904). In both fields, these factors (“*g*” and the “Big One” or General Factor of Personality [GFP], respectively) have been interpreted as reified causal entities that exist independent of the data and, by causing variation in their indicators, give rise to an observed correlational structure (Gould, 1996). This interpretation of general factors has led researchers to search for genetic or biological properties that may play such a causal role (Detterman, 2002; Musek, 2007), a search that is now beginning to take hold in psychopathology (Pettersson, Larsson, & Lichtenstein, 2016). So far, this search has yielded no clear candidates that might instantiate these general factors in the brain.

The goal of research investigating the structure of psychopathology is to come to understand why disorders arise, and thus how to prevent and cure them. Charting the structure of psychopathology is thus a vitally important endeavor, so it is crucial to point out misunderstandings that arise when general factors are over-interpreted. Although the present paper mainly focuses on the analyses performed by Caspi et al. (2014), the arguments apply to a broader line of research that started with Simms, Grös, Watson, and O’hara (2008) and has included multiple research groups since then (Laceulle et al., 2015; Lahey et al., 2012; Patalay et al., 2015; Snyder, Young, & Hankin, 2017). All these researchers fit bi-factor models to psychopathology data to demonstrate support for a general factor. Here, we argue that: (a) factor-analytic resolution, that is, convergence of the literature on a particular factor structure, is unlikely to arise in the presence of multiple nearly-equivalent models; and (b) the true underlying model may not be a factor model at all, because alternative explanations can account for the correlational structure of psychopathology. We follow this argument with several suggestions for future research on the structure of psychopathology.

9.2 THE PROBLEM OF NEARLY-EQUIVALENT MODELS

In advocating for a general factor, whether it is psychopathology, personality, or intelligence, most research relies on comparing models that are nearly indistinguishable at a statistical level, but radically different in terms of interpretation. As a result, small differences in data due to sampling variability result in very different theories on the structure of psychological traits and abilities.

9.2.1 NEAR-EQUIVALENCE OF BI-FACTOR AND HIGHER-ORDER MODELS

The bi-factor model of psychopathology is a special case of a hierarchical factor model¹, in which all indicators load on a general factor, and several specific factors account for the remaining shared variance among subsets of items. Caspi et al. found that a bi-factor model and a correlated three-factor model “fit [their] data similarly well, with [the bifactor model] offering a slightly more parsimonious solution” (Caspi et al., 2014, p.126)², leading them to conclude that the general factor, p , is a “single dimension that represents the tendency to experience psychiatric problems as persistent and comorbid” (Caspi et al., 2014, p.131). The correlated three-factor model is equivalent to a higher-order factor model in which subsets of indicators load onto lower-order factors that, in turn, load onto a general factor. As such, the model comparisons reported in Caspi et al. are equivalent to those that would have resulted from comparing their bi-factor model with a higher-order model.

Model fit in structural equation modeling (SEM) is measured by assessing how similar the observed variance-covariance matrix is to the one implied by a model. Exactly equivalent models lead to the exact same variance-covariance matrix, so fitting two equivalent models to any sample dataset will result in the same fit. Equivalent models cannot be distinguished on statistical grounds, even though they may imply very different interpretations. Nearly equivalent models imply very similar variance-covariance matrices. While, in theory, these models can be

¹Hierarchical factor models are models with multiple layers of orthogonal factors and simple structure within each layer (Yung, Thissen, & McLeod, 1999). The bi-factor model is a hierarchical factor model with only two layers, of which one layer consists of a general factor on which all indicators load.

²Caspi et al. (2014) reported that the correlated three-factor model fit was $X^2(df = 1018) = 1737.159$, CFI = .962, TLI = .958, RMSEA = .027[.024, .029], while their bi-factor model fit was $X^2(df = 1012) = 1652.586$, CFI = .966, TLI = .963, RMSEA = .025[.023, .027]. We note that in terms of degrees of freedom, the bi-factor model appears to be the less parsimonious model.

distinguished statistically, the outcome is highly subject to sampling variability. For some nearly equivalent models, standard statistical fit indices such as the Akaike information criterion (AIC), Bayesian information criterion (BIC), and root mean square error of approximation (RMSEA) frequently fail to identify the true data-generating model. As we will show, the higher-order factor model and bi-factor model are nearly equivalent. Statistical comparisons of these two models lead to biased conclusions (Morgan, Hodge, Wells, & Watkins, 2015; Murray & Johnson, 2013). Higher-order and hierarchical factor models imply very similar covariance matrices because they are mathematically closely related to each other. In fact, the higher-order factor model is equivalent to a special case of the hierarchical factor model (i.e., the class of higher-order factor models is nested in the class of hierarchical factor models, see Yung et al., 1999). Morgan et al. (2015) demonstrated an extensive overlap of fit values when comparing bi-factor models to higher-order factor models and correlated factor models. They even showed that when cases are selected from a population that was generated by a higher-order structure, approximate fit indices tend to incorrectly identify the bi-factor model as the best fitting model. Likewise, Murray and Johnson (2013) noted that studies of cognitive ability tend to find that bi-factor models fit better. They conducted a simulation study to evaluate that tendency, and found a substantial bias favoring the bi-factor model, even when the true structure underlying the data was a higher-order factor model. These and similar simulations have led multiple researchers to argue that, when comparing higher-order factor models and bi-factor models, the decision about which model to adopt should not rely on model fit (Bonifay, Lane, & Reise, 2017; Murray & Johnson, 2013).

9.2.2 THEORETICAL INTERPRETATION OF BI-FACTOR AND HIGHER-ORDER MODELS

While the higher-order and bi-factor model differ only slightly in fit, their theoretical implications are very different. Figure 9.1 (a) shows a higher-order model, in which some general factor, “*g*” (e.g., the *p*-factor), explains the shared variance of the lower-order factors (e.g., Internalizing, Externalizing, and Thought disorder). The lower-order factors, in turn, represent the entirety of the variance that is shared among their indicators. Figure 9.1 (b) shows the bi-factor model, in which the specific factors, Internalizing, Externalizing, and Thought disorder, are no longer caused by the general factor but in contrast are orthogonal to the general factor. In this model, the specific factors explain the shared variance among items that remains *after* the effect of the general factor is partialled out. Thus, the specific factors in the two models account for very

Figure 9.1: (a) Higher-order factor model, in which g represents a general factor; (b) bi-factor model: a specific kind of Hierarchical factor model (Yung et al., 1999), in which g represents a general factor.

different parts of the variance shared among indicators.

Despite this difference in the referent of the specific factors, researchers tend to give them the same labels in both models (e.g., “Internalizing”, “Externalizing”, and “Thought”), suggesting that these specific factors are mapped onto the same psychological constructs. When specific factors are given the same meaning, the general factor means something different as a result. Whereas in the bi-factor model, variance in p is hypothesized to be independent of Internalizing, Externalizing, and Thought disorders; in the higher-order factor model p is hypothesized to *explain* the correlations between Internalizing, Externalizing, and Thought disorders. Moreover, interpreting specific factors as referring to the same constructs in both models leads to different theories of how these constructs relate to each other, as well as to external variables. Thus, although both models contain a “general factor”, the posited mechanisms by which these respective factors relate to other factors in the model (and to external variables) are radically different.

9.2.3 IMPLICATIONS FOR FACTOR-ANALYTIC RESOLUTION

In the preceding sections, we argued that: (a) the bi-factor and higher-order factor models cannot be reliably distinguished on the basis of statistical fit due to their near equivalence and (b) that despite their statistical near-equivalence, their interpretations diverge strongly. Thus, small differences due to sampling variation can lead to completely different theories of what the general factor *is*. Because different models will fit best in different studies, it seems unlikely that the field will arrive at a consensus on which model best reflects the structure of psychopathology.

This lack of consensus is evident in the fields of intelligence and personality research. In the intelligence literature, many studies have been dedicated to finding the “true” structure of general intelligence. Various models have been proposed to explain the structure of correlations found in cognitive ability data. Some of these modeling efforts result in a single general intelligence factor (“*g*”; Gustafsson, 1984), others have several correlated higher-order factors (Flanagan & McGrew, 1998; Lansman, Donaldson, Hunt, & Yantis, 1982), and some feature a bi-factor model structure (Gignac & Watkins, 2013). Although all of these purport to have found the best model of intelligence, the literature has not converged upon a single best model. When the structure of intelligence is modeled as a higher-order model, *g* is conceptualized as a factor that is superordinate to the lower-order factors. By contrast, proponents of the bi-factor model for the structure of intelligence argue that *g* should be conceptualized as a breadth factor that is defined by a larger number of observed variables than the narrow group factors (Gignac, 2008).

In personality psychology, a similar line of research has been pursued; competing factor models ranging from five correlated factors to one general personality factor (“the Big One” or “GFP”), as well as variations on these models, have been purported to explain the correlations between personality test responses (e.g., Digman, 1997; McCrae & Costa, 1987; Musek, 2007; Rushton & Irwing, 2008). This literature has also not converged upon a single best model. Taken together, one should not expect factor-analytic resolution to arise when multiple similar models are fitted to correlational data, due to the highly similar covariance matrices such models imply. Although the issue of equivalent and nearly equivalent models in latent variable models has been recognized for decades (Duncan, 1975; Raykov & Marcoulides, 2001), it is still not widely understood. As we noted above, every paper that has suggested a general factor of psychopathology has based their conclusions on a statistical comparison of nearly equivalent models. While the correlated factor and bi-factor models reflect very different theories on the structure of psychopathology, the models are statistically too similar to rely on model fit to de-

cide which model to adopt.

9.3 THE RISK OF AFFIRMING THE CONSEQUENT

Factor modeling allows researchers to test causal assumptions in the model based on the logic that if the model does not fit the data, some causal assumptions implied by the model must not hold (Bollen & Pearl, 2013). Including a general factor in the model implies the causal assumption that the shared variance among the items is due to a common cause. The logic of model testing has led many researchers to reify a general factor upon finding that a general factor model fits the data. In the following section, we argue that the existence of a general factor is not tested against the data because any dataset that features a positive manifold will necessarily support a general factor model, whether or not a general factor underlies the data.

9.3.1 THE SEARCH FOR GENERAL FACTORS

After the *g*-factor was proposed, many researchers tried to find a neural basis for general intelligence based on the belief that *g* is real and that, if it is real, it must have a biological basis (e.g., Detterman, 2002; Garlick, 2002; Gray & Thompson, 2004). Many cognitive abilities have been proposed to explain individual differences in *g*, including speed and efficiency of processing, working memory, and the capacity to deduce relationships (Detterman, 2002), as well as biological variables such as cortical thickness, the size of specific brain regions, or the overall grey matter in these brain regions (Deary, Penke, & Johnson, 2010). Data from twin and family studies have consistently provided support for the hypothesis that general intelligence is heritable and therefore must have a genetic component (Devlin, Daniels, & Roeder, 1997; Haworth et al., 2010; Jacobs et al., 2001). Alternative models that do not introduce a general factor to explain the covariation among cognitive abilities, however, do not preclude the finding that the shared variance among these cognitive abilities is correlated within twins. After all, as long as the underlying structure giving rise to these cognitive abilities is to some extent heritable, one will find that a general factor that comprises the shared variance among these cognitive abilities correlates within twins. Despite the consistent support for the heritability of general intelligence in twin studies, not a single gene has been identified to be reliably associated with general intelligence (Payton, 2009). In personality research, Musek (2007) introduced the Big One and proposed that genes responsible for biological systems may give rise to this general factor of personality. Since then, the Big One has been re-labeled as the General Factor of Personality (GFP) by Rushon, Bons, and Hur (2008), and a similar search as in the *g*-factor literature has followed for

the genetic underpinnings of the GFP (Loehlin & Martin, 2011; Rushton et al., 2008; Veselka, Schermer, Petrides, & Vernon, 2009). In neither field have researchers succeeded in empirically identifying the abstract factors of factor analysis anywhere other than in factor analysis itself.

As in the *g*-factor and GFP literatures, the nascent *p*-factor literature exhibits a similar tendency to reify the general factor. For example, Sharp et al. (2015) found support for a general factor underlying a small set of personality disorders, and they suggested this factor might be the *same factor* as the broader *p*-factor identified by Caspi et al. (2014). This suggestion only makes sense if one interprets *p* as an entity that is external to the statistical model: If these factors do not refer to entities outside the statistical model, then they cannot refer to the *same* entity. Caspi et al. conclude that *p* “reflects meaningful differences between persons on a single dimension that represents the tendency to experience psychiatric problems as persistent and comorbid” (Caspi et al., 2014, p. 131). They finish their paper by stating that “at a minimum, researchers should no longer assume a specific relation between the disorder they study and a biomarker/cause/consequence/treatment without empirical verification. Rather, [their] finding suggests the default assumption must be that biomarkers/causes/consequences/treatments relate first to *p*” (Caspi et al., 2014, p. 134). Clearly, the suggestion that *p* mediates effects of biological variables on specific disorders does not make sense if *p* is just the shared variance among the disorders.

Just like for *g* and GFP, a similar search for a neural basis for *p* has begun in psychopathology (Lahey, Van Hulle, Singh, Waldman, & Rathouz, 2011; Pettersson, Anckarsäter, Gillberg, & Lichtenstein, 2013; Pettersson et al., 2016). As with *g*, twin studies suggest that *p* is heritable, but a similar argument for *g* can be made: whatever structure explains the correlations between disorders in psychopathology, to the extent that this structure is heritable, it will manifest as a correlation across twins.

9.3.2 GENERAL FACTOR MODEL RELIES ON POSITIVE MANIFOLD

The previous section showed that the statistical support for a *p*-factor has resulted in speculation about what the external referent of *p* would be, and the search for a genetic basis of *p*. Finding a fitting general factor model, however, does not provide support for the existence of a general factor that goes beyond the observation of a positive manifold in the data. To see why this is the case, it is important to observe that *any* general factor model relies statistically on the existence of a positive manifold, and that there are alternative explanations for a positive manifold that do not rely on the logic of factor modeling. A positive manifold simply means that all vari-

ables are positively correlated with each other (or become positively correlated after variables are appropriately re-coded). Given any such positive correlation matrix, all factor loadings and all covariances between factors in a simple structure factor analysis model (i.e., one in which indicators only load onto one latent factor) will be positive (Krijnen, 2004). Because all correlations between factors are positive, the same step can be repeated to model these factors with higher-order factors, again resulting in only positive factor loadings and factor correlations; eventually, therefore, a single general factor will *always* be found at some stratum of the factor hierarchy (Carroll, 1993). In sum, it is a mathematical necessity that whenever there is a positive manifold, factor analysis will result in a general factor.

9.3.3 EXPLANATIONS FOR THE POSITIVE MANIFOLD

Now, *if* the data are caused at some level by a single underlying common cause, the variance-covariance matrix will feature a positive manifold. But the reverse argument is not valid: finding a positive manifold does not entail a common cause. There are many alternative explanations for the existence of a positive manifold, several of which have been proposed in the literature on general intelligence (e.g., Van der Maas et al., 2006). Because alternative explanations exist, finding that a general factor model fits does not provide additional evidence that a general factor exists, over and above the initial observation of a positive manifold.

One such alternative explanation for the positive intercorrelations between cognitive tests is *sampling theory* (Bartholomew, Deary, & Lawn, 2009; Thomson, 1950; Thorndike, 1927). In sampling theory it is hypothesized that cognitive tasks require the use of multiple independent components or elements of the mind (so-called bonds)³, such that each task is a multidimensional measure. Each cognitive test measures a group of these independent bonds, but the groups of bonds will overlap across tests. As a result, the cognitive tasks that are used to measure different cognitive abilities will be positively correlated because they draw on overlapping bonds, resulting in a positive manifold. Sampling theory is likely to explain at least some of the positive correlations among mental disorders, which show a clear pattern of overlap in the symptoms that are used to diagnose them. For example, Major Depressive Episode and Generalized Anxiety Disorder each feature insomnia, fatigue, concentration problems, and psychomotor agitation as diagnostic criteria (American Psychiatric Association, 2013). Such patterns of overlap are present throughout the realm of psychopathology. Unsurprisingly, these patterns of over-

³Thomson does not specify these bonds any further but suggests that these bonds can be “identified by different readers with different entities” (Thomson, 1950, p. 307).

lap explain part of the correlation structure between disorders; for example, Borsboom (2002) reported a correlation of .62 between the number of overlapping symptoms for any two disorders, and the empirical correlation found between them in empirical studies of comorbidity.

Van der Maas et al. (2006) offer a second alternative explanation to the positive manifold in cognitive test scores, based on the biological concept of mutualism. In this perspective, cognitive ability is modeled analogously to the way ecosystems of lakes are modeled in aquatic ecology. Here, researchers seek to explain why some lakes flourish better than others, with a wider variety of life and higher water quality. Measuring variables such as the variety of life and quality of water in multiple lakes will likely lead to a collection of positive correlations – a positive manifold; lakes that have better water quality will have a wider variety of life. Biologists do not model such systems as a common cause structure in which a single general factor of “lake health” is used to explain positive correlations, even though, as explained above, a general factor would no doubt emerge from such data. Instead, biologists rely on a more plausible explanation of such data: having high water quality allows for more variety of life, and more life will improve the quality of the water (Scheffer et al., 2009; Scheffer, Carpenter, Foley, Folke, & Walker, 2001). This concept is termed *mutualism* because it describes a system of mutually beneficial relationships between the causal factors that determine its dynamics. Cognitive systems could be viewed in a similar way; strength in one cognitive system (e.g., reading comprehension) improves one’s ability in another cognitive system (e.g., reasoning). Van der Maas et al. (2006) showed analytically as well as in a simulation study that such a system can lead to a positive manifold, even though no general factor is present.

The analogy to cognitive systems can be extended to systems of psychopathology as well (Borsboom, 2008; Cramer et al., 2010; Schmittmann et al., 2013). Rather than modeling symptoms as consequences of a latent common disease, symptoms can be seen to influence each other; someone might not be able to concentrate as well because he or she is worrying a lot, and that person might worry a lot because a poor concentration lead to problems at work (Borsboom & Cramer, 2013). The term mutualism may be less suitable for psychopathology, because many causal relationships might only be one-directional – insomnia leads to fatigue but fatigue might not lead to insomnia – but the overall interpretation is that problems lead to more problems (e.g., insomnia → concentration problems), and problems hardly ever solve other problems (e.g., feeling depressed !→ fewer panic attacks). These positive causal relations between symptoms may give rise to a positive manifold. Following this line of thought, psychopathological symptoms are modeled as active agents in networks of interacting components rather than passive indicators of latent variables.

Figure 9.2: Three different models that can be equivalent to one another. (a) Bi-factor model; (b) network model; (c) Residual Network model.

In these networks, typically observed variables are represented by nodes that are connected by edges (links) when two variables are conditionally dependent given other nodes in the network. When the variables are normally distributed, these conditional dependencies are reflected in the partial correlations between the variables. A network of such partial correlation coefficients is termed a Gaussian graphical model (GGM), and it has been shown that any SEM model can be characterized by an equivalent GGM model (Epskamp, Rhemtulla, & Borsboom, 2017). Thus, for any bi-factor model (e.g., Figure 9.2 (a)) there will be a statistically equivalent network model (e.g., Figure 9.2 (b)) that explains the covariance matrix with direct relations rather than with a general factor. It is also possible to explain the positive manifold in psychopathology data with a model that includes both latent variables and direct relations between symptoms, but does not include a general factor. The Residual Network Model (RNM; Epskamp, Rhemtulla, & Borsboom, 2017), is a model in which the residuals of a factor model form a GGM. Figure 9.2 (c) shows an example of such a RNM⁴. When the residual network is properly constrained (in this case, it would be expected to be low-rank), the RNM in Figure 9.2 (c) can also be equivalent to the bi-factor model in Figure 9.2 (a). As a result, all three figures can, under certain constraints of the parameter space, be completely equivalent while featuring strikingly different causal interpretations. Network theory therefore, just like sampling theory, provides an alternative explanation for the positive manifold without a general factor.

In sum, extracting a general factor does not adduce novel evidence for the existence of a general factor, if one already knows that the positive manifold holds – a finding that can be ex-

⁴The connections between the residuals in Figure 9.2 (c) should not be mistaken for doubleheaded arrows, which would represent marginal residual correlations in SEM. Instead, these undirected edges indicate conditional residual correlations (partial correlations); also termed concentration (Dempster, 1972).

plained in multiple different ways. In the cases of Caspi et al. (2014) and Sharp et al. (2015), therefore, their model fitting exercises do not produce evidence for general factors that goes beyond the fact that we see a positive manifold in data on DSM disorders; and this fact has been known for many years (Krueger, 1999).

9.4 CONCLUSIONS

We discussed a recent line of research in which bi-factor models are fitted to psychopathology data to demonstrate support for p , the general factor of psychopathology (Caspi et al., 2014; Laceulle et al., 2015; Lahey et al., 2012). Because analogous modeling strategies have been followed for over a century in the context of general intelligence research, we discussed the g -factor literature as well as some of the literature on general factors in personality to point out some of the risks in interpreting a general factor model as identifying a common cause that explains the positive manifold. We focused our discussion not on the specific models found by the above mentioned authors, but rather on their more general – and arguably more important – finding: psychopathology, like cognitive ability and personality, features a positive manifold; correlations between symptoms are almost invariably positive.

As we have argued in this paper, such a positive manifold will allow some form of general factor model to fit the data by mathematical necessity. We emphasize that we do not regard the positive manifold as something trivial; the observation of a positive manifold is certainly a remarkable observation that has proven to be a robust finding in many areas within psychology. Not only the positive manifold, but also the observation that some variables within this positive manifold form clusters of more strongly correlated variables, are robust findings in many fields (e.g., responses on mathematical items correlate more strongly with each other than with responses on language items). Such phenomena mandate the search for an explanatory model, and in many cases the factor model is a very fruitful candidate. There are, however alternative explanations for the emergence of a positive manifold, and factor analysis does not choose between these alternatives. Instead, factor analysis will always explain the positive manifold with a general factor. Therefore, we already knew, before Caspi and colleagues even started their research, that they would come up with a general factor as a matter of mathematical necessity. We suggest that scientists in the field of psychopathology do not rush into the p -factor but also carefully explore alternatives for the structure of psychopathology that do not rely on the logic of factor modeling.

Factor modeling is a powerful method for measuring constructs that are not directly observ-

able – e.g., factor analysis enables researchers to create a measurement model of constructs like working memory by observing the *effects* of working memory, such as performance on memory tasks. The problem we addressed in this paper is not the use of factor analysis for psychological data, but the risks of interpreting general factors of which it is more dubious whether such a construct exists as a causal entity. In such cases factor analysis is not used as a tool to measure unobserved constructs, but instead, factor analysis is used as a method to discover some inscrutable variable that explains as much as possible. Our aim is to show that the same care should be taken in interpreting the general *p*-factor as is now being carried out in intelligence (and, to some extent, personality) research.

To do so, model comparison based on fit measures can be supplemented with theory on how the general factor relates to external factors or specific factors within the model, and preferably experimental or quasi-experimental interventions to test such relations. Additionally, insight into the structure of psychopathology might benefit from the comparison between factor models and models different from factor modeling, instead of only comparing factor models *within* the paradigm of factor analysis. More generally, we propose to use all of the insights that can be gained with these different models (factor models, network models, or other models) and try to find ways to unify these different methods, instead of constraining ourselves to merely one subpopulation of models.

10

Discussion

This dissertation has investigated the meaning of several key concepts in psychometrics. The chapters addressed conceptual questions such as “how should error be interpreted?” and “what is the chance experiment that makes response variables random variables?” (addressed in Chapter 3) and “what does it mean that models are equivalent? Are these models the *same* model?” (addressed in Chapters 6 and 8). In thinking through these concepts, I have drawn a distinction between psychometric models as *statistical models* and as *causal models*, defining causal models as a set of equations that are used to represent the data-generating mechanism (Pearl, 1998, p. 236, see also Chapter 2).

I have discussed several benefits of interpreting psychometric models as causal models. For example, Chapter 2 argued that a causal interpretation legitimizes the assumptions that are tested when fitting a model to data, such as conditional independence, and legitimizes why shared variance but not the unique variance in the response variables is attributed to the latent variable. Chapter 3 provided a causal theory of the error scores and used this framework to consider concepts in modern test theory such as reliability, item bias and local homogeneity. Part II of this dissertation focused on the comparison of models from different modeling frameworks. For example, Chapter 6 compared latent variable models and network models. While the group of latent variable models is mathematically equivalent to the group of network models, interpreting these models causally triggers the question of whether a network model or latent variable model underlies the data. This question does not make sense when the models are mere descriptions of data. Chapter 6 also discusses, given a causal interpretation of the models, which groups of network models and latent variable models can be compared empirically and offers some first steps to do so.

In this discussion, I consider two remaining issues. First, while this dissertation has explored

the meaning of several psychometric concepts, there is one concept that is fundamental to this dissertation that has not been properly specified yet; the concept of a *causal relation*. Second, given that this dissertation argues for a causal interpretation of psychometric models, what are the implications of taking such a stance? In the next two sections, I discuss both these issues.

10.1 THE MEANING OF A CAUSAL INTERPRETATION

In a causal interpretation of psychometric models, some of the parameters (e.g., edge weights in network models or factor loadings in factor models) are interpreted as reflecting a *causal relation* between the variables. In what follows, I specify what I mean by a causal relation in the context of psychometric modeling.

Causality has been a topic of longstanding debates in many academic fields, most notably the philosophy of science. I do not intend to summarize those debates here, nor to propose yet another definition of causality. Instead, I aim to shed light on how a causal relation can be interpreted in psychometric models by discussing two accounts of causality that are particularly apt in the context of psychometric modeling and have shaped how causality is approached in experimental psychology: Lewis' account of causality (Lewis, 1973, 1986), and Pearl's account of causality (Pearl, 2000a). The first uses counterfactuals to express causal relations, whereas the latter uses probabilities to express causal relations. Both are considered difference-making accounts of causality, which are based on the intuition that a cause should make a difference to its effects (Sartorio, 2005; Williamson, 2006). Difference-making accounts are contrasted to mechanistic accounts of causality. While statistics can be used to show that a cause makes a difference to its effect, mechanisms provide an explanation of *how* the effect comes about (Russo & Williamson, 2007). The assumption that a cause must make a difference to the effect is, for example, at the core of Mill's method of difference (Mill, 1843, Bk. III, ch. VIII, sct. 2, p. 454):

“If the effect of $A B C$ is $a b c$, and the effect of $B C$, $b c$, it is evident that the effect of A is a . So again, if we begin at the other end, and desire to investigate the cause of an effect a , we must select an instance, as $a b c$, in which the effect occurs, and in which the antecedents were $A B C$, and we must look out for another instance in which the remaining circumstances, $b c$, occur without a . If the antecedents, in that instance, are $B C$, we know that the cause of a must be A : either A alone, or A in conjunction with some of the other circumstances present.”

The idea of a cause as a difference maker is not only intuitive, it is also deeply ingrained

in how causality is approached in experimental psychology. Mill's method of difference is, for example, applied in controlled experiments (Kment, 2010). The rationale for causal inferences in controlled experiments is as follows: if the experimental group and control group differ in one respect (the experimental condition) but can be assumed to be the same in every other respect, and these groups differ in the effect that we wish to study, then this supports the conclusion that the experimental condition is what caused this difference in effect.

10.1.1 CAUSALITY EXPRESSED IN COUNTERFACTUALS

One way of understanding the cause as a difference maker, is to say that a cause makes a difference to the effect if the effect would not have occurred if the cause had not occurred. This line of reasoning is central to Lewis' counterfactual account of causality (Lewis, 1986, pp. 160-161):

“We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it. Had it been absent, its effects – some of them, at least, and usually all – would have been absent as well.”

In this account, causal dependence is defined as follows. Let c and e denote two distinct actual events, then c is a cause of e , if and only if e would not have occurred if c had not occurred. Here, ‘if c had not occurred, e would not have occurred’ ($\neg c \rightarrow \neg e$), is the counterfactual.

Considering this account of causality, what does it mean to interpret a psychometric model causally? For example, what is the counterfactual that should hold for the latent variable to be a cause of the response variables in a causal interpretation of the latent variable model? Borsboom et al. (2003) distinguish two types of causal statements that articulate a causal effect of the latent variable on the response variables: (1) differences between subjects in their position on the latent variable cause differences between subjects in their item responses, and (2) a subject's position on the latent variable causes that subject's item response. The first statement situates the causal relation between the latent variable and response variables between subjects. The second statement situates the causal relation within subjects. In the following, I zoom in on how these different causal statements can be expressed in counterfactuals.

Let X_j denote a binary response variable with responses to item j . Let C denote the population of people with position θ on the latent variable. Let the expectation of X_j in that population be $P(X_j = 1) = .7$, that is, $P(X_j = 1 \mid \theta) = .7$. Considering the counterfactual account of causality, θ is the cause of the expectation of the responses to item j if the counter-

factual ($\neg \theta \rightarrow \neg (P(X_j = 1) = .7)$) holds, that is, ‘if the position on the latent variable had not been θ , the expectation of X_j would not have been .7’.

This counterfactual does not explicate whether the causal relation is situated within or between subjects. Now consider this adapted version of the counterfactual: (I) ‘if a different population than C had been selected, the expectation of X_j would not have been .7’. This counterfactual merely expresses that populations with different positions on the latent variable have different expectations for the response variable X_j , and thus situates the causal relation between the latent variable and the item responses between subjects.

The counterfactual statement (I) does not posit stochasticity at the level of the individual, in line with the random sampling rationale that was discussed in Chapter 3 of this dissertation. In the random sampling rationale, the conditional probability $P(X_j = 1 \mid \theta) = .7$ refers to the probability of sampling a subject from C who gives a positive response to item j . Counterfactual (I) is thus both consistent with the random sampling rationale and with causal statement (1) in which the causal relation is situated between rather than within subjects.

In the random sampling rationale, stochasticity results from sampling a random subject from population C . In this rationale there is no stochasticity at the level of the individual and thus each subject either gives a positive response with probability 1 or with probability 0. The conditional probability $P(X_j = 1 \mid \theta) = .7$ means that in the population of people with position θ on the latent variable, 70% of the people give a positive response and 30% of the people give a negative response, so that the probability of sampling a person who gives a positive response is .7. In contrast, the stochastic subject rationale does posit stochasticity at the level of the individual. In this rationale, the conditional probability of X_j given θ refers to the expectation of the responses to item j over repeated measurements for any person with position θ on the latent variable. In this rationale, $P(X_j = 1 \mid \theta) = .7$ means that each person in C has probability .7 of giving a positive response to item j (see Holland, 1990, and Chapter 3 of this dissertation for an explanation of these rationales).

While it might seem that the random sampling rationale goes hand in hand with a causal statement at the between subjects level, I will show that while the random sampling rationale does not posit stochasticity within subjects, it is still possible to express a causal relation at the within-subjects level that is consistent with the random sampling rationale. More generally, for both rationales there is both a counterfactual that expresses a causal relation situated within subjects and one that expresses a causal relation situated between subjects.

Consider the stochastic subject formulation of the latent variable model. Let c be a person from the population C , that is, c ’s position on the latent variable is θ , and c has probability

.7 of giving a positive response to item j . In the stochastic subject formulation of the latent variable model, statement (2) states that person c 's position on the latent variable is the cause of person c 's response to item j if the following counterfactual holds: (II) 'if person c 's position on the latent variable had not been θ , the expectation of the item responses for person c would not have been .7 '. The same counterfactual as for population C should thus now hold for person c . This counterfactual situates the causal relation between the latent variable and the item responses within person c .

Borsboom et al. (2003) argue that the first part of counterfactual (II) can be interpreted in the same way as the first part of counterfactual (I), so that we get: (III) 'if a person with a different position on the latent variable than person c had been selected, the expectation of the item responses for this person would not have been .7 '. This expresses a causal relation at the between-subjects level, namely that the difference in position on the latent variable between person c and some other person who could have been selected, causes the difference in their item responses.

In the stochastic subject rationale, it is thus possible to formulate both a counterfactual that situates the causal relation within subjects¹ and one that situates the causal relation between subjects. For the random sampling rationale, which does not posit stochasticity at the level of the individual, it is less clear what the counterfactual statement would be that expresses a causal relation within subjects.

Consider two populations: C , the population of people with position θ , and C^* , the population with position θ^* . The conditional probability of a positive response in C is $P(X_j = 1 \mid \theta) = .7$ and the conditional probability of a positive response in C^* is $P(X_j = 1 \mid \theta^*) = .2$. Let c denote a person from the population C who gave a positive response to item j . We want a counterfactual that expresses that the latent variable, Θ , is a cause of X_j at the individual level, that is, that person c 's position on the latent variable, θ , caused person c 's positive response to item j , within the random sampling rationale. That is, individuals have either probability 0 or probability 1 of giving a positive response to item j . The following counterfactual statement at the individual level does not hold: 'if person c 's position on the latent variable had been θ^* , the probability of a positive response to item j would have been .2 '. After all, in the random sampling rationale, persons have either probability 1 or 0 of giving a positive response (i.e., no stochasticity at the level of the individual). The counterfactual statement 'if person c 's position

¹Borsboom et al. (2003, p.212) argue that counterfactual (II) is actually "a between-subjects conclusion in disguise" and should be interpreted as counterfactual (III). This argument is open to debate (Borsboom, 2015; Weinberger, 2015) and is beyond the scope of this discussion.

on the latent variable had not been θ , the probability of giving a positive response to item j would not have been 1' also does not hold, as for different positions on the latent variable it is possible to have a probability 1 of giving a positive response to item j (e.g., 20% of the people in C^* also have probability 1 of giving a positive response to item j). To solve this, consider the latent response formulation of the IRT model (Takane & de Leeuw, 1987),

$$X_j = \begin{cases} 1 & \text{if } \Theta - b_j + E_j > 0 \\ 0 & \text{if } \Theta - b_j + E_j \leq 0, \end{cases} \quad (10.1)$$

in which b_j is a difficulty parameter and E_j is the error score variable. Equation (10.1) represents a 1-PL model if the variance of E_j is constrained to be the same over different items (which implies equal item discriminations) and represents a 2-PL model if the variance of E_j is free to vary over items. As explained in Chapter 3, the random sampling rationale implies that for person i with position θ on the latent variable, the item score x_{ij} and error score ε_{ij} are constants,

$$x_{ij} = \begin{cases} 1 & \text{if } \theta - b_j + \varepsilon_{ij} > 0 \\ 0 & \text{if } \theta - b_j + \varepsilon_{ij} \leq 0. \end{cases} \quad (10.2)$$

We can define a fixed score for person i , $d_{ij} = b_j - \varepsilon_{ij}$,

$$x_{ij} = \begin{cases} 1 & \text{if } \theta > d_{ij} \\ 0 & \text{if } \theta \leq d_{ij}. \end{cases} \quad (10.3)$$

The score d_{ij} functions as the cut-off point on the latent variable for which person i gives a positive response to item j . d_{ij} is a function of how difficult the item is (for more difficult items, a higher position on the latent variable is needed to result in a positive response than for easy items) and of person i 's position on the characteristic variables that compose E_j . Characteristic variables are perfectly reliable unique influences of the responses to item j (see Chapter 3). Higher values of ε_{ij} indicate that person i has characteristics that help person i to give a positive response to item j , whereas lower values of ε_{ij} indicate that person i has characteristics that prevent person i from giving a positive response to item j . A lower value of ε_{ij} has to be compensated with a higher position on the latent variable to result in a positive response, hence a lower value of ε_{ij} results in a higher value of d_{ij} . Equation (10.3) permits the following counterfactual statement at the level of the individual: (IV) 'if person c 's position on the latent variable had been smaller than d_{cj} , the expectation of the responses to item j for person c would have

been 0'. The result is that we can distinguish four different counterfactuals:

- I 'if a different population than C had been selected, the expectation of the responses to item j for this population would not have been .7'
- II 'if person c 's position on the latent variable had not been θ , the expectation of the responses to item j for person c would not have been .7'
- III 'if a person with a different position on the latent variable than person c had been selected, the expectation of the responses to item j for this person would not have been .7'
- IV 'if person c 's position on the latent variable had been smaller than d_{cj} , the expectation of the responses to item j for person c would have been 0'

Both the stochastic subject rationale and random sampling rationale are thus consistent with two counterfactual statements, one at the between-subjects level (counterfactual (I) is consistent with the random sampling rationale and counterfactual (III) is consistent with the stochastic subject rationale) and one at the individual level (counterfactual (IV) is consistent with the random sampling rationale and counterfactual (II) is consistent with the stochastic subject rationale). The four counterfactuals reflect different possible causal interpretation of the latent variable model, that are consistent with different assumptions in the latent variable model. For example, the counterfactuals that are consistent with the stochastic subject formulation of the latent variable model require the additional assumption of local homogeneity (for an explanation of this assumption see Chapter 3). Also, it is still up for debate whether the counterfactuals at the within-subjects level require an additional assumption, just like the stochastic subject formulation, or whether a causal relation at the within-subjects level is implied by a causal relation at the between-subjects level, because the counterfactuals at the between-subjects level cannot be interpreted meaningfully without positing a causal relation within subjects (Borsboom, 2015; Weinberger, 2015).

This section has enquired how a causal relation between the latent variable and response variables can be expressed using counterfactuals. This resulted in four different counterfactuals that are consistent with different assumptions about the latent variable model. That is, depending on whether one assumes local homogeneity and whether one assumes causality at the within subject level, a slightly different counterfactual can be used to express the causal relation. Next, I will discuss a different account of causality in which probabilities are used to express a causal relation between variables in a psychometric model.

10.1.2 CAUSALITY EXPRESSED IN PROBABILITIES

Probabilistic accounts of causality add to the intuition that the cause should make a difference to its effect, the intuition that this difference making manifests itself in probabilistic dependencies between cause and effect (Suppes, 1970; Williamson, 2009). The intuition to express difference making with probabilities follows from the intuition that there can be uncertainty involved in causal relations. For example, even though when someone drives recklessly that person will not always have a car accident, it is intuitive to claim that reckless driving causes car accidents. Deterministic accounts have a hard time dealing with such uncertainties, while probabilities do so naturally: reckless driving increases the probability of a car accident. One probabilistic account of causality that is of particular interest in the context of psychometric models is Pearl's account of causality (Pearl, 2000a). Since the late 20th century – a time up to which most statisticians refrained from causal talk, being warned that correlation need not imply causation (Pearl, 2000b) – probabilistic causality has been on the rise and, with it, a mathematical treatment of causality (Williamson, 2009). The books of Pearl (1988, 2000a), in which he reasons with probabilities to get a tractable system of causal connections between variables, reached a wide audience, bringing back causal talk in statistics (Williamson, 2009).

Probabilities can be used to express dependence between variables. For example, $P(A, B) \neq P(A)P(B)$ (i.e., the joint probability of A and B is different from the product of the probabilities of A and B) implies that A and B are associated. Similarly, the association between A and B can be expressed using conditional probabilities; $P(B \mid A = a) \neq P(B)$ (i.e., the probability of B conditional on $A = a$ is different from the marginal probability of B) implies that A and B are associated. But, associations or correlations are symmetric, whereas causal relations can be asymmetric (A is a cause of B but B is not a cause of A). Adding the direction of time (the cause must happen before the effect and thus if A and B are associated and A happens before B , A is a cause of B) does not solve the problem. $P(B \mid A = a) \neq P(B)$ does not express a causal relation between A and B because there can be other reasons for A and B to be associated, such as a third variable that causes both. Consider the famous example in which the drop in column of mercury in a barometer is associated with the occurrence of a storm (Jeffrey, 1969). The drop in mercury happens before the storm occurs; however, the drop in mercury does not cause the storm, a drop in atmospheric pressure first causes the drop in mercury and shortly after also causes the storm. Thus, although $P(\text{storm} \mid \text{drop in mercury}) > P(\text{storm})$ and the drop in mercury happens before the storm occurs, the drop in mercury is not a cause of the storm.

Figure 10.1: The latent variable model. Θ represents the latent attribute, and X_1 to X_J represent the response variables that load on the latent variable.

Pearl uses the notion of an intervention to solve this problem; A is a cause of B if the probability of B given an intervention on A is different from the marginal probability of B (Pearl, 2000a). Regular conditional probabilities such as $P(\text{storm} \mid \text{drop in mercury})$ express the probability of a storm when ‘seeing’ a drop in mercury (i.e., $P(\text{storm} \mid \text{see}(\text{drop in mercury}))$). The example of the atmospheric pressure as a cause of both the drop in mercury and the storm to occur shows that $P(B \mid \text{see}(A = a)) \neq P(B)$ does not express a causal relationship between A and B , merely an association. The regular probability language therefore does not allow for expressing causal relationships. Pearl introduces the *do*-operator, which expresses an intervention on a variable. The conditional probability $P(\text{storm} \mid \text{do}(\text{drop in mercury}))$ is the probability of a storm given an intervention on the barometer that results in a drop in mercury. The *do*-operator makes it possible to express a causal relationship with conditional probabilities: A causes B iff $P(B \mid \text{do}(A = a)) \neq P(B)$.

For this account of causality, what does it mean to interpret psychometric models causally? Consider the unidimensional factor model defined in the Introduction of this dissertation. Let $\mathbf{X} = \{X_1, X_2, \dots, X_j, \dots, X_J\}$ denote a vector of J observed variables and let Θ denote a latent variable. Furthermore let E_j denote the error score variable, λ_j denote the factor loading of X_j on the latent variable, and v_j denote the intercept (Jöreskog, 1971). The observed variable X_j is a linear function of Θ ,

$$X_j = v_j + \lambda_j \Theta + E_j. \quad (10.4)$$

This model is represented in Figure 10.1. It is assumed that the error scores are independent of each other, and it is assumed that the observed variables are monotonically nondecreasing in Θ

(Ellis & Junker, 1997).

I have argued in this dissertation that parameters such as edge weights in network models and factor loadings in a latent variable model can be interpreted as *causal hypotheses*. Applying Pearl's account of causality, factor loadings (represented as arrows between the latent variable and response variables in Figure 10.1) can be interpreted as the hypothesis that an intervention on the latent variable changes the probabilities of responses on the response variable. For example, intervening on someone's intelligence (by, for example, giving this person a couple of shots tequila) will result in this person having a lower probability of solving a math problem. The arrows in Figure 10.1 that connect the latent attribute to the response variables represent *predictions* about what would happen to the variable connected to the head of the arrow if one were to intervene on the variable connected to the tail of the arrow, or as Pearl (2000a, p. 22) describes: "a joint distribution tells us how probable events are and how these probabilities would change with subsequent observations, but a causal model also tells us how these probabilities would change as a result of external interventions". In the causal theory of error scores proposed in Chapter 3, the arrows in Figure 10.1 that connect the error scores to the response variables also reflect causal hypotheses. In Chapter 6, I described constraints on the parameter space also as hypotheses that are interpreted in light of the theory about the data-generating process reflected by the model. That is, if the psychometric model is interpreted as a causal model, both free parameters and constraints on the parameter space can reflect causal hypotheses.

For example, the causal model, which we get by interpreting the statistical model formulated in Equation 10.4 as a representation of the data-generating mechanism, assumes that the omitted unique determinants of the response variables that constitute the error score variable, are independent, and that no direct effects exist between any pair of variables in $\mathbf{X}$, as well as that the latent variable causes the response variables. These are what I called causal hypotheses and what Pearl calls *causal assumptions* (Bollen & Pearl, 2013; Pearl, 1998, 2012) that are encoded in the model. Pearl distinguishes between causal assumptions and causal queries (Pearl, 2012). The causal assumptions concern which parameters are freely estimated and which constraints are put on the parameter space. The causal queries concern the magnitude of the structural coefficients. When the causal model is fit to data, the causal assumptions have testable implications that can be compared to the observed data to support or reject the model, whereas the causal queries result in estimated coefficients that can be interpreted as the strength of the assumed causal relationships, conditional on the assumptions encoded in the model.

To summarize, a causal interpretation of psychometric models means that the model represents a data-generating mechanism. In such a causal model, the structural equations reflect

causal hypotheses or assumptions². When considering Lewis' counterfactual account of causality, the equation $x = v + \lambda\theta + \varepsilon$ can be interpreted as 'if the position on the latent variable had not been θ , then the item score would not have been x '³. When considering Pearl's account of causality, the equation $x = v + \lambda\theta + \varepsilon$ can be interpreted as 'increasing Θ (intervention) results in an increase in X '.

Now that I have discussed different ways of interpreting psychometric models as causal models and have provided benefits of doing so, I will use the next and final section of this dissertation to consider the implications of such a view on psychometric models.

10.2 IMPLICATIONS OF A CAUSAL INTERPRETATION

A causal interpretation of psychometric models calls for an assessment of how close the data-generating mechanism represented by the model is to the true data-generating mechanism. This implication is discussed in Chapter 2. While fitting a model to data tests certain assumptions of the model, it is hard to compare models based on model fit when models are mathematically closely related. This problem is discussed in Chapter 9. In the extreme, two models that are statistically equivalent fit equally well to any dataset and thus while the models may represent different causal mechanisms, model fit does not distinguish between the models.

One seemingly straightforward solution follows from Pearl's account of causality. If the arrows in the causal model represent predictions about changes in a dependent variable when intervening on the independent variable, experimentally intervening on the independent variable can serve to test these predictions. However, in many cases experimental intervention is not as straightforward as it may sound. First of all, the data that are available do not always allow for interventions. For example, one assumption that should hold to infer causality from interventions is that one can intervene on autonomous parts of the model. In some cases it might not be possible to intervene on one variable without affecting other variables, for example because the

²The use of the word *hypothesis* versus *assumption* to describe the causal interpretation of parameters in the model may depend on the context. In a causal interpretation of the network model, constraining an edge weight to zero reflects the assumption that there is no direct causal relation between those variables. In a context in which this assumption is tested to see if it holds, the assumption functions as an hypothesis that can turn out true or false.

³This counterfactual is the same as the counterfactuals discussed earlier but then for a continuous response variable. The counterfactual leaves open whether the causal relation is between subjects or within subjects, that is, whether 'if the position on the latent variable had not been θ ' refers to selecting another person or refers to the same person having a different position on the latent variable.

variables draw on the same psychological processes. In that case it is not clear whether changes in a variable that one did not intend to intervene on, are nonetheless a direct result of the intervention or whether these changes are the indirect result of changes in the variable that one intended to intervene on. When the variable one intends to intervene on is a latent variable, matters get even more complicated. After all, the latent variable is only measured via the response variables and thus if the response variables do not change as a result of the intervention, it is not clear whether this is because the intervention did not work, or whether the response variables are not effects of the latent variable. More generally, to infer a causal effect from experimental intervention on a latent variable, the latent variable must be manipulated independently of the observed variables, and other latent variables that influence the observed variables must be held constant, but neither of these conditions can be tested (Trendler, 2009).

Thus, while interventions are essential to causal inferences, and a focus on more intervention data is definitely a way forward in providing insight into the causal mechanisms underlying the data, good intervention data are not always available. In such cases it might be worth spending more time thinking about how the assumptions that are tested when fitting the model to data relate to the data-generating mechanism that is represented by the model. For example, model fit does not distinguish between a one-factor model and a fully connected network that has a rank constraint on the edge weights (see Chapter 6), but the rank constraint is an essential implication of the data-generating mechanism that is represented by a one-factor model (a single underlying common cause implies that the covariance structure is of rank one), while this constraint does not follow from the hypothesis that the variables cause each other (i.e., network theory). In this case, finding that the rank constraint holds in the data can be seen as support for a factor model over a network model, even though equivalent network models could be constructed in which the rank constraint also holds (Gignac, 2016). This is not to say that it is not possible to have a specific theory about the network in a particular context that implies that the direct causal relations result in covariances for which a rank constraint holds, but it does not automatically follow from the hypothesis of direct causal relations between variables. Thus, in order to examine data-generating mechanisms in psychology, it is often necessary to examine the meaning of constraints, and for what theories about the data-generating mechanism these assumptions provide support when they hold in the data. Which constraints are essential in testing the data-generating mechanism and what constraints merely function to further simplify the model?

A causal interpretation of network models necessitates a serious investigation of what ways current estimation methods fail to approach the causal skeleton of the true data-generating

mechanism, and an attempt to improve our estimation methods in order to achieve a closer approximation of this skeleton. A *skeleton* is what results from removing the arrowheads from a causal graph (Pearl, 2000a, p. 12). So, in a skeleton, variables are connected when they have a direct causal relation, but the edges are undirected. The networks that are typically estimated from psychological data are estimations of the population precision matrix (Epskamp et al., 2012; Epskamp, Maris, et al., 2018, see also Chapter 4), which also results in undirected networks. The population precision matrix could be viewed as an approximation of the skeleton, but there are also ways in which the two differ. The precision matrix encodes the conditional dependence structure, that is, zeroes in the precision matrix reflect conditional independencies, and non-zero elements reflect conditional dependencies. Direct causal relations imply conditional dependencies (and thus imply that the corresponding elements in the precision matrix are non-zero), but absence of a direct causal relation does not imply that the variables are conditionally independent (and thus also do not imply that the corresponding element in the precision matrix is zero). For example, conditioning on a collider of two variables that are not directly causally related implies that the two variables become conditionally dependent (Pearl, 2000a), resulting in a non-zero element in the precision matrix. Furthermore a common cause of two variables that are not directly causally related, but that is not conditioned on (e.g., when the common cause is not included in the network) also implies that the two variables are conditionally dependent. These structures result in conditional dependencies that do not correspond to a direct causal relationship. Inferring causal relationships from these conditional dependencies thus results in incorrect conclusions.

These divergences between the skeleton and the conditional dependence structure that is estimated with the current estimation techniques, are no problem if the estimated network is not interpreted as a skeleton. That is, one can choose to not interpret the estimated network causally in which case one also does not make incorrect inferences about causal relations in the network. However, as I have shown throughout this dissertation, refraining from a causal interpretation is hardly a solution. For example, when one wants to explore new treatment strategies in the context of symptom networks, the goal of estimating a network can be to identify variables that have an impact on the rest of the network when intervening on this variable. In such cases, the causal skeleton is of more interest than the conditional dependence structure, so refraining from a causal interpretation defeats the goal of why one estimated a network in the first place.

When the goal is to make inferences about the data-generating mechanism (rather than summarizing the data), interpreting the estimated model as a representation of this mechanism (e.g., the network as an approximation of the skeleton) urges researchers to critically evaluate in what

ways the estimation methods fail to approximate this mechanism. Not interpreting the estimated model causally runs the risk of stagnating this goal of improving methods to get a closer representation of the data-generating mechanism. After all, if the parameters are only interpreted statistically without a referent outside the data (e.g., edges simply represent partial correlations) then there is no need to assess in what ways the model *fails* to identify the true mechanism (e.g., in what ways the partial correlation structure fails to identify true direct effects).

CONCLUSION

I have introduced psychometrics as the field that is concerned with the theory of psychological measurement. Psychometric theories connect psychological constructs to data of observable behaviors. These theories can be formalized in a set of equations and assumptions and the resulting psychometric model can be fit to data. This dissertation shows that for psychometric models to meet the goal of *explaining* the relation between the psychological construct and the observed data, the psychometric model needs to be interpreted as a representation of the mechanism that underlies the observed data; that is, as a causal model.

The field of psychometrics not only has the task of constructing new models, but also to reflect on existing models and explicate the theories about the data-generating mechanism that are represented by these models. This latter goal calls for an investigation into possible interpretations of psychometric models. This includes, among other things, reflecting on the meaning of assumptions in the model, and providing an explanation of how that interpretation fits in the broader framework that one works in. For example, there may be a rationale for putting a rank constraint on the edge weights in a network model. This rationale should make explicit what the constraint means within the network framework. By making this rationale explicit, psychologists who use these models can consider this rationale in the context that they are working in, when making the choice of whether to include such a constraint in the model.

The view on psychometric models laid out in this dissertation shows some of the consequences of choosing a certain interpretation of psychometric models and, regardless of whether one chooses to interpret the models statistically, causally or otherwise (there are definitely other options as well that I did not consider in this dissertation), the dissertation shows the importance of thinking through the meaning of psychometric models and provides insights in what to consider when making this choice.

A

Appendix to Chapter 6

A.1 ITEMS IRRATIONAL PROCRASTINATION SCALE

Irrational Procrastination Scale (IPS; Steel, 2010)

1. I put things off so long that my well-being or efficiency unnecessarily suffers
2. If there is something I should do, I get to it before attending to lesser tasks (R)
3. My life would be better if I did some activities or tasks earlier
4. When I should be doing one thing, I will do another
5. At the end of the day, I know I could have spent the time better
6. I spend my time wisely (R)
7. I delay tasks beyond what is reasonable
8. I procrastinate
9. I do everything when I believe it needs to be done (R)

Note: Items designated with an (R) are reverse scored.

	V ₁	V ₂	V ₃	V ₄	V ₅	V ₆	V ₇	V ₈	V ₉
V ₁	1.000	0.187	0.476	0.517	0.393	0.399	0.520	0.529	0.272
V ₂	0.187	1.000	-0.097	0.304	0.047	0.439	0.264	0.309	0.426
V ₃	0.476	-0.097	1.000	0.436	0.557	0.260	0.394	0.395	0.086
V ₄	0.517	0.304	0.436	1.000	0.476	0.351	0.576	0.520	0.320
V ₅	0.393	0.047	0.557	0.476	1.000	0.422	0.508	0.502	0.178
V ₆	0.399	0.439	0.260	0.351	0.422	1.000	0.492	0.531	0.449
V ₇	0.520	0.264	0.394	0.576	0.508	0.492	1.000	0.740	0.291
V ₈	0.529	0.309	0.395	0.520	0.502	0.531	0.740	1.000	0.368
V ₉	0.272	0.426	0.086	0.320	0.178	0.449	0.291	0.368	1.000

Table A.1: Polychoric correlation matrix of the 9 procrastination items. This correlation matrix is obtained with `cor_auto()` of the R-package `qgraph` (Epskamp et al., 2012).

B

Appendix to Chapter 7

Chapter 7 proposed a bootstrap test to identify partial correlations that are not between zero and the zero-order correlation. This bootstrap test was applied to two examples: (1) data simulated from a UFM and (2) data simulated from a random correlation matrix. We here provide additional information on these examples. We also provide results of the simulation study on the performance of the proposed bootstrap test.

B.1 EXAMPLE I: UFM

As specified in Chapter 7, the model implied covariance matrix of a unidimensional factor model (UFM) is of the form:

$$\Sigma = \lambda\lambda' + \Theta. \quad (\text{B.1})$$

The following provides some additional information on the simulated example, in which data was generated from a UFM with factor loadings close to zero. Factor loadings were sampled from a uniform distribution over $[0.05, 0.2]$, corresponding to 6 observed variables. Each factor loading had a probability of .5 to be multiplied with -1 resulting in both positive and negative factor loadings. Θ is a diagonal matrix with $\theta_{ii} = 1 - \lambda_i^2$, such that the diagonal of Σ is one, resulting in a correlation matrix. We simulated 60 observations. The following includes the R-code we used to create the UFM and sample data with.

```
lambda = sample(c(-1, 1), 6, replace = T)*runif(6, 0.05, 0.2)
Sigma = outer(lambda,lambda) + diag(1 - lambda ^ 2)
X = matrix(rnorm(length(Sigma)*10),,ncol(Sigma)) %%% chol(Sigma)
```

	V ₁	V ₂	V ₃	V ₄	V ₅	V ₆
V ₁	1	0.212	0.139	0.231	0.193	0.079
V ₂	0.129	1	0.296	0.170	0.010	0.042
V ₃	0.128	0.319	1	-0.157	0.065	-0.060
V ₄	0.207	0.210	-0.266	1	0.193	-0.101
V ₅	0.158	-0.074	0.084	0.166	1	-0.088
V ₆	0.122	0.071	-0.110	-0.137	-0.076	1

Table B.1: Upper triangle of matrix represents sample correlation matrix of dataset with 60 observations that is simulated from UFM with both positive and negative factor loadings. The absolute factor loadings come from a uniform distribution over [0.05, 0.2]. The lower triangle of the matrix represents the corresponding partial correlations.

The sample correlation matrix of the simulated dataset is presented in the upper triangle of the matrix in Table B.1. The corresponding sample partial correlation matrix of the data is presented in the lower triangle of this same matrix. The results from the bootstrap test on this dataset can be found in Table B.2 and Figure B.1. All of the CIs include one and the UFM should thus not be rejected.

	V ₁	V ₂	V ₃	V ₄	V ₅	V ₆
V ₁	-	0.028	0.08	0.076	0.045	0.117
V ₂	-	-	0.216	0.129	0.065	0.084
V ₃	-	-	-	0.633	0.063	0.080
V ₄	-	-	-	-	0.068	0.103
V ₅	-	-	-	-	-	0.049
V ₆	-	-	-	-	-	-

Table B.2: Lower bound of the CI obtained with the bootstrap test for each zero-order correlation and corresponding partial correlation in Table B.1.

Unidimensional Factor Model

Figure B.1 (previous page): Bootstrap results for the simulated data example. The results are based on a dataset with 6 variables and 60 observations. The horizontal axis represents $\log(|2\rho_{y_1 y_2 \cdot Z} / \rho_{y_1 y_2} - 1|)$. The vertical dashed red line indicates the point where $\log(|2\rho_{y_1 y_2 \cdot Z} / \rho_{y_1 y_2} - 1|) = 0$ (i.e., $|2\rho_{y_1 y_2 \cdot Z} / \rho_{y_1 y_2} - 1| = 1$). Variables for which the CI of the bootstrapped values only includes values greater than 1 provide evidence against the UFM.

B.2 EXAMPLE II: RANDOM CORRELATION MATRIX

Here we provide a similar example as Example I but with data that are not generated from a unidimensional factor model but from a random correlation matrix with six variables. We created the correlation matrix by taking the cross product of two 6×6 matrices that consisted of random values drawn from a uniform distribution over $[-1, 1]$. We simulated 60 observations from this random correlation matrix. The following includes the R code we used to create these matrices and sample data from it.

```
Sigma = cov2cor(crossprod(matrix(runif(6 ^ 2, -1, 1), 6)))
X = matrix(rnorm(length(Sigma)*10), ncol(Sigma)) %%% chol(Sigma)
```

Here Sigma necessarily does not conform to a UFM. The sample correlation matrix is presented in the upper triangle of the matrix in Table B.3.

	V ₁	V ₂	V ₃	V ₄	V ₅	V ₆
V ₁	1	-0.001	-0.406	-0.039	-0.257	-0.185
V ₂	0.004	1	0.215	0.034	0.233	0.310
V ₃	-0.520	-0.228	1	0.256	0.427	0.408
V ₄	0.480	0.254	0.820	1	0.780	-0.652
V ₅	-0.497	-0.115	-0.646	0.919	1	-0.156
V ₆	0.424	0.349	0.864	-0.950	0.816	1

Table B.3: The upper triangle of the matrix represents a random correlation matrix. The correlations are randomly drawn from a uniform distribution that takes both positive and negative values. The lower triangle of the matrix represents the corresponding partial correlations. Each zero-order correlation with a gray background corresponds to a partial correlation with a gray background and the combination refers to a pair for which the bootstrap test was significant (the CI obtained with the bootstrap test did not include one).

	V ₁	V ₂	V ₃	V ₄	V ₅	V ₆
V ₁	-	0.071	0.557	3.159	0.847	2.439
V ₂	-	-	0.986	0.339	0.14	0.089
V ₃	-	-	-	2.396	2.761	1.99
V ₄	-	-	-	-	1.116	1.436
V ₅	-	-	-	-	-	4.793
V ₆	-	-	-	-	-	-

Table B.4: Lower bound of the CI obtained with the bootstrap test for each zero-order correlation and corresponding partial correlation in Table B.3. The values that have a gray background refer to a combination of a zero-order correlation and partial correlation for which the CI obtained with the bootstrap test does not include one.

The corresponding partial correlation matrix is presented in the lower triangle of this same matrix (Table B.3). The results of the bootstrap test to these data are presented in Table B.4 and Figure B.2. There are eight pairs of variables for which the CI does not include the value one. The sample correlation matrix and sample partial correlation matrix in Table B.3 show that four of these pairs correspond to zero-order correlations that have a different sign than the partial correlation (r_{14} , r_{35} , r_{16} and r_{56}). The other four pairs correspond to partial correlations that are stronger than the zero-order correlation (r_{34} , r_{45} , r_{36} and r_{46}).

Random Correlation Matrix

Figure B.2 (previous page): Bootstrap results when the data were generated from a random correlation matrix with both positive and negative correlations. The results are based on a dataset with 6 variables and 60 observations. The horizontal axis represents $\log(|2\rho_{y_1 y_2 \cdot Z} / \rho_{y_1 y_2} - 1|)$. The vertical dashed red line indicates the point where $\log(|2\rho_{y_1 y_2 \cdot Z} / \rho_{y_1 y_2} - 1|) = 0$ (i.e., $|2\rho_{y_1 y_2 \cdot Z} / \rho_{y_1 y_2} - 1| = 1$). Variables for which the CI of the bootstrapped values only includes values greater than 1 provide evidence against the unidimensional factor model.

B.3 SIMULATION STUDY: PERFORMANCE OF THE EMPIRICAL BOOTSTRAP TEST

We performed a simulation study to obtain the false positive rate and the power of the bootstrap test in rejecting the UFM when the underlying model is not a UFM. We rejected the unidimensional factor model when the bootstrap test identified at least one partial correlations that was not between zero and the zero-order correlation. We tested the power of the test in three different situations. That is, we tested the power of the bootstrap test (1) when the population correlation matrix is generated from a UFM but misfit is created by adding or subtracting a given value from one of the correlations, (2) when the population correlation matrix is generated from a UFM but misfit is created by switching the sign of one of the correlations in the population correlation matrix, and (3) when the population correlation matrix is a random correlation matrix.

For the alternative model in which a given value was added to a correlation, we considered three different values of delta (0.2, 0.4 and 0.6). This results in six models: the UFM and five alternative models. For these six models we considered 3 different numbers of observed variables (5, 10 and 15 variables), 10 different sample sizes (100 to 1000 in steps of 100) and considered both positive manifold data (since it is common in psychology to have solely positive correlations), and data with both positive and negative correlations, resulting in a total of 360 conditions. We used 1000 iterations for each condition.

For conditions in which the UFM is the true data generating model, random factor loadings were drawn from a uniform distribution over $[0.1, 1]$ (or $[-1, 1]$ in conditions with both positive and negative correlations). We specified Θ to be a diagonal matrix with $\theta_{ii} = 1 - \lambda_i^2$, such that the diagonal of Σ is one, resulting in a correlation matrix. For the first type of alternative models a value delta (0.2, 0.4 or 0.6) was added to one randomly selected correlation in Σ . In roughly half of the cases delta was multiplied by -1 so that delta was either added or subtracted. For the second type of alternative models the sign of a randomly selected correlation in

Σ was switched around. For the third type of alternative models we generated random correlation matrices that were inconsistent with a UFM by standardizing the cross-product of a matrix with values randomly drawn from a uniform distribution over $[-0.25, 1]$ (or over $[-1, 1]$ in the conditions with both positive and negative correlations). The reason for drawing values from a uniform distribution over $[-0.25, 1]$ rather than over $[0, 1]$ in conditions with positive correlations, is that in the latter case the resulting correlations are very large (most correlations $> .8$). When drawing values from $[-0.25, 1]$ the correlations have a wider range, including low correlations and sometimes even a negative correlation. We selected the matrices with only positive correlations. The R-code for generating these matrices is included below.

B.3.1 RESULTS FOR CONDITIONS WITH ONLY POSITIVE CORRELATIONS

Results of the simulation study for data with only positive correlations are summarized in Table B.5 and visualized in Figure B.3. The false positive rate is lower than 5% for all conditions except for the condition of 15 variables and 400 observations for which the false positive rate is 5.2%.

model	# variables	N									
		100	200	300	400	500	600	700	800	900	1000
UFM	5	0.3	0.7	0.9	0.7	0.7	0.7	0.6	0.6	0.5	0.6
	10	2.8	2.4	2.9	1.8	2.7	1.9	2.1	2.4	2	1.7
	15	2.7	3.6	4.5	5.2	4.1	4.3	4.4	4.4	4.2	4.3
delta = 0.2	5	19.6	28.1	36.8	42.3	45.1	49.2	51.3	53	53.3	54.8
	10	25.9	42.4	51.9	55.2	53.9	57.9	62.4	64.1	62.6	64.5
	15	25.4	40.2	51.1	53.2	59.1	59.0	60.4	61.4	59.3	59.3
delta = 0.4	5	50.0	64.9	67.1	70.2	70.1	71.2	75.0	76.6	73.7	75.2
	10	51.7	62.1	65.7	68.2	69.5	70.0	69.3	69.6	70.9	72.3
	15	42.4	55.7	61.5	60.2	63.8	64.2	65.1	67.8	67.4	65.6
delta = 0.6	5	70.0	78.1	80.1	79.0	81.7	83.1	85	83.1	85	87.1
	10	55.5	66.2	73.1	73.8	75.2	74.6	76.5	76.9	74.1	77.0
	15	46.3	61.6	64.6	65.9	69.0	69.8	71.3	74.0	73.0	74.8
switch sign	5	62.4	73.9	83.7	85.1	87.8	89.8	91.6	91.9	93.5	94.5
	10	54.1	69.6	77.8	84.1	85.9	90.2	90.4	90.5	93.2	94.3
	15	48.1	67.7	75.5	81.5	83.5	87.4	88.3	91.1	92.0	92.2
random correlations	5	97.8	99.0	99.3	99.7	99.8	99.8	99.9	99.8	99.9	99.9
	10	100	100	100	100	100	100	100	100	100	100
	15	100	100	100	100	100	100	100	100	100	100

Table B.5: Percentage of cases in which the bootstrap test rejects the UFM. The results in this table stem from simulations in which either a UFM was the data-generating model (i.e., percentages reflect the false positive rate) or an alternative model was the data-generating model (i.e., percentages reflect the power of the test). For this set of simulations all models resulted in solely positive correlations except the alternative model in which one of the correlations was switched from a positive to a negative sign.

Performance of bootstrap test: positive correlations

Figure B.3: Performance of the bootstrap test on simulated data with only positive correlations.

One might note that the false positive rate increases slightly with an increasing number of variables. The reason for this is that by increasing the number of variables the number of comparisons grows very fast. The Bonferroni correction corrects for this, but for a smaller α more bootstraps are needed to get the empirical distribution range over the $1 - \alpha$ interval. When the number of bootstraps is increased from 1000 to 10000, the false positive rate for 15 variables ranges between 0.8% and 2% over the different sample sizes. It is therefore important that the number of bootstrap samples is large enough, and a smaller α requires more bootstraps.

For random correlation matrices the power of the test was always 1 except for the condition with 5 variables for which the power ranges from 0.978 to 0.999; that is, with 10 and 15 variables the test always rejected the UFM when the data was simulated from a random correlation matrix. For the other alternative models the power increased as the sample size increased.

B.3.2 RESULTS FOR CONDITIONS WITH BOTH POSITIVE AND NEGATIVE CORRELATIONS

The results of the simulation study for data with both positive and negative correlations is summarized in Table B.6 and visualized in Figure B.4. The results can be interpreted in the same way as for the conditions with only positive correlations.

B.3.3 SPECIFYING ALTERNATIVE MODELS

The following provides R-code for generating the models that are used in the simulation study:

```
# Alternative models in simulation study:
nV <- 5 # number of variables
lambda = runif(nV, 0.1, 1)
Sigma_ufm = outer(lambda, lambda) + diag(1-lambda ^ 2) # Sigma of UFM
# Alternative 1: delta = 0.2 or delta = 0.4 or delta = 0.6
d <- -1
while(d<0){
  delta<-sample(c(-1, 1), 1)*delta
  i <- sample(1:nV, 1)
  j <- sample((1:nV)[-i], 1) # selecting one random off-diagonal element: Sigma[i,j]
  Sigma_alternative <- Sigma_ufm
  Sigma_alternative[i, j] <- Sigma_ufm[i, j]+delta
  Sigma_alternative[j, i] <- Sigma_ufm[j, i]+delta
  d <- det(Sigma_alternative)
}
```

Performance of bootstrap test: positive and negative correlations

Figure B.4: Performance of the bootstrap test on simulated data with both positive and negative correlations.

model	# variables	N									
		100	200	300	400	500	600	700	800	900	1000
UFM	5	0.5	0.9	0.5	0.1	0.5	0.4	0.9	0.6	0.1	0.6
	10	2.2	1.5	1.9	2	1.9	1.9	1.5	1.2	2.3	1.7
	15	2.8	3.5	3.4	3.2	3.7	4.4	3.3	4.6	3.7	3.5
delta = 0.2	5	15.4	27.3	33.3	36.1	40.2	43.1	47.9	45.3	46.2	48.7
	10	19.1	32.1	42.6	46.8	51.7	54.7	53.9	55.7	54.9	58.5
	15	22.5	33.8	43.5	49.4	50.0	54.6	57.9	51.4	57.4	57.6
delta = 0.4	5	46.1	55.1	63.5	66.9	67.5	74.0	72.8	71.1	74.9	73.8
	10	43.9	54.0	61.0	61.8	65.3	69.0	67.8	66.5	68.6	71.5
	15	37.0	50.2	55.3	56.3	61.0	62.4	63.8	65.8	65.6	65.6
delta = 0.6	5	61.4	70.4	76.2	78.4	79.9	82.0	82.5	83.0	82.7	84.4
	10	49.6	65.8	67.1	70.9	72.4	71.0	74.9	74.4	76.0	77.1
	15	42.6	54.9	62.6	65.3	66.4	66.9	68.1	72.4	70.3	71.8
switch sign	5	46.5	56.1	64.3	67.4	70.8	73.7	73.2	75.7	75.9	75.2
	10	37.9	52.2	62.1	62.2	67.8	68.0	69.3	75.1	73.4	71.5
	15	33.7	51.6	55.4	61.1	66.5	66.0	72.2	70.6	73.0	74.8
random correlations	5	97.8	99	99.3	99.7	99.8	99.8	99.9	99.8	99.9	99.9
	10	96.8	99.1	99.4	99.4	99.7	99.8	100	100	99.8	100
	15	100	100	100	100	100	100	100	100	100	100

Table B.6: Percentage of cases in which the bootstrap test rejects the UFM. The results in this table stem from simulations in which either a UFM was the data-generating model (i.e., percentages reflect the false positive rate) or an alternative model was the data-generating model (i.e., percentages reflect the power of the test). For this set of simulations all models resulted in correlation matrices with both positive and negative correlations.

```
# Alternative 2: switch sign
d <- -1
while(d<0){
  i <- sample(1:nV, 1); j <- sample((1:nV)[-i], 1)
  Sigma_alternative <- Sigma_ufm
  Sigma_alternative[i, j] <- -1*Sigma_ufm[i, j]
  Sigma_alternative[j, i] <- -1*Sigma_ufm[j, i]
  d <- det(Sigma_alternative)
}

# Alternative 3: Random correlation matrix
# positive and negative correlations:
Sigma = cov2cor(crossprod(matrix(runif(nV ^ 2, -1, 1), nV)))
# positive correlations:
d <- 1
while(d != nV*nV){
  Sigma = cov2cor(crossprod(matrix(runif(nV ^ 2, -0.25, 1), nV)))
  d <- sum(Sigma > 0)
}
```

Appendix to Chapter 8

C.1 PROOF OF THEOREM 1

The marginal expectation of η is equal to $\mathbb{E}(\eta) = \sum_{i=1}^p \mathbb{E}(X_i)\gamma_i = \boldsymbol{\mu}^\top \boldsymbol{\gamma}$, and its marginal variance is equal to

$$\text{Var}(\eta) = \mathbb{E}(\text{Var}(\eta \mid \mathbf{X})) + \text{Var}(\mathbb{E}(\eta \mid \mathbf{X})) = \sigma^2 + \text{Var}(\mathbf{X}^\top \boldsymbol{\gamma}) = \sigma^2 + \sum_{i=1}^p \tau_i^2 \gamma_i^2.$$

The covariance between η and X_j is equal to

$$\text{Cov}(\eta, X_j) = \text{Cov}\left(\sum_{i=1}^p X_i \gamma_i + \zeta, X_j\right) = \sum_{i=1}^p \gamma_i \text{Cov}(X_i, X_j) = \gamma_j \text{Cov}(X_j, X_j) = \gamma_j \tau_j^2.$$

Thus, the joint distribution of $\mathbf{X}$ and η is given by

$$\begin{pmatrix} X_1 \\ \vdots \\ X_p \\ \eta \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \\ \sum_{i=1}^p \mu_i \gamma_i \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \dots & 0 & \gamma_1 \tau_1^2 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & \tau_p^2 & \gamma_p \tau_p^2 \\ \gamma_1 \tau_1^2 & \dots & \gamma_p \tau_p^2 & \sigma^2 + \sum_{i=1}^p \tau_i^2 \gamma_i^2 \end{pmatrix}\right).$$

The covariance matrix $\boldsymbol{\Sigma}$ can be partitioned in four submatrices,

$$\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array} \right) = \left(\begin{array}{ccc|c} \tau_1^2 & \dots & 0 & \gamma_1 \tau_1^2 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & \tau_p^2 & \gamma_p \tau_p^2 \\ \hline \gamma_1 \tau_1^2 & \dots & \gamma_p \tau_p^2 & \sigma^2 + \sum_{i=1}^p \tau_i^2 \gamma_i^2 \end{array} \right).$$

The conditional distribution of $\mathbf{X}$ given η is also multivariate normal. The conditional expected value of $\mathbf{X}$ given η (Anderson, 1958) is given by

$$E(\mathbf{X} \mid \eta) = \boldsymbol{\mu} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\eta - \mathbb{E}(\eta)),$$

which is equal to

$$\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} + \begin{pmatrix} \gamma_1 \tau_1^2 \\ \vdots \\ \gamma_p \tau_p^2 \end{pmatrix} \times \frac{\eta - \sum_{i=1}^p \mu_i \gamma_i}{\sigma^2 + \sum_{i=1}^p \gamma_i^2 \tau_i^2}.$$

The conditional covariance matrix of $\mathbf{X}$ given η is

$$\boldsymbol{\Sigma}(\mathbf{X} \mid \eta) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21},$$

which is equal to

$$\begin{pmatrix} \tau_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \tau_p^2 \end{pmatrix} - \frac{1}{\sigma^2 + \sum_{i=1}^p \tau_i^2 \gamma_i^2} \times \begin{pmatrix} \tau_1^2 \gamma_1 \\ \vdots \\ \tau_p^2 \gamma_p \end{pmatrix} \begin{pmatrix} \tau_1^2 \gamma_1 \\ \vdots \\ \tau_p^2 \gamma_p \end{pmatrix}^T.$$

All off-diagonal elements of this conditional covariance matrix are negative whenever the γ parameters have the same sign. This means that the variables in $\mathbf{X}$ can always be recoded such that a negative manifold is obtained for the conditional associations between the variables in $\mathbf{X}$ given η .

C.2 PROOF OF THEOREM 2

The conditional distribution $p(\eta \mid x_1, \dots, x_p)$ is the normal distribution,

$$p(\eta \mid \mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(\eta - \mathbf{x}^T \boldsymbol{\gamma})^2\right),$$

and the marginal distribution $p(X_1, \dots, X_p)$ is given in equation (8.4), as

$$p(\mathbf{x}) = \prod_{i=1}^p p(x_i) = \prod_{i=1}^p \frac{\exp(x_i \mu_i)}{\sum_{x_i} \exp(\mu_i)},$$

where $\mu_i \in \mathbb{R}$ is an intercept. Together, these two distributions specify the joint distribution

$$p(\eta, \mathbf{x}) = p(\eta | \mathbf{x}) p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(\eta - \mathbf{x}^\top \boldsymbol{\gamma})^2\right) \prod_{i=1}^p \frac{\exp(x_i \mu_i)}{\sum_{x_i} \exp(\mu_i)}.$$

We derive the conditional distribution $p(\mathbf{x} | \eta)$ from this joint distribution,

$$\begin{aligned} p(\mathbf{x} | \eta) &= \frac{p(\eta, \mathbf{x})}{p(\eta)} \\ &= \frac{p(\eta | \mathbf{x}) p(\mathbf{x})}{\sum_{\mathbf{x}} p(\eta | \mathbf{x}) p(\mathbf{x})} \\ &= \frac{\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(\eta - \mathbf{x}^\top \boldsymbol{\gamma})^2\right) \prod_{i=1}^p \frac{\exp(x_i \mu_i)}{\sum_{x_i} \exp(\mu_i)}}{\sum_{\mathbf{x}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(\eta - \mathbf{x}^\top \boldsymbol{\gamma})^2\right) \prod_{i=1}^p \frac{\exp(x_i \mu_i)}{\sum_{x_i} \exp(\mu_i)}} \\ &= \frac{\exp\left(-\frac{1}{2\sigma^2}(\eta - \mathbf{x}^\top \boldsymbol{\gamma})^2 + \mathbf{x}^\top \boldsymbol{\mu}\right)}{\sum_{\mathbf{x}} \exp\left(-\frac{1}{2\sigma^2}(\eta - \mathbf{x}^\top \boldsymbol{\gamma})^2 + \mathbf{x}^\top \boldsymbol{\mu}\right)} \\ &= \frac{\exp\left(-\frac{1}{2\sigma^2}[\eta^2 - 2\eta \mathbf{x}^\top \boldsymbol{\gamma} + \mathbf{x}^\top \boldsymbol{\gamma} \boldsymbol{\gamma}^\top \mathbf{x}] + \mathbf{x}^\top \boldsymbol{\mu}\right)}{\sum_{\mathbf{x}} \exp\left(-\frac{1}{2\sigma^2}[\eta^2 - 2\eta \mathbf{x}^\top \boldsymbol{\gamma} + \mathbf{x}^\top \boldsymbol{\gamma} \boldsymbol{\gamma}^\top \mathbf{x}] + \mathbf{x}^\top \boldsymbol{\mu}\right)} \\ &= \frac{\exp\left(\frac{1}{\sigma^2} \eta \mathbf{x}^\top \boldsymbol{\gamma} - \frac{1}{2\sigma^2} \mathbf{x}^\top \boldsymbol{\gamma} \boldsymbol{\gamma}^\top \mathbf{x} + \mathbf{x}^\top \boldsymbol{\mu}\right)}{\sum_{\mathbf{x}} \exp\left(\frac{1}{\sigma^2} \eta \mathbf{x}^\top \boldsymbol{\gamma} - \frac{1}{2\sigma^2} \mathbf{x}^\top \boldsymbol{\gamma} \boldsymbol{\gamma}^\top \mathbf{x} + \mathbf{x}^\top \boldsymbol{\mu}\right)} \\ &= \frac{\exp\left(\mathbf{x}^\top \left(\frac{1}{\sigma^2} \eta \boldsymbol{\gamma} + \boldsymbol{\mu}\right) + \mathbf{x}^\top \left(-\frac{1}{2\sigma^2} \boldsymbol{\gamma} \boldsymbol{\gamma}^\top\right) \mathbf{x}\right)}{\sum_{\mathbf{x}} \exp\left(\mathbf{x}^\top \left(\frac{1}{\sigma^2} \eta \boldsymbol{\gamma} + \boldsymbol{\mu}\right) + \mathbf{x}^\top \left(-\frac{1}{2\sigma^2} \boldsymbol{\gamma} \boldsymbol{\gamma}^\top\right) \mathbf{x}\right)}, \end{aligned}$$

which is an Ising model with intercepts $\boldsymbol{\mu} + \frac{1}{\sigma^2} \eta \boldsymbol{\gamma}$ and a fully negative connectivity matrix $-\frac{1}{\sigma^2} \boldsymbol{\gamma} \boldsymbol{\gamma}^\top$.

To see that this is in fact an Ising model note that the model can be rewritten as follows. Let $\boldsymbol{\mu}^* = \boldsymbol{\mu} + \frac{1}{\sigma^2} \eta \boldsymbol{\gamma}$ denote the vector of conditional intercepts so that $\mu_i^* = \mu_i + \frac{1}{\sigma^2} \eta \gamma_i$ denotes the intercept for X_i conditional on η . Let $\boldsymbol{\Omega} = -\frac{1}{\sigma^2} \boldsymbol{\gamma} \boldsymbol{\gamma}^\top$ denote the conditional connectivity matrix, so that $\omega_{ij} = -\frac{1}{\sigma^2} \gamma_i \gamma_j$, represents the edge weight connecting X_i and X_j conditional on η . We can rewrite the above model as follows:

$$\begin{aligned} p(\mathbf{x} | \eta) &= \frac{\exp\left(\mathbf{x}^\top \left(\frac{1}{\sigma^2} \eta \boldsymbol{\gamma} + \boldsymbol{\mu}\right) + \mathbf{x}^\top \left(-\frac{1}{2\sigma^2} \boldsymbol{\gamma} \boldsymbol{\gamma}^\top\right) \mathbf{x}\right)}{\sum_{\mathbf{x}} \exp\left(\mathbf{x}^\top \left(\frac{1}{\sigma^2} \eta \boldsymbol{\gamma} + \boldsymbol{\mu}\right) + \mathbf{x}^\top \left(-\frac{1}{2\sigma^2} \boldsymbol{\gamma} \boldsymbol{\gamma}^\top\right) \mathbf{x}\right)} \\ &= \frac{\exp\left(\mathbf{x}^\top \boldsymbol{\mu}^* + \mathbf{x}^\top \boldsymbol{\Omega} \mathbf{x}\right)}{\sum_{\mathbf{x}} \exp\left(\mathbf{x}^\top \boldsymbol{\mu}^* + \mathbf{x}^\top \boldsymbol{\Omega} \mathbf{x}\right)} \\ &= \frac{\exp\left(\sum_{i=1}^p x_i \mu_i^* + \sum_{\langle ij \rangle} x_i x_j \omega_{ij}\right)}{\sum_{\mathbf{x}} \exp\left(\sum_{i=1}^p x_i \mu_i^* + \sum_{\langle ij \rangle} x_i x_j \omega_{ij}\right)}. \end{aligned}$$

Bibliography

- Agresti, A. (2007). *An introduction to categorical data analysis*. Hoboken, NJ: John Wiley & Sons, Inc.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: American Psychiatric Association.
- Anderson, T. W. (1958). *An introduction to multivariate statistical analysis* (Vol. 2). New York, NY: Wiley.
- Asmundson, G. J. G., Frombach, I., McQuaid, J., Pedrelli, P., Lenox, R., & Stein, M. B. (2000). Dimensionality of posttraumatic stress symptoms: A confirmatory factor analysis of DSM-IV symptom clusters and other symptom models. *Behaviour Research and Therapy*, 38, 203–214.
- Bartholomew, D. J., Deary, I. J., & Lawn, M. (2009). A new lease of life for Thomson's bonds model of intelligence. *Psychological Review*, 116, 567–579.
- Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.). London, England: Arnold. (Kendall's Library of Statistics 7).
- Bickel, P. J., & Doksum, K. A. (2015). *Mathematical statistics: Basic ideas and selected topics* (Vol. 1). Boca Raton, FL: CRC Press.
- Bilodeau, M., & Brenner, D. (1999). *Theory of multivariate statistics*. New York, NY: Springer-Verlag.
- 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, MA: Addison-Wesley.
- Bollen, K. A. (1989a). A new incremental fit index for general structural equation models. *Sociological Methods & Research*, 17, 303–316.

- Bollen, K. A. (1989b). *Structural equations with latent variables*. New York, NY: John Wiley and Sons.
- Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, 53, 605–634.
- Bollen, K. A. (2011). Evaluating effect, composite, and causal indicators in structural equation models. *MIS Quarterly*, 35, 359–372.
- Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods*, 16, 265–284.
- Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, 110, 305–314.
- Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 301–328). Dordrecht, the Netherlands: Springer.
- Bollen, K. A., & Ting, K.-F. (1993). Confirmatory tetrad analysis. *Sociological Methodology*, 23, 147–175.
- Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science*, 5, 184–186.
- Borsboom, D. (2002). The structure of the DSM [Letter to the editor]. *Archives of General Psychiatry*, 59, 569–570.
- Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology*, 64, 1089–1108.
- Borsboom, D. (2015). What is causal about individual differences?: A comment on Weinberger. *Theory & Psychology*, 25, 362–368.
- Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, 9, 91–121.
- Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, 110, 203–219.

- Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111, 1061–1071.
- Borsboom, D., & Molenaar, D. (2015). Psychometrics. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (pp. 418–422). Amsterdam, the Netherlands: Elsevier.
- Boschloo, L., van Borkulo, C. D., Rhemtulla, M., Keyes, K. M., Borsboom, D., & Schoevers, R. A. (2015). The network structure of symptoms of the diagnostic and statistical manual of mental disorders. *PLoS One*, 10, e0137621.
- Bridgman, P. W. (1927). *The logic of modern physics* (Vol. 3). New York, NY: Macmillan.
- Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21, 230–258.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY: Cambridge University Press.
- Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., ... others (2014). The p factor: One general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science*, 2, 119–137.
- Chen, B., & Pearl, J. (2013). Regression and causation: A critical examination of six econometrics textbooks. *Real-World Economics Review*, 65, 2–20.
- Chen, B., & Pearl, J. (2014). Graphical tools for linear structural equation modeling. Technical Report R-432 (Department of Computer Science, University of California, Los Angeles, CA). *Psychometrika*, in press. Available at <http://ftp.cs.ucla.edu/pub/stat_ser/r432.pdf>.
- Christofferson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, 40, 5–32.
- Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. New York, NY: Cambridge University Press.
- Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7, 309–319.

- Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine*, 119, 166.e7-166.e16.
- Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., & Cramer, A. O. J. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality*, 54, 13–29.
- Cramer, A. O. J., & Borsboom, D. (2015). Problems attract problems: A network perspective on mental disorders. *Emerging Trends in the Social and Behavioral Sciences*, 1–15, doi:10.1002/9781118900772.etrds0264.
- Cramer, A. O. J., van Borkulo, C. D., Giltay, E. J., van der Maas, H. L. J., Kendler, K. S., Scheffer, M., & Borsboom, D. (2016). Major depression as a complex dynamical system. *PLoS One*, 11, e0167490.
- Cramer, A. O. J., van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., ... Borsboom, D. (2012). Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality*, 26, 414–431.
- Cramer, A. O. J., Waldorp, L. J., van der Maas, H. L. J., & Borsboom, D. (2010). Comorbidity: A network perspective. *Behavioral and Brain Sciences*, 33, 137–150.
- Cramér, H. (1999). *Mathematical methods of statistics* (Vol. 9). Princeton, NJ: Princeton University Press.
- Cronbach, L. J. (1947). Test “reliability”: Its meaning and determination. *Psychometrika*, 12, 1–16.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Dalege, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M., & van der Maas, H. L. J. (2016). Toward a formalized account of attitudes: The causal attitude network (CAN) model. *Psychological Review*, 123, 2–22.

- Deary, I. J., Penke, L., & Johnson, W. (2010). The neuroscience of human intelligence differences. *Nature Reviews Neuroscience*, 11, 201–211.
- de Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer Science+Business Media, LLC.
- de Fátima Salgueiro, M., Smith, P. W. F., & McDonald, J. W. (2008). The manifest association structure of the single-factor model: Insights from partial correlations. *Psychometrika*, 73, 665–670.
- Dekker, K., Blanken, T. F., & van Someren, E. J. (2017). Insomnia and personality — A network approach. *Brain Sciences*, 7, doi:10.3390/brainsci7030028.
- Dempster, A. P. (1972). Covariance selection. *Biometrics*, 28, 157–175.
- Deserno, M. K., Borsboom, D., Begeer, S., & Geurts, H. M. (2017a). Multicausal systems ask for multicausal approaches: A network perspective on subjective well-being in individuals with autism spectrum disorder. *Autism*, 21, 960–971.
- Deserno, M. K., Borsboom, D., Begeer, S., & Geurts, H. M. (2017b). Relating ASD symptoms to well-being: Moving across different construct levels. *Psychological Medicine*, 48, 1–14, doi:10.1017/S0033291717002616.
- Detterman, D. K. (2002). General intelligence: Cognitive and biological explanations. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The general factor of intelligence: How general is it* (pp. 151–182). Mahwah, NJ: Erlbaum.
- Devlin, B., Daniels, M., & Roeder, K. (1997). The heritability of IQ. *Nature*, 388, 468–471.
- Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research*, 61, 1203–1218.
- Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology*, 73, 1246–1256.
- Duncan, O. D. (1975). *Introduction to structural equation models*. New York, NY: Academic Press.

- Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, 5, 155–174.
- Ellis, J. L. (2014). An inequality for correlations in unidimensional monotone latent variable models for binary variables. *Psychometrika*, 79, 303–316.
- Ellis, J. L., & Junker, B. W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika*, 62, 495–523.
- Ellis, J. L., & Van den Wollenberg, A. L. (1993). Local homogeneity in latent trait models. A characterization of the homogeneous monotone IRT model. *Psychometrika*, 58, 417–429.
- Elshout, J. J., & Akkerman, A. E. (1975). *Vijf persoonlijkheidsfactoren test 5PFT: Handleiding (The Five Personality Factor Test (5PFT): Manual)*. Nijmegen, the Netherlands: Berkhout B.V.
- Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40, 31–53.
- Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, 50, 195–212.
- Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, 48, 1–18.
- Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*, doi:10.1037/met0000167.
- Epskamp, S., Kruis, J., & Marsman, M. (2017). Estimating psychopathological networks: Be careful what you wish for. *PLoS One*, 12, e0179891.
- Epskamp, S., Maris, G., Waldorp, L. J., & Borsboom, D. (2018). Network psychometrics. In Irwing P., Hughes D., & Booth T. (Eds.), *The Wiley Handbook of Psychometric Testing, 2 Volume Set: A Multidisciplinary Reference on Survey, Scale and Test Development*. New York, NY: Wiley.

- Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika*, 82, 904–927.
- Epskamp, S., Waldorp, L. J., Möttus, R., & Borsboom, D. (2018). The gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research*, doi:10.1080/00273171.2018.1454823.
- Eysenck, H. J. (1983). Psychophysiology and personality: Extraversion, neuroticism and psychoticism. In A. Gale & J. A. Edwards (Eds.), *Physiological correlates of human behaviour: Individual differences and psychopathology* (pp. 13–30). London, England: Academic Press Inc.
- Ferguson, T. S. (1996). *A course in large sample theory*. London, England: Chapman and Hall.
- Fischer, G. (1995). Derivations of the rasch model. In G. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 15–38). New York, NY: Springer.
- Flanagan, D. P., & McGrew, K. S. (1998). Interpreting intelligence tests from contemporary Gf-Gc theory: Joint confirmatory factor analysis of the WJ-R and KAIT in a non-white sample. *Journal of School Psychology*, 36, 151–182.
- Foygel, R., & Drton, M. (2010). Extended bayesian information criteria for gaussian graphical models. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), *Advances in neural information processing systems* 23 (pp. 604–612).
- Foygel, R., & Drton, M. (2011). Bayesian model choice and information criteria in sparse generalized linear models. *arXiv preprint arXiv:1112.5635*.
- Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9, 432–441.
- Frigg, R., & Hartmann, S. (2006). Models in science. In S. Sarkar & J. Pfeifer (Eds.), *The philosophy of science: An encyclopedia* (pp. 397–479). New York, NY: Routledge.
- Garlick, D. (2002). Understanding the nature of the general factor of intelligence: The role of individual differences in neural plasticity as an explanatory mechanism. *Psychological Review*, 109, 116–136.

- Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: g as superordinate or breadth factor? *Psychology Science Quarterly*, 50, 21–43.
- Gignac, G. E. (2016). Residual group-level factor associations: Possibly negative implications for the mutualism theory of general intelligence. *Intelligence*, 55, 69–78.
- Gignac, G. E., & Watkins, M. W. (2013). Bifactor modeling and the estimation of model-based reliability in the WAIS-IV. *Multivariate Behavioral Research*, 48, 639–662.
- Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology*, 42, 139–167.
- Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 215–231.
- Gould, S. J. (1996). *The mismeasure of man*. New York, NY: W. W. Norton.
- Gray, J. R., & Thompson, P. M. (2004). Neurobiology of intelligence: Science and ethics. *Nature Reviews Neuroscience*, 5, 471–482.
- Greenland, S. (2003). Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology*, 14, 300–306.
- Gustafsson, J. E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, 8, 179–203.
- Guttman, L. (1940). Multiple rectilinear prediction and the resolution into components. *Psychometrika*, 5, 75–99.
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255–282.
- Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stoufer, L. Guttman, E. A. Suchman, P. L. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Studies in social psychology in world war II: Measurement and prediction* (Vol. 5, pp. 60–90). Princeton, NJ: Princeton University Press.
- Guttman, L. (1953). Image theory for the structure of quantitative variates. *Psychometrika*, 18, 277–296.

- Haig, B. D. (2005). An abductive theory of scientific method. *Psychological Methods*, 10, 371–388.
- Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. (1978). Developments in latent trait theory: Models, technical issues, and applications. *Review of Educational Research*, 48, 467–510.
- Harrell, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*. Cham, Switzerland: Springer.
- Haworth, C. M. A., Wright, M. J., Luciano, M., Martin, N. G., de Geus, E. J. C., Van Beijsterveldt, C. E. M., ... Plomin, R. (2010). The heritability of general cognitive ability increases linearly from childhood to young adulthood. *Molecular Psychiatry*, 15, 1112–1120.
- Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, 55, 577–601.
- Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, 14, 1523–1543.
- Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods*, 12, 205–218.
- Irwing, P., Booth, T., & Hughes, D. (Eds.). (2018). *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*. New York, NY: Wiley.
- Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31, 253–258.
- Isvoranu, A. M., van Borkulo, C. D., Boyette, L., Wigman, J. T., Vinkers, C. H., Borsboom, D., & Investigators, G. (2016). A network approach to psychosis: Pathways between childhood trauma and psychotic symptoms. *Schizophrenia Bulletin*, 43, 187–196.
- Jacobs, N., Van Gestel, S., Derom, C., Thiery, E., Vernon, P., Derom, R., & Vlietinck, R. (2001). Heritability estimates of intelligence in twins: Effect of chorion type. *Behavior Genetics*, 31, 209–217.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York, NY: Springer.

- Jeffrey, R. C. (1969). Statistical explanation vs. statistical inference. In N. Rescher (Ed.), *Essays in honor of Carl G. Hempel: A tribute on the occasion of his sixty-fifth birthday* (pp. 104–113). Dordrecht, the Netherlands: D. Reidel Publishing Company.
- Jeffreys, H. (1961). *The theory of probability* (3rd ed.). Oxford, England: Oxford University Press.
- Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.
- Johnson, M. S. (2007). Marginal maximum likelihood estimation of item response models in R. *Journal of Statistical Software*, 20, 1–24.
- Jonas, K. G., & Markon, K. E. (2016). A descriptivist approach to trait conceptualization and inference. *Psychological Review*, 123, 90–96.
- Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36, 109–133.
- Junker, B. W., & Ellis, J. L. (1997). A characterization of monotone unidimensional latent variable models. *The Annals of Statistics*, 25, 1327–1343.
- Kac, M. (1968). Mathematical mechanisms of phase transitions. In M. Chréten, E. P. Gross, and S. Deser (Eds.), *Statistical physics: Phase transitions and superfluidity, Brandeis university summer institute in theoretical physics*. (Vol. 1, pp. 241–305). New York, NY: Gordon and Breach Science Publishers.
- Klauer, K. C. (1995). The assessment of person fit. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 97–110). New York, NY: Springer.
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: The Guilford Press.
- Kment, B. (2010). Causation: Determination and difference-making. *Noûs*, 44, 80–111.
- Knight, K. (2000). *Mathematical statistics*. Boca Raton, FL: Chapman & Hall/CRC.
- Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. Cambridge, MA: MIT press.

- Krijnen, W. P. (2004). Positive loadings and factor correlations from positive covariance matrices. *Psychometrika*, 69, 655–660.
- Krueger, R. F. (1999). The structure of common mental disorders. *Archives of General Psychiatry*, 56, 921–926.
- Kruis, J., & Maris, G. (2016). Three representations of the Ising model. *Scientific Reports*, 6(34175), 1–11, doi: 10.1038/srep34175.
- Laceulle, O. M., Vollebergh, W. A. M., & Ormel, J. (2015). The structure of psychopathology in adolescence: Replication of a general psychopathology factor in the TRAILS study. *Clinical Psychological Science*, 3, 850–860.
- Lahey, B. B., Applegate, B., Hakes, J. K., Zald, D. H., Hariri, A. R., & Rathouz, P. J. (2012). Is there a general factor of prevalent psychopathology during adulthood? *Journal of Abnormal Psychology*, 121, 971–977.
- Lahey, B. B., Van Hulle, C. A., Singh, A. L., Waldman, I. D., & Rathouz, P. J. (2011). Higher-order genetic and environmental structure of prevalent forms of child and adolescent psychopathology. *Archives of General Psychiatry*, 68, 181–189.
- Langford, E., Schwertman, N., & Owens, M. (2001). Is the property of being positively correlated transitive? *The American Statistician*, 55, 322–325.
- Lansman, M., Donaldson, G., Hunt, E., & Yantis, S. (1982). Ability factors and cognitive processes. *Intelligence*, 6, 347–386.
- Lauritzen, S. L. (1996). *Graphical models* (Vol. 17). New York, NY: Oxford University Press.
- Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis* (Vol. 1). Boston, MA: Houghton Mill.
- Lewis, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.
- Lewis, D. (1986). Causation. In *Philosophical papers II* (pp. 158–269). Oxford, NY: Oxford University Press.
- Liao, T. F. (1994). *Interpreting probability models: Logit, probit, and other generalized linear models*. Thousand Oaks, CA: Sage.

- Loehlin, J. C., & Martin, N. G. (2011). The general factor of personality: Questions and elaborations. *Journal of Research in Personality*, 45, 44–49.
- Lord, F. M. (1952). *A theory of test scores*. New York, NY: Psychometric Society.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- MacCallum, R. C., & Browne, M. W. (1993). The use of causal indicators in covariance structure models: Some practical issues. *Psychological Bulletin*, 114, 533–541.
- MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, 114, 185–199.
- Markus, K. A. (2002). Statistical equivalence, semantic equivalence, eliminative induction, and the raykov-marcoulides proof of infinite equivalence. *Structural Equation Modeling*, 9, 503–522.
- Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L., ... Maris, G. (2018). An introduction to network psychometrics: Relating Ising network models to item response theory models. *Multivariate Behavioral Research*, 53, 15–35.
- Marsman, M., Maris, G., Bechger, T., & Glas, C. (2015). Bayesian inference for low-rank Ising networks. *Scientific Reports*, 5(09050), 1–7, doi: 10.1038/srep09050.
- McCrae, R. R. (2015). A more nuanced view of reliability: Specificity in the trait hierarchy. *Personality and Social Psychology Review*, 19, 97–112.
- McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52, 81–90.
- McCullagh, P. (2002). What is a statistical model? *Annals of Statistics*, 30, 1225–1267.
- McDonald, R. P. (2003). Behavior domains in theory and in practice. *Alberta Journal of Educational Research*, 49, 212–230.

- McNally, R. J., Robinaugh, D. J., Wu, G. W. Y., Wang, L., Deserno, M. K., & Borsboom, D. (2014). Mental disorders as causal systems: A network approach to posttraumatic stress disorder. *Clinical Psychological Science*, 3, 1–14. doi: 10.1177/2167702614553230.
- Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, 7, 105–118.
- Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, 13, 127–143.
- Mellenbergh, G. J. (1994a). Generalized linear item response theory. *Psychological Bulletin*, 115, 300–307.
- Mellenbergh, G. J. (1994b). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, 29, 223–236.
- Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, 1, 293–299.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525–543.
- Mill, J. S. (1843). *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation* (Vol. 1). London, England: John W. Parker, West Strand.
- Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling*, 2, 255–273.
- Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods*, 2, 248–260.
- Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, 72, 461–473.
- Mokken, R. J. (1971). *A theory and procedure of scale analysis with applications in political research*. The Hague, the Netherlands: Mouton.

- Molenaar, I. W. (1995). Some background for item response theory and the rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models* (pp. 3–14). New York, NY: Springer.
- Moneta, A., & Russo, F. (2014). Causal models and evidential pluralism in econometrics. *Journal of Economic Methodology*, 21, 54–76.
- Morgan, G. B., Hodge, K. J., Wells, K. E., & Watkins, M. W. (2015). Are fit indices biased in favor of bi-factor models in cognitive ability research?: A comparison of fit in correlated factors, higher-order, and bi-factor models via monte carlo simulations. *Journal of Intelligence*, 3, 2–20.
- Mõttus, R. (2016). Towards more rigorous personality trait–outcome research. *European Journal of Personality*, 30, 292–303.
- Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, 65, 391–411.
- Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence*, 41, 407–422.
- Musek, J. (2007). A general factor of personality: Evidence for the Big One in the five-factor model. *Journal of Research in Personality*, 41, 1213–1233.
- Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, 43, 551–560.
- Muthén, B., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus web notes*, 4, 1–22.
- Newsom, J. T. (2015). *Longitudinal structural equation modeling: A comprehensive introduction*. New York, NY: Routledge.
- Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 32, 1–13.
- Patalay, P., Fonagy, P., Deighton, J., Belsky, J., Vostanis, P., & Wolpert, M. (2015). A general psychopathology factor in early adolescence. *The British Journal of Psychiatry*, 207, 15–22.

- Payton, A. (2009). The impact of genetic research on our understanding of normal cognitive ageing: 1995 to 2009. *Neuropsychology Review*, 19, 451–477.
- Pe, M. L., Kircanski, K., Thompson, R. J., Bringmann, L. F., Tuerlinckx, F., Mestdagh, M., ... Gotlib, I. H. (2015). Emotion-network density in major depressive disorder. *Clinical Psychological Science*, 3, 292–300.
- Pearce, N., & Richiardi, L. (2014). Commentary: Three worlds collide: Berkson's bias, selection bias and collider bias. *International Journal of Epidemiology*, 43, 521–524.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.
- Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27, 226–284.
- Pearl, J. (2000a). *Causality: Models, reasoning and inference*. New York, NY: Cambridge University Press.
- Pearl, J. (2000b). Epilogue: The art and science of cause and effect. In *Causality: Models, reasoning and inference* (pp. 331–358). Cambridge, England: Cambridge University Press.
- Pearl, J. (2012). The causal foundations of structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling*. New York, NY: Guilford Press.
- Pettersson, E., Anckarsäter, H., Gillberg, C., & Lichtenstein, P. (2013). Different neurodevelopmental symptoms have a common genetic etiology. *Journal of Child Psychology and Psychiatry*, 54, 1356–1365.
- Pettersson, E., Larsson, H., & Lichtenstein, P. (2016). Common psychiatric disorders share the same genetic origin: A multivariate sibling study of the Swedish population. *Molecular Psychiatry*, 21, 717–721.
- Popper, K. R. (1959). *The logic of scientific discovery*. New York, NY: Basic Books.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute of Educational Research.

- Raykov, T., & Marcoulides, G. A. (2000). *A first course in structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Raykov, T., & Marcoulides, G. A. (2001). Can there be infinitely many models equivalent to a given covariance structure model? *Structural Equation Modeling*, 8, 142–149.
- Reckase, M. D. (2009). *Multidimensional item response theory models*. New York, NY: Springer.
- Reichenbach, H. (1956). *The direction of time*. Los Angeles: University of California Press.
- Residual. (n.d.). Entry: “residual”. In *Oxford English Dictionary*. Retrieved from <http://www.oed.com/view/Entry/163583>.
- Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, 49, 425–435.
- Rozeboom, W. W. (1966). *Foundations of the theory of prediction*. Homewood, IL: The Dorsey Press.
- Rozental, A., Forsell, E., Svensson, A., Forsström, D., Andersson, G., & Carlbring, P. (2014). Psychometric evaluation of the Swedish version of the pure procrastination scale, the irrational procrastination scale, and the susceptibility to temptation scale in a clinical population. *BMC Psychology*, 2, doi:10.1186/s40359-014-0054-z.
- Rue, H., & Held, L. (2005). *Gaussian Markov random fields: Theory and applications*. Boca Raton, FL: Chapman and Hall/CRC.
- Rushton, J. P., Bons, T. A., & Hur, Y.-M. (2008). The genetics and evolution of the general factor of personality. *Journal of Research in Personality*, 42, 1173–1185.
- Rushton, J. P., & Irwing, P. (2008). A general factor of personality (GFP) from two meta-analyses of the Big Five: Digman (1997) and Mount, Barrick, Scullen, and Rounds (2005). *Personality and Individual Differences*, 45, 679–683.
- Russo, F., & Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, 21, 157–170.

- Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, 16, 561–582.
- Sartorio, C. (2005). Causes as difference-makers. *Philosophical Studies*, 123, 71–96.
- Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., ... Sugihara, G. (2009). Early-warning signals for critical transitions. *Nature*, 461, 53–59.
- Scheffer, M., Carpenter, S., Foley, J. A., Folke, C., & Walker, B. (2001). Catastrophic shifts in ecosystems. *Nature*, 413, 591–596.
- Schmittmann, V. D., Cramer, A. O. J., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology*, 31, 43–53.
- Sharp, C., Wright, A. G. C., Fowler, J. C., Frueh, B. C., Allen, J. G., Oldham, J., & Clark, L. A. (2015). The structure of personality pathology: Both general ('g') and specific ('s') factors? *Journal of Abnormal Psychology*, 124, 387–398.
- Shealy, R., & Stout, W. (1993a). An item response theory model for test bias and differential test functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197–240). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Shealy, R., & Stout, W. (1993b). A model-based standardization approach that separates true bias/dif from group ability differences and detects test bias/dif as well as item bias/dif. *Psychometrika*, 58, 159–194.
- Sherman, J., & Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21, 124–127.
- Siegler, R. S., & Alibali, M. W. (2005). *Children's thinking* (4th ed.). Upper Saddle River, NJ: Prentice Hall.
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of cronbach's alpha. *Psychometrika*, 74, 107–120.
- Sijtsma, K., & van der Ark, L. A. (2015). Conceptions of reliability revisited and practical recommendations. *Nursing Research*, 64, 128–136.

- Silk, A. J. (1977). Test-retest correlations and the reliability of copy testing. *Journal of Marketing Research*, 14, 476–486.
- Simms, L. J., Grös, D. F., Watson, D., & O'hara, M. W. (2008). Parsing the general and specific components of depression and anxiety with bifactor modeling. *Depression and Anxiety*, 25, E34–E46.
- Smits, I. A. M., Dolan, C. V., Vorst, H. C., Wicherts, J. M., & Timmerman, M. E. (2013). Data from 'cohort differences in big five personality factors over a period of 25 years'. *Journal of Open Psychology Data*, 1, 1–3, doi:10.5334/jopd.e2.
- Snyder, H. R., Young, J. F., & Hankin, B. L. (2017). Strong homotypic continuity in common psychopathology-, internalizing-, and externalizing-specific factors over time in adolescents. *Clinical Psychological Science*, 5, 98–110.
- Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, 27, 229–239.
- Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, 15, 201–292.
- Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York, NY: Macmillan.
- Steel, P. (2010). Arousal, avoidant and decisional procrastinators: Do they exist? *Personality and Individual Differences*, 48, 926–934.
- Suppes, P. C. (1970). *A probabilistic theory of causality*. Amsterdam, the Netherlands: North-Holland Publishing Company.
- Svartdal, F. (2017). Measuring procrastination: Psychometric properties of the Norwegian versions of the irrational procrastination scale (IPS) and the pure procrastination scale (PPS). *Scandinavian Journal of Educational Research*, 61, 18–30.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361–370.

- Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52, 393–408.
- Terracciano, A., Sutin, A. R., McCrae, R. R., Deiana, B., Ferrucci, L., Schlessinger, D., ... Costa Jr., P. T. (2009). Facets of personality linked to underweight and overweight. *Psychosomatic Medicine*, 71, 682–689.
- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51, 567–577.
- Thomson, G. H. (1950). *The factorial analysis of human ability* (4th ed.). London, England: University of London Press.
- Thorndike, R. L. (1927). *The measurement of intelligence*. New York, NY: Teachers College.
- Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory & Psychology*, 19, 579–599.
- Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika*, 11, 1–13.
- Van der Vaart, A. W. (1998). *Asymptotic statistics*. New York, NY: Cambridge University Press.
- Van Bork, R., Grasman, R. P. P. P., & Waldorp, L. J. (2018). Unidimensional factor models imply weaker partial correlations than zero-order correlations. *Psychometrika*, 83, 443–452.
- Van Bork, R., Rhemtulla, M., & Borsboom, D. (2015). Latent variable and network model implications for partial correlation structures. Presented at the 80th Annual Meeting of the Psychometric Society (IMPS), Beijing, July 2015.
- Van Bork, R., Rhemtulla, M., Waldorp, L. J., & Borsboom, D. (2016). Distinguishing latent variable models and network models. Presented at the 28th Annual Convention of the Association for Psychological Science (APS), Chicago, May 2016.
- Van Borkulo, C., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A., & Waldorp, L. J. (2014). A new method for constructing networks from binary data. *Scientific Reports*, 4(5918), doi:10.1038/srep05918.

- Van Borkulo, C., Boschloo, L., Borsboom, D., Penninx, B. W. J. H., Waldorp, L. J., & Schoevers, R. A. (2015). Association of symptom network structure with the course of longitudinal depression. *JAMA Psychiatry*, 72, 1219–1226. doi:10.1001/jamapsychiatry.2015.2079.
- Van der Linden, W. J., & Hambleton, R. K. (Eds.). (2013). *Handbook of modern item response theory*. New York, NY: Springer Science & Business Media.
- Van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, 113, 842–861.
- Van der Maas, H. L. J., & Kan, K. J. (2016). Comment on ‘Residual group-level factor associations: Possibly negative implications for the mutualism theory of general intelligence’ by Gilles E. Gignac (2016). *Intelligence*, 57, 81–83.
- Van der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, 118, 339–356.
- Van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford, England: Oxford University Press.
- Veselka, L., Schermer, J. A., Petrides, K., & Vernon, P. A. (2009). Evidence for a heritable general factor of personality in two studies. *Twin Research and Human Genetics*, 12, 254–260.
- Weinberger, N. (2015). If intelligence is a cause, it is a within-subjects cause. *Theory & Psychology*, 25, 346–361.
- Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength & Conditioning Research*, 19, 231–240.
- Williamson, J. (2006). Causal pluralism versus epistemic causality. *Philosophica*, 77, 69–96.
- Williamson, J. (2009). Probabilistic theories of causality. In H. Beebe, C. Hitchcock, & P. Menzies (Eds.), *The Oxford handbook of causation* (pp. 185–212). Oxford, England: Oxford University Press.

- Wood, D., Gardner, M. H., & Harms, P. D. (2015). How functionalist and process approaches to behavior can explain trait covariation. *Psychological Review*, 122, 84–111.
- Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.
- Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, 64, 113–128.

Nederlandse samenvatting

De psychologie bestudeert concepten die niet direct waarneembaar zijn, zoals karaktereigenschappen (bijvoorbeeld extraversie of zorgvuldigheid), cognitieve vaardigheden (bijvoorbeeld rekenvaardigheid of geheugencapaciteit) als ook mentale stoornissen (bijvoorbeeld depressie of een angststoornis). Hoewel deze mentale eigenschappen niet direct waarneembaar zijn, worden ze weerspiegeld in observeerbaar gedrag. Bijvoorbeeld, het observeerbare gedrag dat iemand vaak naar feestjes gaat, spraakzaam is en veel vrienden heeft, kan worden gezien als een indicatie dat deze persoon extravert is. Een ander voorbeeld is dat, hoewel rekenvaardigheid niet direct kan worden waargenomen, het wel mogelijk is iemand een test te geven met verschillende rekenpuzzels en te observeren welke van deze puzzels worden opgelost. Zulk waarneembaar gedrag kan gebruikt worden om de onderliggende eigenschap te *meten*. De psychometrie richt zich op het ontwikkelen van theorieën over hoe mentale eigenschappen kunnen worden gemeten aan de hand van observeerbaar gedrag.

Een belangrijke vraag hierbij is *hoe* mentale eigenschappen worden gekoppeld aan waarneembaar gedrag. Hier zijn verschillende *psychometrische theorieën* over. Psychometrische theorieën hebben betrekking op het mechanisme dat de geobserveerde gedragingen aan elkaar verbindt en aan het mentale concept verbindt dat men via deze observeerbare gedragingen probeert te meten. Deze psychometrische theorieën kunnen worden uitgedrukt in statistische modellen die vervolgens gebruikt kunnen worden om data van geobserveerd gedrag te analyseren. De modellen waarin deze theorieën uitgedrukt kunnen worden dwingen echter geen theorie af over het mechanisme dat aan de data ten grondslag ligt. Het model kan in plaats van een psychometrische theorie over het psychologische concept, ook worden geïnterpreteerd als een manier om de relaties tussen variabelen in de data zo goed mogelijk te beschrijven. Ik beargumenteer in dit proefschrift dat er voordelen zijn van een interpretatie van psychometrische modellen als representaties van causale mechanismen en gebruik deze interpretatie om twee typen modellen met elkaar te vergelijken die bij twee verschillende psychometrische theorieën passen.

De eerste psychometrische theorie waar ik me in dit proefschrift op richt is de *latente-trektheorie*. In de latente-trektheorie is de mentale eigenschap een onderliggende, gemeenschappelijke oorzaak van het geobserveerde gedrag: zij gaat graag naar feestjes/is veel aan het woord/heeft veel

Figuur 10.1: Vierkanten geven observeerbare gedragingen aan, de cirkel geeft een latente trek aan. (a) representeert een latente-trektheorie van extraversie, waarin extraversie een oorzaak is van het gedrag. (b) representeert een netwerktheorie van extraversie, waarin het gedrag elkaar versterkt.

vrienden *omdat* zij extravert is. En *omdat* al deze gedragingen door extraversie worden veroorzaakt, hangen de gedragingen met elkaar samen: mensen die graag naar feestjes gaan hebben ook veel vrienden en zijn spraakzaam. De samenhang tussen deze gedragingen weerspiegelt de eigenschap extraversie omdat extraversie een gemeenschappelijke oorzaak is. Een voorbeeld hiervan is weergegeven in Figuur 10.1 (a).

De tweede psychometrische theorie waar dit proefschrift zich op richt is de *psychologische netwerktheorie*. In de netwerktheorie hangt gedrag met elkaar samen omdat het elkaar versterkt en is het psychologische concept een naam voor dit netwerk aan gedragingen. Bijvoorbeeld, omdat iemand graag naar feestjes gaat ontmoet deze persoon makkelijk nieuwe mensen, en omdat zij op een feestje makkelijker in de gelegenheid komt om een praatje aan te knopen maakt ze makkelijker nieuwe vrienden. Deze nieuwe vrienden geven ook feestjes en zo komt ze in de gelegenheid om naar meer feestjes te gaan waar ze weer makkelijk praatjes aanknoopt en nieuwe mensen leert kennen. Dit cluster van gedragingen die elkaar beïnvloeden, en we daarom vaak samen zien gaan, geven we vervolgens de naam ‘extraversie’. Het psychologische concept is hierin dus een netwerk van gedragingen die elkaar versterken of afzwakken. Dit voorbeeld is weergegeven in Figuur 10.1 (b).

De latente-trektheorie kan worden uitgedrukt in een *latente variabele model* waarin de latente variabele die geschat wordt alle samenhang tussen de geobserveerde variabelen in de data verklaart. De netwerktheorie kan worden uitgedrukt in een *netwerkmodel* waarin variabelen

Figuur 10.2: Een causale interpretatie van psychometrische modellen legt een relatie tussen het statistische model en de psychometrische theorie. De linker afbeelding weergeeft deze relatie in abstracte vorm. De rechter afbeelding past deze relatie toe op extraversie.

worden verbonden in het netwerk als de samenhang tussen deze variabelen niet volledig verklaard kan worden door de andere variabelen in het netwerk.

Hoewel psychometrische theorieën kunnen worden aangedragen als een onderbouwing om voor een bepaald statistisch model te kiezen, dwingt het statistische model geen psychometrische theorie af. Zo hoeft men bij het gebruik van een netwerkmodel niet te geloven in directe causale invloeden tussen de gedragingen en hoeft men bij het gebruik van het latente variabele model niet te geloven in de mentale eigenschap als een gemeenschappelijke oorzaak van het gedrag. Het model kan ook worden geïnterpreteerd als een samenvatting van de data in plaats van als een representatie van het onderliggende mechanisme waardoor de data, en de geobserveerde relaties tussen variabelen in de data, tot stand zijn gekomen. Het latente variabele model vat samen wat alle geobserveerde variabelen in de data met elkaar delen, terwijl een netwerkmodel beschrijft wat ieder paar van twee variabelen nog met elkaar deelt nadat er wordt gecorrigeerd voor het deel van de samenhang dat door andere variabelen in het netwerk wordt verklaard. In deze opvatting van psychometrische modellen zijn de modellen twee verschillende manieren om de relaties tussen variabelen in de data samen te vatten met een focus op verschillende vormen van samenhang tussen de variabelen.

Het interpreteren van het model als een representatie van het mechanisme dat aan de data ten grondslag ligt, noem ik een causale interpretatie van het model. Door het toevoegen van een causale interpretatie wordt het model aan een psychometrische theorie gekoppeld zoals weergegeven in Figuur 10.2. Hoewel een causale interpretatie van het model niet noodzakelijk is, zet dit proefschrift verschillende voordelen van een causale interpretatie van psychometrische modellen uiteen.

Het proefschrift bestaat uit twee delen. Het eerste deel van het proefschrift gaat in op een

causale interpretatie van het latente variabele model. In hoofdstuk 1 en 2 worden een aantal voordelen besproken van een causale interpretatie van psychometrische modellen. Ten eerste, een causale interpretatie van het model verbindt het model aan een theorie over de werkelijkheid die aan de werkelijkheid getoetst kan worden en dus gefalsificeerd kan worden. Een enkel statistische interpretatie van het model impliceert dat het model alleen betrekking heeft op de data, zonder dat er een relatie met een theorie over de werkelijkheid buiten de verzamelde data wordt gelegd. Neem als voorbeeld een latente variabele model dat geschat is uit data van depressiesymptomen. Het statistische model als beschrijving of samenvatting van de data is niet falsificeerbaar. Als het model niet goed op de data past kun je concluderen dat de data niet goed in een enkele dimensie samen te vatten is, maar het zegt niets over de aannemelijkheid van een theorie.

Wanneer je echter het latente variabele model interpreteert als een representatie van een mechanisme waarin deze symptomen een gemeenschappelijke oorzaak hebben, dan zegt hoe goed het latente variabele model op de data past iets over de aannemelijkheid van de theorie dat er een gemeenschappelijke oorzaak aan de data ten grondslag ligt. Als het model slecht past dan schaaft dit de aannemelijkheid van de theorie. Als het model goed past geeft dit ondersteuning voor de theorie, maar het betekent niet dat de theorie die het representeert waar moet zijn; misschien is er een alternatief mechanisme dat de geobserveerde structuur in de data ook kan verklaren. In sommige gevallen kan een experiment dan extra ondersteuning voor of tegen de theorie leveren. De theorie dat alle samenhang tussen de symptomen door een gemeenschappelijke oorzaak wordt verklaard, impliceert dat de samenhang niet het gevolg is van directe causale invloeden tussen de symptomen. Men kan in een experiment invloed uitoefenen op een symptoom en observeren of andere symptomen hierdoor ook veranderen. De conclusie uit dit experiment heeft dan betrekking op de plausibiliteit van de theorieën die door de modellen gerepresenteerd wordt, niet op de modellen als samenvatting van de data. Dat wil zeggen, misschien kunnen een latente variabele model en netwerkmodel even goed de data samenvatten, maar pleit de uitkomst van het experiment voor de aannemelijkheid van het ene mechanisme over het andere mechanisme.

Een ander argument voor een causale interpretatie van psychometrische modellen is dat door een model te verbinden aan een theorie over de werkelijkheid onderzoekers die met verschillende data werken toch dezelfde theorie kunnen onderzoeken. Neem bijvoorbeeld twee onderzoekers die beide data hebben verzameld van depressiesymptomen maar niet van dezelfde symptomen, of de data zijn in verschillende regio's verzameld. Als de modellen die zij uit deze data schatten een theorie over depressie representeren, dan kan de kennis uit beide studies gecombineerd worden en op elkaar voortbouwen. Maar, wanneer de modellen die zij schatten enkel een samenvatting van de data voorstellen en geen betrekking hebben op een mechanisme *buiten* de

data, dat zij *door middel van* de data proberen te begrijpen, dan heeft de kennis uit ieder van deze onderzoeken ook geen betrekking op *hetzelfde* mechanisme, waardoor de kennis ook niet op elkaar voortbouwt.

Verder kan een causale interpretatie ook een rechtvaardiging bieden voor keuzes die worden gemaakt bij het schatten van een model (zoals bepaalde restricties die aan het model worden opgelegd). Zo biedt de theorie dat er een gemeenschappelijke oorzaak aan de data ten grondslag ligt en je geïntereiseerd bent in het meten van deze gemeenschappelijke oorzaak, een onderbouwing voor het enkel meenemen van gedeelde variantie tussen geobserveerde variabelen en niet de unieke variantie in de geobserveerde variabelen. Een gemeenschappelijke oorzaak wordt immers weerspiegeld in wat de effecten delen. Wanneer men het model als een beschrijving van de data interpreteert kan nog steeds de keuze worden gemaakt een latente variabele te schatten die de gemeenschappelijke maar niet de unieke variantie verklaart, maar ontbreekt een onderbouwing van deze keuze op basis van theorie.

In deel 2 van dit proefschrift vergelijk ik het latente variabele model met het netwerkmodel. Zoals eerder beschreven, passen deze twee modellen bij verschillende theorieën over hoe de relaties tussen geobserveerde variabelen tot stand komen (respectievelijk door een gemeenschappelijke oorzaak of door directe invloeden tussen de variabelen). Ondanks dat de modellen hierin zo verschillen zijn de modellen wiskundig equivalent aan elkaar. Dat betekent dat een netwerkmodel altijd om te schrijven is naar een latente variabele model en andersom. Hoewel men hier uit zou kunnen concluderen dat de modellen hetzelfde zijn en het dus niet zinvol is om de modellen met elkaar te vergelijken, maakt een causale interpretatie de modellen verschillend van elkaar. Twee modellen die wiskundig in elkaar om te schrijven zijn kunnen immers een hele andere theorie over het achterliggende causale mechanisme representeren. Als een netwerk, geschat uit data van symptomen, de theorie representeert dat de symptomen elkaar beïnvloeden, terwijl het latente variabele model de theorie representeert dat de symptomen een gemeenschappelijke oorzaak hebben en elkaar niet beïnvloeden, dan kan een experiment waarin een symptoom wordt beïnvloed en er wordt gekeken naar het effect op andere symptomen, ondersteuning bieden voor een van deze theorieën, ondanks dat de twee theorieën uitgedrukt kunnen worden in twee modellen die wiskundig in elkaar om te schrijven zijn.

Ik beargumenteer dat wanneer je het netwerkmodel en latente variabele model als causale modellen (in plaats van beschrijvingen van de data) met elkaar vergelijkt, het belangrijk is om de interpreteerbaarheid van restricties aan het model te beoordelen. Bijvoorbeeld, de aanname dat de geobserveerde variabelen niet meer correleren als de latente variabele constant wordt gehouden (dit heet *lokale onafhankelijkheid*) leidt tot bepaalde restricties in het latente variabele

model. Wanneer deze restricties er toe leiden dat het model niet goed op de data past, dan wordt daarmee de aanname van lokale onafhankelijkheid geschonden en kan dit resulteren in het verworpen van dit model. De aanname van lokale onafhankelijkheid volgt uit de theorie dat er een gemeenschappelijke oorzaak is. Immers, als extraversie de oorzaak is voor het gedrag ‘spraakzaam zijn’ en ‘graag naar feestjes gaan’, dan correleren de mate van spraakzaamheid en hoeveel men naar feestjes gaat *omdat* mensen die extravert zijn meer van beide gedragingen laten zien dan mensen die minder extravert zijn. Wanneer alleen mensen in beschouwing worden genomen met precies dezelfde mate van extraversie, dan zal meer naar feestjes gaan niet meer correleren met de mate van spraakzaamheid; de variabelen zijn onafhankelijk gegeven dezelfde mate van extraversie.

Een latente variabele model met deze restricties is equivalent aan een netwerkmodel waarbij de restricties in het latente variabele model vertaald zijn naar restricties in het netwerkmodel. Hoewel dit netwerkmodel en latente variabele model wiskundig equivalent zijn, vormen de restricties een interpreteerbare hypothese in het latente variabele model maar is het niet duidelijk welke hypothese deze restricties representeren in het netwerkmodel. Het is bijvoorbeeld niet duidelijk waarom lokale onafhankelijkheid verwacht zou worden bij variabelen die elkaar direct beïnvloeden. Als het bezoeken van feestjes directe invloed heeft op het aantal nieuwe vrienden dat je maakt, dan correleren deze variabelen ook binnen een groep mensen met dezelfde mate van extraversie. Bij het vergelijken van netwerkmodellen en latente variabele modellen is het dus ook van belang om mee te nemen of de restricties die getoetst worden wanneer de modellen op de data worden gepast betekenis hebben in beide theorieën.

Vervolgens laat ik zien dat als men specifieke groepen netwerkmodellen (alleen netwerkmodellen waarin niet alle variabelen met alle andere variabelen verbonden zijn) vergelijkt met specifieke groepen latente variabele modellen (alleen latente variabele modellen met een enkele latente variabele en waar lokale onafhankelijkheid niet geschonden wordt), het wel mogelijk wordt om deze te onderscheiden aan de hand van geobserveerde data.

In de discussie specificeer ik ten slotte wat ik bedoel met *causale relaties* tussen variabelen in een psychometrisch model en ga ik in op een aantal implicaties van de causale interpretatie van psychometrische modellen. De belangrijkste implicatie is dat een causale interpretatie van het psychometrische model een kritische beoordeling behoeft van in hoeverre het causale mechanisme dat door het model gerepresenteerd wordt, overeenkomt met het ware causale mechanisme dat aan de data ten grondslag ligt. Als het model een samenvatting van de data is, dan is een dergelijke overeenkomst niet nodig; het ware mechanisme hoeft immers geen reden te zijn om voor een bepaalde samenvatting te kiezen (bijvoorbeeld als netwerkmodel of als latente va-

riabele model). Als het model echter causaal geïnterpreteerd wordt, dan is het belangrijk om na te gaan of de causale relaties die in het model worden weergegeven overeenkomen met causale relaties tussen deze variabelen in de werkelijkheid.

Dit proefschrift laat zien dat het belangrijk is om na te denken over de keuze om het psychometrische model als representatie van een causaal mechanisme te interpreteren of als beschrijving van de data en verschaft inzicht in de implicaties van deze keuze.

Summary

The field of psychology is concerned with constructs that cannot be observed directly, such as personality traits (e.g., extraversion or conscientiousness), cognitive abilities (e.g., math ability or memory capacity), and mental disorders (e.g., depression or anxiety disorders). Although these constructs cannot be observed directly, they are reflected in observable behavior. For example, the observable behavior of going to parties, being talkative, and having lots of friends can be taken as an indication that a person is extraverted. Similarly, although math ability cannot be observed directly, it is possible to subject a person to a test with a number of math problems and observe which math problems are solved. Such observable behavior can be used to *measure* the psychological construct. The field of psychometrics aims to develop theories on how to measure psychological constructs through observable behavior.

An important question that follows is *how* the psychological construct is related to observed behavior. Psychometric theories are concerned with the mechanism that connects observable behaviors to each other and to the psychological construct at hand. These psychometric theories can be expressed in statistical models, which can then be used to analyze data of observed behavior. However, the statistical models in which these theories can be expressed do not necessarily stipulate a theory on the data-generating mechanism. That is, the model can be interpreted as a representation of a psychometric theory, but it can also be understood as merely a way to summarize the relations between variables in the data. In this dissertation, I consider the interpretation of psychometric models as representations of causal mechanisms and provide several benefits of such a causal interpretation of psychometric models. I build on this interpretation to compare two types of models that correspond to two different psychometric theories.

The first psychometric theory I focus on in this dissertation is the *latent trait theory*. Latent trait theory understands psychological constructs as underlying common causes of observed behavior that explain the associations between certain behaviors: she often goes to parties, is talkative and easily makes new friends *because* she is extraverted. And *because* all of these behaviors are caused by extraversion, these behaviors are associated: people who often go to parties tend to easily make new friends and are more talkative. This example is presented in Figure 10.3(a).

Figure 10.3: Squares represent observable behaviors, the circle represents a latent trait. Panel (a) represents a latent trait theory of extraversion, in which extraversion causes the behaviors. Panel (b) represents a network theory of extraversion, in which the behaviors reinforce each other.

The second psychometric theory I focus on in this dissertation is the *psychological network theory*. In this theory, behaviors correlate because they mutually reinforce each other and the psychological construct refers to the resulting cluster of associated behaviors. For example, because someone often goes to parties, she easily meets new people, and because she is talkative she easily makes new friends. These new friends also invite her for parties, and so she has the opportunity to go to more parties and meet more people. This cluster of behaviors that impact each other and therefore tend to co-occur is called extraversion. The psychological construct is a network of behaviors that reinforce or weaken each other. This example is presented in Figure 10.3(b).

The latent trait theory can be expressed in a latent variable model in which the estimated latent variable explains all correlations between observed variables in the data. The network theory can be expressed in a network model in which variables are connected if their correlation cannot be fully explained by other variables in the network.

Psychometric models can be used as a rationale for using a certain statistical model, but statistical models do not stipulate a psychometric theory. For example, one can use a network model without committing to the theory of direct causal relations between behaviors, and one can use a latent variable model without committing to the theory of a mental trait as a common cause. Instead, these models can be interpreted as mere descriptions of the data rather than a representation of the data-generating mechanism. A latent variable model summarizes the shared

Figure 10.4: A causal interpretation of psychometric models relates the statistical model to a psychometric theory. The left image represents this relation abstractly. The right image applies this relation between a statistical model and psychometric theory to the extraversion.

variance of the variables in the data while a network model describes the association between any pair of variables that is not explained by the other variables in the network. According to this interpretation, different psychometric models refer to different ways of summarizing the correlation structure in the data. In this dissertation, I instead argue for what I call a *causal interpretation* of psychometric models, in which psychometric models are interpreted as a representation of the data-generating mechanism. A causal interpretation relates the psychometric model to a psychometric theory as is visualized in Figure 10.4. This causal interpretation is not necessary, but in this dissertation, I outline several benefits of it.

The dissertation consists of two parts. The first part zooms in on a causal interpretation of the latent variable model. Chapter 1 and 2 identify several benefits of a causal interpretation. A causal interpretation of a psychometric model connects the model to a theory about a mechanism in the external world that can be tested and thus can be falsified. A mere statistical interpretation implies that a model only relates to the data, and proposes no theory on the relation between the data and the external world. Take for example a unidimensional latent variable model estimated from data of depression symptoms. A statistical model as a description of these data is not falsifiable. If the model does not fit the data this implies that the data cannot be summarized in a single dimension, but the fit does not provide evidence for or against the plausibility of a theory. If, on the other hand, the model is interpreted as a representation of a mechanism in which the symptoms have a common cause, then model fit tells something about the plausibility that the data is truly generated by a common cause. A poor model fit makes this theory less plausible. A good model fit supports this theory but does not mean that the data-generating mechanism represented by the model is true, as there might be different data-generating mechanisms that imply the same correlational structure in the data. In some cases,

experimental intervention can provide additional evidence for or against the theory about the data-generating mechanism represented by the model. The theory that all associations between variables in the data are due to their shared reliance on a common cause implies that these variables do not cause each other directly, and thus an experimental intervention on one of the variables should not affect the other variables. Such an experiment is concerned with the plausibility of the theory that is represented by the model, not with the model as a mere description of the data. So, even though two theories may summarize data equally well, an experiment could speak to the plausibility of one mechanism over the other.

A causal interpretation of psychometric models also allows for the analysis of two different datasets to build on the same theory. Consider for example two researchers who both gathered data of depression symptoms, but who gathered the data in a different region or for a different set of depression symptoms. If the models that are estimated from these data refer to a theory of depression, then insights from both data sets complement each other. If, on the other hand, the models are only a description of the data and do not relate to a causal mechanism external to the observed data, then the insights of the two models do not speak to the same theory and do not build on each other.

A causal interpretation of psychometric models also offers a justification for why one makes certain choices in estimating the model (e.g., choices on which restrictions to impose on the model). For example, the theory of a common cause offers a justification for choosing to only take into account the shared variance among observed variables and not the unique variance in each of the observed variables. After all, a common cause is reflected in what is shared between its effects, and thus if the psychological construct is a common cause of the variables in the data, then the shared variance between these variables reflects this psychological construct. A descriptive interpretation of the model still allows such restrictions, but does not provide a rationale for imposing them.

In part 2 of this dissertation I compare the latent variable model to the network model. As described before, these two models correspond to different theories about how the relations between observed variables are generated (through a common cause or through direct effects respectively). Despite this difference, these models are mathematically equivalent. This means that a network model could always be reformulated into a latent variable model and vice versa. This might be taken to suggest that these models are *the same* and thus that comparing between the two models is pointless. However, two models that are mathematically equivalent might still imply a very different causal mechanism. A causal interpretation thus distinguishes the two models from one another. If a network, estimated from data of symptoms, represents the

theory that symptoms cause each other while the latent variable model represents the theory that the symptoms share a common cause, then an experiment in which one of the symptoms is manipulated could provide support for one of these theories even if the models, in which these theories are expressed, are mathematically equivalent.

When comparing the network model and latent variable model as causal models (instead of different ways to summarize the data), it is important to evaluate the meaning of the constraints in the model. Take for example the assumption that the observed variables no longer correlate conditioning on the latent variable (i.e., local independence). This assumption imposes restrictions on the latent variable model. When these restrictions result in poor fit of the model, one can conclude that the assumption of local independence is violated, which can be a reason to reject the model. These restrictions have meaning in the latent trait theory as they express the hypothesis that the observed variables are conditionally independent given the latent variable, which follows from the theory of a common cause underlying these variables. After all, if extraversion is a cause of ‘being talkative’ and ‘often going to parties’, then these variables correlate because people that are more extraverted are both more talkative and also go to more parties than people that are less extraverted. In a group of people with the exact same level of extraversion (i.e., conditioning on extraversion), being talkative and often going to parties no longer correlates.

This latent variable model, in which local independence is included as a constraint on the model, is mathematically equivalent to a network model in which the constraint on the latent variable model is translated to a constraint on the network model. This constraint in the equivalent network model, however, lacks a clear interpretation in the network theory. If going to parties directly affects making new friends then these variables can still be correlated within a group of people with the same level of extraversion. I conclude that when comparing network models and latent variable models as representations of a theory about the data-generating mechanism, it is important to consider what the constraints on the models mean within these theories.

Subsequently, I show that it is possible to distinguish specific groups of latent variable models and network models empirically. I compare sparse network models (i.e., network models that are not fully connected) to unidimensional latent variable models and show that these groups of models have diverging implications for empirical data.

In the discussion, I identify different understandings of how to interpret a causal relation between variables in the context of psychometric models and further elaborate on some of the implications of a causal interpretation of psychometric models. The most important implication is that a causal interpretation calls for a critical assessment of the degree to which the data-

generating mechanism represented by the model corresponds to the causal mechanism that actually generated the data. If the model is a description of the data, then there is no need to assess such a correspondence, but if the model is interpreted as a causal model, it is important to evaluate to what extent this causal model approximates the causal relations between these variables in the real world.

This dissertation shows the importance of thinking through the choice of interpreting psychometric models either as a representation of a causal mechanism or as a description of the data and provides insights in the implications of that choice.

Publications

Chapter 2 accepted for publication as:

Van Bork*, R., Wijsen*, L. D., & Rhemtulla, M. (in press). Toward a causal interpretation of the common factor model. *Disputatio*.

RVB and LDW wrote the draft of the article and prepared the article for submission. MR provided feedback on the manuscript and wrote parts of the article. RVB, LDW and MR revised the article.

Chapter 3 in preparation as:

Van Bork, R., Rhemtulla, M., Sijtsma, K., & Borsboom, D. (in preparation). A causal theory of error scores.

RVB wrote the draft of the manuscript. MR, KS and DB have revised several versions of the manuscript.

Chapter 4 published as:

Van Bork, R., Rhemtulla, M., & Borsboom, D. (2016). Composites can be causal too. *European Journal of Personality*, 30, 307 – 308. doi:10.1002/per.2060

RVB, MR and DB deliberated on the argument that is made in the article. RVB wrote the draft of the article. MR and DB provided feedback on the article.

Chapter 5 published as:

Van Bork*, R., van Borkulo*, C. D., Waldorp, L. J., Cramer, A. O. J., & Borsboom, D. (2018). *Network models for clinical psychology*. In E.J. Wagenmakers (Ed.) *Stevens' Handbook of Experimental psychology and Cognitive Neuroscience* (Fourth Edition, Volume 5), doi: 10.1002/9781119170174

Chapter 5 of this dissertation is part of a bookchapter that was published as above. The part that was included in this dissertation is the part of that bookchapter that was written by me, except for the section 'Approximations of the covariance matrix to identify connections'. For

* Authors contributed equally.

the published bookchapter the work was divided as follows. All authors wrote parts of the chapter. RVB and CDVB combined the parts and made it into one coherent chapter. RVB and CDVB revised the chapter. All authors provided feedback on the chapter.

Chapter 6 submitted as:

Van Bork, R., Rhemtulla, M., Waldorp, L. J., Kruis, J., Rezvanifar, S., & Borsboom, D. (under review). Latent variable models and networks: Statistical equivalence and testability. Submitted to *Multivariate Behavioral Research*

RVB, MR and DB deliberated on the design of the study. RVB performed the analyses and wrote the article. MR and DB supervised the analyses and revised several versions of the article. SR provided the data. LJW and JK provided feedback on the article.

Chapter 7 published as:

Van Bork, R., Grasman, R.P.P.P., & Waldorp, L. J. (2018). Unidimensional factor models imply weaker partial correlations than zero-order correlations. *Psychometrika*, 83, 443–452.

RVB, LJW and RPPPG proved the main result in the article and deliberated on the test that is included in the article. RVB performed the analyses and wrote the article. LJW and RPPPG revised the article.

Chapter 8 submitted as:

Van Bork, R., Marsman, M., Rhemtulla, M., Epskamp, S., Kruis, J., & Borsboom, D. (under review). Common effect models: Positive or negative manifold? Submitted to *British Journal of Mathematical and Statistical Psychology*

RVB, MM, MR and DB deliberated on the argument in the article. RVB wrote the article. MM, SE and RVB proved the main result that is included in the appendix. MM, MR, SE, JK and DB provided feedback on several versions of the article.

Chapter 9 published as:

Van Bork, R., Epskamp, S., Rhemtulla, M., Borsboom, D., & van der Maas, H. L. J. (2017). What is the p -factor of psychopathology? Some risks of general factor modeling. *Theory & Psychology*, 27, 759–773.

RVB, SE, MR, DB and HLJVDM deliberated on the argument in the article. RVB and SE wrote the article. MR, DB and HLJVDM provided feedback on several versions of the article. All authors revised the article.

OTHER PUBLICATIONS:

- Rhemtulla, M., **van Bork, R.**, & Borsboom, D. (in revision). Worse than measurement error: Consequences of inappropriate latent variable measurement models. (Submitted to *Psychological Methods*).
- Deserno, M. K., Borsboom, D., Begeer, S., **van Bork, R.**, & Geurts, H. M. (in revision). Highways to happiness for ASD adults? Perceived causal relations among clinicians. (Submitted to *Clinical Psychological Science*).
- Van Borkulo, C. D., **van Bork, R.**, Boschloo, L., Kossakowski, J. L., Tio, P., Schoevers, R. A., Borsboom, D., & Waldorp, L. J. (in revision). Comparing network structures on three aspects: A permutation test. (Submitted to *Psychological Methods*).
- Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., **van Bork, R.**, Waldorp, L. J., & Maris, G. (2018). An introduction to network psychometrics: Relating Ising network models to item response theory models. *Multivariate Behavioral Research*, 53, 15–35.
- Rhemtulla, M., Borsboom, D., & **van Bork, R.** (2017). How to Measure Nothing. *Measurement: Interdisciplinary Research and Perspectives*, 15, 95–97.
- Rhemtulla, M., **van Bork, R.**, & Borsboom, D. (2015). Calling models with causal indicators measurement models implies more than they can deliver. *Measurement*, 13, 59–62.
- Open Science Collaboration (2015). Estimating the reproducibility of Psychological science. *Science*, 349(6251), doi:10.1126/science.aac4716
- Langley, D. J., Wijn, R., Epskamp, S., & **van Bork, R.** (2015). Should I get that jab? Exploring influence to encourage vaccination via online social media. *ECIS 2015 Research-in-Progress Papers*, Paper 64.

CONTRIBUTIONS TO BOOKS:

- Van Bork, R.** & Cramer, A. O. J. (2017). *Sample size*. SAGE Encyclopedia of Abnormal and Clinical Psychology.
- Van Bork, R.** & Cramer, A. O. J. (2017). *Sample selection*. SAGE Encyclopedia of Abnormal and Clinical Psychology.

Cramer, A. O. J. & **van Bork, R.** (2017). *Internal validity*. SAGE Encyclopedia of Abnormal and Clinical Psychology.

Wijsen, L. D., **van Bork, R.**, Haig, B. D., & Borsboom, D. (under review). *Latent variable models and their alternatives: Three different approaches to the causal analysis of psychometric models*. (Bookchapter in The SAGE Handbook of Theoretical Psychology)

SHORT CV

Riet van Bork was born in Amsterdam on the 9th of November 1990. After the Montessori school (Montessori Lyceum Amsterdam), she studied Psychology at the University of Amsterdam, obtaining her Bachelor's degree (cum laude), including a minor in Philosophy of Language, Logic and Philosophy of Science (cum laude), and Master's degree (cum laude), including elective courses at the Institute for Logic, Language and Computation at the University of Amsterdam. She received the Master thesis prize of Psychology for her thesis titled 'On the identity of variables: An intensional approach in understanding the latent variable model'. She started her PhD project in November 2014 under supervision of prof. dr. Denny Borsboom and dr. Mijke Rhemtulla. During the third year of her PhD she visited dr. Mijke Rhemtulla at the University of California, Davis for six months. She returned to the University of California, Davis in the fourth year of her PhD for another two months.

Dankwoord

Tot slot richt ik mij graag tot de vele collega's, vrienden en familie die aan dit proefschrift hebben bijgedragen. Dank jullie wel. Dank voor jullie bijdragen aan de verschillende hoofdstukken, dank voor de inspirerende gesprekken, en dank dat ik altijd bij jullie terecht kon als ik ergens tegenaan liep.

First and foremost, I want to thank my supervisors Mijke and Denny. I have been so lucky with you as a team. Our meetings started with the three of us in Denny's office, but when Mijke moved to Davis we started skype meetings, either Denny and me on one side of skype and Mijke at the other side, or Mijke and me from Davis and Denny from Amsterdam. These meetings were always fun and gave me many new ideas. I have really enjoyed the full four years of my PhD and I am aware that that was for an important part because of your great supervision.

Mijke, so many times since I have started my master's thesis with you, I have thought of you as an example of how I want to do things. I admire how you teach, present, review, and how you argue in discussions. There have been many meetings, also in Davis, in which I was arguing something that I couldn't clearly put into words, and while it would confuse people, you seemed to always understand my point and explain it in a way that captured exactly what I was trying to say. I have learned so much from you, and I really appreciate all your thorough feedback and that I could always walk in for a coffee and a thought-provoking conversation.

Denny, ik ben zo vaak even bij je binnen gestapt om mijn standpunt in een methodologische of filosofische kwestie op jou te testen, en ik liep altijd weer naar buiten met vele tegenargumenten die me weer dagen aan het denken zetten. Kan causaliteit nou wel of niet op populatieniveau bestaan zonder op het niveau van het individu te bestaan? Moet je een psychometrisch model nou wel of niet causaal interpreteren? En waar zit nou precies de stochastiek in het datagenererende proces? Dit soort gesprekken gaven niet alleen heel veel inspiratie voor de projecten die ik heb gedaan, het was ook een onderdeel van mijn PhD tijd waar ik erg veel plezier en energie uithaalde en wat mijn denken sterk gevormd heeft. Daarnaast was ook jouw optimistische houding erg leerzaam. In plaats van alle problemen op te willen lossen is het ook heel waardevol om sommige van deze dingen niet als een probleem te zien maar als een kans.

Lieve paranimfen, ik ben ontzettend blij dat jullie 14 juni naast mij staan. Tessa, wat ben

ik blij dat ik in mijn methodenleerjaar bij jou in de groep kwam. Ik wist heel snel dat jij een goede vriendin zou gaan worden en we werden al snel een onafscheidelijk duo dat jaar. Hoe wij zo ontzettend gemakkelijk samen werkten, iedere opdracht en ieder nieuw vak weer, dat is heel tekenend geworden voor hoe ik samenwerken met jou ervaar. Ik heb het gevoel dat we elkaar heel goed begrijpen en elkaar altijd uitdagen en tot nieuwe inzichten brengen. Maar ook buiten studie en werk om ben je een hele waardevolle vriendin geworden, bij wie ik altijd terecht kan en met wie ik ontzettend veel leuke dingen heb meegemaakt. Ik ben dan ook heel erg blij dat jij tijdens de verdediging naast mij staat. Marie, wij zijn samen begonnen en precies tegelijk waren we ook weer klaar. Ik vond het ontzettend leuk dat ik deze vier jaar met jou samen ben doorgegaan. Vanaf de eerste week promoveren tot aan de laatste week hebben we samen onze pauzes doorgebracht. Koffie halen, broodjes halen, gewoon even buiten zitten of natuurlijk in de pantry. Ik ben heel blij dat ik aan deze tijd zo een leuke vriendin heb overgehouden en het voelt dan ook heel natuurlijk dat jij naast mij staat op de dag waarop deze periode wordt afgerond, en we weer een nieuwe periode ingaan.

Han, ik waardeer het heel erg dat je altijd betrokkenheid laat zien en vaak bent langsgelopen om te vragen hoe het gaat. Je stelt altijd creatieve vragen die me aan het denken zetten en ook discussies met jou geven mij veel inspiratie. Bovendien ben jij degene die mij de beste raadsels vertelt. Helaas heb ik het raadsel met het schaakbord nog steeds niet opgelost, maar ik geef niet op.

Adela, Lisa, Claudia, Gaby, Ria, Jonas and Angelika, we were the fun room! Although I am not sure our plants agree with that. There is one thing we all share, and that is office yoga. No-one escaped office yoga, and I really enjoyed these fun moments when we locked the door with books and put up a yoga video. Gaby en Lisa ik kan altijd alles met jullie bespreken en of dat nou op werk is of met wijn en pasta in een café, ik vind het fijn dat we zoveel met elkaar kunnen delen. Claudia, jij bent de enige kamergenoot met wie ik ieder kantoor tot nu toe heb gedeeld. Eerst op de diamantbeurs, toen in Go.40 en nu weer. Heel gezellig!

Psychosystems group, thank you all for the weekly conversations and inspiration.

Sacha, mijn allereerste artikel heb ik met jou geschreven. Ik weet nog goed hoe we de avond na mijn eerste keer submitten meteen daarna biertjes zijn gaan drinken om het te vieren, en we maar net op tijd bleken te zijn omdat het die nacht al afgewezen werd.

Claire, Maarten, Gaby en Don, onze schrijf-bijeenkomsten vormen een fijn moment in de week om problemen door te spreken die we tegenkomen, niet alleen in het schrijven maar ook meer algemeen bij het werken in de wetenschap. Ik kijk er naar uit deze afspraken voort te zetten.

Maarten heel erg bedankt dat ik altijd bij je binnen kan lopen en voor al onze gesprekken en

discussies die mij inspiratie geven en jouw enthousiasme dat altijd nieuwe motivatie geeft. Ik vind het dan ook ontzettend leuk dat jij in mijn commissie zit.

Ik wil ook mijn andere PML collega's bedanken voor de leuke tijd. Alexander met jou kan er op ieder PML feestje of avondje stappen tijdens een IOPS conferentie goed gedanst worden. Ravi, in de pauzes zorgde jouw geklier voor vertier, en Jonas de enige cappuccino tussen de lattes.

IOPS, I have enjoyed the conferences and the courses I took, and all the colleagues from other universities I got to know via IOPS.

Colleagues from Davis, thank you for being so welcoming. I have really enjoyed my time at the department there, and have learned a lot from the weekly DIPS meetings and Brown-Bag meetings. Tim, Nate and Madeline, thank you for making me feel at home.

Lieve vrienden buiten academia, jullie hielpen mij relativeren en jullie gezelligheid is voor mij de beste manier van ontspannen. Roos en Lilian, ik geniet ervan dat we na 15 jaar vriendinnen te zijn nog zoveel met elkaar delen. Ik kijk altijd erg uit naar onze etentjes en gesprekken waaruit telkens weer blijkt dat, hoewel we andere paden in zijn gegaan, we toch nog zoveel gedeelde ervaringen hebben. Nikki, dankjewel voor het ontwerpen van de voorkant, daar ben ik ontzettend blij mee. Rasila, Inga en Janna, ik hoop op nog vele maandelijks ontbijtjes. Mari, Max en Belle ik voel me volledig thuis bij jullie. Jullie zijn ontzettend belangrijk voor me en een onmisbaar onderdeel geweest van de vier jaar dat ik aan het promoveren was. Bram, Sam en Eva, met jullie er bij zijn we een huis vol, maar een heel gezellig huis en dat hoop ik de komende jaren nog vele weekenden mee te maken. Devi, Iris, Joris en Jiska, de vrijdagmiddagborrels zijn een hoogtepunt in mijn week: de ultieme ontspanning van nacho's en goed gezelschap. Joris en Devi, het heeft mij ook altijd veel geholpen om ervaringen uit te wisselen met vrienden die hetzelfde doorlopen maar dan op een andere plek. Boris, in het methodenleerjaar werden we hechte vrienden, en dat is zo gebleven. Heel gezellig om binnenkort bij je om de hoek te wonen! Sarah, bijna 10 jaar geleden begonnen we aan psychologie, en binnenkort zijn we burens. Ik kijk daar zo naar uit! Jonathan, hoe vaak is het de afgelopen jaren wel niet gebeurd dat ik nog heel even iets af moest maken op werk en jij ondertussen een heerlijke maaltijd aan het bereiden was. Het oefenwebkookboek is dan misschien uit, het is altijd een feestmaal als jij gekookt hebt.

En ten slotte mijn familie. Remko en Machteld, ik heb vaak mijn dilemma's bij jullie kunnen bespreken. Zo herinner ik me nog goed ons gesprek rondom mijn keuze tussen filosofie of methodenleer. Ilze, hoewel je het moeilijk voor te stellen vond dat iemand het leuk zou vinden om vier jaar over modellen na te denken, was je wel altijd heel benieuwd naar wat ik nou eigenlijk precies deed in al die tijd. Maar naast jullie meedenken en meeleven, zijn jullie vooral een plek waar ik me altijd thuis voel en ontzettend gelukkig mee ben. Ik heb altijd een sterk en

onvoorwaardelijk vertrouwen vanuit jullie ervaren en ik heb van jullie geleerd om mijn keuzes op de eerste plaats te laten leiden door waar ik plezier uit haal. Klaas en Olivier, mijn broertjes, jullie grapjes over hoe saai het wel niet moet zijn om een artikel te lezen, laat staan om er een te schrijven, helpen om het allemaal wat te relativeren. Ik heb altijd heel veel lol met jullie, en dat is altijd zo geweest, maar is ook nu nog net zoveel het geval. En Klaas, ik geloof nog steeds dat jij statistiek leuker vindt dan je zelf wilt toegeven.

Lieve Wiebe, er is heel veel waarvoor ik je wil bedanken, zoals het meelesen van stukken, het doorspreken van mijn gedachtes en me af en toe vertellen dat alles goed komt. Dankjewel voor alles. Ik was net een jaar aan het studeren toen we bij elkaar kwamen. Geen van ons beide had toen voorzien dat we onderzoek doen zo leuk zouden gaan vinden. We zijn samen naar de wetenschap toegegroeid en jij houdt mij scherp door mijn gedachtes vanuit een heel andere kant te benaderen dan ik in mijn veld gewend ben. Jouw enorme vertrouwen in mij maakt me zekerder op alle onzekere momenten. Ik kijk uit naar alles dat we nog samen mee gaan maken, laat het volgende avontuur maar komen.