Faculty of Medicine
University of Helsinki

PSYCHOMETRICS OF DRIVER BEHAVIOR

Doctoral Programme in Psychology, Learning and Communication,
Department of Psychology and Logopedics,
Faculty of Medicine,
University of Helsinki, Finland

Markus Mattsson

DOCTORAL DISSERTATION

To be presented for public discussion with the permission of the Faculty of Medicine of the University of Helsinki, in Hall 107, Athena, (Siltavuorenpenger 3 A), Helsinki, on December 11th, 2020, at 12 o'clock.

Helsinki 2020

Supervisors	Professor emeritus Heikki Summala Traffic Research Unit University of Helsinki Finland
	Docent Kimmo Vehkalahti Centre for Social Data Science Faculty of Social Sciences University of Helsinki Finland
	Docent Otto Lappi Department of Digital Humanities Faculty of Arts University of Helsinki Finland
Reviewers	Professor Richard Rowe Department of Psychology Faculty of Science The University of Sheffield United Kingdom
	Professor Reijo Sund School of Medicine Institute of Clinical Medicine University of Eastern Finland Finland
Opponent	Docent Anders af Wåhlberg Driving Reserch group Cranfield University United Kingdom

ISBN 978-951-51-6790-3 (pbk.)
ISBN 978-951-51-6791-0 (PDF)

The Faculty of Medicine uses the Urkund system (plagiarism recognition) to examine all doctoral dissertations

Unigrafia
Helsinki 2020

ABSTRACT

The role of human factors in crash causation is a central theme in traffic psychology. Human factors are often roughly categorized into cognitive errors and a tendency to break rules. In data analysis, these psychological properties are treated as measurable, continuous quantities, quite alike weight, length and temperature. Their existence is inferred based on covariation among individual traffic behaviors, which for their part function as measurements of the level of these properties: for instance, driving under the influence of alcohol and speeding are thought to reflect the tendency to break traffic rules.

The thesis examines joint variation among traffic behaviors and compares two competing explanations for the phenomenon: 1) The latent variable view of errors and violations, according to which covariation among traffic behaviors is explained by latent, unobservable psychological properties that cause variation in them and 2) The network view, according to which traffic behaviors interact directly with one another, which makes it unnecessary to posit unobservable psychological properties as explanations of behavior.

Within traffic psychology, questions such as these are usually not explicitly raised; rather, latent variable models are used as the default tool in data analysis. This practice entails certain assumptions, such as that of the latent variable models measuring the same unobservable properties in the same way across groups of respondents. Moreover, more fundamental questions, such as the theoretical status of latent variables in terms of realist vs. constructionist commitments and the nature of the relationship between latent and observed variables are seldom considered. The present thesis addresses these issues.

Studies I and II examine a central property of latent variable models of driver behavior: whether the same psychological properties can be measured in the same way across different subgroups of drivers that are defined based on age, sex and nationality. Both studies utilize rigorous latent variable measurement equivalence analyses. Study I concludes that if the latent variable view is adopted, patterns of covariation among self-reported traffic behaviors are sufficiently different across subgroups of Finnish respondents formed based on age and gender that the latent variables may well be specific to the group in question. Study II reaches a similar conclusion concerning social behavior (breaking rules in traffic) based on a comparison of young Finnish and Irish drivers. On the other hand, it shows that cognitive errors can more readily be interpreted as being related to similar – but not identical – latent variables across countries.

Study III assumes a novel point of view, and examines interactions among individual traffic behaviors using psychological network models. This shifts the focus from abstract psychological properties to potentially causal

relationships between traffic behaviors: drivers who are more likely to exceed speed limits are also more likely to end up driving close to another vehicle, for instance. In other words, edges in the network models are interpreted as causal hypotheses. Study III also presents Poisson regression models that predict crashes from self-reported traffic behaviors instead of latent variables. This enables various self-reported traffic behaviors to have differential associations with crashes, which is intuitively plausible as, for instance, the violations range from driving under the influence of alcohol to honking at others. The models are built and tested in independent sets of data, making it possible to avoid overfitting the predictive models to data at hand. This procedure, together with selecting variables based on regularized regression, is argued to have useful properties in predicting crashes in traffic psychology.

As a whole, the thesis presents two new interpretations for the relationship between individual traffic behaviors and the psychological properties investigated within traffic psychology. First, the psychological properties may reduce to nametags for behaviors that co-occur in certain kinds of contexts and have no causal power of their own. Second, they may prove to be emergent properties arising from the interaction among the behaviors. These alternatives are discussed together with an intermediate view that combines the latent variable view and the network view. The thesis, then, positions itself as a part of recent psychometric discussion in which psychological properties are seen as being formed through the interaction of different behaviors, thoughts and emotions without necessarily treating psychological properties as unidimensional, measurable quantities.

ACKNOWLEDGEMENTS

Among all the important individuals who have made the present work possible, I would first like to mention professor emeritus Heikki Summala who introduced me to the field of traffic psychology. Heikki has done groundbreaking theoretical work and his zero-risk theory is already a classic within the field. I consider it a great honor of having had a creative researcher of his calibre as a supervisor of my work. Docent Otto Lappi has had a similarly important role to play in my research career. Otto has the rare gift of being able to acutely perceive what is relevant to solve a scientific puzzle. This has certainly something to do with the immense breadth of his knowledge base: Otto is a philosopher by training and an original researcher and an inspiring teacher by profession. I have learned a lot from him, and he has always been an acute and constructive critic of my work. I want to thank him for being both a supervisor and a supportive friend. Finally, Kimmo Vehkalahti has functioned as a teacher and mentor in the use of statistical methods, urging me toward learning more about their underlying mathematics and providing practical opportunities in doing so. I am most grateful to have had such a wonderful trio of supervisors for the work.

Among my co-authors, Fearghal O'Brien has a special place in contacting me after the publication of Study I and suggesting that we might have common research interests. This lead to writing Study II with him as the second author, and the present thesis would not have become what it is without him – so thank you Fearghal! I am similarly grateful for the work of the researchers who did all the hard work of collecting the data I got to use in the present thesis: Timo Lajunen, Heikki Summala, Michael Gormley, British colleagues working in the Cohort II project – I wish to express my gratitude equally to all of you.

Among fellow PhD researchers and colleagues, I wish to mention Esko Lehtonen and Jami Pekkanen. Esko is a curious mind, always into learning new things and trying out new ideas. Together with his technical expertise, this amounts to being a productive researcher. I feel that my relationship to him is best characterized by great mutual respect, and discussions with him have helped me develop as an independent thinker. Jami, for his part, is the technical mastermind and programming genious in the Traffic Research Unit. He has taken but a perfunctory glance at my meager programming puzzles and suggested a way forward while seemingly effortlessly creating new analysis software and ingenious statistical models of his own. Besides this, Jami has delightfully strong opinions on the philosophical questions of choosing a fruitful framework for research, and discussions with him have encouraged me to finish the present work.

People outside the immediate academic circle have played at least as great a role in making the present work possible. First, I wish to thank my parents for being there for me. My father, a mathematician and lately also a social psychologist, has encouraged me to take the academic path, and has ever since been a fellow wonderer and among the best of my discussion partners. My mother has no academic training, but what she lacks in credentials, she makes up in love and pride. She has always wanted to understand what I am working with, and talking with her about the abstract and mathemathical has been a valuable exercise even academically. Patte, my brother, thank you for always believing in me. Pappa, olen äärettömän onnellinen, että ehdit nähdä tämän päivän – kiitos kannustuksesta! Maxine, tack för allting: kärleken, din ödmjuka och lugna “förstås klarar du det här” –attityden som gav mig styrka särskilt vid de sista krejsiga veckor av skrivandet och din närhet som påminde mig om vad som är viktigt. Tack även för våra fina diskussioner.

Last but not least, thank you VALT for funding my work. Your support made it possible to look deeply into the hard problems of measurement. I also wish to thank the University of Helsinki for the grant that made it possible to finish writing the present work.

CONTENTS

1 Introduction 11

1.1 A review of relevant DBQ literature 19

1.1.1 Studies examining the measurement properties of the DBQ 23

1.2 Motivation for the studies of this thesis 27

2 Methods 30

2.1 Data 30

2.2 Questionnaires used 31

2.3 Statistical methods 33

2.3.1 Structural equation models 33

2.3.2 Analyses of measurement equivalence 38

2.3.3 Introduction to network analysis 41

2.3.4 Statistical methods for Study I 45

2.3.5 Statistical methods for Study II 47

2.3.6 Statistical methods for Study III 49

3 Results 52

3.1 Study I 52

3.2 Study II 56

3.2.1 Dimensionality of the DBQ 56

3.2.2 Measurement equivalence of the DBQ across countries 58

3.3 Study III 59

3.3.1 Network analyses 59

3.3.2 Regression analyses 63

3.3.3 Summary of the findings of Study III 65

4 Discussion 66

4.1 Relationships between latent variables and self-reported driving behaviors ... 71

4.1.1 The structure of violations ... 74

4.1.2 The structure of errors ... 77

4.2 Predicting crashes from individual driver behaviors ... 78

4.3 The measurability of psychological properties ... 80

4.4 Limitations of the studies ... 84

4.5 Open questions and future directions ... 90

4.6 Conclusions ... 92

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following publications:

I Mattsson, M. (2012). Investigating the factorial invariance of the 28-item DBQ across genders and age groups: an exploratory structural equation modeling study. *Accident Analysis & Prevention* 48, 379-396.

II Mattsson, M., O'Brien, F., Lajunen, T., Gormley, M., Summala, H. (2015). Measurement invariance of the driver behavior questionnaire across samples of young drivers from Finland and Ireland. *Accident Analysis & Prevention* 78, 185-200.

III Mattsson, M. (2019). Network models of driver behavior. *PeerJ* 6, e6119.

The publications are referred to in the text by their roman numerals.

ABBREVIATIONS

CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
EBIC	Extended Bayesian Information Criterion
EFA	Exploratory Factor Analysis
ESEM	Exploratory Structural Equation Modeling
DBQ	Driver Behavior Questionnaire
GEMS	Generic Error Modeling System
GGM	Graphical Gaussian Model
LASSO	Least Absolute Shrinkage and Selection Operator
LNM	Latent Network Model
LVM	Latent Variable Model
NM	Network Model
RMSEA	Root Mean Square of Approximation
RNM	Residual Network Model
SRMR	Standardized Root Mean Square Residual
WRMR	Weighted Root Mean Square Residual
WLSMV	Weighted least squares estimator with mean and variance correction

1 INTRODUCTION

This thesis presents three studies on the psychological properties – such as *the tendency to commit violations* or *proneness to cognitive errors* – that are frequently taken to underlie unsafe traffic behavior within traffic psychology. It is motivated by two questions: 1) whether these properties are measurable and 2) whether individual self-reported traffic behaviors can be thought of as measurements of them. Specifically, it asks the psychometric question of whether these properties can be measured in the same way across subgroups of drivers using self-report instruments. It answers largely in the negative and argues that this is because individual traffic behaviors are determined by multiple psychological properties instead of being reducible to a small number of very general ones. It then builds on an alternative view of psychometrics – known as network psychometrics (Borsboom, 2017; Epskamp, 2017) – that focuses on the interplay of traffic behaviors instead of treating them as measurements of a small number of underlying general psychological properties. The network perspective enables viewing the psychological properties as emerging as a consequence of this interplay rather than explaining it by functioning as latent causes. On the other hand, under the network view, the psychological properties can also be viewed as classificatory categories (i.e. nametags). In addition to discussing the relationship between psychological properties and observable behaviors, the thesis also utilizes methods of statistical learning theory to build a predictive model of accidents based on individual traffic behaviors. This is done because it is plausible that the different driving behaviors that are commonly treated as measurements of the same psychological property have differential relationships with crash risk.

Traffic psychology is a practical enterprise. Much research in the field is motivated by an interest in traffic safety, and the success of safety-oriented research judged by its ability to produce effective interventions. This is especially true of research into human errors and violations in traffic, which are seen as important determinants of crashes. Because of this, it is of central importance to understand the nature of human errors and violations: are there different kinds of errors and violations? What causes them? Are they measurable, unidimensional phenomena?

Much current research on the relationships of errors and violations to crashes is based on self-report instruments such as the Driver Behavior Questionnaire (DBQ; Reason, Manstead, Stradling, Baxter, & Campbell, 1990). The DBQ is perhaps the most widely used such instrument in traffic psychology with the *locus classicus* publication (Reason et al., 1990) having been cited 716 times by the end of the year 2019 according to a search carried out in the Web of Science portal. The DBQ is based on a theory in cognitive ergonomics, the Generic Error Modeling System (GEMS, Reason, 1990), which describes human errors in safety-critical situations. It differentiates skill-based errors (attentional *slips* and memory-related *lapses*, also referred to as *monitoring failures*) from problem-solving failures (rule-based mistakes and knowledge-based mistakes). *Slips* are defined as skill-related errors in that the persons committing slips have a plan that they intend to follow, but fail to do so due to paying either too little or too much attention to the task; turning on the windscreen wipers instead of the blinker serves as an example. *Lapses* are otherwise similar to slips, but are related to forgetting something along a sequence of actions, such as not remembering to turn in an intersection when driving somewhere. *Rule-based mistakes* are called problem-solving errors because they involve a person following a normally well-functioning plan that turns out not to work. An example of a rule-based mistake would be a person following the rule (in right-hand traffic): "If another road user is turning left, overtake them from the right-hand side" without realizing that the road is too narrow, which causes the person to drive off the road. Knowledge-based mistakes are related to solving completely novel problems to which a pre-existing rules cannot be applied; an (admittedly slightly far-fetched) example would be a driver hearing weird noises from the motor, ignoring them and thinking that it will be fine, whereupon the motor catches fire because of lack of oil. The categories are not hard-and-fast, as rule-based mistakes have more in common with skill-based errors than knowledge-based mistakes.

In addition, Reason (1990) considers *violations*, deliberate deviations from safe practices. While *slips*, *lapses* and *mistakes* are derived from an analysis of cognitive processes, *violations* are characterized using interpersonal, social concepts in that they are related to breaching implicit or explicit social agreements between people. Most of the time, violations are related to trying to achieve a well-intentioned outcome, such as speeding to get to work on time, rather than having an outright malicious intention, such as sabotaging a car to avenge something to another person. Still, *mistakes* and *violations* are seen as similar in an important respect: both are types of *intended actions*, whereas *slips* and *lapses* are actions that deviate from intention. Since the seminal publication (Reason et al., 1990), the DBQ has functioned as an operationalization of the central concepts of the GEMS; however, the concepts have been characterized in various and partly conflicting ways in the DBQ tradition (Table 1).

Table 1. *(Reproduced with permission from Study I Table 8): classification and characteristics of aberrant behaviors in the DBQ research tradition*

	Error
Definition	"Failure of planned actions to achieve their intended consequences" (Reason et al., 1990)
	Dangerous, declines with age (Reason et al., 1990)
	Associated with the way in which individuals deploy their limited attentional resources in response to competing task demands (Reason et al., 1990)
	"Manifestation of over-engagement [with driving]", something that men do (Reason et al., 1990)
	Mistake with potentially dangerous consequences (Parker et al., 2000)
	Lapse
Definition	"Unwitting deviation of action from intention" (Reason et al., 1990) , also a
	"Covert memory failure" (Reason et al., 1990)
	"Associated with 'attentional capture' by things other than the driving task", something that women do (Reason et al., 1990)
	Attentional failure, is a mistake (Parker et al., 2000)
	Associated with the way in which individuals deploy their limited attentional resources in response to competing task demands (Reason et al., 1990)
	Slip
Definition	"Unwitting deviation of action from intention" (Reason et al., 1990), also
	"Potentially observable as externalized actions-not-as-planned" (Reason et al., 1990)
	Associated with the way in which individuals deploy their limited attentional resources in response to competing task demands (Reason et al., 1990)
	Mistake
Definition	"Departure of planned actions from some satisfactory path toward a desired goal" (Reason et al., 1990)
	Divided into
	Rule-based (applying inappropriate condition rule) vs.
	Knowledge-based (on-line reasoning) in relation to an imperfect mental model of the problem situation (Reason et al., 1990)
	Something not intended (Parker et al., 2000)
	Violation
Definition	Deliberate deviation from safe practices (Reason et al., 1990)
	Social in nature (Reason et al., 1990)
	Associated with maladaptive motives, such as a preference of speed over safety (Reason et al., 1990)
	Something done by drivers who consider themselves skillful (Reason et al., 1990)
	Risky driving done intentionally, potentially adverse consequences (Parker et al., 2000)
	Can be divided into
	Interpersonal aggression, emotional vs.
	"Ordinary", non-emotional (Parker et al., 2000)

The traffic-safety aspect of DBQ studies comes in through presenting the questionnaire (or, to be specific, some version of it, see section 1.1 below) to a group of drivers and correlating the underlying psychological properties, operationalized as latent variables, with whether the drivers have been involved in an accident or not.

Depending on the study, these groups of drivers may be a random sample of everyone with a driver's license in a country (e.g. Lajunen, Parker, & Summala, 2004; Parker, Reason, Manstead, & Stradling, 1995) or members of subgroups of drivers, such as young (Biederman et al., 2012; Roman, Poulter, Barker, McKenna, & Rowe, 2015) or old (Parker, McDonald, Rabbitt, & Sutcliffe, 2000) drivers, professional drivers (Maslać, Antić, Lipovac, Pešić, & Milutinović, 2018; Öz, Özkan, & Lajunen, 2010), users of specific vehicles such as motorcycles (Sakashita et al., 2014) etc. Further, accident liabilities of different groups of drivers can be compared (for instance, men vs. women, drivers of different ages etc.; Parker et al., 1995; Sullman, Stephens, & Taylor, 2019). Importantly, carrying out such comparisons based on the DBQ presupposes that the instrument works in the same way and measures the same underlying psychological properties, such as *errors* and *violations*, or *slips*, *lapses*, *mistakes* and *violations*, in the groups to be compared.

Within the factor analytic tradition, the GEMS variables are treated as *measurable properties* similar to, say, intelligence or personality traits. Further, they are assumed to possess quantitative structure quite similarly to prototypical examples of quantitities such as height, weight and temperature. This, in itself, is a remarkably strong assumption and is discussed in some detail in Section 4.3. When using the DBQ, the GEMS variables are then operationalized as questionnaire items. Ontologically, the latent variables are considered stable psychological traits that can be compared across genders, age groups, traffic cultures etc. This is evidenced by statements related to the nature of the latent variables in the DBQ literature such as

> *"As each type of behavior has a distinct psychological underpinning (Reason et al., 1990), different interventions are required to reduce their frequency and also associated crash risk"*
>
> *(Stephens & Fitzharris, 2016)*

In Study III this approach is called *the latent variable view of violations and errors* and it amounts to assuming that the latent variables can be measured (in a technical sense, see Sections 2.3.1. and 2.3.2.) based on intercorrelations among self-reported traffic behaviors. The central properties of the latent variable view are as follows:

- there exist fundamentally different types of "aberrant behavior" that need to be targeted by different types of interventions that
- have different relationships with the drivers' accident risk, and
- these different types of behavior (*unintentional errors* and *intentional violations*) are not directly observable, but can be measured by correlating questionnaire items related to individual traffic behaviors and statistically estimating a low-dimensional structure of underlying latent variables

The idea is given formal expression by assuming that the observed item scores X are composed of true scores and error ($X = T + E$). The central idea is that the variance shared by the observed variables X is due to these latent variables. For instance, the question items *How often do you disregard the speed limits?* and *How often do you race away from the traffic lights with the intention of beating the driver next to you?* are seen as measurements of a driver's tendency to violate traffic rules. Similarly, the items *How often do* *you fail to notice that pedestrians are crossing when turning into a side street?* and *How often do you fail to check your rear-view mirror before pulling out, changing lanes, etc.?* are seen as measurements of a driver's tendency to commit attention-related errors (i.e. *slips*). More fine-grained distinctions can be made: For instance, *rule violations* can be divided into *aggressive violations* and *traffic rule violations*.

The causal assumptions underlying the measurement exercise are formalized as the reflective measurement model (Howell, Breivik, & Wilcox, 2007), a schematic representation of which is given in Figure 1. There, the latent variables – *the tendency to violate rules* and *the tendency to commit errors* – are shown as causing variation in the observed variables – the behaviors represented by the questions *How often do you exceed speed limits? How often do you drive even though you suspect you may be drunk?* etc. It is noteworthy both theoretically and in terms of potential traffic safety interventions that all assumed causal paths run through the latent variables. For instance, variables related to driver characteristics, such as enjoying speed, affect the drivers' violation-proneness, which then affects individual driving behaviors such as speeding. Notably, the individual behaviors are modelled as independent of each other once the drivers' position on the latent variables is known. In other words, *speeding* correlates with *overtaking dangerously*, *drunk driving* and *showing aggression* only because these behaviors reflect the underlying latent variable *violation-proneness*. Further, *speeding* correlates with *missing signs* only because their respective causes (*violation-proneness* and *error-proneness*) are correlated.

Under this interpretation, strictly taken, there is no other reason for the observed variables to correlate than the fact that the latent variable underlying them causes variation in all of them. In particular, the observed variables are assumed not to be causally related to one another. Further, the observed variables are assumed to be qualitatively similar, interchangeable indicators of the latent variable in that any given observed variable can be dropped or exchanged with another one (Bollen & Bauldry, 2011). This conceptualization allows, though, the reliabilities of the observed variables and strengths of the relationship between a given observed variable and the latent variable (factor loadings) to vary; the point is that dropping any given observed variable would not affect the relationships between the latent variable and the other observed variables (Bollen & Lennox, 1991). Further, under these causal assumptions, manipulating the observed variables should have no effect on the latent variables or other observed variables. For instance, under the reflective measurement model, making it impossible to drive under the influence of alcohol (by, say, installing an alcolock) has no effect on the other violations (speeding, overtaking etc.) or the latent variable *violation-proneness*.

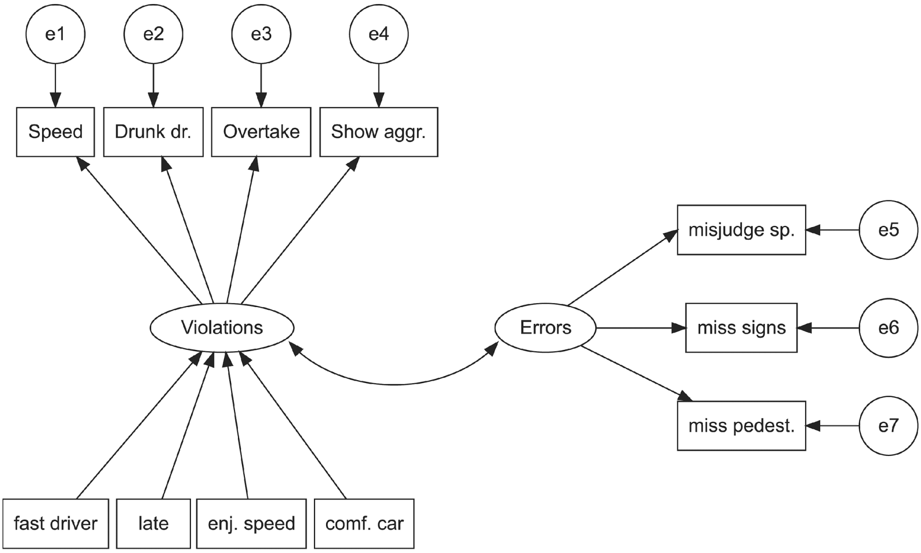


Figure 1 A reflective measurement model of violations and errors. The unidirectional arrows refer to assumed causal relationships and bidirectional arrows to covariation between variables. The ovals and circles represent latent variables and the rectangles observed variables. Reproduced from Study III under the CC-BY-4.0 licence.

The observation that the reflective measurement model seems to offer an inadequate causal representation of the actual relationships among the traffic behaviors is a central motivation for the present thesis. Indeed, it seems plausible that the behaviors represented by the observed variables may be causally related without being mediated by the specific mechanisms envisioned in the GEMS. For instance, looking at the driving behaviors in Figure 1, it would seem commonsensical and plausible to think that the tendency to exceed speed limits would have a direct causal connection to the tendency to overtake other drivers irrespective of the driver's position on the latent dimension.

Similar suggestions have recently been made within clinical and personality psychology (see, e.g. Borsboom, 2017; Robinaugh, Hoekstra, Toner, & Borsboom, 2019), where *the network approach* to psychopathology, personality and psychometrics has been developed since 2008. Within this approach, clinical phenomena such as depression, PTSD and psychosis as well as personality traits have been modelled as psychological networks (Costantini et al., 2015; Cramer et al., 2012a; Cramer et al., 2012b; Fried & Cramer, 2017; Fried et al., 2017; Robinaugh et al., 2019). The network approach is based on the premise that psychological properties are formed and maintained through the interactions of their parts; for instance, syndromes as networks of symptoms and personality traits as networks of thoughts, emotions and behaviors. In the case of depression, for example, it is thought that interactions among sleeplessness, concentration difficulties and difficulties in social relations give rise to the phenomenon and also maintain it (for instance, more sleeplessness → more concentration problems → more social problems).

In network models, the relationship between the phenomenon under investigation (e.g. depression) and its symptoms is seen as one of *emergence*: interactions among the symptoms bring about the phenomenon, which exists because of these interactions, not independently of them. This view enables the use of different tools of network science and also viewing the phenomenon as a complex system (Borsboom, 2017), which 1) occupies different stable states (e.g. the depressed and healthy state) depending on the patterns of connectivity among the symptoms, 2) contains feedback loops among the nodes (e.g., in the previous example, social problems feeding back to sleeplessness through a positive connection), 3) contains central symptoms, also referred to as hubs, in addition to peripheral symptoms, 4) adapts to external influences by aiming toward holding onto the stable state through a homeostatic mechanism, 5) is dependent on the history of the system in that the same end state (e.g. healthy or depressed state) can be reached through different developmental pathways. In addition, such systems can be viewed as consisting of nested structures both across time and physical organisation (e.g. neurons – brain systems – behavioural symptoms – social phenomena), embodying non-linear dynamics among their parts and developing over time.

Within psychopathology research, different states of the network are characterized by different connection strengths among the symptoms. Using depression as an example, in the *depressed* state of the symptom network activation readily spreads through the network because of strong interconnections among the symptoms. Then, when a certain node, such as *social problems*, becomes activated, it activates the other symptoms, which are in turn connected to *social problems* by self-sustaining feedback loops (Borsboom, 2017). Moreover, differences in susceptibility to depression can be characterized by differences in initial connection strengths among symptoms: people are more likely to become depressed at some stage of their lives if the symptoms of depression share strong associations in their personal symptom networks. In short, the network approach, together with closely associated neighboring approaches, offers a rich conceptual framework for research in psychopathology, personality psychology, and – as is suggested in the present thesis – traffic psychology. A schematical network model of traffic behavior is showed below in Figure 2. The model is based on no data, and is shown solely for the purpose of illustrating a psychological network model.

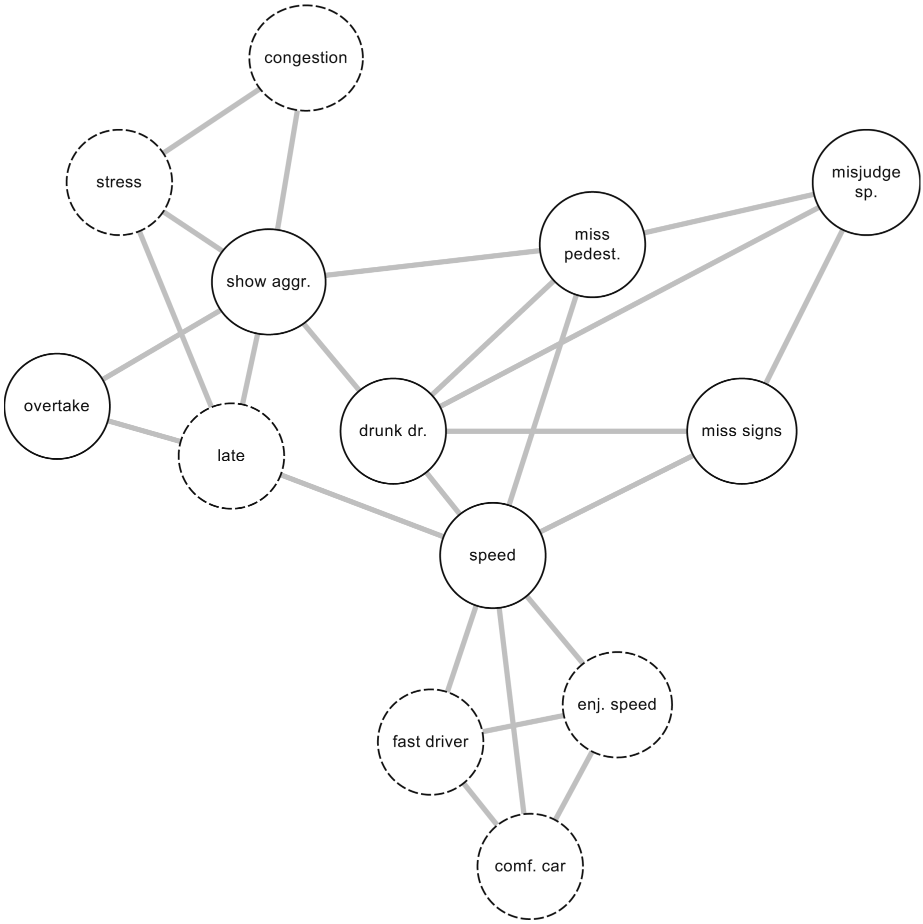


Figure 2 A hypothetical unweighted network model of traffic behavior. Traffic behaviors are shown as nodes drawn using solid lines and background factors assumed to influence them using dashed lines. Reproduced from Figure 2, Study III, under the CC-BY-4.0 licence

In short, network models and latent variable models are motivated by different concerns: examining the internal structure (and dynamics) of phenomena and measuring something that cannot be directly observed, respectively. Because of this, they are perhaps best understood as complementary approaches to psychometrics. In the present thesis, these issues are discussed in Sections 4.1.1. and 4.1.2., and an approach combining latent variable models and network models known as generalized network psychometrics, is briefly discussed in Section 4.4.

One of the central questions in the present thesis concerns the relationship between individual traffic behaviors, such as *speeding* and *tailgating*, and the latent variables, such as *error-proneness* and *violation-proneness*. Under the latent variable view, their relationship is one of measurement, with self-reports of the individual behaviors functioning as measurements of the level of the latent variable. Under the network view, their relationship is less clear. This thesis raises the question of whether the relationship should be seen as one of emergence or constituency similarly to the proposals given in network models of psychopathology and personality, or whether *violations* and *errors* function as descriptive labels without explanatory power of their own. As Study III of the present thesis is the first contribution that applies the network approach to psychometrics within traffic psychology, the thesis needs to be seen as the starting point of a discussion rather than a definitive answer to the questions.

1.1 A review of relevant DBQ literature

There exist a wide variety of different versions of the DBQ. They have been developed based on the original 50 self-report items study (Reason et al., 1990), which were intended to capture variation in five theoretical constructs: *slips*, *lapses*, *mistakes*, *unintended violations* and *intended violations*. The classification derives partly from GEMS (Reason, 1990). As described in Section 1, in the GEMS, actions that deviate from intention can be related to problems in attention (slips) or memory (lapses), while errors may also result from bad planning (mistakes). In addition, people sometimes violate social rules either intentionally (intended violations) or by accident (unintended violations). The study of Reason et al. (1990), however, resulted in a three-factor structure of *silly errors*, *dangerous errors* and *violations*, which the authors deemed, in essence, to be close enough to the intended structure. As the DBQ research tradition began with the study, it is of interest that the intended factor structure was not uncovered. This might have to do with imperfectly formulated or chosen self-report items, but the authors did not discuss the issue. Nonetheless, both the intended factor structure and the obtained factor structure were interpreted as reflecting the functioning of different psychological processes. In much subsequent research based on the DBQ, the latent factors have been similarly interpreted as reflecting the functioning of distinct psychological processes.

Before introducing subsequent research based on using the DBQ, a couple of methodological notes are in order. Reason et al. (1990) used Principal Component Analysis (PCA) with the orthogonal varimax rotation as an analysis method and chose the number of components to retain based on examining the scree plot. Rotation, simply put, is a mathematical procedure used for obtaining a simple and interpretable result in an Exploratory Factor Analysis (EFA) or a PCA (Brown, 2009). In an oblique rotation, the resulting factors / components are allowed to correlate, whereas in an orthogonal rotation, this is not the case. Such analytic choices shape the results that are obtained and the interpretations that are made. It remains unclear how

Reason et al. (1990) ended up with the analytical choices that were made and yet their publication launched the use of the DBQ; because of this, the consequences of the different analytical options are briefly introduced below.

First, by carrying out a PCA instead of an EFA the analysis produced, in a strict sense, a statistical summary of the data rather than information on the latent factors that might underlie the data (Mattsson, 2014, Section 5). In practice, PCA tends to produce slightly higher loadings than FA when the same rotation method is used. This is because in a PCA, all variation in the observed variables is analyzed, whereas in an EFA, variation that is unique to a given observed variable is excluded from the analysis; in EFA, this variation is considered measurement error. It is important to keep this difference in mind when comparing the results of PCA and FA (de Winter & Dodou, 2016). Nonetheless, because differences between the results of PCA and EFA are often in practice small (Velicer & Jackson, 1990) and because Reason et al. (1990) interpreted their results as evidence of underlying factors, this distinction is glossed over in what follows despite its theoretical importance (Mattsson, 2014).

Second, the question of choosing a rotation method is a complex one, and arguments related to simplicity and interpretability can be presented in favor of either an oblique or an orthogonal rotation (Brown, 2009). It remains unknown how Reason et al. (1990) ended up with the orthogonal rotation method that they used rather than a method of oblique rotation. This is relevant, since later DBQ research has found high intercorrelations among obliquely rotated factors (see, for instance, Lajunen et al., 2004; Mattsson, Lajunen, Gormley, & Summala, 2015; Mattsson, 2012; Stephens & Fitzharris, 2016), and it has been recommended that "if the researcher does not know how the factors are related to each other, there is no reason to assume that they are completely independent" (Preacher & MacCallum, 2003). Thus, it is possible that different results would have been obtained by Reason et al. (1990) had they used an oblique rotation method.

Third, thorough reviews of methods of choosing the correct number of components / factors to retain have repeatedly recommended to refrain from using the scree plot as the sole method in making this decision; rather, it is recommended that it be used as an adjunct to more accurate methods such as parallel analysis or the Minimum Average Partial Correlation method (Preacher & MacCallum, 2003; Velicer, Eaton, & Fava, 2000; Zwick & Velicer, 1986). Because of this, it is unclear whether the three-component structure obtained by Reason et al. (1990) in fact fit the data optimally or not.

Due to the three concerns mentioned above, the DBQ-based research tradition was built on a partially shaky foundation. Still, to reiterate the central findings, Reason et al. (1990) ended up with the three-factor structure of *silly errors*, *dangerous errors* and *violations*, which they interpreted as reflecting the functioning of different psychological processes (different kinds of error-proneness and violation-proneness, respectively). Reason et al. (1990) also calculated mean score variables based on the items that loaded on the three factors and concluded that men commit more *violations* than women, that committing *violations* decreases with age, and that women commit more *silly errors* than men. The preconditions that need to be met in order for such comparisons to be permissible are examined in Study I and Study II of the present thesis.

Another early study (Blockey & Hartley, 1995) obtained a different three-factor structure for the question items used by Reason et al. (1990). The authors performed a PCA with varimax rotation and based their conclusions on the three PCs having eigenvalues > 1. The authors referred to the PCs as *general errors* (with items intended to measure *slips*, *mistakes* and *unintentional violations* loading on the factor), *dangerous errors* (with items intended to measure *slips* and *mistakes* loading on it) and *violations*. The factor structure differed somewhat from that obtained by Reason et al. (1990) and the authors speculated that this was due to demographical factors such as differences in age and gender distributions between the two studies. On the other hand, the original three-factor (PC) structure of the DBQ was more or less exactly replicated by Åberg & Rimmo (1998) based on a 44-item version of the instrument in a sample of Swedish drivers.

The next major development of the DBQ took place with Parker et al. (1995) who picked the 8 items having the highest loadings on the three original factors of Reason et al. (1990) and ended up with 24 questionnaire items. Parker et al. (1995) referred to the results of their PCA (varimax rotation, eigenvalue > 1 criterion for retaining PCs) as *errors*, *lapses* and *violations*. This is slightly confusing because *lapses* was originally intended as a sub-category of *errors* (Reason, 1990; Reason et al., 1990) and the *errors* that are not *lapses* would then be categorized as *slips* according to the original nomenclature. Because of this, the present thesis refers to the *errors* that are not *lapses* as *slips*. The study was a typical example of research based on DBQ in that it involved predicting *error* and *violation* scores from demographic factors and the drivers' self-image as drivers, and then used the DBQ factors (together with the demographic variables) for predicting accidents. Further, the principal component structure was nearly perfectly replicated by Westerman & Haigney (2000) using the 24-item version of the instrument in a large sample of UK drivers (PCA with varimax rotation, criterion for number of PCs to retain not reported).

Even though the questionnaire used by Parker et al. (1995) contained items related to aggressive driving, it was Lawton, Parker, Manstead, & Stradling (1997) that explicitly modelled aggressive violations as a factor (PC) of its own after adding items related to *aggressive violations* and *highway code violations*. In the study, *violations* comprised three factors / PCs: *fast driving*, *maintaining progress* and *anger / hostility* that were predicted by various demographics and the drivers' affective evaluations of committing the violations, i.e. whether doing so would make them feel good or bad.

The 27- or 28-item "standard" version of the DBQ is a result of combining the versions of the instrument used by Parker et al. (1995) and Lawton et al.

(1997). The items related to *errors* and *lapses* are taken from the instrument reported in the first-mentioned study, and the 12 violation items from Lawton et al. (1997) and Parker, Lajunen, & Stradling (1998). The resulting 28-item version of the instrument was first reported by Mesken, Lajunen, & Summala (2002), who obtained a factor structure consisting of *errors*, *interpersonal violations*, *speeding violations* and *lapses* after performing a principal axis factor analysis using an unspecified oblique rotation and choosing the number of factors to retain based on examining the scree plot.

Bianchi & Summala (2002) obtained a different four-component structure of *errors*, *ordinary violations*, *aggressive violations* and *lapses* using PCA with unspecified oblique rotation and choosing the number of components to retain based on examining the scree plot. Lajunen et al. (2004) investigated the same 28-item instrument and the 27-item version obtained by dropping the item related to drinking and driving because of its low correlations with the other items. The authors used principal axis factoring with oblimin rotation and chose the number of factors based on examining the scree plot, using the eigenvalue > 1 criterion and by performing a parallel analysis. As the methods produced different results, the number of factors to extract was based on considerations of interpretability. The results of the analysis in Lajunen et al. (2004) resulted in a similar factor structure to the one reported by Bianchi & Summala (2002); the correlations among these first-order factors were further explained by performing a second-order factor analysis in which the two violation-related factors loaded on a second-order factor (*violations*) and the other two factors on a factor that was dubbed *mistakes*. This version of the instrument is used in the studies reported in the current thesis, as well.

The 28-item DBQ has subsequently been used in the Spanish context (Eugenia Gras et al., 2006). In that study a four-component structure was obtained after dropping one item ("misread signs, exit on wrong road") and performing a PCA based on oblique and orthogonal rotation (the results of the orthogonal rotation were reported because the results did not differ markedly). The choice of the number of components to extract was made based on parallel analysis. The factor structure that was obtained comprised *errors* (with items intended to measure *errors*, *lapses* and *aggressive violations* loading on the factor), *violations* (with items intended to measure *violations*, *aggressive violations* and *errors* loading on it), *interpersonal violations* (with the three items commonly referred to as *aggressive violations* loading on it) and *lapses* (with three out of seven items intended to measure *lapses* loading on it).

In addition, administering the 28-item DBQ to different driver groups has resulted in different factor structures. Dimmer & Parker (1999) expected to uncover a four-factor structure (*errors*, *lapses*, *aggressive violations*, *violations*) on data collected from company car drivers, but ended up with six factors that were labelled *errors*, *aggressive violations*, *violations / speeding violations*, *action slips*, *inattention lapses* and *not caring about the vehicle*.

Similarly, Sullman, Meadows, & Pajo (2002) administered the 28-item DBQ to Australian truck drivers and began by extracting an eight-component solution, which they subsequently dropped. They also dropped certain items and extracted four principal components (which they dubbed *errors*, *violations*, *lapses* and *aggressive violations*) based on 22 items, dropping the rest of the items. The authors performed PCAs with varimax rotation and chose the number of components by examining the scree plot.

Besides the 27 / 28-item DBQ, several versions of the instrument, differing in the number and nature of latent variables and items, have been developed. For instance, Åberg & Rimmo (1998) constructed a Swedish version of the DBQ to measure *violations*, *mistakes*, *inattentional errors* and *inexperience errors* using 104 items. In addition, Kontogiannis, Kossiavelou, & Marmaras (2002) constructed an instrument, also naming it the DBQ, that measures *mistakes, highway code violations*, *negligence*, *aggressive violations*, *lapses*, *social disregard* and *parking violations* using 112 items, while Özkan & Lajunen (2005) suggested adding a positive behaviors subscale to the instrument so that it would measure *violations*, *errors* and *positive behaviors* using 38 items. Similarly, culture-specific versions of the instrument have been created for individual studies: for instance, Sümer (2003) formulated a Turkish version of the instrument with 28 items specific to the Turkish traffic context while Xie & Parker (2002) constructed a 29-item Chinese version of the DBQ containing specifically Chinese traffic behaviors such as driving on a bicycle lane when the road is congested. These different versions of the instrument are not directly relevant for the concerns of the present thesis, but are mentioned here for the sake of completeness.

1.1.1 Studies examining the measurement properties of the DBQ

Early DBQ studies such as Blockey & Hartley (1995) raised the question that the DBQ factor structures might differ across countries, traffic cultures, drivers of different ages and across sexes. The similarity of DBQ factor structures across groups of respondents has subsequently been investigated using different methods.

The first study to investigate the measurement properties of some version of the DBQ was a Swedish study that focused on the 32-item version of DBQ-SWE that aims to measure *violations*, *mistakes*, *inattention errors* and *inexperience errors* (Rimmö, 2002). The model fit roughly equally well across sexes and in data from new and inexperienced drivers (in all groups the RMSEA was < 0.05, for instance). Model fit to data from young drivers and experienced drivers was slightly worse.

The above-mentioned study of Lajunen et al. (2004) was close in spirit to Study II of the present thesis in that both studies compared factor structures across traffic cultures. Lajunen et al. (2004) based their analysis on comparing the similarity of EFA loadings across three countries (Great Britain, Finland and the Netherlands) based on several descriptive statistics

(Pearson correlations, Tucker's Phi coefficients, additivity and identity coefficients). Many of these indices received quite high values when comparing the factor structures; for instance, the correlations between the factors across samples ranged from 0.86 to 0.91 and the Tucker's phi values from 0.94 to 0.98, respectively. It seems clear, then, that the factors examined by Lajunen et al. (2004) were quite similar across samples. Translated into the language of measurement equivalence testing (Section 2.3.2.), the result most closely corresponds with testing the *configural equivalence* of the factor solutions across countries; in other words, assessing whether the factor loading patterns were similar across countries. Still, Lajunen et al. (2004) describe several important differences in these patterns across the countries. The present thesis builds on these results by 1) teasing apart different forms of similarity of factor structures and 2) presenting rigorous statistical tests on the similarities of factor structures across groups of drivers. Similarly, Özkan, Lajunen, Chliaoutakis, Parker, & Summala (2006) investigated the cross-cultural similarity of the DBQ factor structures across data sets obtained from Finland, Great Britain, Greece, Iran, the Netherlands and Turkey. The study was based on a 19-item version of the DBQ that was obtained by dropping the 8 items related to lapses from the 27-item version of the instrument. A confirmatory factor analysis indicated that the model had at best a moderate fit to data from the six countries (for instance, the values of the CFI fit index ranged from 0.79 to 0.87 and the RMSEA ranged from 0.05 to 0.09). The authors also investigated the similarity of EFA patterns across countries using the same indices as Lajunen et al. (2004). Unlike Lajunen et al. (2004), they state that the factors *aggressive violations* and *errors* were quite dissimilar across countries. The remaining factor, *ordinary violations*, was more similar across countries.

The stability of 2–6 factor solutions of a 21-item version of the DBQ across time was investigated by Özkan, Lajunen, & Summala (2006). The authors found that only the two- and four-factor solutions were interpretable across time. Among these, only the two-factor solution showed adequate stability across time, leading the authors to conclude that "In spite of its good cross-cultural validity, DBQ showed surprisingly low test–retest factor stability over three years in the present study". In addition, at least two other studies have assessed the longitudinal measurement equivalence of the DBQ. Roman et al. (2015) concluded that longitudinal scalar equivalence (equivalence of factor loadings and item intercepts, see Section 2.3.2) of the 27-item DBQ holds in a sample of young drivers (same data as used in Study III of the present thesis), while longitudinal scalar equivalence holds for a 47-item version of the DBQ after dropping certain items in a sample of old drivers (Koppel et al., 2018).

The fit of two-, three-, and four-factor models in different subgroups of Danish respondents that were constructed based on age, sex and annual mileage was tested by Martinussen, Hakamies-Blomqvist, Møller, Özkan, & Lajunen (2013), who administered a 27-item version of the DBQ that they had derived from the original 50-item DBQ. The two-factor model of *errors* and *violations* fit the data quite poorly, while the three-factor model of *errors*, *lapses* and *violations* and the four-factor model of *unfocused errors / lapses*, *emotional violations*, *reckless violations / lapses* and *confused errors / lapses* had a better fit to data. Nonetheless, none of the three models had adequate fit in terms of the CFI index in any of the 15 subgroups in which the model was fit.

Various studies on the measurement equivalence of the DBQ have been carried out since Study I of the present thesis was published. Stephens & Fitzharris (2016) carried out a rigorous study assessing the measurement equivalence of the 28-item DBQ across age groups and genders in a representative sample of Australian drivers. The study employed a largely similar experimental design as Study I, even though Stephens & Fitzharris (2016) used using Confirmatory Factor Analysis (CFA) instead of Exploratory Structural Equation Modeling (ESEM) as an analysis method. Stephens & Fitzharris (2016) fit a four-factor model derived from earlier research (*violations, aggressive violations, lapses and errors*) and found that the model had a "tolerable" fit to the whole sample of drivers once the error variances of the speeding-related items were allowed to covary. Full strong (scalar) equivalence was found when comparing genders, while partial strong equivalence was obtained when comparing the age groups of 26–39-year-olds and 40–64-year-olds. The four-factor model fit only after dropping several items and correlating certain error variances in the youngest and oldest age groups (drivers of ages 17–25 and 65–75 years, respectively), and it was deemed inappropriate in a group of professional drivers. Further, the four-factor model employed by Stephens & Fitzharris (2016) has been shown to work quite well in an Italian sample (Spano et al., 2019).

In Stanojević, Lajunen, Jovanović, Sârbescu, & Kostadinov (2018), none of the commonly used factor solutions (ones with either two, three or four factors) fit adequately when comparing model fit across three South-East European countries (Bulgaria, Romania and Serbia) based on the 27-item DBQ. Adding higher-order factors to the model did not produce adequate model fit, either. Because of this, the authors ended up with two factors (or PCs, *errors* and *violations*) that were qualitatively roughly similar across the three countries. The authors also compared the frequencies of individual traffic behaviors across the countries to better understand the differences between the countries.

Sullman et al. (2019) fit an EFA in one sample of drivers in New Zealand and used CFA in an independent sample to test model fit. After running the EFA, the authors deleted two items (*drunk driving* and *overtaking on the inside*). The factor loadings of the EFA indicated that the factors were different in nature from those reported by e.g. Lajunen et al. (2004) and Stephens & Fitzharris (2016) but still similar enough that their original names were retained. The authors concluded the configural model fit well (even though some of the fit indices, such as the CFI at 0.90 had not entirely satisfactory values) and proceeded to comparing factor means across gender, age groups and drivers who had vs. had not been involved in a crash.

Other studies carried out in different traffic cultures have indicated that the four-factor model fails to offer a universally applicable factor solution for the 27 / 28 –item DBQ. For instance, in a study based on the 28-item DBQ in China (Chu, Wu, Atombo, Zhang, & Özkan, 2019) a solution with three factors was deemed appropriate. The first factor was named *errors* even though it comprised also different behaviors traditionally considered as *violations* (*drink & drive*, *disregard speed limit*, *close following*, *forcing your way on another lane*), while the *violations* factor comprised mostly items that are in one way or another related to aggressive behavior. The authors also removed several items from the instrument "to optimize its psychometric properties"; the items included, among others, two items related to exceeding speed limits and one related to driving when drunk.

The relationships between the DBQ and being involved in a car crash have mainly been examined by correlating the latent variables with accident data. These correlations have been examined in two meta-analyses (de Winter, Dodou, & Stanton, 2015; de Winter & Dodou, 2010). Due to different numbers of items and factors in the various DBQ studies, the analyses in both studies were based on the *errors* / *violations* dichotomy. According to the first meta-analysis (de Winter & Dodou, 2010), errors and violations have roughly similar zero-order correlations with being involved in a car crash (0.10 and 0.13, respectively). The second meta-analysis (de Winter et al., 2015) updated the correlations to 0.09 and 0.13, respectively.

The use of the DBQ as a tool for predicting accidents has, however, also been questioned on several grounds. First, the correlation between the DBQ scales and self-reported crashes has been argued to arise due to common method variance, since it has been shown that the original 50-item version of the scale predicts self-reported crashes, but not those that have been shown to occur to bus drivers according to company records or to drivers according to police records (af Wåhlberg, Dorn & Kline, 2011). Similarly, another meta-analysis (af Wåhlberf, Barraclough & Freeman, 2015) indicated that the correlation between the violations scale and self-reported accidents was much higher ($r = 0.147$) than that between the violations scale and recorded crashes ($r = 0.023$); the meta-analysis also argued that the higher correlation between the self-reported crashes and the violation scale may have been due to common-method bias and the confounding effects of exposure.

Interestingly for the present thesis, a previous study (Wallén Warner, Özkan, Lajunen, & Tzamalouka, 2011) has also examined the relationships between individual driver behaviors and being involved in a crash. The study found regression coefficients in the range of 0.12 – 0.32 for individual traffic behaviors in a Poisson regression analysis that controlled for age, gender and annual mileage. In an analysis incorporating data from different countries, crashes were predicted by *getting angered*, *disregarding the speed limit on the motorway* (interestingly with a negative coefficient, i.e. that speeding protected the drivers from crashes), *disregarding the speed limit within residential areas*, *overtaking on the inside*, *pulling out of a junction dangerously* and *getting into a wrong lane after a roundabout* (again with a negative coefficient). In analyses that considered different countries separately, *overtaking on the inside* was found to be predictive of crashes in Greece and *becoming angered* and *disregarding the speed limit* within residential areas in Turkey.

Overall, a remarkable number of instruments, all referred to as “the DBQ”, have been published. The instruments differ in the number and content of items, number and nature of latent variables and the methods of data analysis used to arrive at the latent variables. The measurement properties of the various versions of the DBQ have been examined using descriptive statistics such as correlations among factor loadings across groups, but also using modern measurement equivalence analyses (see also Section 2.3.2). No universally well-fitting model has been found, even though the common 2- and 4-factor solutions for the 27- / 28-item version of the DBQ have shown at best even strong measurement equivalence across groups of drivers or across time; then again, other studies performed in different traffic cultures have shown that the same models do not fit data at all.

1.2 Motivation for the studies of this thesis

When operating within the latent variable view of errors and violations, being able to correctly specify the relationships between the observed and latent variables (in this case, traffic behaviors and the GEMS variables, respectively) is a critical prerequisite for using the latent variables in further analyses, such as predicting accidents. More technically: correctly specified measurement models are a prerequisite for formulating intelligible structural models (see e.g. Kline, 2011, ch. 7). Likewise, for between-group comparisons on the latent variables to make sense (for instance, for asking whether men and women are equally violation-prone in traffic), the measurement models must be correctly specified and the same measurement model must apply across groups. In other words, the observed variables must be connected to the correct latent variables in models such as the one given in Figure 1, and the same measurement model must apply to all the groups to be compared. Self-report studies of traffic behavior have since the seminal publication (Reason et al., 1990) been motivated by an interest in comparing groups on the latent variables and in predicting accidents, but as described in Section 1.1., rigorous studies investigating the measurement properties of the DBQ have been scarce.

Studies I and II in this thesis address the measurement properties of the DBQ. Study I is based on data previously collected in Finland; study II based on data previously collectd in Finland and Ireland. More specifically, Studies I and II focus on whether the psychological properties outlined in the GEMS can be similarly measured across subgroups of respondents that are formed based on age, gender and nationality. The conclusions of these articles are largely critical in that they highlight shortcomings in the measurement properties of the DBQ. From the practical point of view this is important, since the operationalized GEMS variables have been widely used as mediators when relating various background variables to accident risk. For instance, error- and violation-proneness have been regressed on background variables such as driving experience, age and sex while using them as predictors of accidents. Such models make sense only when the latent variables have the same structure (or "mean the same thing") across subgroups of people. Thus, even though sum scores have been widely used in various analyses, the assumption that the instrument functions similarly across (groups of) respondents has often been taken for granted instead of being rigorously assessed.

Study III offers a constructive contribution in terms of presenting a network model of traffic behavior. It is based on publicly available data collected in the United Kingdom. The contribution can be understood by making a comparison with the causal assumptions of the latent variable view of errors and violations, according to which the individual driving behaviors (speeding, drunk driving, misjudging speeds or distances etc.) function as causally passive reflections of the underlying latent variables. In a network model, on the other hand, direct pairwise relationships among the traffic behaviors are the phenomenon of main interest. For instance, it is assumed that drivers who are more likely than the average driver to speed may also be more likely to overtake others dangerously, miss observing traffic signs or other road users and so on. Such direct associations are problematic for the latent variable view of violations and errors because according to that view, speeding and missing observing something function as measurements of different latent variables (violations and errors, respectively).

One of the essential problems of comparing sum scores of observed DBQ variables can be illustrated by the following thought experiment from Study III (the same logic applies - *mutatis mutandis* - to comparisons of latent means):

> *"Consider two imaginary persons filling in the DBQ: John, known for his quick temper, answers the three items related to aggressive behavior with the option "nearly all the time," and reports performing no other violations, thus obtaining the sum score of 21. Bill, on the other hand, known for his careful nature, chooses the option "never" to the aggression-related items and the option "hardly ever" or "occasionally" to the other violation items. As there are many more items related to non-aggressive violations than to aggressive ones, both respondents receive identical scores, even though their behavioral profiles are quite different."*
>
> *Study III*

Study III also presents a predictive model of crashes based on the individual traffic behaviors. This is in contrast to most of the predictive models based on the DBQ: it has been common practice to first create sum variables to represent the latent dimensions of interest and then use the sum variables in predicting accidents. The model is motivated by the idea that an individual traffic behavior may function as an excellent predictor of accidents, completely irrespective of how much it correlates with other traffic behaviors. For instance, driving under the influence of drugs may well have a low correlation with other traffic behaviors, which would lead to the corresponding variable being rejected in a factor analysis – even though based on a clinical consideration, it is, on the contrary, important to include it as a predictor. Stated more technically, such analyses carry the benefit of capitalizing on the unique variation in the question items related to these behaviors. Still, it has been common practice to leave out items related to such infrequent driver behaviors precisely due to their low correlations with other items; see, e.g., Lajunen et al. (2004) and the distinction between the 27- and the 28-item versions of the DBQ.

Another novel contribution of the predictive models presented in Study III is that the models are built and tested in independent sets of data, thus offering a novel perspective to the generalizability of predictive models widely used in traffic psychology. Traditionally, such models have been fitted and tested in the same sample of data. This leads to good fit in that particular data set, but carries a risk of not generalizing to other data sets.

The overarching theme across the three studies is that of measurement. Studies I and II assume the measurability of psychological properties such as *error-proneness* and *violation-proneness* and proceed to test a central characteristic of their measurement models – that of measurement equivalence. An important precondition must, however, be met in order for the measurement models to make sense: the psychological properties being measured must have quantitative structure. The issue is discussed at some length in Section 4.3. As the network models (Study III) do not necessarily entail the existence of quantitative psychological properties, they provide an interesting alternative to the measurement models traditionally used in traffic psychology.

2 METHODS

Various kinds of Structural Equation Models (SEMs) were used when assessing the measurement properties of the DBQ. Study I was based on using Exploratory Structural Equation Models (ESEMs) to investigate the measurement equivalence of the DBQ across subgroups of Finnish respondents. ESEMs combine the flexibility of EFAs with the statistical tests commonly employed with Confirmatory Factor Analyses (CFAs). Study II utilized CFAs for the same purpose and also implemented a rigorous procedure for testing partial measurement equivalence. CFAs were used instead of ESEMs, as ESEMs were not yet at that time (year 2015) implemented in the open-source R programming environment, and the use of open-source software was seen as a value in itself. In addition, various graphical methods were used for making it easier to understand the multidimensional data. In Study III, psychological network models were constructed as representations of direct interactions among driving behaviors; in addition, regression models were built for predicting accidents according to the principles of statistical learning theory.

2.1 Data

The current dissertation is based on a re-analysis of previously collected data. In Finland, ethical review is not required for studies that are based on public documents, registries or archival data (National Advisory Board on Research Ethics, 2009). When it comes to the original studies where the data were collected, informed consent was inferred from returned postal or online questionnaires.

The Finnish data used in Study I is a sample of 2000 Finnish car owners, stratified by age and gender and with an equal number of men and women. The total number of responses was 1126 (for details, see Mesken et al., 2002). After removing cases without data in any of the DBQ variables, 1017 cases were retained. The dataset has previously been used also by Lajunen et al. (2004) and Özkan et al. (2006).

Study II was based on data on the driving behavior of young drivers (18–25 years of age) from Finland and Ireland. The Finnish data comprised a stratified random sample from the driving license register (Lajunen & Summala, 2004). The overall response rate was 35.3 % and the sample size 1051. The mean age of the Finnish respondents was 20.6 years, and median age 20. 62.5 % of the respondents were female, 37.5 % male. The Irish data (N = 816) comprised a convenience sample collected using an online questionnaire. The mean age of respondents was 20.3 and median age 20 years. 53.6 % of respondents were female, 46.4 % male. The respondents were college students at Trinity College, Dublin and visitors of online car forums or car sections of other online forums. The participants were motivated to participate by offering them a possibility of winning a gift voucher. The Irish data set contained no missing values as the online system used in data collection required the respondents to answer all the questions.

Study III is based on data from the longitudinal Cohort II study from 2001–2005 on new and novice drivers in the United Kingdom (Wells, Tong, Grayson, & Jones, 2008). The total sample size was 20,512 and the study comprised four waves of data collection with the following numbers of responses and response rates: 10,064 at 6 months (49%), 7,450 at 12 months (36%), 4,189 at 24 months (26%) and 2,765 at 36 months (26%) after licencure. The data comprised mostly young drivers: 59 % of respondents were under the age of 20 at the first wave of data collection, while 76 % were under the age of 25. This is representative of the population of newly licensed drivers in the UK. On the other hand, female drivers were slightly overrepresented, with 64 % of respondents (first wave) being female.

The so-called between-person network model reported in Study III was formed based on average responses across the four waves of data collection. Only cases without any missing data at any time point were included, resulting in 1,173 observations. The respondents had a mean age of 24.04 years (SD = 9.62) and 71% of them were female. In addition, a cross-sectional network model describing connections between driving behaviors and background variables was formed based on data collected at the first wave. The sample size was 8,858 when cases with no missing data in any of the variables were included. The respondents had a mean age of 22.51 (SD = 7.95), and 64 % were female. The regression analyses were similarly based on data with no missing values on the independent variables or the dependent variable (number of crashes), resulting in a sample size of 1152, 69 % female.

2.2 Questionnaires used

The 28-item version of the DBQ (Table 2) was used in Study I of the present thesis. In Study II, the 27-item version was used. In Study III, the 28-item version served as the starting point, and certain additional items that were judged as potentially relevant determinants of other traffic behaviors were included. These additional items included *driving after taking drugs* and *using a mobile phone while driving*. In addition, variables related to the *drivers' self-image*, *self-perceived improvement needs* and *attitudes* were included in the cross-sectional network analysis that is reported in Study III but not reproduced in the current thesis.

Table 2. *The questionnaire items in the 28-item version of the DBQ with their intended factor loadings in the 2-, 3-, and 4-factor solutions. Dropping the last item in the table results in the 27-item version of the DBQ.*

	Latent variable in the solution with		
Item	2 factors – Errors and Violations	3 factors – Slips, Lapses and Violations	4 Factors – Slips, Lapses, Violations and Aggressive Violations
Hit something when reversing that you had not previously seen	Error	Lapse	Lapse
Intending to drive to destination A, you “wake up” to find yourself on the road to destination B	Error	Lapse	Lapse
Get into the wrong lane approaching a roundabout or a junction	Error	Lapse	Lapse
Switch on one thing, such as the headlights, when you meant to switch on something else, such as the wipers	Error	Lapse	Lapse
Attempt to drive away from the traffic lights in third gear	Error	Lapse	Lapse
Forget where you left your car in a car park	Error	Lapse	Lapse
Misread the signs and exit from a roundabout on the wrong road	Error	Lapse	Lapse
Realise that you have no clear recollection of the road along which you have just been travelling	Error	Lapse	Lapse
Queuing to turn left* onto a main road, you pay such close attention to the main stream of traffic that you nearly hit the car in front	Error	Slip	Slip
Fail to notice that pedestrians are crossing when turning into a side street from a main road	Error	Slip	Slip
Fail to check your rear-view mirror before pulling out, changing lanes, etc.	Error	Slip	Slip
Brake too quickly on a slippery road or steer the wrong way in a skid	Error	Slip	Slip
On turning left* nearly hit a cyclist who has come up on your inside	Error	Slip	Slip
Miss “Give Way” signs and narrowly avoid colliding with traffic having right of way	Error	Slip	Slip
Attempt to overtake someone that you had not noticed to be signalling a right* turn	Error	Slip	Slip

Table 2. (continued) *The questionnaire items in the 27-item version of the DBQ with their intended factor loadings in the 2-, 3-, and 4-factor solutions*

Underestimate the speed of an oncoming vehicle when overtaking	Error	Slip	Slip
Sound your horn to indicate your annoyance to another road user	Violation	Violation	Aggression
Become angered by another driver and give chase with the intention of giving him/her a piece of your mind	Violation	Violation	Aggression
Become angered by a certain type of a driver and indicate your hostility by whatever means you can	Violation	Violation	Aggression
Pull out of a junction so far that the driver with right of way has to stop and let you out	Violation	Violation	Violation
Disregard the speed limit on a residential road	Violation	Violation	Violation
Stay in a motorway lane that you know will be closed ahead until the last minute before forcing your way into the other lane	Violation	Violation	Violation
Overtake a slow driver on the inside	Violation	Violation	Violation
Race away from traffic lights with the intention of beating the driver next to you	Violation	Violation	Violation
Drive so close to the car in front that it would be difficult to stop in an emergency	Violation	Violation	Violation
Cross a junction knowing that the traffic lights have already turned against you	Violation	Violation	Violation
Disregard the speed limit on a motorway	Violation	Violation	Violation
Drive when you suspect you may be over the legal alcohol limit	Violation	Violation	Violation

* The items marked with an asterisk are worded as indicated in the table in studies conducted in countries that use left-hand traffic (in the present thesis United Kingdom and Ireland). When referring to right-hand traffic, words "left" and "right" are interchanged in these items.

2.3 Statistical methods

2.3.1 Structural equation models

Two types of SEMs were used in Studies I and II: Exploratory Structural Equation Models (ESEMs, Asparouhov & Muthén, 2009) and Confirmatory Factor Analysis (CFA), respectively.

CFA and ESEM, being variants of factor analysis, build on the principle that variation in observed variables consists of reliable variance that is due to the latent variable(s) and of random error that is due to unmodelled causes. The error terms are assumed to be uncorrelated with the latent variables and the error terms of the other observed variables. CFA is confirmatory in that the analyst needs to specify, in advance, which model parameters to estimate. ESEM, on the other hand, comprises an EFA as a measurement model, and a central difference between CFA and ESEM is that in the latter, the relationships between the latent and observed variables do not need to be specified in advance (Asparouhov & Muthén, 2009).

In a CFA, a matrix of factor loadings (i.e. regression equations regressing the observed variables on the latent variables) is always among the parameters that must be specified. The loadings are most usually given as a matrix of ones and zeroes: a one indicates that the loading is to be estimated, a zero that it will be constrained to zero (Gunzler & Morris, 2015). CFA is, however, a flexible method and it is possible to specify the factor loadings to any value of interest: for instance, it can be tested whether factor loadings obtained in a previous study offer a good fit to subsequently collected data (Kline, 2011). Further, in a multi-group analysis the values of factor loadings can be constrained to equality across groups as discussed below in Section 2.3.2. Further parameters, such as correlations among the error terms of the observed variables, can also be specified.

The general idea of a CFA can be introduced by looking at Figure 1, where the observed variables *speeding*, *drunk driving* etc. are specified to load on the latent variable *violations*, while the observed variables *misjudge speed*, *miss signs* and *miss observing pedestrians* load on *errors*. The error terms related to, e.g., *drunk driving* (e2) and *showing aggression* (e4) could be specified to correlate if the analyst believed that they share variation that is not in its entirety explained by the position that the respondent occupies on the latent variable *violations*.

Figure 1 actually shows a full structural equation model that has a CFA as a part. The CFA is known as a *measurement model* as it specifies certain observed variables as measurements of the latent variables (the arrows emanating from the latent variables marked by the ellipses). In addition, the arrows emanating from the background variables (*being a fast driver*, *being late* etc.) and pointing to the latent variable *violations* describe a *structural model* that specifies the ways the latent variables are related to other variables; in this case, those other variables are their putative causes.

The general idea of both CFA and the measurement model in an ESEM can be described by equation (1); see also Asparouhov & Muthén (2009)

$$x_i = \nu_i + \Lambda_{ik}\eta_k + \varepsilon_i \qquad (1)$$

Here, the p × 1 vector x_i refers to the observed variables (i = 1...p), the p × 1 vector ν_i to the intercept terms, the m × 1 vector η_k to the latent variables (k =

1...m), the p × m matrix Λ_{ik} to the influence the latent variables η_k are posited to have on the observed variables and the p × 1 vector ε_i to the random error terms. It is standardly assumed that the error terms are normally distributed with mean 0 and a diagonal variance-covariance matrix Θ, even though the error terms can also be specified to be correlated with one another. Under the standard model the observed variables are conditionally independent of one another given the latent variables, i.e. $\mathrm{cov}(x_i,x_j \mid \eta_k) = 0$ for $i \neq j$.

Referring again to Figure 1, $\eta_{violations}$ is the value of the latent variable *violations* and $\lambda_{violations,speeding}$ is the influence that *violations* has on *speeding* (here, the lowercase λ refers to an individual regression coefficient whereas the uppercase Λ above refers to a matrix of such coefficients). $\nu_{speeding}$ refers to the intercept term when regressing *speeding* on *violations*. The intercept terms express the values of the observed variables when the values of the latent variables are zero; i.e. in the case of speeding, the level of speeding that is unrelated to the level of the tendency to violate traffic rules. The intercepts are sometimes interpreted as indications of the level of acquiescent responding, i.e. the tendency to agree with the items irrespective of their content (Gregorich, 2006).

The central idea of ESEMs can also be illustrated by referring to Figure 1. The ESEM is a general method, and both Exploratory Factor Analysis (EFA) and CFA can be considered as special cases of ESEM. Similarly to other structural equation models, an ESEM can contain both a measurement model and a structural model. In what follows, only the measurement models will be considered, as the studies included in the present thesis are based on using only measurement models.

The conceptual idea of formulating the measurement model as an EFA – as is commonly done when constructing an ESEM – amounts to allowing all observed variables to load on all latent variables. In terms of Figure 1, this would mean allowing factor loadings such as *errors → speeding*, *errors → drunk driving* etc. and *violations → misjudging speed, violations → missing traffic signs* and *violations → missing pedestrians*. Obtaining unique solutions for the model parameters (i.e. identifying the model) necessitates specifying certain restrictions on the general model described in equation (1), since the total number of model parameters may not exceed the number of unique elements in the covariance matrix of data. In the ESEM context, this is accomplished by first estimating the model and then applying a rotation similarly to any other EFA; the question of rotations is addressed below. The estimation proceeds by 1) specifying the variance-covariance matrix of latent variables as an identity matrix (which gives m(m+1)/2 restrictions) and 2) fixing all entries in the p × m factor loading matrix above the main diagonal to zero (which gives m(m-1)/2 restrictions); Asparouhov & Muthén (2009). These restrictions enable obtaining the initial, unrotated model. Specifying the measurement model as an EFA is useful because even though the analyst might have a clear idea on which observed variables should be related to which latent variables, the possibility of cross-loadings (minor loadings on other factors) remains. Further, this way of specifying the measurement model leaves room for surprises: perhaps the intended structure of factor loadings fails to fit in the data at hand for one reason or another. On the other hand, this strength of the ESEM can also become its weakness in that the EFA is a data-driven method, and may, because of that, result in factor loadings that are specific to the sample of data at hand that may not replicate in another sample.

As noted above, specifying a CFA model proceeds differently from specifying an EFA model: the analyst needs to specify which factor loadings (and other potential parameters) to estimate. In a typical CFA, only certain loadings are estimated, which frees up degrees of freedom for estimating other parameters (e.g. correlated error terms). Because of this flexibility, it is difficult to give an algebraic formula for describing when a CFA is identified. Nonetheless, what the analyst needs to do is to 1) choose a method for scaling the latent variables (usually either by fixing their variances to unity or by fixing a given factor loading to unity), 2) have a sufficient number of observed variables per latent variable, 3) think about the correlations between constructs (only a given number of correlated measurement errors can be specified for observed variables that load on different latent variables, or the latent variables can be specified to be uncorrelated), 4) estimate factor loadings such that for each observed variable, there is at least one other observed variable that it does not share a correlated measurement error with and 5) avoid specifying an excessive number of multiple loadings for any given observed variable to avoid problems with the so-called empirical idenfiability of the model (Kline, 2011, Chapter 6).

As mentioned above, in an ESEM the initial estimation of the model is followed by rotating the factor matrix. Factor rotation refers to, as the name indicates, rotating the m-dimensional coordinate system with the m factors as the axes so that the loadings (points in the coordinate space) will be divided optimally among the factors (Browne, 2001). What is considered optimal in this case is a *simple structure* of factor loadings, i.e. roughly that items have a high loading on a given factor and low loadings on other factors.

The rotations can be either orthogonal (latent variables are not allowed to correlate) or oblique (they are allowed to correlate). There are many different rotation methods that optimize different mathematical criteria, but among them, a method known as *target rotation* has benefits that the other methods lack. As the name indicates, target rotation aims to rotate the factor space to reproduce a structure of loadings specified by the analyst. The target values are typically zeroes (cf. CFA in which certain loadings are typically fixed to zero) representing substantive considerations such as prior knowledge of the relationships between the observed and latent variables. In practice, the analyst specifies a $p \times m$ target matrix B such as

$$B = \begin{bmatrix} ? & 0 & 0 \\ ? & 0 & 0 \\ ? & 0 & 0 \\ 0 & ? & 0 \\ 0 & ? & 0 \\ 0 & ? & ? \\ 0 & 0 & ? \\ 0 & 0 & ? \end{bmatrix}$$

The rotation function that is then applied is of the form:

(2) $$f(\Lambda) = \sum_{i=1}^{p} \sum_{j=1}^{m} a_{ij}(\lambda_{ij} - b_{ij})^2$$

Where a_{ij} equals 1 if b is specified and 0 otherwise (Browne, 2001). In the sum terms, p = number of observed variables and m = number of latent variables (factors). Identification conditions related to using target rotation in the ESEM context are described by Asparouhov & Muthen (2009, pp. 409-411). Target rotation is flexible in that it allows the analyst to specify the expected pattern of loadings, but it nonetheless allows the loadings to differ from this structure in a data-driven manner. In this way, ESEM with target rotation can be considered to occupy a middle ground between CFA and EFA and to combine both their strengths. In a CFA, no rotation method needs to be used as the pattern of factor loadings of interest is specified by the analyst.

A structural equation model is said to fit the data if it is able to reproduce the data accurately. One way to assess model fit is to compare the observed covariance matrix of the indicators of the latent variables (in this case, the DBQ items) with the covariance matrix reproduced based on the model. These can be compared using the χ^2-test, which, however, tends to produce a significant result whenever the test is applied in practice. This happens because a large sample size is required to apply the method in the first place, and because large sample sizes lead to even minuscule differences between the covariance matrices to become statistically significant. Because of this, a host of *indices of approximate model fit* have been developed (Kline, 2011, Chapter 8).

In Studies I and II, different kinds of fit indices are reported. First, the Root Mean Square of Approximation (RMSEA) is used as a parsimony-corrected badness-of-fit index. What this means is that the index penalizes for model complexity, and receives the value of zero when the model fits the data in the best possible manner. The Bentler Comparative Fit Index (CFI) measures model fit in reference to a baseline model, which is usually chosen to be a model where the observed variables are independent of each other. The Akaike Information Criterion (AIC) used in Study II is a comparative measure of fit, which means that it can be used for comparing several models, even though it is quite meaningless by itself. The AIC is based on the

χ^2-statistic, the number of parameters to be estimated and the degrees of freedom, and it penalizes for the complexity of the model. Finally, the Standardized Root Mean Square Residual (SRMR) and the Weighted Root Mean Square Residual (WRMR) describe the size of the residual correlations, i.e. the correlations among the observed variables that remain after fitting the model to data. More information on the fit indices can be found, e.g., in Kline (2011, Chapter 8).

2.3.2 Analyses of measurement equivalence

In Studies I and II, the SEMs are used as tools in analyses of measurement equivalence, i.e. investigating whether the same latent variables are being measured in the same way across subgroups of respondents. The current standard practice in measurement equivalence analysis is based on research carried out by Meredith (1993). These analyses begin by investigating whether the observed variables are related to the same latent variables across groups, and then setting a series of increasingly tightening constraints on the models. These constraints are described in equation (3) and figure 3.

Figure 3 shows a two-factor measurement model being fit in two groups, which might be chosen based on sex, age, traffic culture etc. The measurement model consists of two latent variables, η_1 and η_2 in both groups. In other words, fitting the models presupposes that the same two-factor latent variable structure actually is the correct one in both groups. Any of the parameters of the model, such as the factor loadings λ or the intercepts ν can either be freely estimated in both groups or constrained to equality across groups. Conceptually, analyses of measurment equivalence consist of comparing the fit of models with and without such constraints; for instance, if a model that constrains the factor loadings λ to equality across groups fits roughly equally well as a model without such constraints, the factor loadings can be judged to be equal across groups.

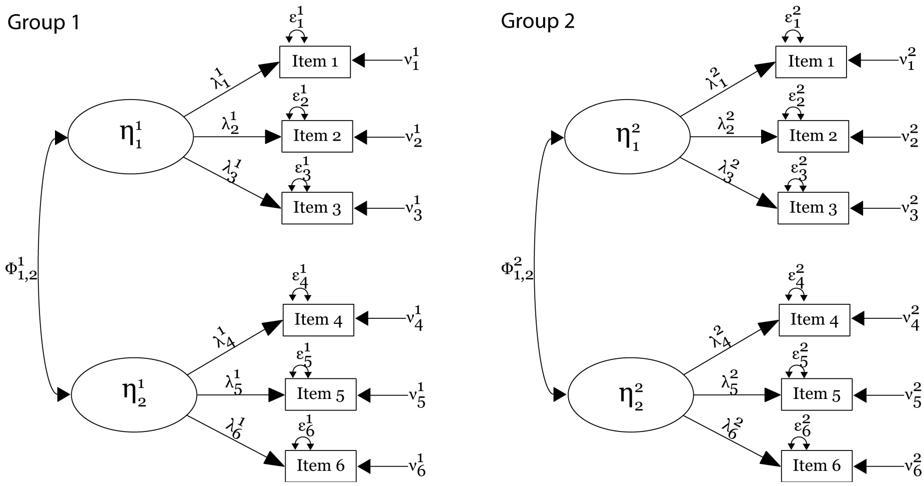


Figure 3 An illustration of analysis of measurement equivalence with two groups. The bidirectional arrows refer to correlations, the directed arrows to regression relationships. The superscripts represent the groups, i.e. η_1^1 refers two latent variable η_1 in group 1, η_1^2 to latent variable η_1 in group 2 and so on. ϕ = covariance, η = latent variable, λ = factor loading, ε = error variance, υ = intercept term (the value of the observed variable when the value of the latent variable = 0).

Measurement equivalence analyses begin by fitting the same model separately in the different groups and assessing model fit. This stage is called *analysis of configural equivalence*. Reaching configural equivalence requires that the same subsets of items are related to the same latent variables. In figure 3, this is the case, since items 1-3 are related to latent variable η_1 and items 4-6 to latent variable η_2 in both groups.

When following the procedure suggested by Meredith (1993), establishing configural equivalence is followed by a sequence of equivalence models known as weak (or metric) equivalence model, strong (or scalar) equivalence model and strict equivalence model. In these models, factor loadings λ, item intercepts ν and item error variances ε are, respectively, constrained to equality across groups in a multigroup CFA / ESEM model.

The measurement equivalence constraints can be illustrated by equation (3), which expresses the value of an observed variable x_i as a function of the latent variables η_k

(3) $$x_i^g = \nu_i^g + \Lambda_{ik}^g \eta_k^g + \varepsilon_i^g$$

where x_i^g is the $p \times 1$ vector of observed variables (i = 1...p) in group g, ν_i^g is the the $p \times 1$ vector of intercept terms when regressing x_i^g on η_k^g, the $m \times 1$ vector η_k of latent variables (k = 1...m) in group g, Λ_{ik}^g is the $p \times m$ matrix describing the influence that the latent variables η_k^g are posited to have on the observed variables in group g and ε_i^g is the p × 1 vector of random error terms. The standard assumption in the model is that the error terms are normally distributed with mean 0 and a diagonal variance-covariance matrix Θ (even though the error terms can also be specified to be correlated with one another). Under the standard model, given the latent variables, the observed variables are conditionally independent of one another, i.e. $cov(xi,xj \mid \eta k) = 0$ for $i \neq j$. Equation (3) applies to CFA analyses based on continuous observed variables x_i and is used here to illustrate the conceptual logic of measurement equivalence analyses.

If the configural equivalence model is deemed to fit equally well across groups, a constraint is introduced into the models: the respective factor loadings are constrained to equality across groups (for instance, $\lambda_1^1 = \lambda_1^2$, $\lambda_2^1 = \lambda_2^2$, $\lambda_3^1 = \lambda_3^2$ and so on in equation (3) and figure 3). This constraint is known as weak measurement equivalence (or metric equivalence; Meredith, 1993; Gregorich, 2006).

As the factor loadings are taken to be causal influences under the reflective measurement model, the test of weak equivalence is often interpreted as showing that the latent variables have the same meaning across groups, i.e. that they are interpreted in the same way across groups. The latent variable view of errors and violations is built on the framework of reflective measurement models (Howell, Breivik, & Wilcox, 2007).

If the weak / metric equivalence model does not fit significantly worse than the configural equivalence model, an additional constraint is introduced: the item intercepts are constrained to equality (e.g. $\nu_1^1 = \nu_1^2$, $\nu_2^1 = \nu_2^2$, $\nu_3^1 = \nu_3^2$, and so on, equation (3) and figure 3). Passing this test – known as the test of strong (or scalar) equivalence – enables the researcher to compare factor means across groups, as in that case systematic biases in item means can be judged to be similar enough in the groups under comparison (Meredith, 1993; Gregorich, 2006).

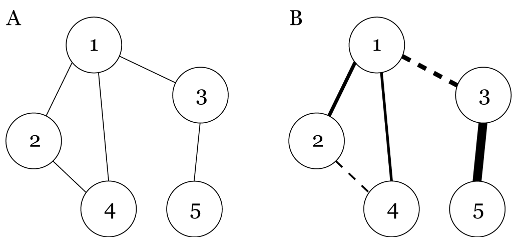
If strong measurement equivalence is established, item errors can be constrained to equality (e.g. $\varepsilon_1^1 = \varepsilon_1^2$, $\varepsilon_2^1 = \varepsilon_2^2$, $\varepsilon_3^1 = \varepsilon_3^2$ and so on, equation (3) and figure 3). This constraint is known as that of strict equivalence, and can be interpreted as showing that the observed variables have similar reliabilities across groups (Meredith, 1993; Gregorich, 2006). It is also noteworthy that this test needs to be passed for comparisons of sums of observed variables to be warranted. Comparisons of such sum variables have been of interest in studies of self-report traffic behavior since the seminal study of Reason et al. (1990), but to the best of my knowledge, no study has ever showed the strict equivalence to hold for the models in question.

Statistical tests on whether the constrained models fit the data are based on the idea of nested models. For instance, the weak equivalence model is nested within the configural equivalence model and the strong equivalence model in the weak equivalence model; in other words, the set of parameters constrained to equality is a subset of the set of parameters estimated in the previous model. This makes it possible to use the χ^2-test to compare the models and to calculate various fit indices to assess how much model fit worsened through introducing the constraint.

2.3.3 Introduction to network analysis

The network approach to data analysis can be seen as a practical application of mathematical graph theory. Network methods have been applied in domains ranging from different branches of physics to biology, computer science, economics, sociology and traffic research (Borgatti, Mehra, Brass, & Labianca, 2009; Holme, 2003; Mason & Verwoerd, 2007; Newman, 2008).

In network analysis, the phenomenon of interest is described using nodes connected by edges, which can be either weighted or unweighted. For instance, when analyzing air traffic, the nodes might correspond to airports and the weighted nodes to traffic volumes between them. Further, depending on the analysis question, the edges may be signed in addition to being weighted. For instance, when analyzing social networks, the intensity of friendship relations might be represented as positively signed weighted edges and the intensity of adversarial relations as negatively signed weighted edges. Figure 4 illustrates these ideas. The unweighted network in the left panel could correspond to airline routes between cities or social connections between people. The weighted and signed network in the center might represent social relations, with nodes (persons) 3 and 5 having a strong positive connection, nodes (persons) 1 and 3 having quite a strong negative connection and so on.



C

$$\begin{pmatrix} 0 & 4 & -5 & 3 & 0 \\ 4 & 0 & 0 & -2 & 0 \\ -5 & 0 & 0 & 0 & 10 \\ 3 & -2 & 0 & 0 & 0 \\ 0 & 0 & 10 & 0 & 0 \end{pmatrix}$$

Figure 4 Examples of network graphs. Panel A: an unweighted graph. Panel B: a weighted and signed graph, with continuous lines representing positive edges and dashed lines negative edges between nodes. Panel C: Network from panel B represented as a matrix.

The network perspective has recently been applied in modeling psychological phenomena. This approach has come to be known as *network psychometrics* (Epskamp, 2017), and the resulting models have been referred to as *psychological networks*. In such models, nodes represent psychological variables such as symptom scores, self-reported behaviors, experiences or attitudes etc. The edges represent statistical relationships between the variables quantified in some way. Cross-sectional psychological networks have usually been based on co-occurrence data of one sort or another, either correlations (Cramer et al., 2012a) or – as has become more common – partial correlations controlling for the rest of the nodes of the network (Epskamp & Fried, 2018). As such, the networks would only be visualizations of partial correlations. They become models of the underlying phenomena through low partial correlations being constrained to exactly zeroes by using a regularization method such as LASSO estimation (Least Absolute Shrinkage and Selection Operator, introduced in more detail in section 2.3.6. below and in Epskamp & Fried, 2018). Network models can also be constructed based on longitudinal data, where the vector autoregression method has most commonly been applied (Epskamp et al., 2018). Longitudinal data makes it possible to estimate both contemporaneous (relationships between nodes at the same time point) and temporal (an earlier observation predicts a latter one) relationships. The temporal dimension enables the analyst to construct directed networks where the edges are asymmetric. The issue is mentioned here mainly for the sake of completeness, since intensive longitudinal data of this type was not used in the present thesis.

The psychological network models build on the idea of Gaussian Graphical Models (Lauritzen, 1996), which, as the name implies, assume multivariate normally distributed data. In the cross-sectional case, the network structure is related to the inverse of the covariance matrix (also known as the precision matrix). Given a p-dimensional multivariate normal vector X with covariance matrix Σ and precision matrix Ω, a key property of the precision matrix is that zero entries show that pairs of variables are conditionally independent given all other variables. Partial correlations are then obtained by standardizing the off-diagonal elements of the precision matrix and reversing their signs. These issues, together with estimation methods, are described in more detail in, e.g., Kuismin & Sillanpää (2017).

The phenomena of main interest in network models conceived as GGMs are the conditional (in-)dependence relationships between pairs of variables x_i and x_j when conditioned on all other variables x_k ($k \neq i,j$). The edges that remain between nodes x_i and x_j in such networks are interpreted as pairwise interactions that can be given substantive interpretations. These direct relationships are assumed to represent either uni- or bidirectional causal connections, logical entailments (for instance, if questionnaire items contain "*I am able to walk 100 m*" and "*I am able to walk 1 km*", answering in the affirmative to the second entails an affirmative answer to the first, Kossakowski et al., 2016), semantic relationships of the items having similar contents or indeed, in some cases the effects of an unmodeled latent variable affecting the nodes (Costantini et al., 2015).

The role and interpretation of partial correlations in cross-sectional network models can be contrasted with how they are used in factor analysis, where partial correlations (conditioning on all other variables) are treated mainly as a nuisance. For instance, measures such as the KMO measure of sampling adequacy and the examination of anti-image correlation matrices indicate that data is suitable for factor analysis when the off-diagonal elements of partial correlation matrices are small (see, e.g. Hauben, Hung, & Hsieh, 2017, for a practical application of these ideas). This practice reflects the causal assumptions made in latent variable models: the observed variables correlate because they are affected by underlying latent variables. So, even though cross-sectional psychological network models (NMs) and latent variable models are both based on assessing the covariance of the observed variables, the underlying causal assumptions and the role of pairwise interactions are markedly different.

A hypothetical cross-sectional network model of traffic behavior was shown in Figure 2 in Section 1. First, the different causal assumptions of CFAs and NMs can be understood by comparing Figure 2 with Figure 1. While the items *showing aggression*, *overtaking on the inside*, *drunk driving*, and *speeding* are assumed to reflect the influence of a latent variable – proneness to violate rules – in the reflective measurement model shown in Figure 1, they are shown as being connected by direct edges in the NM shown in Figure 2. Such edges can be readily given a causal interpretation: being drunk can be hypothesized to lead to aggressive behavior and to exceeding speed limits, for instance. With cross-sectional data, such causal hypotheses can naturally not be proven or disproven. While such direct connections can in principle be represented in CFAs as correlated error terms of the observed variables, this is possible for only a few observed variables at a time (due to reasons related to the identifiability of the model, see, e.g. Chapter 6 in Kline, 2011 or Gunzler & Morris, 2015); stated technically, local independence is said to be a fundamental assumption of SEMs / CFAs. Further, the fundamental assumption of CFAs remains: in them, traffic behaviors are causally determined by underlying psychological properties.

Second, insofar as the edges of the NM are drawn based on partial correlations, the NM represents relationships of conditional independence among the traffic behaviors. Looking at Figure 2, this would mean, for instance, that *speeding* and *showing aggression* have been shown to be conditionally independent when conditioning on the other nodes of the network. Importantly, this would not mean that *speeding* and *showing aggression* would need to be uncorrelated in the raw data; rather, their raw correlation would be interpreted as being explained by the (potentially causal) relationships they have with the other nodes of the network. Even though Figure 2 is drawn using unweighted edges, the partial correlation network would naturally include weighted edges.

Finally, the NM shown in Figure 2 includes direct connections between background variables and the traffic behaviors. For instance, *enjoying speed* and *owning a comfortable car* are shown as directly linked to *speeding* instead of the latent variable *violations* as in Figure 1. Even though such connections would be possible to specify even in SEMs, they would not accord with the fundamental motivation of using reflective measurement models as in them, the observed variables function as causally passive indicators (or measurements) of the underlying latent variables, such as *violation-proneness* and *error-proneness*. All in all, in a psychological network model, individual observed variables are seen not as indicators of latent factors, but rather as components of a network, with a component referring to a part of the network having unique causal relationships with the rest of the network (Borsboom, 2017; Cramer et al., 2012a).

Various descriptive indices have been developed to characterize networks. The indices have a long history, and they have been developed outside the context of network psychometrics. The indices belong into two main classes, *centrality indices* and *clustering indices*. The former describe the importance of nodes in a network in different ways, the latter the redundancy of the nodes. The perhaps simplest centrality index indicates the number of nodes that the focal node is connected and is known as the *degree* of the node. A generalization to weighted networks is known as the *strength* of a node, and it indicates the sum of the absolute values of the connection weights of the focal node. Both degree and strength are concerned with the connections that the focal node has with its immediate neighbors, i.e. nodes that it is directly connected with (Opsahl, Agneessens, & Skvoretz, 2010).

There are, in addition, centrality indices that consider the relationships that a focal node has with all other nodes. Closeness centrality is, for instance, defined as the reciprocal of the summed distances to other nodes from a focal node (Opsahl et al., 2010), with distance referring to the reciprocal of connection strength in weighted networks. Thus, a node with high closeness centrality can reach other nodes quickly; for instance, in a social network, an individual with high closeness centrality would easily be able to affect the opinions of other people. In addition, betweenness centrality indicates the number of times that the focal node lies on the shortest path between two other nodes (Opsahl et al., 2010).

In psychological network models, the centrality indices have been used in identifying nodes (symptoms, behaviors, attitudes etc.) that play an important role in the network in some sense. It has been suggested, for instance, that central symptoms in a depression network may be useful for predicting the probability of relapse and the treatment response (Fried, Epskamp, Nesse, Tuerlinckx, & Borsboom, 2016). The edges in psychological networks are often interpreted through the idea of activation: central nodes are seen as important determinants of whether other nodes become activated or not. There are numerous further indices to describe node centrality in networks, and more information on them can be found in, e.g., Koschützki et al. (2005).

Clustering indices (Saramäki, Kivelä, Onnela, Kaski, & Kertesz, 2007) come in two flavors, local and global. In the present thesis, only local clustering indices are examined. The local clustering coefficient of a focal node describes how strongly its neighbors are connected with each other; the stronger these connections are, the more redundant the focal node in the network.

Many of these indices have been first developed for describing unweighted networks and then generalized to apply to weighted networks. Some clustering indices, such as Zhang's index used in Study III have also been generalized to the weighted and signed case (Costantini & Perugini, 2014). This is not the case for the centrality indices, though, so absolute values of connection weights were used when calculating them.

2.3.4 Statistical methods for Study I

Study I was based on combining the use of CFA (referred to in the article as a Structural Equation Model) and an ESEM. The models were built for comparing the factor structures of the 27-item DBQ across age groups and genders in a sample of Finnish drivers.

The ESEM combines features of Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) in that the measurement model is estimated similarly to an EFA, while model fit is described similarly to CFA (Asparouhov & Muthén, 2009). Further, and importantly for present purposes, the ESEM enables multi-group measurement equivalence analyses. Article I refers to them as analyses of "measurement invariance", but this nomenclature is unnecessarily complicated, as the lack of measurement invariance would be referred to "measurement non-invariance". The terms "measurement equivalence" and "measurement non-equivalence" are thus used in the present thesis.

In Study I, in which the observed variables are treated as ordinal, the model specified in equation (1) is modified such that X is replaced by X^* according to equation (4).

(4) $$X = k \Leftrightarrow \tau_k < X^* < \tau_{k+1}$$

In other words, it is acknowledged that the ordinal variables X provide an imperfect approximation of underlying continuous variables X*, and that it would in principle be possible to obtain more accurate estimates of the values of these properties. For instance, if x refers to frequency self-reported speeding on a Likert scale from 0 to 5, x^* might be actual time spent speeding, cut into discrete categories at thresholds $\tau_0 \ldots \tau_c$, where c = 5. In Study I, the Likert variables had highly skewed distributions, and were consequently recoded as follows: 0 = 0, 1 = 1, 2–5 = 2.

In Study I, factor loadings were estimated as in an exploratory factor analysis. Model identifiability was ensured by placing several constraints on the parameters of the model. These constraints are not essential for understanding the conceptual logic of the present study, but they can be found in Asparouhov & Muthén (2009), equations (5) – (14). An important consideration in performing an ESEM analysis is that of choosing a suitable rotation method. In Study I, an oblique target rotation to structure derived from an earlier study (Lajunen et al., 2004) was applied; the target was chosen to reflect a clean simple structure without cross-loadings.

The equation for target rotation is shown above in equation 2. It is based on minimizing the squared differences between the actual loadings λ_{ij} and the prespecified targets b_{ij}. Treating the observed variables as ordered categorical influenced the analyses of measurement equivalence, as well. This had two consequences: 1) When testing strong measurement equivalence, item thresholds τ rather than item intercepts ν were constrained to equality and 2) weak and strong measurement equivalence were tested at the same time for model identification purposes (Muthén & Asparouhov, 2002). Even though tests of weak and strong equivalence could not be teased apart, the essential conceptual logic of measurement equivalence testing remained the same. The analysis plan for Study I is shown in Figure 5.

The fit of the models used in Study I was assessed using the following fit indices: χ^2-test, RMSEA, CFI, TLI and WRMR.

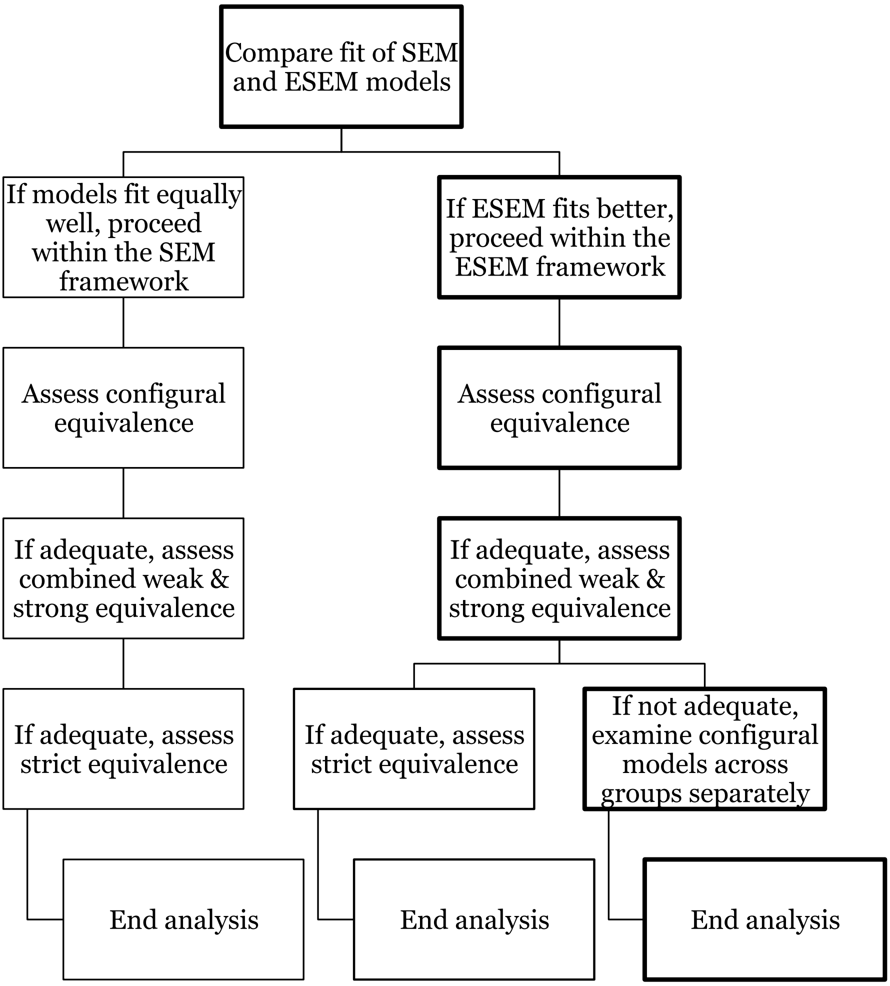


Figure 5 Analysis plan for Study I. The rectangles circled in bold line show the analysis path that was followed.

2.3.5 Statistical methods for Study II

Study II was based on the use of CFAs instead of ESEMs. CFA analyses followed the conceptual logic of fitting SEMs (section 2.3.1.) and performing analyses of measurement equivalence (section 2.3.2.) presented above. In Study II, more fine-grained distinctions related to measurement equivalence were made than in Study I. First, treating the observed variables as continuous enabled teasing apart analyses of weak and strong measurement equivalence. Second, analyses of partial measurement equivalence were performed. The analysis plan followed in Study II is presented in Figure 6.

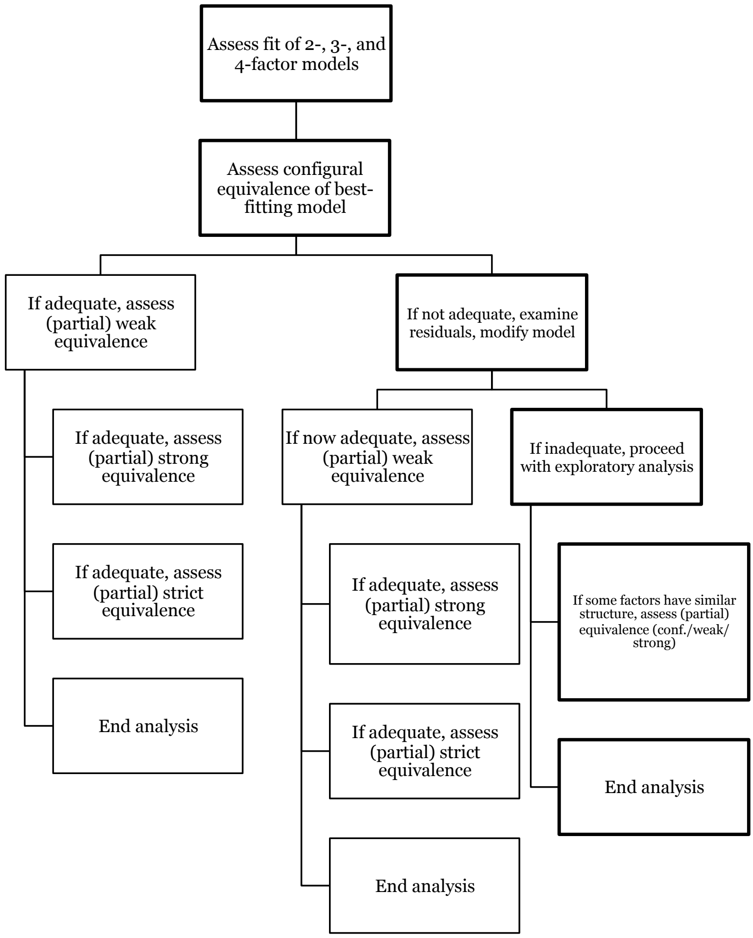


Figure 6 Analysis plan for Study II. The rectangles circled in bold line show the analysis path that was followed in Study II.

Partial measurement equivalence refers to constraining only a subset of parameters to equality at each stage of equivalence testing. The idea can be illustrated by an example. Let us assume that an analyst has concluded that the weak equivalence model does not fit her data. After examining various statistics related to model fit, she may conclude that the problem lies in the loading of the second variable on the first latent variable, which differs across groups ($\lambda_2^1 \neq \lambda_2^2$ in figure 3 and equation 3). After freeing the loading in question to be freely estimated in both groups but holding other loadings of the latent variable equal across groups (i.e. $\lambda_1^1 = \lambda_1^2$, $\lambda_3^1 = \lambda_3^2$), model fit can be tested again. If the model fits the data, the analyst can conclude that her data passes the test of partial weak equivalence. Partial strong equivalence can be tested similarly, just constraining a subset of item intercepts to equality.

When assessing the fit of the models in Study II, the following fit indices were used: χ^2-test, RMSEA, CFI, the Akaike Information Criterion (AIC) and SRMR.

2.3.6 Statistical methods for Study III

2.3.6.1 Network models

In Study III, the network edge weights represent partial correlations between variables x_i and x_k, controlling for all other observed variables X:

$$(5) \qquad Cor(x_i, x_k | X^{-(i,k)}) = \omega_{ik} = \omega_{ki}$$

Partial correlations are affected by sampling variation, which has the effect that no edge weight becomes exactly zero, $|\omega_{ik}| > 0$ for all *i,k*. This is so even if the nodes *i* and *k* would in reality be conditionally independent at the population level when conditioned on all other nodes in the network. To account for this, the partial correlation networks were estimated using the LASSO procedure, which adjusts the absolute values of all correlations slightly toward zero. This is the main principle of the statistical technique of *shrinkage*: trading a slight increase in bias with a reduction in variance (Tibshirani, 1996). Using LASSO, partial correlations with low absolute values become exactly zero (resulting in *sparse networks*), which has the desirable effect of producing more generalizable results that are less likely to overfit sample data (as raw partial correlations would).

The LASSO controls the level of sparsity based on the value of the tuning parameter λ. When $\lambda = 0$, all edges remain in the network, and when $\lambda = max(|\omega_{ik}|)$, i.e. equal to the largest absolute value of the partial correlations, no edges remain. The optimal value of λ is chosen by fitting a large number of different network models with different values of λ (typically 100) to the data. The fit of these models is typically assessed based on the Extended Bayesian Information Criterion (EBIC, Foygel & Drton, 2010) whose value is a function of the likelihood for the LASSO-estimated inverted covariance matrix, the number of edges retained, sample size and the number of parameters in the model (Williams & Rast, 2020). The value of λ that minimizes the EBIC is chosen. The EBIC penalizes for model complexity, and results in a model with high specificity but varying sensitivity depending on sample size and network density (Epskamp, Borsboom, & Fried, 2017). A variant of the LASSO known as the graphical lasso was used in Study III (Friedman, Hastie, & Tibshirani, 2008), and the value of the EBIC hyperparameter γ set to 0.5 as per the recommendation of Foygel & Drton

(2010) to produce sparse models (high specificity) and to err toward parsimony or caution rather than discovery and possibly spurious results.

The accuracy of the edge weight estimates and the stability of the centrality index estimates were assessed using bootstrap analyses. A non-parametric bootstrap procedure was used for calculating 95 % confidence intervals for the edge weight estimates because 1) the input variables were ordinal in nature and 2) the LASSO procedure, which biases estimates downward, was used; under such circumstances, a parametric bootstrap would have produced biased results. The stability of centrality index estimates was assessed using a so-called *m* out of *n* bootstrap, which gradually drops a larger and larger number of observations and re-estimates the values of the index. The correlation between the original estimates and the average of the bootstrapped estimates was then calculated. If the correlation dropped rapidly with dropping observations, the centrality index was considered unstable and was not reported. The Correlation Stability coefficient (CS-coefficient) was also reported. It is defined through the proportion of observations that can be dropped such that the correlation between the original and the average of the bootstrapped estimates does not fall below 0.7. For more details, see Epskamp et al. (2017).

2.3.6.2 Poisson regression models

In addition to the network models, Study III presented several regression models for predicting crashes from individual DBQ variables. The models were truly predictive in that the independent variables used in them were collected at an earlier time point (6 months post-licensure) than the dependent variables (number of crashes at 7–36 months post-licensure). The regression models were first fit in training data and then tested in an independent hold-out data (75 / 25 ratio). Three Poisson regression models were fit: 1) a naive Poisson regression model, 2) a ridge regression model, 3) an elastic net model (Zou & Hastie, 2005).

Poisson regression is used when the dependent variable is a count – in this case, the number of crashes that the driver had been involved in. In Poisson regression, the logarithm of the count is modelled as a linear function of the independent variables according to equation 6 (referred to as naive Poisson regression in Study III).

(6) $$log(\hat{\mu}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

where $\hat{\mu}$ is the crash count and $x_1 \ldots x_p$ are the predictor variables.

The ridge regression model and the elastic net model and are two different variants of regularized regression. In regularized regression, the regression weights (β) are adjusted downward to avoid overfitting the model to the data at hand and to produce more generalizable results. The ridge penalty involves constraining the sum of the squares of regression coefficients $\beta_1 \ldots \beta_p$ such that the sum does not exceed the value of a tuning parameter t. In other words,

$$(7) \qquad \sum_{j=1}^{p} \beta^2 \leq t$$

The optimal values for the tuning parameter were chosen using 10-fold cross-validation (Picard & Cook, 1984) in the training sample. 10-fold cross-validation is based on the idea of first fitting the model in 9/10 of the data and testing it in the remaining part of the data (the validation set). This procedure is repeated for all 10 sets of data such that each set functions once as the validation set and nine times as a part of the training set. In this case, the deviance statistic was used to describe model error. The tuning parameter minimizing model error was chosen.

To understand the elastic net, an equation for the LASSO penalty term needs to be introduced. The LASSO constrains the sum of absolute values of the regression coefficients, or in other words

$$(8) \qquad \sum_{j=1}^{p} |\beta_j| \leq t$$

Unlike the ridge penalty, the LASSO penalty can constrain regression coefficients so that they come to equal zero. It can thus perform variable selection. Its weakness, though, is that it tends to choose one variable at random among many that are correlated. The elastic net is a combination of the ridge and lasso penalties, (Hastie, Tibshirani, & Friedman, 2009, p.73):

$$(9) \qquad \lambda \sum_{j=1}^{p} (\alpha \beta_j^2 + (1-\alpha)|\beta_j|)$$

The elastic net penalty contains two parameters, λ and α. The former adjusts the amount of regularization, the latter the balance between LASSO and ridge penalties. The strength of the elastic net is that it performs variable selection similarly to the LASSO, but includes correlated variables into the model as a group similarly to the ridge penalty.

The optimal values for the parameters α and λ were chosen using 10-fold cross-validation in the training sample. A grid search on the optimal values of α and λ parameters was performed, and the combination of the parameters that minimized error was chosen.

Fitting the Poisson regression models was based on the central idea of statistical learning theory: instead of aiming for a model that would optimally fit the sample of data at hand, the objective was to maximize predictive power out of sample, i.e. in a data set that is independent from the data set used for building the model (Chapman, Weiss, & Duberstein, 2016).

3 RESULTS

3.1 Study I

The objective of Study I was to assess whether the same psychological properties could be measured across groups of Finnish drivers defined based on age and gender. Data analysis began by comparing the fit of a confirmatory factor model ("SEM" in Table 3) and an ESEM. The ESEM proved to fit the data considerably better (see the entries on "Whole sample" in Table 4). The models mainly differ in that items are only allowed to cross-load on several factors in the ESEM. For instance, in the CFA items 18 (*stay on a lane until the last minute, forcing your way onto the neighboring lane*) and 20 (*overtake a slower driver on the inside*) were specified as *rule violations*, while in the ESEM they had roughly equal loadings on *aggressive violations*, *rule violations* and *slips*. Such results can – among other things – be due to several latent variables influencing the observed variables or to different factor loading patterns being present in subgroups of respondents. When the latent variables are interpreted as psychological properties, the importance of the seemingly technical difference of specifying cross-loadings becomes apparent: navigating traffic situations is such a complex endeavor that traffic behaviors are rarely determined by a single psychological mechanism. In the previous example, this would mean that drivers partly overtake on the inside as an act of aggression, partly knowingly breaching a rule without an aggressive intention and partly due to not paying attention to the driving task.

However, the correlational structure of the observed variables naturally remains identical across the two kinds of models. In the confirmatory factor model, the only way of representing the intercorrelations that lead to cross-loadings in the ESEM is through inflating inter-factor correlations. This is shown in Table 3. On the right-hand side (SEM), slips and lapses are highly correlated ($r = 0.8$), while their correlation is lower in the ESEM ($r = 0.61$). Correlations of this magnitude would, as such, call the discriminant validity of the DBQ into question; in other words, when the factors correlate this strongly, it is unclear whether they should be interpreted as two separate latent variables at all. On the other hand, the alternative interpretation based on the results of the ESEM would be better compatible with the existence of two latent variables. Further, differences in inter-factor correlations are high when comparing the correlations of rule violations and aggressive violations ($r_{SEM} - r_{ESEM} = 0.35$) and rule violations and slips ($r_{SEM} - r_{ESEM} = 0.36$). And, in any case, the nonsatisfactory fit of the confirmatory factor model showed that the high intercorrelations between the factors did not succeed in capturing all of the relevant linear covariation in the data.

Table 3. *Factor loadings of the confirmatory factor model (SEM) and ESEM models (reproduced with permission from table 4, Study I)*

ESEM	Factor loadings				SEM	Factor loadings			
Items	Aggress.	Rule V.	Slips	Lapses		Aggress.	Rule V.	Slips	Lapses
Backing up, hit something (1)	0.26	-0.29	0.16	0.37		0.00	0.00	0.00	0.42
Intending to drive to A, find yourself on way to B (2)	0.14	-0.04	-0.16	0.62		0.00	0.00	0.00	0.47
Drink and drive (3)	0.26	0.07	0.12	0.24		0.00	0.50	0.00	0.00
Choose wrong lane before intersection (4)	-0.07	0.09	0.11	0.48		0.00	0.00	0.00	0.57
Queuing, nearly hit car in front (5)	-0.07	0.19	0.54	0.02		0.00	0.00	0.61	0.00
Turning to side-street, miss pedestrian (6)	-0.11	0.24	0.70	-0.07		0.00	0.00	0.68	0.00
Sound horn to indicate your annoyance (7)	0.61	0.17	-0.15	0.12		0.67	0.00	0.00	0.00
Fail to check your rear-view mirror (8)	-0.02	0.03	0.45	0.14		0.00	0.00	0.55	0.00
Brake too quickly on a slippery road (9)	0.03	-0.08	0.38	0.28		0.00	0.00	0.56	0.00
Pass "Give Way" sign, push onto main road (10)	0.15	0.02	0.49	0.06		0.00	0.55	0.00	0.00
Disregard speed limit on a residential road (11)	0.00	0.81	-0.06	0.05		0.00	0.69	0.00	0.00
Mistakenly turn on wrong control (12)	-0.09	-0.01	0.18	0.48		0.00	0.00	0.00	0.55
Turning right nearly hit cyclist (13)	-0.05	0.04	0.67	0.02		0.00	0.00	0.64	0.00
Miss "Give Way" sign, almost collide (14)	0.06	0.04	0.70	0.01		0.00	0.00	0.72	0.00
Attempt to take off on too big gear (15)	-0.03	-0.01	0.24	0.41		0.00	0.00	0.00	0.58
Attempt to overtake someone turning left (16)	0.16	-0.15	0.56	0.20		0.00	0.00	0.68	0.00
Get angry, give chase (17)	0.60	0.31	0.10	-0.15		0.81	0.00	0.00	0.00
Push in at last minute (18)	0.35	0.29	0.22	0.05		0.00	0.69	0.00	0.00
Forgot where parked car (19)	-0.11	0.16	-0.20	0.64		0.00	0.00	0.00	0.46
Overtake a slow driver on the inside (20)	0.28	0.31	0.20	0.00		0.00	0.60	0.00	0.00
Race from lights (21)	0.28	0.53	0.06	-0.06		0.00	0.64	0.00	0.00
Read traffic sign wrong, turn on wrong road (22)	-0.03	0.00	0.21	0.47		0.00	0.00	0.00	0.61
Close following (23)	-0.05	0.59	0.23	0.08		0.00	0.68	0.00	0.00
Shooting lights (24)	0.11	0.43	0.12	0.20		0.00	0.67	0.00	0.00
Lose temper at other driver (25)	0.68	0.24	-0.05	0.04		0.82	0.00	0.00	0.00
No memory of road (26)	-0.14	0.37	-0.01	0.41		0.00	0.00	0.00	0.55
Underestimate the speed of a vehicle (27)	-0.05	0.11	0.42	0.25		0.00	0.00	0.65	0.00
Disregard the speed limit on motorway (28)	0.11	0.72	-0.11	0.05		0.00	0.65	0.00	0.00

	Factor correlations					Factor correlations			
	Aggress.	Rule V.	Slips	Lapses		Aggress.	Rule V.	Slips	Lapses
Aggress.	1.00				Aggress.	1.00			
Rule V.	0.40	1.00			Rule V.	0.75	1.00		
Slips	0.32	0.26	1.00		Slips	0.37	0.62	1.00	
Lapses	0.24	0.30	0.61	1.00	Lapses	0.32	0.58	0.80	1.00

Table cells have been color coded to aid interpretation. White background translates to a loading of ≤ .15, light gray to a loading of > .15 and < .3 and dark gray to a loading of ≥ .3. Aggress. = Aggressive violations; Rule V. = Rule violations.

Continuing the ESEM analysis showed that the configural model had an adequate fit both across genders and age groups (for definitions, see Section 2.3.2). In an ESEM analysis, this result is in itself not yet all that strong, as the factor loadings were allowed to differ across groups; in other words, the configural equivalence models are actually nothing but EFAs run separately in the subgroups. Interpreted through the example given in Figure 1, this would amount to all the variables specified as loading on violations as also loading on errors and vice versa, leaving ample room for intergroup differences. Still, target rotation was used as a rotation method, which increases the similarity of patterns of factor loadings across groups, other things being equal.

The analysis proceeded by examining the combined weak and strong equivalence across genders and age groups, and showed that these constrained models fit worse across genders and age groups (Table 4) than the unconstrained models (see especially ΔCFI and the other fit indices that indicate worse model fit for the constrained models across the board).

Table 4. *Model fit for all models in Study I (reproduced from table 3, Study I)*

	$p(\chi^2)$	RMSEA	CFI	TLI	WRMR	ΔCFI
Whole sample, original scale						
SEM	< 0.001	0.069	0.861	0.929	1.670	--
ESEM	< 0.001	0.036	0.960	0.981	0.805	--
Whole sample, recoded scale[b]						
SEM	< 0.001	0.054	0.886	0.943	1.506	--
ESEM	< 0.001	0.029	0.968	0.984	0.797	--
Men/women, recoded scale[b], measurement non-invariance						
ESEM	< 0.001	0.031	0.969	0.983	0.999	--
Men/women, recoded scale[b], measurement invariance (equal factor loadings & thresholds)						
ESEM	< 0.001[c]	0.038	0.950	0.973	1.330	0.019
Age groups, recoded scale[b], measurement non-invariance						
ESEM	< 0.001	0.034	0.968	0.979	1.327	--
Age groups, recoded scale[b], measurement invariance (equal factor loadings & thresholds)						
ESEM	< 0.001[c]	0.042	0.948	0.967	1.879	0.020

[a]cut-off limits. RMSEA: 0.05; CFI: 0.95; TLI: 0.93; WRMR: 1.1

[b]recoded scale: DBQ variables recoded according to (0 --> 0, 1 --> 1, 2-5 --> 2)

[c]the p-values refer to DIFFTEST-results

After concluding this, Study I proceeded with examining differences in factor loadings between the genders and age groups in a painstaking manner. The analysis concluded that the profiles of factor loadings and cross-loadings were different enough to warrant naming the latent variables differently across the groups. For instance, in the youngest age group, three out of nine items that were expected to measure *rule violations* had an unexpectedly low (*push in at last minute*) or non-existing loading on it (*drink and drive*, *push on to a main road irrespective of a "give way" sign*). The expected pattern is shown in Table 3 above (see the right-hand panel, "SEM model"). On the other hand, several items that were intended to measure other latent variables had an unexpectedly high loading on the *rule violations* factor (*having no memory of the road*; *sound horn to indicate annoyance*; *turning to side street, barely missing pedestrians*). The resulting factor thus differed from the one that was intended to be measured to such an extent that it was named *violations of social norms*. Similarly, as items potentially related to inexperience loaded on one factor in the youngest age group, it was suggested that it might be appropriate to call the factor *inexperience* rather than *lapses*.

Similarly, in the oldest age group several items that were expected to measure *rule violations* loaded most strongly on other factors (*drunk driving* on lapses; *pushing on to a main road* and *pushing in at last minute* on slips) or had a high cross-loading on another factor (*close following* on slips and *driving through a traffic light on red* on lapses). Because of observations such as these, Study I suggested that perhaps drivers belonging to the oldest age group unwittingly committed certain driving behaviors that were originally considered as *rule violations*.

The results of Study I can further be illustrated by an example of comparing two age groups. Table 5 shows ESEMs separately for two age groups (36–50-year olds and over 50-year-olds). Let us examine the items that were expected to measure *rule violations*. Item 3 (*drunk driving*) loaded on *rule violations* in age group 36-50, but not in age group 51-. Item 10 (*pass "give way" sign, push onto main road*) did not have its main loading on the intended factor (*rule violations*) in either group, even though it cross-loaded on it in the younger age group. Item 18 (*pushing in at the last minute*) loaded mostly on *rule violations* in age group 36-50 and on *slips* in age group 51-.

Table 5. *ESEMs with no equivalence constraints, 36-50 year-olds and over 50-year-olds (reproduced with permission from table 7, Study I)*

Age group 36-50	Factor loadings				Age group 51-	Factor loadings		
Items	Aggress.	Rule V.	Slips	Lapses	Aggress.	Rule V.	Slips	Lapses
Backing up, hit something (1)	0.33	-0.12	0.42	0.06	0.28	-0.22	0.26	0.25
Intending to drive to A, find yourself on way to I	0.14	0.07	0.08	0.46	0.20	-0.01	-0.06	0.64
Drink and drive (3)	0.15	0.42	0.04	0.09	0.52	-0.07	0.21	0.12
Choose wrong lane before intersection (4)	-0.08	0.27	0.23	0.31	0.05	0.19	0.30	0.29
Queuing, nearly hit car in front (5)	0.12	0.19	0.39	0.14	0.09	0.16	0.45	0.11
Turning to side-street, miss pedestrian (6)	0.17	-0.11	0.54	0.26	-0.16	0.37	0.77	-0.11
Sound horn to indicate your annoyance (7)	0.74	-0.03	-0.13	0.20	0.38	0.23	-0.06	0.12
Fail to check your rear-view mirror (8)	0.02	0.02	0.57	0.14	-0.07	-0.03	0.35	0.37
Brake too quickly on a slippery road (9)	0.11	0.17	0.42	0.01	0.30	-0.34	0.43	0.41
Pass "Give Way" sign, push onto main road (10)	0.06	0.29	0.60	-0.22	0.22	-0.02	0.54	0.15
Disregard speed limit on a residential road (11)	-0.15	0.79	-0.06	0.13	0.18	0.60	0.06	0.05
Mistakenly turn on wrong control (12)	0.10	0.02	0.27	0.38	-0.10	0.11	0.23	0.56
Turning right nearly hit cyclist (13)	0.06	-0.05	0.61	0.24	0.04	0.18	0.45	0.15
Miss "Give Way" sign, almost collide (14)	-0.20	0.37	0.66	-0.02	0.26	0.05	0.72	0.00
Attempt to take off on too big gear (15)	0.05	0.17	0.32	0.23	-0.08	-0.12	0.55	0.29
Attempt to overtake someone turning left (16)	0.09	0.01	0.68	0.17	-0.08	-0.03	0.85	0.06
Get angry, give chase (17)	0.78	0.12	0.11	-0.08	0.61	0.21	0.10	-0.13
Push in at last minute (18)	0.16	0.52	0.23	-0.10	0.14	0.36	0.44	0.00
Forgot where parked car (19)	-0.04	-0.04	-0.08	0.72	-0.10	-0.02	-0.24	0.81
Overtake a slow driver on the inside (20)	0.06	0.33	0.24	0.06	0.10	0.29	0.26	0.30
Race from lights (21)	0.21	0.70	-0.13	-0.11	0.28	0.64	-0.12	-0.03
Read traffic sign wrong, turn on wrong road (22	0.01	0.12	0.43	0.11	0.05	0.37	0.06	0.40
Close following (23)	0.05	0.43	-0.03	0.39	-0.04	0.50	0.37	0.01
Shooting lights (24)	0.10	0.43	0.18	0.24	-0.02	0.51	0.05	0.27
Lose temper at other driver (25)	0.65	0.31	-0.19	-0.08	0.56	0.31	-0.12	0.04
No memory of road (26)	-0.01	0.14	0.07	0.60	-0.20	0.22	-0.05	0.68
Underestimate the speed of a vehicle (27)	-0.08	0.29	0.40	0.18	-0.06	0.23	0.50	0.27
Disregard the speed limit on motorway (28)	0.08	0.82	-0.21	-0.02	0.14	0.63	-0.16	0.19

	Factor correlations					Factor correlations			
	Aggress.	Rule V.	Slips	Lapses		Aggress.	Rule V.	Slips	Lapses
Aggress.	1.00				Aggress.	1.00			
Rule V.	0.41	1.00			Rule V.	0.20	1.00		
Slips	0.31	0.43	1.00		Slips	0.30	0.27	1.00	
Lapses	0.24	0.39	0.39	1.00	Lapses	0.27	0.32	0.50	1.00

Table cells have been color coded to aid interpretation. White background translates to a loading of ≤ .15, light gray to a loading of > .15 and < .3 and dark gray to a loading of ≥ .3. Aggress. = Aggressive violations; Rule V. = Rule violations

Based on examining the results in this manner, Study I suggests that the items were interpreted differently enough by the different subgroups of respondents that the latent variables should be named differently. Further, the strong cross-loadings among the items attest that several latent variables affected the variation in the individual observed variables. All in all, the results of Study I casted doubt on the ability of the DBQ to measure the same psychological properties across subgroups of Finnish drivers.

3.2 Study II

Study II assessed the measurement equivalence of the 27-item DBQ across groups of young drivers from Finland and Ireland. In other words, the study aimed at answering the question of whether the instrument measures the same latent variables in the same way across countries. It showed that the socially oriented latent variables (*aggressive violations* and *traffic rule violations*) had clearly different structures across the countries, while the cognitively oriented latent variables (*slips* and *lapses*) were more similar.

3.2.1 Dimensionality of the DBQ

Study II took part in the discussion concerning the dimensionality of the DBQ when modelled using latent variables. In other words, the study aimed at answering the question of which number of latent variables to use for representing the correlational structure of the 27-item version of the DBQ. This discussion started with the seminal DBQ publication (Reason et al., 1990) and a definitive answer to the question is still lacking. Study II showed that out of the three main competing factor models used in the DBQ literature, the four-factor model fit the data best in both samples (Table 6).

Table 6. *Fit indices for testing the dimensionality of the DBQ (reproduced with permission from table 2, Study II)*

Model	χ^2	df	p	RMSEA	90% CI	CFI	AIC	SRMR
Finnish data								
two factors (2F)	1572.97	323	<.001	.061	.058-.063	0.766	60412	0.066
three factors (3F)	1492.78	321	<.001	.059	.056-.062	0.781	60307	0.065
four factors (4F)	1346.65	318	<.001	.055	.053-.058	0.808	60094	0.062
$\Delta\chi^2$(3F - 2F)	58.66	2	<.001			0.015*		
$\Delta\chi^2$(4F - 3F)	59.04	3	<.001			0.027*		
Irish data								
two factors	1262.21	323	<.001	.060	.057-.063	0.788	50268	0.071
three factors	1179.89	321	<.001	.057	.054-.060	0.806	50166	0.070
four factors	1003.20	318	<.001	.051	.048-.055	0.845	49945	0.068
$\Delta\chi^2$(3F - 2F)	77.88	2	<.001			0.018*		
$\Delta\chi^2$(4F - 3F)	168.29	3	<.001			0.039*		

The $\Delta\chi^2$ -values are Satorra-Bentler scaled chi square values.

*ΔCFI

Even though the four-factor model had the best fit out of the models that were compared, the approximate fit indices (RMSEA and, especially, CFI) indicated that it failed to reach a desirable fit to data. Sources for the lack of fit were examined by calculating modification indices and by examining residual correlation graphs (Figure 7). The modifications that would have improved model fit were partly different across samples (cf. the patterns of residuals in Figure 7), but certain commonalities could be found: specifying item 9 to load on slips and allowing the residual correlations of the speeding-related items 11 and 27 to correlate improved model fit in both samples.

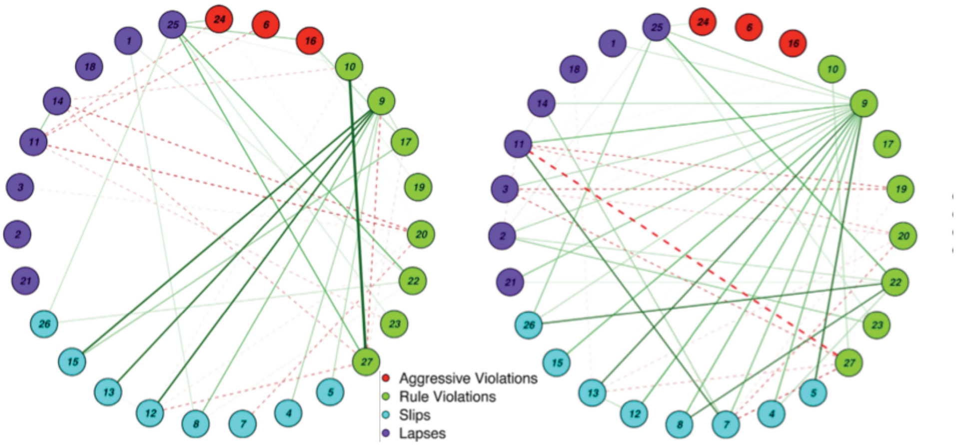


Figure 7 Residual correlations (|r| > 0.10) among the DBQ items after fitting the four-factor model to the Finnish (left) and the Irish (right) sample. The color and type of the lines indicates whether the correlation is positive (solid green) or negative (dashed red), while the width and the level of transparency of the line indicate the strength of the correlation. Reproduced with permission from Figures 3 and 4, Study II.

After modifying the model and still not reaching a satisfactory fit to data, the exploratory mode of analysis was adopted as suggested by Browne (2001):

> *"Confirmatory factor analysis procedures are often used for exploratory purposes. Frequently a confirmatory factor analysis, with prespecified loadings, is rejected and a sequence of modifications of the model is carried out in an attempt to improve fit. The procedure then becomes exploratory rather than confirmatory . . . In this situation the use of exploratory factor analysis, with rotation of the factor matrix, appears preferable . . . The discovery of misspecified loadings, however, is more direct through rotation of the factor matrix than through the examination of model modification indices."*
>
> *(Browne, 2001, p.113)*

EFAs (Figure 8) indicated why the structures differed across samples: the socially-oriented latent variables (the violations) comprised different item sets across samples. However, the cognitively-oriented latent variables (*slips* and *lapses*) involved the same items. Because of this, analyses of (partial) measurement equivalence were carried out across the samples.

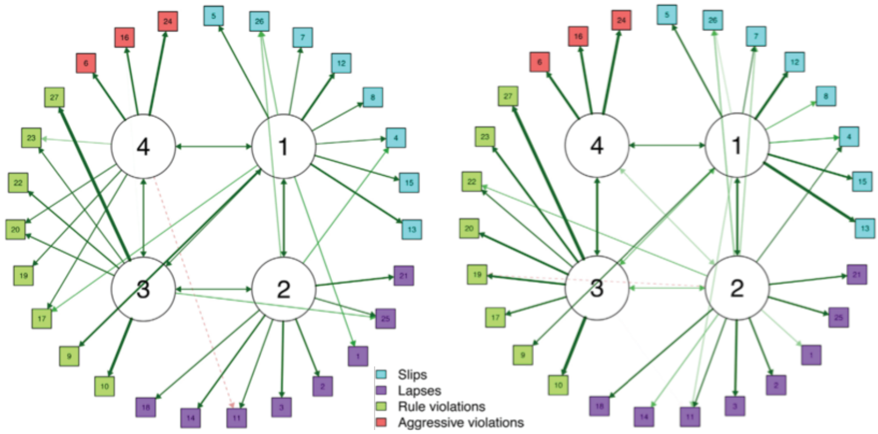


Figure 8 Results of the exploratory factor analysis in the Finnish (left) and the Irish (right) sample. Factor loadings (|>0.2|) implied by the original four-factor model are shown in the legend. Reproduced with permission from Figures 6 and 7 (Study II).

3.2.2 Measurement equivalence of the DBQ across countries

The cognitively-oriented latent variables, *slips* and *lapses*, were deemed sufficiently similar across samples so that analyses of measurement equivalence could be performed. The analyses were carried out one factor at a time, and they proceeded as described in section 2.3.2, i.e. testing whether equal subsets of items load on the same factors (configural equivalence), followed by tests of equality of factor loadings (weak equivalence) and of item intercepts (strong equivalence). Unlike Study I, Study II was able to tease apart analyses of weak and strong measurement equivalence. In addition, differential item functioning was assessed using tests of partial equivalence, i.e., the equivalence constraints were relaxed for individual items.

When examining the latent variable *slips*, configural and weak equivalence were established, but the strong equivalence model failed to adequately fit the data. Tests of partial strong equivalence showed that the intercepts of items 7 (*fail to check the rear-view mirror*), 8 (*brake too quickly*) and 15 (*attempt to overtake someone turning left*) could be constrained to equality, while those of the remaining five indicators of *slips* needed to be freely estimated. Looking at *lapses*, the configural model fit the data adequately, whereas the weak equivalence model fit significantly worse. Partial weak equivalence was established with loadings of items 2 (*intending to drive to A, find yourself on your way to B*) and 18 (*forget where you left your car*) estimated freely and the loadings of the other five indicators of lapses constrained to equality.

In summary, Study II showed that the socially-oriented latent variables (different types of violations) differed in nature across young drivers from Finland and Ireland, while the cognitively-oriented latent variables were more similar in nature. Study II recommended against comparing sum scores calculated based on any of the latent variables across the countries (or traffic cultures), even though latent means of *slips* could be compared after fitting the appropriate partial equivalence models to data.

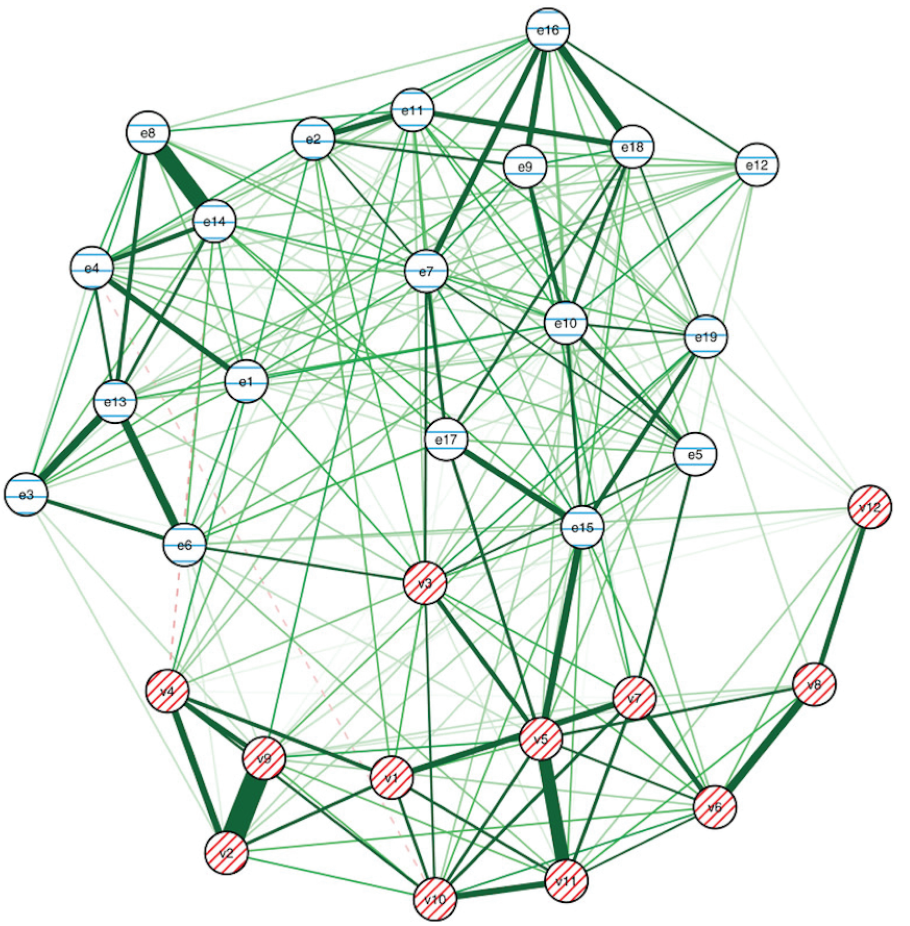
3.3 Study III

Study III viewed the traffic behaviors encoded in the DBQ as psychological networks (Sections 2.3.3. and 2.3.6.1.). The statistical model and the causal assumptions were thus essentially different from those of Studies I and II. In brief, the study focused on direct interactions between the nodes of the network (either DBQ variables or DBQ variables together with background variables), whereas in Studies I and II, the traffic behaviors were seen as causally inefficacious reflections of the latent variables (*rule violations*, *aggressive violations*, *slips* and *lapses*). Study III, then, examined the structure of psychological constructs instead of attempting to measure them.

3.3.1 Network analyses

The results of Study III were based on a longitudinal data comprising four time points (the first after 6 months and the rest at 12, 24, and 36 months of obtaining a driver's licence). The between-person model (Figure 9) represents relationships among the respondents' average scores across the four time points. The benefit of forming a network model based on average scores is that doing so reduces the effects of reporting biases and spurious effects such as mood-congruent recall (Shiffman, Stone, & Hufford, 2008).

The psychological network shown in Figure 9 comprises traffic behaviors labeled as violations (v) and errors (e) in a LASSO-estimated partial correlation network. Weak edges are constrained to zero as per the logic of LASSO-estimation (Section 2.3.6.1). The network has certain conspicuous properties: 1) The errors are most clearly connected with other errors and violations with other violations, but there exist several driving behaviors that connect the two, functioning as bridge behaviors; for instance, v5 (*speeding*), v3 (*driving through a red light*), e7 (*failing to notice people*). 2) Many of the edges are quite weak (thin and transparent), while certain nodes are connected by clearly stronger edges (thick and opaque); 3) Within violations and errors, thematically related nodes are connected by strong edges, the aggression-related nodes (v2, v4 and v9) and speeding-related nodes (v5 and v11) being a case in point, but 4) there are other thematically related nodes that share only weak edges, such as the two nodes related to not perceiving traffic signs (e8 and e16), 5) at least certain edges can be readily interpreted as causal hypothses, e.g. drivers who speed more than average (v5) tailgating more than average (e15).



- e1: drive away from traffic lights at too high a gear
- e2: attempt to overtake and hadn't noticed signalling right
- e3: forget where left car in carpark
- e4: switch on one thing when meant to switch on other
- e5: pull out of junction so far that driver has to let you out
- e6: realised have no recollection of road been travelling
- e7: failed to notice people crossing when turning
- e8: misread signs and taken wrong turning off roundabout
- e9: turning left nearly hit cyclist on inside
- e10: when queuing to turn left nearly hit car in front
- e11: misjudged speed of oncoming vehicle when overtaking
- e12: hit something when reversing that hadn't seen
- e13: noticed ending up on a different road than intended
- e14: get into wrong lane approaching roundabout/junction
- e15: drive so close to car that would not be able to stop*
- e16: missed giveway signs and avoided colliding with traffic
- e17: failed to check rear-view mirror before manoeuvring
- e18: brake too quickly on slippery road / steer wrong in skid
- e19: had to brake or swerve to avoid accident

- v1: overtake a slow driver on inside
- v2: sound horn to indicate annoyance
- v3: crossed junction knowing lights have turned against you
- v4: become angered by driver and given chase
- v5: disregarded speed limit on residential road
- v6: used mobile phone without hands free kit
- v7: stay in motorway lane you know will be closed
- v8: drive when suspect over legal alcohol limit
- v9: become angered by driver and indicated hostility
- v10: raced away from traffic lights to beat other driver
- v11: disregarded speed limit on motorway
- v12: drove after taking drugs which affected you

*NB: Item e15 was erroneously classified as an *error*. It tends to load on *violations* in factor analyses of the DBQ, so it should have been classified as a violation in Figure 9.

Figure 9 The between-person network model. The colors of the nodes correspond with the errors / violations dichotomy. Green edges signify positive associations, dashed red edges negative ones. The wider and the more opaque the edge, the stronger the association. Reproduced under the CC-BY-4.0 licence from Figure 3, Study III.

The accuracy of the edge weight estimates was assessed using the bootstrap procedure described in Section 2.3.6.1. The edge weights had for the most part quite narrow confidence intervals, which indicates that their values were accurately estimated. This is understandable as Study III was based on a large sample of new drivers.

The importance of the individual nodes as well as their redundancy can be better understood by examining indices of centrality and clustering (Figure 10). Strength centrality reflects the connection strengths that a given node has with its immediate neighbors. Panel B in Figure 11 shows a smoothly declining curve with most of the nodes near to the center of the curve. Six nodes, from e10 (*nearly hit car in front*) and v5 (*speeding on residential roads*) to v3 (*crossed junction knowing lights had turned against you*) were clearly more strength central than the rest of the nodes. Some of the connections between the most strength central nodes can be given a causal interpretation (for instance v5 *speeding* – e15 *tailgating*), rendering these nodes potential candidates for traffic safety measures. Closeness centrality is less sensitive to single strong edges than strength centrality, so it is perhaps a useful index in this context. As stated in Study III: "*In general, the nodes along the path connecting speeding with various errors (v11-v5-v3-e7-e17-e15-e10 or v11-v5-e15, etc.) were central.*" Connections between some of these nodes can be interpreted as causal hypotheses, as well. On the other hand, nodes v12 (*driving under the influence of drugs*) and e12 (*hitting something while reversing*) were clearly least central among the nodes. They share only weak connections with the other nodes in Figure 9.

The stability of the centrality index estimates was assessed using the bootstrap procedure described in Section 2.3.6.1. The bootstrap analyses showed that strength centrality and closeness centrality were largely unaffected by the composition of the sample. For instance, after dropping 70 % of the observations in the sample, the correlation between the original value of the closeness centrality index and the average of the bootstrapped values was still roughly 0.7; for strength centrality, the corresponding correlation was still higher.

Nodes with high values of the clustering coefficient can be interpreted as redundant in that they add little unique information to the network. Node v4 (*become angered, give chase*) received a high value for Zhang's clustering coefficient largely because its neighbors v2 and v9 share an extremely strong connection. Similar logic applies to the redundancy of node e3 (*forget where left car in carpark*) with respect to its neighbors e6 and e13.

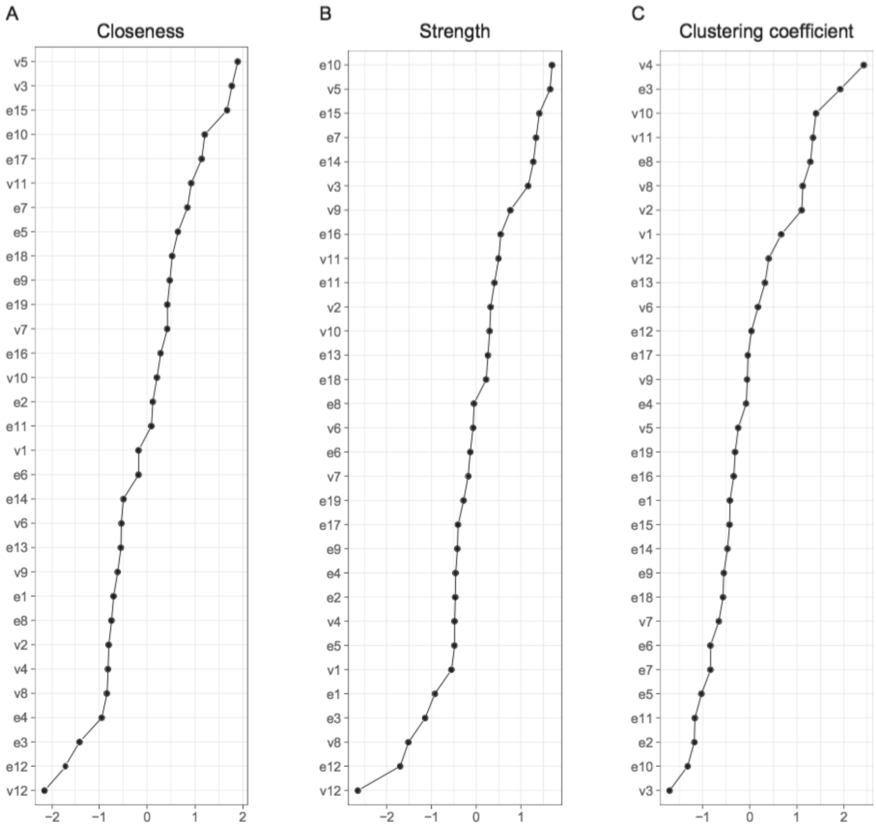


Figure 10 Indices of centrality and clustering for the between-person network. All indices are shown in a standardized metric. (A) Closeness centrality. (B) Strength centrality. (C) Zhang's local clustering coefficient. Reproduced under the CC-BY-4.0 licence from Figure 4, Study III.

The network model with background variables included as nodes of their own is only described here, and shown as Figure 5 in Study III. In short, the nodes related to *perceiving certain improvement needs in one's own driving* shared negative edges with various violations. For instance, *perceiving a need to improve lane changing skills* shared a strong negative edge with nodes related to *staying in a lane that will be closed* and *forcing one's way on a free lane*. Further, *positive attitude towards decreasing speed limits* shared a strong negative edge with the node related to *speeding on a motorway*. Similarly, a *negative attitude toward overtaking on the inside* shared a strong negative edge with reporting actually doing so. While these associations may seem trivial, it is noteworthy that questionnaires assessing attitudes and self-reported behavior were filled in on different occasions, thus perhaps reducing common method variance and providing validity evidence of sorts. Also, as in all network models discussed herein, the results remained after controlling for the effects of all other variables in the network.

As stated above, such results can be interpreted as hypothetical causal relations between the variables connected by an edge. Furthermore, it is of interest that if the relations indeed prove to be causal in nature, causation may flow either from attitudes to behavior or vice versa – a possibility that is seldom considered in SEMs investigating the relationships of attitudes and traffic behavior.

3.3.2 Regression analyses

Separate regression analyses were performed for predicting the number of self-reported crashes based on the individual driving behaviors. The DBQ variables were collected at 6 months post-licensure, while the number of crashes was aggregated from data collected at 12 months, 24 months and 36 months post-licensure, making the models truly predictive. Three kinds of regression models were fit as described in Section 2.3.6.2. The models were first fit in the training sample and then tested in an independent test sample.

Model fit in both samples for all three models is shown below in Table 7. The naive Poisson model (i.e. the "normal way" of fitting linear models) provided the best fit to the training data (as it should), but a remarkably poor fit to test data, actually fitting worse than the null model with no predictors as evidenced by the negative value for McFadden's pseudo R^2 statistic. The elastic net and the ridge regression models fit both the training and test data roughly equally well, which favors the elastic net model because of its greater parsimony. The regression weights are shown below in Table 8.

The regression weights in table 8 are low partly because they have been adjusted downward according to the principle of regularized regression (Section 2.3.6.2). Variables with the highest regression weights were those related to *failing to notice pedestrians*, *using gears* and different forms of *aggressive behavior*. *Hitting something while reversing*, which in itself is a minor crash, also predicted future crashes.

Table 7. *Regression model fit in the training and hold-out samples. Reproduced under the CC-BY-4.0 licence from Table 1, Study III*

Variable	Elastic net	Ridge	Naive Poisson
Training sample			
Pearson r	0.246	0.260	0.317
Mean square error	0.625	0.622	0.590
Min-max index	0.121	0.122	0.128
McFadden pseudo R2	0.030	0.033	0.057
Deviance	968.4	964.8	924.8
Null model deviance	1018.1		
Hold-out sample			
Pearson r	0.155	0.160	0.086
Mean square error	0.627	0.626	0.745
Min-max index	0.158	0.157	0.156
McFadden pseudo R2	0.010	0.011	-0.044
Deviance	317.9	317.2	348.4
Null model deviance	323.6		

Table 8. *Regression weights for predicting accidents (elastic net regression analysis). Reproduced under the CC-BY-4.0 licence from Table 2, Study III.*

Variable	Regression weight
Mileage	0.04
Gender	
Age	
Drive away from traffic lights at too high a gear	0.07
Overtake a slow driver on inside	0.04
Have to confirm youre in right gear	
Attempt to overtake and hadnt noticed signalling right	
Forget where left car in carpark	
Sound horn to indicate annoyance	0.07
Switch on one thing when meant to switch on other	
Change into wrong gear when driving along	
Pull out of junction so far that driver has to let you out	0.04
Have used plates to warn drivers you are a new driver	
Realised have no recollection of road been travelling	
Crossed junction knowing lights have turned against you	0.01
Failed to notice people crossing when turned into sidestreet	0.08
Become angered by driver and given chase	
Misread signs and taken wrong turning off roundabout	
Drive in either too low or high gear for conditions	
Disregard speed limit on residential road	0.04
When turning left have nearly hit cyclist on inside	
Used mobile phone without hands free kit	
Stay in motorway lane know will be closed	
When queueing to turn left nearly hit car in front	
Drive when suspect over legal alcohol limit	
Forget to take handbrake off before moving off	0.03
Become angered by driver and indicate hostility	0.06
Misjudge speed of oncoming vehicle when overtaking	
Hit something when reversing that hadn't seen	0.06
Raced away from traffic lights to beat other driver	0.01
Used a hands free kit	
Selected wrong gear when wanting to go in reverse	
Noticed on different road to destination want to go	-0.01
Get into wrong lane when approaching roundabout/junction	
Drive so close to car that wouldn't be able to stop	
Forget headlights were on full beam	
Missed give-way signs and barely avoided collision	0.02
Disregarded speed limit on motorway	0.01
Failed to check rear-view mirror before manouevring	
Brake too quickly on slippery road or steer wrong in skid	
Drove after taking drugs which think affected you	
Brake/swerve to avoid accident	0.01

3.3.3 Summary of the findings of Study III

Study III had two aims: firstly, to construct and analyze network models of driver behavior and secondly, to report the results of cross-validated, regularized regression analyses of predicting crashes from individual driver behaviors. The data used in the study was longitudinal in nature, consisting of four different time points.

Regarding the first aim, two network models were constructed: a between-subjects model that was based on the averages of the responses across the four time points and a cross-sectional model with background variables that was based on data from the first time point. The between-subjects model showed that even though the errors had strongest connections with other errors and violations with other violations, no sharp boundary existed between errors and violations. Centrality analyses showed that nodes along a path connecting speeding-related violations to errors related to car control *(v11-v5-v3-e7-e17-e15-e10 or v11-v5-e15)* were central in the network. Study III suggested that certain edges along this path could be interpreted as causal hypotheses. The network models with background variables showed that driving-related attitudes and the self-image of the drivers were related to the driving behaviors in an understandable manner (for instance, a positive attitude toward reducing speed limits shared a negative edge with self-reported speeding). Insofar as the relationships are interpreted as causal hypotheses, it is of interest that the network models allow interpreting causation to flow in either direction – or both directions – between attitudes and behavior.

The regression models offered several novel methodological contributions to the field. First, the models were built and tested in separate subsets of data in order to minimize the risk of over-fitting the models to data. Second, regularized regression methods were used for performing variable selection and model estimation at the same time. Traditionally, methods that rely on calculating p-values that require correcting for multiple comparisons have been used in similar contexts, even though the problem of multiple comparisons has simply been ignord (see, e.g., Wallén Warner et al., 2011). Third, the optimal models were searched for using ten-fold cross-validation, which further reduced the risk of over-fitting; cross-validation was also used for choosing the values of the so-called hyperparameters of the regularized regression models. Variables with the highest regression weights were those related to *failing to notice pedestrians*, *using gears* and different forms of *aggressive behavior*.

4 DISCUSSION

Self-report-based traffic research is a lively field that has the potential to produce results that contribute to the well-being of people and to increasing traffic safety. One particular tradition of carrying out research in the field emerged with the introduction of the DBQ in the early 1990s: self-report data is collected, factor analysis performed, sum scores – that serve as operationalizations of the assumed latent variables – calculated and accidents predicted. The tradition has its theoretical background in the Generic Error Modeling System (GEMS; Reason, 1990), which describes different types of human errors in safety-critical situations, and the DBQ latent variables are often identified with the psychological processes described in the GEMS (e.g. Mesken et al., 2002; Reason et al., 1990; Stephens & Fitzharris, 2016). Even though the description is a simplification and does not do justice to more nuanced ideas within the field, it still describes the current state of *normal science* (Kuhn, 1996) within self-report-based traffic psychology.

According to Kuhn, during a period of normal science, experiments and observations serve several purposes, such as articulating the paradigm theory, revealing the nature of things the theory refers to and solving problems the theory has drawn attention to (Kuhn, 1996). When applied to self-report based traffic research, the idea translates to, among other things, searching for the optimal factor solution for the DBQ, examining the correlates of the factor scores and using them to predict crashes.

The "normal science" approach has served the field well for 30 years, producing important results regarding accident risk (de Winter & Dodou, 2010), the social and and emotional determinants of traffic behavior (Lawton et al., 1997) and the traffic behavior of different special groups (Biederman et al., 2012; Parker et al., 2000; Sakashita et al., 2014). The approach to self-report studies in traffic psychology was named the *latent variable view of violations and errors* in the present thesis. In short, it is the view that violation and error proneness are stable psychological traits that are similar across age groups, genders, traffic cultures etc., need to be targeted by specific interventions targeted at just them, have different relationships with the drivers' accident risk and can be measured using the DBQ. The view naturally encompasses also other factor structures of the DBQ, such as the four-factor structures examined in Studies I and II.

According to Kuhn, not questioning the underlying assumptions of the current theory or methodology pertains to the nature of normal science. On the other hand, in order for the field to make methodological and theoretical progress, it is timely to critically assess the underlying – often tacit – causal assumptions that are made when adopting the latent variable view of violations and errors. That is something the current thesis attempts to do in order to provide the field with new methods and conceptual tools to be applied in the future.

Studies I and II took the latent variable view as a starting point and assessed the assumption that the DBQ measures the same latent variables in the same way across different groups of drivers. Study III, on the other hand, departed from the latent variable view and made two general suggestions: 1) traffic behaviors influence each other instead of, or in addition to, being influenced by latent psychological properties and 2) individual traffic behaviors can be fruitfully used as predictors of crashes without positing latent variables.

Study I concentrated on different subgroups of Finnish drivers that were defined based on age and gender. It tested whether the latent variable model comprising *aggressive violations*, *rule violations, slips* and *lapses* applies to all these groups in the same way. The conclusion of Study I was that the factor structures were different enough to warrant being named differently. Previous studies have examined similar questions within the DBQ tradition. Rimmö (2002) concluded that the 32-item Swedish version of the DBQ functions roughly equally well within subgroups of men and women. On the other hand, the study was about a different instrument with different item content, measurement model and latent variables than the 27- / 28-item version of the DBQ examined in the present thesis. An additional difference was that the study was based on building separate SEMs in the different subgroups rather than performing rigorous tests of measurement equivalence across them. Further, after Study I was published, Stephens & Fitzharris (2016) replicated the study using CFA in a sample of Australian drivers and found that the model had a tolerable fit to data after correlating the error variances of the speeding-related items. In addition, Stephens & Fitzharris (2016) obtained partial strong equivalence when comparing two groups of middle-aged drivers after freely estimating the intercepts of two items. The four-factor model fit only after dropping several items and correlating certain error variances in the youngest and oldest age groups (drivers of ages 17–25 and 65–75 years, respectively). The results were, then, similar to those of Study I in that the model had the best fit in the middle-aged age groups. Further, the study elegantly built on Study I in separating analyses of weak and strong equivalence and in performing analyses of partial equivalence. Still, the essential conclusions were quite similar: adequate model fit was obtained only after heavily modifying the original model – particularly in the youngest and oldest age groups – calling it into question whether the same psychological properties were being measured across all groups. Similarly, Martinussen et al. (2013) constructed yet another version of the DBQ and compared model fit using CFA in different subgroups of Danish drivers. Their four-factor model had the best fit to data, but as the latent variables and items were largely different than those used in the present thesis, the results are of little direct relevance for the present concerns.

Study II had a similar research question as Study I: it compared the fit of the same measurement model in groups of young Finnish and Irish drivers based on confirmatory factor analyses. Its general finding was that especially the socially-oriented latent variables (*rule violations* and *aggressive violations*) were different in nature across the two traffic cultures, whereas the cognitively-oriented latent variables (*slips* and *lapses*) were more similar – even though not quite enough so to warrant, for instance, calculating sum scores and comparing them across the two countries (or traffic cultures). Previous studies have examined similar questions, but mainly based on more descriptive and therefore more ambiguous methods. Lajunen et al. (2004) found that EFA solutions obtained for the 28-item DBQ were quite similar (had high correlations) across three countries: Great Britain, Finland, and the Netherlands. Still, even high correlations are compatible with significant differences in the configural models, and Lajunen et al. (2004) indeed found the loading patterns to differ across the countries in important respects: the loading patterns of *aggressive violations* and *errors* (*slips* in the present thesis) were most similar across countries, while those of *violations* and *lapses* were more varied. Özkan et al. (2006), for their part, examined a similar research question in data obtained from Finland, Great Britain, Greece, Iran, the Netherlands and Turkey using CFA and a 19-item version of the DBQ without items related to *lapses*. Their results were in a sense opposite to those of Lajunen et al. (2004) in that they found *aggressive violations* and *errors* to be quite dissimilar across countries and *violations* to be more similar. Further, a quite different three-factor structure was obtained in Chinese data for the 27 / 28-item DBQ (Chu et al., 2019), indicating that the four-factor model does not offer a universal solution across different traffic cultures.

All in all, Studies I and II together with the studies cited above call into question the widely held practice of forming sum variables based DBQ items and comparing them across different groups of drivers in order to investigate their accident risk. This is because the correlational patterns among the DBQ variables differ in important respects from sample to sample. Further, rigorous studies based on analyses of measurement equivalence (Study I, Study II, Stephens & Fitzharris, 2016) obtained well-fitting models only after handcrafting the original model to fit the data at hand. Also, even in highly similar groups of drivers such as the two groups of middle-aged drivers examined by Stephens & Fitzharris (2016), only partial strong equivalence for the four-factor model of the 27-item DBQ was obtained, and even that after modifying the original model. The result is thus similar to that of Study II when it comes to *slips*, where partial strong equivalence was also obtained after modifications to the original model. In practice, then, it seems safe to conclude that it is the exception rather than the rule that the levels of the latent variables can be compared across groups of drivers, and even then only after carefully examining and modifying the model structures to fit the data at hand.

Study III, for its part, was based on data from novice drivers from the Great Britain and offered an alternative to the latent variable view by employing methods of *network psychometrics* (Epskamp, 2017). When applied to traffic psychology, it states that 1) there exist direct relationships among individual traffic behaviors, 2) these relationships can be modelled based on correlational data and 3) analysed using tools developed within mathematical graph theory and the analysis of social networks, such as indices of centrality and clustering.

The network models can be understood as tools for generating novel testable hypotheses concerning the causal relations among the traffic behaviors in a data-driven manner. Because of that, they offer a useful point of view to associations among DBQ variables that are problematic for the latent variable models. For instance, it is easy to understand that *exceeding speed limits*, *tailgating* and *having to brake abruptly* are associated as soon as they are assumed to be causally related. On the other hand, under the latent variable view, the first two of these behaviors are taken to be generated by the tendency to violate rules and the remaining one by the error-proneness of the driver, so the former two should not be correlated with the third one. The causal assumption made in the network models naturally remains as nothing more than an assumption, but at least one that is compatible with the results and with what we know independently about how people behave in traffic. In latent variable models, such associations can be explained by allowing the error variances of the observed variables to correlate, but doing so is necessarily more of an exception than a rule: the models are able to accommodate only a certain number of such associations before becoming overly complex to be estimated based on the observed correlations (see e.g. Chapter 6 in Kline, 2011, or Gunzler & Morris, 2015).

The application of centrality indices for identifying intervention targets is an actively discussed topic within the network psychometric literature (see, for instance, Fried et al., 2018 and Bringmann et al., 2019). If high values of the centrality indices identify nodes that have strong causal connections with the rest of the network, then introducing a change in the central nodes might have a marked effect on the rest of the network (for a thoughtful discussion of this interpretation, see Fried et al., 2018). Study III suggests that *speeding*, *crossing an intersection on red* and *tailgating* could be among such central behaviors. This can be contrasted with conclusions from studies based on the latent variable view of violations and errors: such studies 1) identify groups of behaviors, based on the factor loadings of individual observed variables, that are putatively affected by the same psychological processes, 2) analyze subgroup differences in the level of these processes – e.g. the tendency to deliberately violate rules or a general tendency to be commit errors – and finally 3) suggest interventions targeting the process in the subgroups that have high values on the respective sum scores. As argued above, studies I and II identify critical problems in this practice, as identical latent variables are seldom uncovered across groups. Further, a remark from Study III deserves to be repeated:

> *"If we take the latent variable view seriously, we can only influence individual behaviors through manipulating the latent variables: whether we want to reduce drunk driving or speeding, we should aim at the drivers' rule-breaking tendencies, because influencing an individual violation has no effect on other behaviors under the latent variable view".*
>
> *Study III*

In these respects, then, network models of driving behaviors seem to offer a more intuitive basis for possible traffic safety interventions: if certain driving behaviors are directly related to one another, influencing their interactions becomes meaningful. Still, one must be cautious when drawing data-driven conclusions and making recommendations based on statistics such as the centrality indices, since even if the network models would succeed in capturing true causal associations between variables, the models are compatible with several different causal structures (Fried et al., 2018); indeed, the correct interpretation and use of centrality indices in network psychometrics is an active current research topic (Bringmann et al., 2019). In addition, the centrality indices seem blind to certain traffic behaviors that are independently known to be critical to traffic safety, such as *driving under the influence of drugs*, which appears as the least central node in the network model reported in Study III (Figures 9 and 10). Uncritical use of any statistics, be them factor loadings or centrality indices, will certainly lead to poor conclusions.

Similar considerations in other domains of psychology have led researchers to discuss *sufficient causal variables*; for instance, in the domain of quality of life studies, intense pain may by itself be a sufficient cause for lowered quality of life (Fayers & Hand, 2002), even though there may be other similarly important sufficient causes, such as intense vomiting. This has implications on constructing self-report instruments: all predictors of clinically relevant outcomes need to be included, irrespective of their correlational patterns with other variables. This is relevant also when it comes to the DBQ: for instance, the 27- and the 28-item versions of the instrument differ in including vs. leaving out the item related to drunk driving, and the decision to do so is motivated by the item having low correlations with the rest of the items. Further, when adopting the idea of sufficient causal variables, it may be wise to perform analyses by calculating the maximum value of the sufficient causal variables rather than calculating a sum of all indicator scores, which may dilute the effects of the important variables (Fayers & Hand, 2002). This, for its part, might mean picking certain DBQ variables based on their independently known relevance in accident causation and examining responses to them, irrespective of their correlations with the rest of the variables.

On a more general level, the network view offers promise of bringing different traditions of traffic research – self-report-based research on the one hand, theoretical and experimental research on the other hand – closer to each other. This is because the network view makes it possible for self-report-based research to adopt a starting point that has been self-evident within the other traditions: recognizing the importance of individual driver behaviors and their interrelationships. For instance, rather than viewing speeding as an indicator of the rule-breaking tendency of a driver, theories of driver motivation (Fuller, 2005; Summala, 2007) and engineering models of accidents (Abdel-Aty & Radwan, 2000) view it as an important determinant of crashes and driver errors in itself. Similarly, the network view is easily reconciled with studies that aim at determining the reasons for individual traffic behaviors such as speeding (Lawton et al., 1997; Parker et al., 1992; Wallén Warner & Åberg, 2006).

4.1 Relationships between latent variables and self-reported driving behaviors

Factor analyses of the DBQ are often motivated by the idea that they are useful for identifying psychological traits or mechanisms that underlie observed traffic behaviors (e.g. Mesken et al., 2002; Reason et al., 1990; Stephens & Fitzharris, 2016). The present thesis as a whole takes a sceptical view toward this motivation, and argues that relationships between psychological mechanisms and individual traffic behaviors are likely much more complicated than suggested by studies based on factor analyses of the DBQ. Not only is it plausible that certain behaviors are affected by different psychological mechanisms in different groups of drivers (for instance, an older driver may drive through a red light by mistake whereas a younger one may do it on purpose), individual differences in the frequency of performing the driving behaviors are likely to depend on the functioning of multiple psychological processes. In addition, it is important to remember that analyses of cross-sectional data produce results related to differences between individuals. As is well known, psychological mechanisms functioning within individuals cannot be studied by such methods, and careful modelling, laboratory work and, indeed, case studies of individuals are needed instead (Borsboom et al., 2003). This observation is not new: William Stern, often hailed as the father of differential psychology, actually spoke fervently for *personalistic inquiry* as a method for understanding psychological processes in the beginning of the 20th century; Lamiell (2003) is a book-length treatise of his ideas. Further, the distinction has been acknowledged within mainstream psychology at least since Cronbach's classic article *The two disciplines of scientific psychology* (Cronbach, 1957).

The question of the relationship between differential and experimental psychology is a deep one, as it relates to what it is possible to investigate within scientific psychology. At one extreme it has been argued that the psychic structures of no two individual are alike (Lykken, 1991). If this were indeed the case, scientific psychology would reduce to studying individuals one person at a time and no finding could be generalized to others. In the other extreme it can be assumed that the dimensions along which individuals differ from each other are identical to the dimensions that explain variation within individuals and that the psychological structures underlying the mental lives of any two individuals are essentially alike. It is possible to take different kinds of intermediate positions, as well. The first two studies in the present thesis imply that one way of integrating the perspective of interindividual differences and intraindividual processes would be to construct theories and models that account for the behavior of suitably chosen subgroups of drivers – even though see the discussion on ergodicity, below, for central reservations that call this interpretation into question.

The third study takes a different perspective and concentrates on direct associations among individual traffic behaviors, interpreting them as causal hypotheses. Such hypotheses would naturally need to be tested either in laboratory conditions or based on time-series data (Costantini & Perugini, 2018). Adopting this approach has the effect of downplaying the importance of latent variables, and the present thesis considers the possibility of viewing the latent dimensions as emergent properties of interactions among the individual driving behaviors. Ideas from the study of complex systems are also tentatively applied: it is suggested that the network of driving behaviors may occupy different states which depend on the status of the driver (stressed, tired, etc.), the driving situation and the (social) context in which the driver is embedded.

The discussion related to complex systems is motivated by a central methodological question:

> *Given a particular set of selected variables, under which conditions will an analysis of interindividual variation—an analysis in which information is pooled across subjects—yield the same results as an analysis of intraindividual variation?*
>
> *Molenaar & Campbell, 2009*

The question is known as that of *ergodicity* (Molenaar & Campbell, 2009) and the answer is, in a nutshell: extremely rarely. For a process to qualify as ergodic, the conditions of homogeneity of the population and stationarity of the phenomenon must be fulfilled. The former refers to the applicability of the same statistical model across individuals. In the present case, the assumption would be fulfilled if the same latent variable model or the same network model of the DBQ would fit data from each individual comprising the sample. Assessing the assumption would naturally necessitate collecting intensive longitudinal data. The same goes for testing the stationarity assumption, according to which the statistical parameters of the model (such as factor loadings or network edge weights) remain invariant over time. Questions related to ergodicity and interpretations relying on ideas from complex systems research cannot, then, be tested based on the data used in Studies I – III, and they are mainly mentioned as possible directions for future research.

Studies I and II partitioned the samples of drivers into subgroups that shared a property such as age, gender or nationality. Molenaar & Campbell (2009) describe a study that went much further and constructed within-person models of the personality of individual respondents based on Big Five data. Perhaps surprisingly, the within-person models comprised two to four factors instead of five, calling into question whether the Big Five model functions as a description of the personality of any given individidual across time. The example encourages considering whether the factor models and network models of the DBQ might prove to be ergodic in the same sense. When it comes to factor models, the lack of measurement equivalence of the DBQ reported in Studies I and II casts preliminary doubt on the assumption.

On the other hand, could traffic be such a sphere of human life where the ergodicity assumption holds? Consider, for the sake of a thought experiment, edges connecting speeding and tailgating in a between-person network model and (as of yet non-existent) within-person network models. For example, it may be that if John speeds more than the average driver, he ends up tailgating more than the average driver; similarly, if he on a certain day speeds more than he usually does, he may end up tailgating more than he usually does. In other words, because traffic behavior is regulated by rather strict rules, it seems at least conceivable that edges formed in within-person and between-person network models of driver behavior might behave similarly, thus supporting the assumption of the homogeneity of population. When it comes to the assumption of stationarity, could it be that after the relationships between the driving behaviors of a given individual have stabilized, i.e. when their driving has become automatized, their relationships could be described by statistical parameters that are invariant across time?

The preceding discussion is admittedly speculative, but it is important to consider such questions because in the DBQ tradition, the latent variables have commonly been identified with psychological processes. The linkages between psychological processes and latent variables are, then, likely to be much more complex than currently assumed within DBQ studies. The following discussion is organized around the broad themes of *violations* and *errors*, and it examines the causal assumptions of different kinds of statistical models for DBQ data.

4.1.1 The structure of violations

Studies I and II concluded that the internal structure of the latent variables related to *violations* differed across subgroups of respondents, perhaps according to what was considered a violation of social rules in the different groups of drivers. Network models of traffic behavior (Study III) offered a novel point of view and a potential explanation for these differences when violations are considered from the point of view of network science and as complex systems. This approach allows us to ask questions such as whether the networks of traffic behaviors can be thought to occupy different states similarly to the symptom model of depression that can be said to be either in the healthy state or the depressed state (Borsboom, 2017). If so, what would these states be? Answering the questions would necessitate collecting intensive longitudinal data on individual drivers, but the results of Study III can at least be interpreted from this point of view.

Study III tentatively suggests that the states of the networks of driving behaviors might be related to the drivers' status (such as *tired*, *under the influence of substances* or *intensely emotional*) or the presence of other people in the car: for instance, young drivers are likely to behave differently when accompanied by their peers than when they are in the company of older relatives (Alver, Demirel, & Mutlu, 2014). The peers may encourage the drivers to try their limits, which perhaps translates to the nodes related to *speeding* becoming activated in the context of driving with peers. Similarly, the crash risk of young mothers has been shown to be elevated when they drive with an infant passenger vis-à-vis alone (Maasalo, Lehtonen & Summala, 2017); it is likely that the dynamics of their behaviors are different across these two contexts.

Figure 9 of the present thesis indicates that drivers who are more likely than average to *exceed speed limits on motorways* are also more likely to *race from traffic lights* and somewhat more likely to *overtake other drivers from the inside*. High activation of the node related to *speeding on motorways* would then activate the other two nodes, which are themselves connected to nodes related to *aggressive behavior* and *staying in lanes that will be closed*. Perhaps, then, the *tendency to violate traffic rules* and *the tendency to behave aggressively* could be interpreted as emergent properties of the network of traffic behaviors: Rather than being unobservable properties that cause people to behave in a certain manner, they are states of high activity in the network of driving behaviors and properties that arise from patterns of strongly interconnected behaviors in suitable contexts such as the presence of other people in the car, the driver being stressed or tired etc.

Study III also suggests that errors and violations cannot be categorically differentiated from each other. This seems intuitively plausible, since violations related to *speeding* or *driving through intersections against a red light* may well increase the probability of committing errors as shown in, for instance, in the zero-risk theory of Summala (2007): when drivers approach the limit of their skills, they have less cognitive resources available for considering everything that happens in their surroundings. It must be noted, however, that one of the nodes in Figure 9 of the present thesis is categorized erroneously: node e15 (*tailgating*) usually loads on *violations* rather than *errors*, and even though it often has a strong cross-loading on an *error* factor, it should have been categorized as a violation. Still, the essential observation holds: the node in question has strong associations with prototypical error nodes such *failing to check rear-view mirror* and *having to brake abruptly*; especially the latter is a plausible candidate for a causal consequence of *tailgating*. Similarly, node v3 (*crossing a junction on red*) is connected to *missing observing pedestrians* and *not having a clear recollection of the road,* which in itself may function as an indicator of an unmodelled latent variable *absent-mindedness* as argued below in Section 4.1.2. Interpreted as a causal hypothesis, this association can be taken as indicating the possibility that *absent-mindedness* might increase the probability of *driving through an intersection against a red light*.

Further, it is possible that the personal histories of the drivers determine the initial strengths of the interconnections between the individual traffic behaviors, such as speeding and showing aggression. This would explain why certain drivers become more aggressive than others when their progress is impeded as per the quote from Björklund (2008) below. Indeed, one motivation for performing network analyses of traffic behavior is that they are well compatible with the idea of human behavior being highly context-dependent and likely affected by multiple factors, with the individual behaviors being in dynamic relationships with each other (Mischel & Shoda, 1998). The importance of considering such dynamic relationships has been voiced, among others, when investigating aggression in traffic:

> *"Drivers who enjoy a somewhat faster speed than other drivers will more often be obstructed by other traffic, and therefore they will become irritated more often and be more likely to educate other road users. They probably also will become more irritated than other drivers when obstructed, because they want a faster progress"*
>
> *(Björklund, 2008)*

The above quote illustrates the interplay of individual differences in traffic behaviors and character traits. Similar relationships were in fact observed in Study III even though the study targeted the whole population of new drivers instead of those with anger-management issues. The background variable of *perceiving oneself to be a fast driver* shared a strong edge with *racing away from traffic lights to beat other drivers*, which was for its part related to *becoming angered by other drivers and giving them chase.*

Incidentally, the *becoming angered* node shared another strong edge with the node *overtaking others when they are turning* (traditionally classified as an error), which was, for its part, related to *overtaking slower drivers on the* *inside*. These behaviors can also be interpreted as being related to the dynamics described in the quote from Björklund (2008), above.

Further, the network models may help to explain how character traits and behaviors become aligned, for instance for avoiding cognitive dissonance (Dalege et al., 2016). Indeed, in Study III, the node describing attitudes *decreasing motorway speed limit is a good idea* shared a moderately strong negative edge with *perceiving oneself to be a fast driver* and a considerably strong negative edge through one of the nodes related to *speeding*. These relationships are seen by inspecting Figure 5 in Study III. The interesting property of network models in this context is that they encourage the researcher to consider different causal (and non-causal) relationships between attitudes and behaviors. For instance, it may well be that drivers who usually drive according to the speed limits develop anti-speeding attitudes as a consequence, i.e. behaving in a certain manner may influence the attitudes of a driver instead of or in addition to the opposite.

The network conceptualization of the relationships between attitudes and behavior offers both theoretical and methodological benefits when compared to structural equation models of these phenomena. For instance, Lucidi et al. (2019) present a SEM that consists of directed relationships between latent variables; the causal chain encoded in the SEM runs from personality traits to attitudes, from there to GEMS variables (violations, lapses and errors) and from there to crashes and traffic fines. Network models leave room for such relationships to be non-directed, which allows the researcher more freedom in interpreting the results. Furthermore, in a typical SEM, a great multitude of equivalent models (models that fit the data equally well) can be specified, even though typically only one model is reported (Kline, 2011, Chapter 8). This property is known as the issue of underdetermination of latent variable models. Network models are more transparent in this respect, as the estimation procedures that are commonly used lead to a single model (Epskamp, 2017).

This section presented a network perspective to traffic rule violations and cognitive errors in traffic; it was suggested that many of the relationships that have traditionally been analyzed as reflecting the level of a driver's tendency to break traffic rules could equally well be interpreted as direct causal links between the behaviors. Still, the whole picture seems more complicated than that. For instance, the aggression-related nodes form a strong clique and one of them (v4; *become angered and give chase*) is the most redundant node in the network (Figure 10). The aggression-related variables could perhaps indeed be modelled by a latent variable of their own; on the other hand, it might well suffice to include only one of the three aggression-related nodes in a network model of traffic behavior. Models integrating latent variables and network models are discussed below in section 4.4.

4.1.2 The structure of errors

The common theme across the three studies included in the present thesis is that *errors* are not a homogenous group of traffic behaviors. Studies I and II indicated that the four-factor model fits better than the two-factor model, while the network analyses of Study III indicate that not all errors share strong associations with one another. This is perhaps to be expected, as the errors are not intended as a homogenous category: for instance, the GEMS characterizes *slips* as a failure in the execution stage of an action sequence, and *lapses* as a failure in the storage stage (Reason et al., 1990, p. 43). Moreover, as discussed above (Section 4.1.1.), there is no sharp empirical boundary between errors and violations – something also attested by the strong cross-loadings and different factor structures observed across subgroups of drivers in Studies I and II.

The results related to errors can partly be interpreted from the point of view of network psychometrics, looking at the network of driving behaviors as a complex system occupying different states, as argued above (Section 4.1.1.). Still, some of the associations are perhaps most naturally interpreted as reflecting the effects of unmodelled latent variables. The driving behaviors related to *lapses* in the four-factor model of the DBQ occupy the upper left-hand side in Figure 9. They do not, however, form a single homogenous *clique* in the network, i.e. they are not all interconnected; rather, node e13 (*noticed ending up on a different road than intended*) shares strong connections with the rest of the nodes. Particularly, errors apparently related to *absent-mindedness* (e3: *forget where left car*, e13: *ending up on a different road than intended* and e6: *no recollection of road*) form a strong clique, and could perhaps be modelled by a latent variable of their own. Further, the association of *misreading signs* and *getting into wrong lane* can rather naturally be interpreted as a causal hypothesis whereas the strong association between e4 (*switch on a wrong thing*) and e1 (*take off in too high a gear*) may reflect the presence of an unmodelled latent variable such as *inexperience with car controls*. Looking at behaviors that are categorized as *lapses* (in the four-factor model) in Figure 9, it seems that no parsimonious explanation for the results is to be found: the results are likely related both to unmodeled latent variables and potentially causal associations among the items.

Further, it is possible, and perhaps even likely, that the frequency of different kinds of cognitive errors depends on the age and experience of the driver, thus complicating psychometric studies of driver behavior. Study I suggested that different subgroups of drivers might be homogenous enough so that a common denominators for different categories of error could be found within the groups, if not across groups. For instance, it might be that the items thought of as measurements of *lapses* could be related to *inexperience* in the youngest age group, to *inadequate planning of the driving task* in slightly older drivers and to *problems in updating a mental model of the driving situation* in the oldest age group. In Study II, *lapses* were also interpreted as reflecting the inexperience of the young drivers. Even though such interpretation may seem at odds with the considerations related to ergodicity (Section 4.1.), it may be that phenomena such as inexperience occupy a middle ground between fully individual psychological processes and latent variables intended as psychological descriptions of the whole population: it may be that they are formulated at a suitable level of granularity to describe a given subgroup of respondents if not the whole sample.

Moreover, as the network models reported in Study III show, doing research in the real world presents complications not accounted for in the ideal picture painted in GEMS. Even if *slips* or *lapses* are based on similar malfunctions in cognitive processing, the individual driving behaviors that are taken as measurements of them under the latent variable view are still related to other driving behaviors – some of which are classified as violations – in ways that are shown in Figure 9 of the present thesis. That is, certain violations share potentially causal associations with certain errors as discussed above.

In closing, it may be of interest to note that recent research in mathematical psychology has developed models that combine an intraindividual information-processing model with a latent variable model of interindividual differences. In the first stage, intraindividual processes are explicitly and mathematically modelled, and the parameters of these models used as input to latent variable models concerned with interindividual differences (van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011; Vandekerckhove, 2014). The first stage aims at modelling the cognitive processes that underlie the production of a response to a given item in the test or self-report instrument and can consist of, for instance, a model of the component processes that influence the reaction time to an item. Van der Maas et al. (2011) refer to their models as proofs-of-concept rather than fully developed analytical solutions, but such models offer a promise of uniting the two disciplines of scientific psychology. Perhaps the cognitive errors encoded in the DBQ would benefit of being modelled based on a related method: the GEMS variables are, after all, related to intraindividual processes, and it is unlikely that interindividual variation would succeed in directly mirroring the functioning of intraindividual processes (Molenaar & Campbell, 2009).

4.2 Predicting crashes from individual driver behaviors

Study III reported several regression analyses in which crashes were predicted from driving behaviors. In all models, age, sex and annual mileage were controlled for. *Missing observing other road users*, *having problems with car controls* and *behaving aggressively* were the most important predictors together with *hitting something when reversing*, which incidentally is a minor crash in itself. The results can be compared to those of

Wallén Warner et al. (2011), who also predicted self-reported crashes from individual driving behaviors. Incidentally, five out of six of the predictors they ended up with were also among the predictors of the elastic net model reported in the present study. The predictive model reported in the present study included, however, also several other predictors. It is not possible to draw firm conclusions based on only two studies, so further studies assessing individual driving behaviors as predictors of crashes are in order.

Nonetheless, the approach taken to model building in the present thesis was certainly as important (if not more so) than the actual results of the regression analyses. First, from the substantive point of view, it seems obvious that the different kinds of violations encoded in the DBQ must have differential associations with accidents; they range, after all, from quite harmless (honking at other road users) to clearly irresponsible (driving under the influence of alcohol). Stated slightly more technically, when DBQ variables are used as predictors of accidents, it is quite likely that they contain unique information when it comes to predicting accidents; such unique information is lost when a sum variable representing a putative underlying latent variable is calculated. The regression analyses reported in Study III can also be interpreted as representing an approach to constructing psychometric scales known as *criterion-keyed scale development* (Chapman, Weiss & Duberstein, 2016). Under this framework, the internal consistency (reliability) of the scale plays no role; indeed, the precision of the estimates of the value of the predicted variable is maximized when the collinearity of the scale variables is zero (Chapman, Weiss & Duberstein, 2016). From this perspective, the variables with non-zero regression coefficients in Table 8 above represent a criterion-keyed DBQ, with the self-reported accidents serving as the criterion. For the sake of clarity, the version of the DBQ with those variables could be called $DBQ_{accident\text{-}keyed}$, and it would be of interest to see whether future studies embodying the same analytical framework end up including the same variables in the accident-keyed DBQ scale.

When it comes to methodological details, Study III had certain benefits not shared by previous models that aim at predicting accidents based on the DBQ: In Study III, the regression models were truly predictive in that the independent variables and the dependent variables were collected at different time points, the models were built and tested in separate subsets of data, and model fit was tested based on cross-validation of models. Regularized regression was used for avoiding overfitting the model to data and for selecting the predictors. Being able to employ the hold-out data was deemed especially useful, since doing so enabled showing that the naïve Poisson model including all predictors fit the hold-out data worse than a null model with no predictors. Overfitting is, then, a serious problem when creating models of accident prediction based on self-report data. It is hoped that the methods used presently would benefit future self-report based studies in traffic psychology.

4.3 The measurability of psychological properties

Before considering the limitations of the present work and possible future directions afforded by the network perspective, an interlude to the measurability of psychological properties is made. Interest in such philosophical questions has motivated both the empirical studies related to the measurement properties of the DBQ (Studies I and II) and the search for alternative conceptualizations for the correlational structure of the instrument such as the network models presented in Study III.

As argued above, the latent variable view of violations and errors assumes the existence of stable and measurable psychological traits as properties of individual drivers. Speaking of psychological traits amounts to adopting a realist interpretation of latent variables, i.e. assuming that they are things that exist independently in the world and have causal power. The realist view is compatible with the correspondence theory of truth, but hardly with the coherence theory of truth associated with a social constructivist view of psychological properties (Borsboom, Mellenbergh & Van Heerden, 2003). As the issue has not (to my knowledge) been explicitly discussed within the DBQ tradition, the argument of Borsboom et al. (2003) is briefly presented below.

Borsboom et al. (2003) argue that latent variables need to be given a realist interpretation on three grounds: First, evaluating the position of a subject on the latent variable is possible only if there is something to be right or wrong about, i.e. a continuum along which subjects are placed; otherwise, it seems unsensible, for instance, to speak about having made an error in such an evaluation, even though much of latent variable theory is about the probabilities of making such mistakes given the positions that two individuals occupy on the continuum. Second, and on similar grounds, estimating the population values of parameters would seem to require that there is something to be estimated, and again, something to be either right or wrong about. The third argument is a bit more involved. It revolves around the ideas misspecification ("all models are false even though some of them are useful"), underdetermination of models and the relationships these ideas have with estimating the values of parameters. In short, the concept of misspecification of models would seem to presuppose giving latent variables a realist interpretation because there needs to be a true model to which the misspecified model is compared. On the other hand, equating truth with empirical adequacy would lead to counterintuitive conclusions when it comes to, for instance, equivalent models such as a model with two correlated factors vs. a single factor. This is because empirically, both models fit data equally well and under the definition *truth = empirical adequacy*, the assumption that model B is true can be replaced with the assumption that model A is true. This, then, leads to such potent absurdities as estimating the correlation between two latent variables even though the model only contains one. These themes are more fully discussed in Borsboom, Mellenbergh, & Van Heerden, (2003, pp. 209–211). The DBQ research tradition seems to be in line with these arguments, since the common practice of equating the latent variables (*errors* and *violations*) with psychological properties has a realist undertone to it. In what follows, a realist interpretation of the latent variables (*errors* and *violations*) is assumed.

A central consequence of adopting the latent variable view is that the *measuranda* (things to be measured) are – perhaps implicitly – assumed to possess quantitative structure (Michell, 2008). This assumption has remarkable consequences, even though it is seldom problematized within psychology (Michell, 2008). Theories and measurements of intelligence function as a prime example of purportedly quantitative research within psychology: within its history, psychologists have considered problems such as the unit of intelligence and whether there exists a well-defined zero point of intelligence etc. (Eysenck, 1973). Considering questions such as these is important in itself, as measurement involves deep philosophical questions that are to date without equivocal answers (Tal, 2017), but also because the measurability of attributes has its own important role to play in determining whether the results of inquiry can be replicated (Hanfstingl, 2019).

Within psychology, the question of whether attributes are quantitative is famously sidestepped by defining measurement as the "assignment of numerals to objects or events according to rules" (Stevens, 1946). When the assignment is done to classify things, one speaks of *nominal measurement* and when to order things, of *ordinal measurement*. *Measurement on the interval scale* supports ordering things such that consecutive points on the scale are equidistant, while *measurement on the ratio scale* has the additional property of having a meaningful zero point. The first two categories are often treated as nonquantitative, the latter two quantitative. Although this may seem unproblematic, it is important to note that the question "*When is an attribute quantitative?*", i.e. when is it appropriate to apply one of the latter two scales, is rarely raised. This is because the theory of Stevens was *operationalist* through and through, i.e. it assumed that the meanings of scientific concepts are defined by the operations used in identifying them (Michell, 1997). Further, this amounted to measurement becoming synonymous with the operations used when assigning the numerals to objects, resulting in, among others, the famous "definition":

> *"Intelligence is what the test tests"*
>
> *(Boring, 1923)*

This, however, is obviously unsatisfactory for a researcher that assumes a realist stance towards psychological attributes (Borsboom et al., 2003). As remarked above, in traffic psychology a realist stance towards errors and violations is, in fact, assumed; further, errors and violations are treated as continuous latent variables that have by definition quantitative structure. Because of this, it would appear that an independent definition for an attribute to possess quantitative structure is needed.

The first formal definition for an attribute to be considered as a quantity was given in the beginning of the 20th century by Hölder in the form of seven axioms, the first four of which are reproduced below for the sake of an example in the form presented by Michell (1999). Below, *Q* refers to an attribute such as length, intelligence or the tendency to violate rules. In order to be measurable, a property needs to fulfill these conditions for a quantity (with analogous conditions applying to positive real numbers):

> *1. Given any two magnitudes,* a *and* b*, of* Q*, one and only one of the following is true:*
> *(i)* a *is identical to* b (a = b, b = a);
> *(ii)* a *is greater than* b *and* b *is less than* a (a > b, b < a)*; or*
> *(iii)* b *is greater than* a *and* a *is less than* b (b > a, a < b)
> (any two magnitudes of the same quantity are either identical or different and if the latter, one is always greater than the other).
>
> *2. For every magnitude,* a*, of* Q*, there exists a* b *in* Q *such that* b < a (for every magnitude of a quantity there is another that is less).
>
> *3. For every ordered pair of magnitudes,* a *and* b*, from* Q*, there exists* c *in* Q *such that* a + b = c (for every pair of magnitudes, there exists another, their sum).
>
> 4. For all a and b in Q, a + b > a and a + b > b (every sum of two magnitudes is greater than each of those summed).
>
> *Michell (1999, pp. 52–53)*

As the present thesis is not an exercise in mathemathical psychology, the axioms will not be discussed in detail. They are, however, reproduced here to highlight the fact that – contra Stevens (1946) – it is not trivial for a property to qualify as a quantity. Nonetheless, a couple of remarks are made to illustrate the kinds of questions that are at stake. Consider axiom 4, according to which the magnitudes of quantities must be greater than zero. Intuitively, it would seem possible for someone to be not at all depressed so that their depression would be equal to zero, or to think that someone is not at all likely to break rules so that their tendency to violate rules would be zero; however, this is not possible under axiom 4. Importantly, these examples concern ideal cases; in the real world, measuring such vague properties as depression – even if it were possible – involves so many uncertainties related to conceptualization, observation and measurement that it would be impossible to know if a measured zero value of depression corresponds to an actual zero level of depression[1].

Or consider axiom 3, according to which magnitudes are additive. Is it correct to say: "If Brian's tendency to violate rules is 5 and Joel's tendency to

[1] I am grateful to professor Reijo Sund, who acted as the pre-examiner of the present thesis, for pointing out this qualification.

violate rules is 10, taken together, their tendency to violate rules is 15" or that "On Tuesday, Brian's tendency to commit errors was 12 and on Friday it was 24, thus on Friday he was twice as likely to commit errors"? It seems at least possible that psychological properties do not behave in this manner, that is, it seems conceivable that they are properties that do not possess additive and / or multiplicative structure.

Meanwhile, sophisticated (even though perhaps in practice infeasible) methods ascertaining the nature of psychological properties as quantities have been developed. Additive conjoint measurement (ACM, Luce & Tukey, 1964) is sometimes hailed as the greatest of such developments (Michell, 1997), and some scholars have argued that modern psychometrics should be replaced by ACM to ensure that what is being measured indeed are quantities (for a review of the discussion, see Sijtsma, 2012). The ACM model builds on the idea that the thing to be measured, P, is independently influenced by two other factors, A and X. For instance, test performance might be influenced by motivation (A) and ability (X)

$$P = f(A,X) = A + X$$

Axioms such as single cancellation and double cancellation are central to ACM. They involve independently manipulating ordered pairs of A and X that provide a partial ordering on P, and from this it can be inferred that P has a quantitative structure. The presentation given here is obviously limited, and a fuller account of ACM can be found, for instance, Michell (1999, ch. 8). Others argue that so-called Rasch models developed within Item Response Theory (IRT) function as stochastic versions of ACMs, with the important difference that they take random error explicitly into account (for a review of the discussion, see Sijtsma, 2012).

These issues are the subject of ongoing philosophical and methodological discussion. It has been argued, for instance, that measurement in psychology cannot be possible even in principle, because being able to measure something would necessitate ruling out the influence of systematic error factors, and because this cannot be done in psychology, it is not even possible to determine that a property has equal levels across subjects or across time within a subject (Trendler, 2009). Nonetheless, painstaking, decades-long experimentation based on the ACM has now tentatively showed that certain fundamental psychological properties such as perceived loudness and brightness can perhaps be treated as quantities (Luce & Steingrimsson, 2011). As a sidenote, such research is as unpopular as it is necessary: the Luce & Steingrimsson (2011) article has acquired 15 citations in 9 years according to Google Scholar.

The point I wish to make with these examples is that the measurability of properties cannot be unproblematically assumed and taken for granted within traffic psychology. Constructive reactions to these problems have included underscoring the importance of well-developed theories as the basis of measurement and urging researchers to consider whether classification or ordering of phenomena would be feasible objectives in their field (Sijtsma, 2012). These questions must be addressed within traffic psychology, as well. When it comes to an underlying theory, self-report-based traffic research is in a good place in that the GEMS exists as a carefully developed background theory for the self-report instruments. The following quote may well be applicable in the present case, as well – in Section 1.1. it was argued that the original version of the DBQ failed to operationalize all the constructs referred to in GEMS:

> *"Even for a correct attribute theory, an unfortunate operationalization may undermine meaningful measurement."*
>
> *(Sijtsma, 2012)*

Further, whether our focus is on latent variables or individual traffic behaviors, their status as quantities cannot be continued to be taken for granted. Because of this, considering research methods that explicitly treat the subject matter as having not quantitative but rather nominal or ordinal structure, such as Observation Oriented Modeling (OOM) might be in order within traffic psychology as well (Grice, Barrett, Schlimgen, & Abramson, 2012; Grice, 2014).

This discussion can be simultaneously taken as a severe consideration of the limitations of the present thesis. In case the properties examined in Studies I and II are deemed not to be quantities, the results of these studies are essentially devoid of meaning – or perhaps the findings are trivially true. Still, the studies of the present thesis are naturally not alone in facing these challenges, which are shared by all self-report based studies into traffic behavior.

The network models of traffic behavior discussed in the present thesis face the same challenges concerning measurability and the nature of the phenomena as quantities. However, the co-occurrence of phenomena can be examined using different kinds of indices of mutual information, which do not entail that the properties under comparison are quantities or their relations linear. Network analyses based on partial mutual information have been employed in, for instance, investigating the structure of stock markets (You, Fiedor, & Hołda, 2015) – such indices may perhaps prove useful also within traffic psychology.

4.4 Limitations of the studies

The current section considers the theoretical and methodological limitations of Studies I–III. It, however, abstains from philosophical discussion concerning the measurability of psychological properties (Section 4.3.) and takes for granted that the enterprise of treating psychological properties as quantities stands on solid ground – an admittedly bold assumption given the discussion in Section 4.3.

In this thesis, network models have been put forward as a significant advance compared to latent variable models, even though the two models have complementary strengths and weaknesses. Study III, in particular, failed to discuss this theme, so it is briefly considered in what follows. Latent variable models are an elegant method to account for measurement error in item responses, which is precisely the weak spot of network models. If an estimate of the reliability of the measures exists, it can be input directly as a parameter of latent variable models, or it can be estimated using, e.g., CFA or ESEM. On the other hand, creating data-driven hypotheses concerning the (causal) relationships among a group of variables is what network models excel at, whereas latent variable models face the problem of choosing among equivalent models in such a setting.

It may well be that the structures of the network models of traffic behavior discussed in the present thesis are partly due to the influence of unmodelled latent variables (see Sections 4.1.1. and 4.1.2.). On the other hand, it seems plausible that at least some of the associations among the traffic behaviors are due to direct (causal) connections among the traffic behaviors. Because of this, an optimal research method would perhaps combine the strengths of latent variable models and network models. Fortunately, such methods have been developed: Latent Network Models (LNMs) contain a latent variable model as a measurement model and then model relationships among the latent variables as a network, while Residual Network Models (RNMs) model the residuals in a Structural Equation Model as network (Epskamp, Rhemtulla, & Borsboom, 2017). Both could be productively used in the context of psychometrics of driver behavior. For instance, the potential unmodelled latent variables discussed above (Sections 4.1.1. and 4.1.2.) could have been explicitly modelled using a LNM, while a "poor man's version" of a RNM was already used in Study II in visualizing the residual correlations of the measurement models.

The usefulness of LNMs is naturally not restricted to explicitly accounting for unmodelled latent variables. Rather, it is a reasonable requirement that measurement error must be accounted accounted for in all variables that are input into a network model. Indeed, it has been suggested that *all* observed variables should be interpreted as multidimensional latent variables:

> *"It is not inconceivable that experience and judgment (positive latent outcomes of age) are positively related to task performance, while decreased physical and cognitive capacity (again latent outcomes positively related to age) may relate negatively to task performance. If these latent outcomes of age happen to cancel each other out, how does one interpret the finding that age is not related to task performance?"*
>
> *(Howell, 2008)*

The example is delightful: While the unidimensionality of latent variables is assumed and receives a great deal of attention in methodological literature, it is a refreshing thought that even the seemingly simplest observed variable can be interpreted as carrying information about several latent variables. Study III might, then, have benefited from modelling all variables that were used in the network models as latent variables, with measurement error input as a parameter for the variables whose reliability could not be estimated otherwise.

When it comes to more purely methodological limitations of the studies, the estimation method used in Study I did not allow teasing apart analyses of weak and strong measurement equivalence. This can be considered a major flaw, since Study II showed that such differences may well be relevant in DBQ data. A further limitation of Study I is that it did not follow up the initial measurement equivalence analyses with analyses of partial measurement equivalence. This might have qualified the results in important ways: perhaps some of the latent variables would have proved to be at least similar in nature across the age groups or genders. Study II, for its part, was based on comparing the measurement properties of the DBQ across two countries using pre-existing archival datasets that were originally collected for other purposes. This may well have introduced bias into the results, but the question could not be investigated for the same reason: the data sets included only few common background variables.

Further, Study I has been criticized for being based on overly small sample sizes, for overextraction of latent variables and for unrealistic expectations (de Winter, 2013). A reply to the arguments has been published (Mattsson, 2014), and the arguments and counterarguments are summarized in what follows. First, de Winter (2013) made the strikingly strong argument that the results of Study I were “likely ... artifacts, caused by failing to recover stable factors”. The argument was supported by a simulation study (de Winter, Dodou, & Wieringa, 2009) according to which hundreds of more observations would have been needed in order to reliably estimate population parameters in a sample such as the one used in Study I. However, the simulation study was based on principal axis factor analyses of continuous, normally distributed variables, whereas in Study I, the input variables were categorical (and thus obviously non-normal), and a weightd least squares estimator with mean and variance correction (WLSMV), based on polychoric correlations, was used instead of principal axis factoring. Because of this, there are more appropriate simulation studies to cite in order to judge the adequacy of the sample sizes of Study I. For instance, Flora & Curran (2004) stated that when the WLSMV estimator is used with polychoric correlations, it produces accurate results even with the most skewed and curtotic variables (much like those used in Study I) with sample sizes exceeding 200, i.e. less than the within-group sample sizes in Study I. Mattsson (2014) cites also several other simulation studies that point in the same direction, supporting the conclusion that sample sizes in Study I were in fact quite adequate.

De Winter (2013) continued the critique of Study I by discussing the overextraction of factors. In other words, he claimed that in Study I, a factor model with too many factors was specified. To put this criticism into perspective, it is good to remember that the model specified the number of factors that the instrument was built to measure, and Study I showed that the four-factor model fit the data better than the two-factor model (errors and violations) advocated by de Winter (2013). Further, it is naturally possible to specify a factor model that combines these perspectives: if the two factors are specified as higher-order factors, the four factors discussed in Study I can be thought of as first-order factors (a proposal also made by Lajunen et al., 2004).

On a more theoretical level, Mattsson (2014) argued that as long as the latent variables are given a realist interpretation, i.e. as long as it is assumed that there exists a cognitive process that the latent variable reflects (something that has always been assumed at least implicitly in the DBQ literature), it would rather be desirable to extract a far larger number of latent variables: the world is, after all, a complicated place and several cognitive processes are likely to determine whether we commit an error in traffic or not, and the DBQ items could be grouped according to the cognitive processes employed when carrying out the behavior or committing the error described in the item. My understanding of these issues has since developed, and if I were to write the article today, I would probably refer to more complex models such as those combining a within-person information-processing model with an individual differences model that were briefly introduced in Section 4.3 and to issues such as ergodicity discussed in Section 4.1.

Finally, de Winter (2013) criticized Study I for unrealistic expectations when it demanded that strong factorial invariance should be demonstrated before comparing factor means. De Winter (2013) cites Meredith & Teresi (2006) as support for his statement that "arguing against comparing means of these subgroups of respondents is an unnecessarily skeptical standpoint." The argument is difficult to understand, since Meredith and Teresi (2006) clearly state that

> *If strong factorial invariance is observed, then one can legitimately compare groups in terms of factor and observed means; without strong factorial invariance, observed group differences will not correspond to differences in underlying factor means, but will be confounded by differences in item-specific intercepts, which are typically not of substantive importance.*
>
> *Meredith & Teresi (2006)*

The intercepts, for their part, are used for modelling differential effects of acquiescence bias, different response styles and social desirability bias etc. across groups. Because of this, stating that strong equivalence is too much to ask when comparing factor means amounts to having the opinion that it does not matter whether the means reflect the above-mentioned biases or the phenomenon of interest. This clearly cannot be the case, especially since within the DBQ research tradition, concrete suggestions related to traffic safety measures have often been made based on such differences in factor means.

Study III had its own limitations that ranged from pre-processing data to interpreting descriptive indices of network models. First, Study III treated missing data in a wasteful manner (listwise deletion). This may have created selective effects in the data. Still, more elegant methods of dealing with missing data, such as Full Information Maximum Likelihood have not yet been implemented within the network analysis framework (Epskamp, 2017). In addition, network models were not compared across driver groups: it is possible that systematic differences would have been found between groups of drivers, such as drivers of different ages – similarly to what was found in Studies I and II using the latent variable methodology. Comparing the structure of the network models across subgroups would amount to a test of measurement equivalence of sorts even though the network models are not measurement models in the traditional sense of the word. Further, it is naturally possible that individual drivers' personal network structures differ from other drivers' respective network structures; this question could, however, be investigated only by creating individual and group-level network models based on longitudinal data. When it comes to the Poisson regression models used in Study III, the suitability of the models was not assessed thoroughly: for instance, zero-inflated Poisson models might have been a more principled choice for modeling rare events such as the crashes[2].

The use of network centrality indices for identifying targets for traffic safety interventions in a data-driven manner is another theme in the data analysis of Study III that can be contested. This is because it is currently an open theme whether and to which extent the indices that have been developed for characterizing social networks are applicable to psychological networks (Bringmann et al., 2019). For instance, the betweenness index and the closeness index refer to flow of information between the nodes of the network, even though it is unclear whether it can sensibly be stated that something "flows" when a node related to a certain traffic behavior is connected to the node related to another behavior. On the other hand, the strength centrality index that was also reported in Study III does not suffer of this difficulty. Further, any given network model is compatible with several

[2] I would like to thank professor Reijo Sund, who acted as a pre-examiner of the present thesis, for this observation.

causal structures, and a recent caution (from a study reporting network models of PTSD) is relevant here, as well:

> *It is important to highlight that centrality does not automatically translate to clinical relevance and that highly central symptoms are not automatically viable intervention targets. Suppose a symptom is central because it is the causal endpoint for many pathways in the data: Intervening on such a product of a causal chain would not lead to any changes in the system.*
>
> *Fried et al. (2018)*

Other possibilities exist, as well: a central node can be involved in several feedback loops, and "switching it off" would perhaps not be a viable intervention target because the other nodes might "switch it back on again" easily (Fried et al., 2018). In Study III, the speeding-related node v5 might be a case in point: for instance, even if we succeeded in discouraging drivers from speeding while *racing from the traffic lights* (v10) still belongs to their behavioral repertoire, it might be that when the drivers end up in a traffic situation that encourages racing from the lights, they would still end up speeding. Further, certain behaviors are important targets for traffic safety interventions quite irrespective of the values of centrality indices in network models; for instance, even though *driving under the influence of drugs* (v12) is the least central node in the network model presented in Figure 9, it is certainly an important target for traffic safety interventions. Such issues related to the limitations of using centrality indices were not presented clearly enough in Study III.

On a more general level, all three studies were based on self-report questionnaires. This is a clear weakness of the present thesis, because based on the Studies I–III, common method bias cannot be teased apart from variance related to the phenomena of interest. For instance, the present results are in principle compatible with the possibility that traffic behavior is indeed determined by a small number of general psychological properties, such as the tendency to break rules or the tendency to commit errors. For instance, it is conceivable that people do indeed exceed speed limits and overtake on the inside because they have a tendency to break rules, even though the self-report questionnaire might not be a sufficiently sophisticated instrument for measuring such properties. If this is so, thorough open-ended interviews on the reasons that drivers give to their behavior, might, for instance, reveal the fact better than self-report questionnaires. Instead of returning to deep questions of measurement in psychology (see Section 4.3.) or considering the myriad of reasons for a given individual to break rules on a given occasion, two general points of view are offered to counter arguments related to method bias:

First, explanations that have the form "people behave in manner X because they have a tendency to do so" are blatantly circular, and do not increase our understanding of the behavior in the least. Because of this, the main point of the present thesis – that human behavior is complex and should be modeled as such – stands irrespective of self-report methodology. Second, because of the circularity of such explanations, it might be more fruitful to consider the psychological properties as taxonomical nametags rather than explanations. It appears, indeed, that this is what Reason (1990) does when using the GEMS concepts: they are a way of classifying factors that contribute to accidents rather than explanations; on the other hand, Reason (1990) does provide thorough explanations of accidents that did take place, such as the Challenger explosion and the Chernobyl accident. These descriptions of the accidents can be considered true explanations as they involve causal chains such as disregarding known problems with o-rings leading to them failing catastrophically in the Challenger explosion. The present thesis argues that such causal chains are what counts when explaining accidents, and that network analysis can prove to be a helpful tool when formulating hypotheses concerning them.

4.5 Open questions and future directions

The major novel question that the present thesis raises is: "What is the constructive contribution that psychometric network models can make in the field of traffic psychology?" It is an open question whether the major difference between network models and latent variable models is related to the causal determinants of traffic behaviors or to the networks' ability to function as natural representations of dynamic phenomena that happen in time and develop over a longer time scale (Bringmann & Eronen, 2018). Further, the network models are built based on pairwise relationships between nodes while statistically controlling for the effects of the remaining nodes. Is this a relevant starting point for modelling traffic behavior, or should the models include dependencies spanning a larger number of relationships? Could these interactions (pairwise or otherwise) be represented using some version of latent variable models as plausibly as with network models? Latent variable models and structural equation models are after all an enormously flexible technique for modelling multivariate data.

On the whole, the status of latent variables in network models of driving behavior is another future challenge to be tackled. For instance, will the drivers' tendency to violate rules and to commit errors still be seen as a relevant research topic in the future? If so, should the focus be on general tendencies (independent of driving or traffic as a context) or on driving-related tendencies? How should the latent variable view and the network view be integrated in the future? Should the tendencies be included as latent variables that influence the driving behaviors, or should they rather be seen as emergent properties of the patterns of interaction in networks of driver behavior?

Moreover, psychometric network models can only prove to be useful in traffic psychology and traffic safety work if they are constructed based on variables that are relevant for traffic safety. Section 4.2. referred to the variables reported in Table 8 forming a criterion-keyed version of the DBQ; those variables served as the best independent predictors of self-reported crashes. What if this idea were to be taken a bit further and network models of driving behavior constructed based on variables that are independently known to function as parts of causal chains leading to crashes? The network models reported in Study III were, after all, constructed based on self-report items that were originally included in the questionnaire because of their maximal intercorrelations (Reason et al., 1990). Such highly redundant items are surely not optimally enlightening if the objective is to uncover causal chains leading to crashes.

Indeed, some of the edges in the network models of traffic behavior were suggested as plausible causal hypotheses, while others were interpreted as indicating the presence of unmodelled latent variables. If the network view is adopted in future studies, it would seem beneficial to build the models based on a carefully selected set of variables with likely causal associations among them. Models of accident analysis, such as the Driving Reliability and Error Analysis Model (DREAM, Wallén Warner et al., 2008) could perhaps be used as inspiration for developing such self-report instruments. The DREAM was constructed based on in-depth analyses of serious crashes and it contains variables that have been identified as parts of causal chains leading to actual crashes. The DREAM contains background variables such as personality traits and stress that are used in explaining why drivers might choose an inappropriately high speed that then lead to them missing observing something etc and to eventually crashing. Network models based on self-report of such phenomena might prove interesting for traffic safety work.

The previous considerations naturally lead to considering the relevance of complex systems theories to network models of traffic behavior. Such theories are useful for understanding change over time across multiple time scales (Richardson, Dale & Marsh, 2014). Traffic behavior is complex in just this way: it involves a large number of interacting elements that operate on different time scales. For instance, the personalities of drivers are taken to influence the way they behave in traffic from one moment to another; still, the dynamics of change in personality traits take place on an extremely long time scale (years), while the moment-to-moment traffic behavior unfolds on a time scale expressed in seconds. Such differences in time scales and the granularity of variables would surely need to be explicitly taken into account when combining background variables such as personality traits into network models of traffic behavior.

Traditionally, research on traffic behavior has been based on different kinds aggregate statistics such as self-reported frequencies of behaving in a certain manner. While the network models reported in the present thesis are based on such aggregate data, they still comprise interacting parts that form an entity with potentially emergent properties – both hallmarks of complex dynamic systems. It would be of great interest to take these ideas further and to examine changes in traffic behaviors modelled as a complex system with interacting parts; this might prove fruitful for understanding the (in-)efficacy of traffic safety interventions (Nóvoa, Pérez & Borrell, 2009), for instance. Intensive longitudinal data would naturally be needed in such studies.

Another potential direction for future research would be investigating the influence that the state of the driver (tired, stressed, in a hurry etc.) has on the network of driving behaviors; it is conceivable that the network could occupy qualitatively different states of interconnectivity in each case. Similar methods could perhaps be used in assessing how novice drivers' driving behaviors interact and solidify into driving habits across a longer period of time. It is an open question whether such developmental studies can be carried out based on the DBQ but it might again be desirable to develop an instrument separately for this purpose in order to better represent the relevant causal connections.

Collecting intensive longitudinal data would make it possible to assess the similarity (or lack thereof) of within-person and between-person models of traffic behavior – something that the current thesis could only point out as an outstanding question. The observations related to ergodicity that were raised in Section 4.1. raise an interesting question: what if we were able to first construe person-specific models describing the dynamical relationships among psychological processes within individuals, and only then aggregate data from those persons whose person-specific models are similar in nature (Molenaar & Campbell, 2009)? This would enable sidestepping the difficult questions of ergodicity, homogeneity of the population and the stationarity of parameters.

4.6 Conclusions

The current thesis has provided a first example of using psychometric network analysis in the context of traffic behavior, and future studies can build on this basis both theoretically and methodologically. Understanding direct interactions among traffic behaviors, together with associated thoughts and emotions, provides a novel point of view to psychometrics of traffic behavior, and brings the field closer to other fields of traffic psychology, such as overarching theories of traffic behavior (Fuller, 2005; Summala, 2007), models of accident causation (Abdel-Aty & Radwan, 2000; Wallén Warner et al., 2008) and modern taxonomies of human error (Stanton & Salmon, 2009). Network analyses feed back to the subject matter theory by considering pairwise relationships between variables and the nature of cliques in the network: as a visual method, network analysis is able to summarize a large amount of information in a small space, which enables the researcher to consider where and how the associations between variables are in line with expectations.

In the presently reported network models, certain potential hubs – behaviors of central importance in terms of determining other behaviors – were suggested. Among these are speeding within or outside residential areas. The present thesis showed an example of how causal hypotheses concerning the assocations among traffic behaviors can be reached through network modelling. Further, the thesis suggests that the relationship between individual traffic behaviors and that of the overarching category of *violations* can be understood as one of emergence. Future studies can build on this basis by examining the dynamics of driving behaviors, such as how certain associations form and change. One direction for future studies would be to employ time-series data to examine the factors that affect changes of network state between the rule-abiding and the violation-prone state.

The present thesis as a whole offers important reminders to researchers wishing to operationalize theoretical concepts using self-report instruments. The history of the DBQ, as developed on the basis of GEMS, highlights the extreme difficulty of coming up with self-report items that would be unequivocally related to only the properties being measured. The extremely large number of different instruments, all going by the name of "the DBQ", encourages considering whether the DBQ should return to its roots: perhaps items should carefully be crafted such that all the important GEMS variables will be accounted for. On the other hand, the present thesis raised the question of whether it might be of more interest to construe self-report instruments that are more consistent with the network view, such as one based on variables that are independently known to affect crash risk. The starting point in network models is, after all, the idea of a *component*, i.e. a node having unique causal associations with the rest of the nodes. Because of this, variables that are known to be associated with crash risk might function as a more principled starting point than those that are psychometrically optimized to have maximal intercorrelations – and consequently low predictive power of their own.

The present thesis also included a method of modelling crashes based on the principles of statistical learning theory that include building and testing the predictive model in different subsets of data and testing model fit based on cross-validation. Further, regularized regression was used both for avoiding overfitting the model and for selecting the predictors. Traffic psychology would perhaps benefit of wider application of such methods and analysis procedures in predicting the occurrence of crashes.

Still, the most original contribution of the present thesis lies in the application of network models and interpreting traffic behaviors as a complex, interacting system. Adopting the network view was based on observing a central contradiction in the foundation of DBQ studies: the studies are motivated by arguing that latent variables – violation- and error-proneness that are obviously related to properties of individual drivers – can be studied using large cross-sectional samples of drivers from the whole population. The present thesis argues that the exercise is likely to fail due to problems related to ergodicity (Section 4.1.) and proposes that studies utilizing network models constructed based on intensive longitudinal data would be needed in order to examine within-person psychological processes such as violation- and error-proneness. Because the network models reported in the present thesis are not based on this kind of intensive longitudinal data, they can be interpreted only as suggesting a hopefully fruitful way forward for self-report-based studies in traffic psychology. Still, the models that were reported can, in the future, be compared with models based on such longitudinal data: it is in itself an interesting question whether and to what extent the within-person and between-person models correspond with each other when studying such a clearly bounded sphere of human life as traffic. I hope future studies build on this basis and examine whether traffic behavior can fruitfully be analysed as a complex system with all that this entails.

REFERENCES

Abdel-Aty, M. A., & Radwan, A. E. (2000). Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention, 32(5),* 633–642.

Aberg, L., & Rimmo, P. (1998). Dimensions of aberrant driver behaviour. *Ergonomics, 41(1),* 39–56.

af Wåhlberg, A., Dorn, L., & Kline, T. (2011). The Manchester Driver Behaviour Questionnaire as a predictor of road traffic accidents. *Theoretical Issues in Ergonomics Science, 12(1),* 66-86.

af Wåhlberg, A., Barraclough, P., & Freeman, J. (2015). The Driver Behaviour Questionnaire as accident predictor; A methodological re-meta-analysis. *Journal of Safety Research, 55,* 185-212.

Alver, Y., Demirel, M., & Mutlu, M. (2014). Interaction between socio-demographic characteristics: traffic rule violations and traffic crash history for young drivers. *Accident Analysis & Prevention, 72,* 95–104.

Asparouhov, T., & Muthén, B. (2009). Exploratory Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16(3),* 397–438.

Bianchi, A., & Summala, H. (2002). Moral judgment and drivers' behavior among Brazilian students. *Psychological Reports, 91(3),* 759–766.

Biederman, J., Fried, R., Hammerness, P., Surman, C., Mehler, B., Petty, C. R., . . . Meller, B. (2012). The effects of lisdexamfetamine dimesylate on driving behaviors in young adults with ADHD assessed with the Manchester driving behavior questionnaire. *Journal of Adolescent Health, 51(6),* 601–607.

Björklund, G. M. (2008). Driver irritation and aggressive behaviour. *Accident Analysis & Prevention, 40(3),* 1069–1077.

Blockey, P. N., & Hartley, L. R. (1995). Aberrant driving behaviour: errors and violations. *Ergonomics, 38(9),* 1759–1771.

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: causal indicators, composite indicators, and covariates. *Psychological Methods, 16(3),* 265–284.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110(2),* 305–314.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science, 323(5916),* 892–895.

Boring, E. G. (1923). Intelligence as the tests test it. *New Republic, 35(6),* 35–37.

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

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

Bringmann, L. F., & Eronen, M. I. (2018). Don't blame the model: Reconsidering the network approach to psychopathology. *Psychological Review, 125(4),* 606–615.

Bringmann, L. F., Elmer, T., Epskamp, S., Krause, R. W., Schoch, D., Wichers, M., Wigman, J.T. & Snippe, E. (2019). What do centrality measures measure in psychological networks?. J*ournal of Abnormal Psychology, 128(8),* 892–903.

Brown, J. (2009). Choosing the right type of rotation in PCA and EFA. *JALT Testing & Evaluation SIG Newsletter, 13(3),* 20–25.

Browne, M. W. (2001). An Overview of Analytic Rotation in Exploratory Factor Analysis. *Multivariate Behavioral Research, 36(1),* 111–150.

Chapman, B. P., Weiss, A., & Duberstein, P. R. (2016). Statistical learning theory for high dimensional prediction: Application to criterion-keyed scale development. *Psychological methods, 21(4),* 603–620.

Chu, W., Wu, C., Atombo, C., Zhang, H., & Özkan, T. (2019). Traffic climate, driver behaviour, and accidents involvement in China. *Accident Analysis & Prevention, 122,* 119–126.

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., & Cramer, A. O. (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.

Costantini, G., & Perugini, M. (2014). Generalization of clustering coefficients to signed correlation networks. *PloS One, 9(2),* e88669.

Costantini, G., & Perugini, M. (2018). A framework for testing causality in personality research. *European Journal of Personality, 32(3),* 254–268.

Cramer, A. O., Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., . . . Borsboom, D. (2012a). 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(4),* 414–431.

Cramer, A. O., Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., . . . Borsboom, D. (2012b). Measurable like temperature or mereological like flocking? On the nature of personality traits. *European Journal of Personality, 26(4),* 451–459.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12(11),* 671–684.

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(1),* 2–22.

de Winter, J. (2013). Small sample sizes, overextraction, and unrealistic expectations: A commentary on M. Mattsson. *Accident Analysis & Prevention, 50,* 776–777.

de Winter, J., & Dodou, D. (2016). Common factor analysis versus principal component analysis: a comparison of loadings by means of simulations. *Communications in Statistics-Simulation and Computation, 45(1),* 299–321.

de Winter, J., Dodou, D., & Stanton, N. (2015). A quarter of a century of the DBQ: some supplementary notes on its validity with regard to accidents. *Ergonomics, 58(10),* 1745–1769.

de Winter, J., & Dodou, D. (2010). The Driver Behaviour Questionnaire as a predictor of accidents: A meta-analysis. *Journal of Safety Research, 41(6),* 463–470.

de Winter, J., Dodou, D., & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research, 44(2),* 147–181.

Dimmer, A., & Parker, D. (1999). The accidents, attitude and behaviour of company car drivers. In Traffic Research Laboratory (Ed.), *Behavioural Research in Road Safety IX*. Crowthorne, Berkshire: Traffic Research Laboratory.

Epskamp, S. (2017). *Network psychometrics*. (Unpublished Doctoral thesis). University of Amsterdam. Retrieved from http://sachaepskamp.com/dissertation/

Epskamp, S., van Borkulo, C. D., van der Veen, D. C., Servaas, M. N., Isvoranu, A. M., Riese, H., & Cramer, A. O. (2018). Personalized network modeling in psychopathology: The importance of contemporaneous and temporal connections. *Clinical Psychological Science, 6(3),* 416–427.

Epskamp, S., Borsboom, D., & Fried, E. I. (2017). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods, 50(1),* 1–18.

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

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

Eugenia Gras, M., Sullman, M. J. M., Cunill, M., Planes, M., Aymerich, M., & Font-Mayolas, S. (2006). Spanish drivers and their aberrant driving behaviours. *Transportation Research Part F: Traffic Psychology and Behaviour, 9(2),* 129–137.

Eysenck, H. (1973). *The measurement of intelligence*. Lancaster: MTP.

Fayers, P. M., & Hand, D. J. (2002). Causal variables, indicator variables and measurement scales: an example from quality of life. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 165(2),* 233–253.

Flora, D. B., & Curran, P. J. (2004). An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis With Ordinal Data. *Psychological Methods, 9(4),* 466–491.

Foygel, R., & Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. *Advances in Neural Information Processing Systems, 23,* 604–612.

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

Fried, E. I., Eidhof, M. B., Palic, S., Costantini, G., Huisman-van Dijk, H. M., Bockting, C. L., Engelhard, I., Armour, C., Nielsen, A.B., & Karstoft, K. I. (2018). Replicability and generalizability of posttraumatic stress disorder (PTSD) networks: A cross-cultural multisite study of PTSD symptoms in four trauma patient samples. *Clinical Psychological Science, 6(3),* 335–351.

Fried, E. I., Epskamp, S., Nesse, R. M., Tuerlinckx, F., & Borsboom, D. (2016). What are 'good' depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. *Journal of Affective Disorders, 189,* 314–320.

Fried, E. I., van Borkulo, C. D., Cramer, A. O., Boschloo, L., Schoevers, R. A., & Borsboom, D. (2017). Mental disorders as networks of problems: a review of recent insights. *Social Psychiatry and Psychiatric Epidemiology, 52(1),* 1–10.

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

Fuller, R. (2005). Towards a general theory of driver behaviour. *Accident Analysis & Prevention, 37(3),* 461–472.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care, 44(11 Suppl 3),* S78–S94.

Grice, J. W. (2014). Observation oriented modeling: preparing students for research in the 21st century. *Comprehensive Psychology, 3,* Article 3.

Grice, J. W., Barrett, P. T., Schlimgen, L. A., & Abramson, C. I. (2012). Toward a brighter future for psychology as an observation oriented science. *Behavioral Sciences, 2(1),* 1–22.

Gunzler, D. D., & Morris, N. (2015). A tutorial on structural equation modeling for analysis of overlapping symptoms in co-occurring conditions using MPlus. *Statistics in Medicine, 34(24),* 3246–3280.

Hanfstingl, B. (2019). Should we say goodbye to latent constructs to overcome replication crisis or should we take into account epistemological considerations? *Frontiers in Psychology, 10,* Article 1949.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. New York, NY: Springer.

Hauben, M., Hung, E., & Hsieh, W. (2017). An exploratory factor analysis of the spontaneous reporting of severe cutaneous adverse reactions. *Therapeutic Advances in Drug Safety, 8(1),* 4–16.

Holme, P. (2003). Congestion and centrality in traffic flow on complex networks. *Advances in Complex Systems, 6(2),* 163–176.

Howell, R. D. (2008). Observed variables are indeed more mysterious than commonly supposed. *Measurement: Interdisciplinary Research & Perspective, 6(1–2),* 97–101.

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

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: The Guilford Press.

Kontogiannis, T., Kossiavelou, Z., & Marmaras, N. (2002). Self-reports of aberrant behaviour on the roads: errors and violations in a sample of Greek drivers. *Accident Analysis & Prevention, 34(3),* 381–399.

Koppel, S., Stephens, A., Charlton, J., Di Stefano, M., Darzins, P., Odell, M., & Marshall, S. (2018). The Driver Behaviour Questionnaire for older drivers: Do errors, violations and lapses change over time? *Accident Analysis & Prevention, 113,* 171–178.

Koschützki, D., Lehmann, K. A., Peeters, L., Richter, S., Tenfelde-Podehl, D., & Zlotowski, O. (2005). Centrality indices. Lecture Notes in Computer

Science, vol 3418. In U. Brandes, & T. Erlebach (Eds.), *Network analysis* (pp. 16–61). Berlin & Heidelberg: Springer.

Kossakowski, J. J., Epskamp, S., Kieffer, J. M., van Borkulo, C. D., Rhemtulla, M., & Borsboom, D. (2016). The application of a network approach to Health-Related Quality of Life (HRQoL): introducing a new method for assessing HRQoL in healthy adults and cancer patients. *Quality of Life Research, 25(4),* 781–792.

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed.). Chicago and London: University of Chicago press.

Kuismin, M. O., & Sillanpää, M. J. (2017). Estimation of covariance and precision matrix, network structure, and a view toward systems biology. *Wiley Interdisciplinary Reviews: Computational Statistics, 9(6),* e1415.

Lajunen, T., & Summala, H. (2004). *The effects of young drivers' life-style and values on traffic safety attitudes, driving behavior and accident involvement* (LINTU Research Programme). Helsinki: Ministry of Transport and Communications.

Lajunen, T., Parker, D., & Summala, H. (2004). The Manchester Driver Behaviour Questionnaire: A cross-cultural study. *Accident Analysis and Prevention, 36(2),* 231–238.

Lamiell, J. T. (2003). *Beyond individual and group differences: Human individuality, scientific psychology, and William Stern's critical personalism*. Thousand Oaks, CA: Sage.

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

Lawton, R., Parker, D., Manstead, A. S., & Stradling, S. G. (1997). The role of affect in predicting social behaviors: The case of road traffic violations. *Journal of Applied Social Psychology, 27(14),* 1258–1276.

Luce, R. D., & Steingrimsson, R. (2011). Theory and tests of the conjoint commutativity axiom for additive conjoint measurement. *Journal of Mathematical Psychology, 55(5),* 379–385.

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

Lucidi, F., Girelli, L., Chirico, A., Alivernini, F., Cozzolino, M., Violani, C., & Mallia, L. (2019). Personality traits and attitudes toward traffic safety predict risky behavior across young, adult, and older drivers. *Frontiers in Psychology, 10,* 536.

Lykken, D. T. (1991). What's wrong with psychology anyway. *Thinking Clearly about Psychology, 1,* 3–39.

Maasalo I., Lehtonen E., & Summala H. (2017). Young females at risk while driving with a small child. *Accident Analysis & Prevention, 108,* 321–331.

Martinussen, L. M., Hakamies-Blomqvist, L., Møller, M., Özkan, T., & Lajunen, T. (2013). Age, gender, mileage and the DBQ: The validity of the Driver Behavior Questionnaire in different driver groups. *Accident Analysis & Prevention, 52,* 228–236.

Maslać, M., Antić, B., Lipovac, K., Pešić, D., & Milutinović, N. (2018). Behaviours of drivers in Serbia: Non-professional versus professional drivers. *Transportation Research Part F: Traffic Psychology and Behaviour, 52,* 101–111.

Mason, O., & Verwoerd, M. (2007). Graph theory and networks in biology. *IET Systems Biology, 1(2),* 89–119.

Mattsson, M., Lajunen, T., Gormley, M., & Summala, H. (2015). Measurement invariance of the Driver Behavior Questionnaire across samples of young drivers from Finland and Ireland. *Accident Analysis & Prevention, 78,* 185–200.

Mattsson, M. (2012). Investigating the factorial invariance of the 28-item DBQ across genders and age groups: An Exploratory Structural Equation Modeling Study. *Accident Analysis & Prevention, 48,* 379–396.

Mattsson, M. (2014). On testing factorial invariance: A reply to J.C.F. de Winter. *Accident Analysis & Prevention, 63,* 89–93.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44(11 Suppl 3),* S69–S77.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58(4),* 525–543.

Mesken, J., Lajunen, T., & Summala, H. (2002). Interpersonal violations, speeding violations and their relation to accident involvement in Finland. *Ergonomics, 45(7),* 469–483.

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

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge, United Kingdom: Cambridge University Press.

Michell, J. (2008). Is psychometrics pathological science? *Measurement, 6(1–2),* 7–24.

Mischel, W., & Shoda, Y. (1998). Reconciling processing dynamics and personality dispositions. *Annual review of psychology, 49(1),* 229–258.

Molenaar, P., & Campbell, C. (2009). The new person-specific paradigm in psychology. *Current directions in psychological science*, 18(2), 112-117.

Muthén, B., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus Web Notes, 4(5),* 1–22.

National Advisory Board on Research Ethics. (2009). Ethical principles of research in the humanities and social and behavioural sciences and proposals for ethical review, 26.12.2019. Retrieved from http://www.tenk.fi/sites/tenk.fi/files/ethicalprinciples.pdf

Newman, M. (2008). The physics of networks. *Physics Today, 61(11),* 33–38.

Nóvoa, A., Pérez, K., & Borrell, C. (2009). *Evidence-based effectiveness of road safety interventions: a literature review*. Gaceta sanitaria, 23(6), 553-e1.

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks, 32(3),* 245–251.

Öz, B., Özkan, T., & Lajunen, T. (2010). An investigation of the relationship between organizational climate and professional drivers' driver behaviours. *Safety Science, 48(10),* 1484–1489.

Özkan, T., & Lajunen, T. (2005). A new addition to DBQ: Positive driver behaviours scale. *Transportation Research Part F: Traffic Psychology and Behaviour, 8(4–5),* 355–368.

Özkan, T., Lajunen, T., Chliaoutakis, J. E., Parker, D., & Summala, H. (2006). Cross-cultural differences in driving behaviours: A comparison of six countries. *Transportation Research Part F: Traffic Psychology and Behaviour, 9(3),* 227–242.

Özkan, T., Lajunen, T., & Summala, H. (2006). Driver Behaviour Questionnaire: A follow-up study. *Accident Analysis & Prevention, 38(2),* 386–395.

Parker, D., Lajunen, T., & Stradling, S. (1998). Attitudinal predictors of interpersonally aggressive violations on the road. *Transportation Research Part F: Traffic Psychology and Behaviour, 1(1),* 11–24.

Parker D., Manstead A., Stradling S., Reason J., & Baxter J. (1992). Intention to commit driving violations: an application of the theory of planned behavior. *Journal of Applied Psychology, 77(1),* 94–101.

Parker, D., Reason, J., Manstead, A., & Stradling, S. (1995). Driving errors, driving violations and accident involvement. *Ergonomics, 38(5),* 1036–1048.

Parker, D., McDonald, L., Rabbitt, P., & Sutcliffe, P. (2000). Elderly drivers and their accidents: the Aging Driver Questionnaire. *Accident Analysis & Prevention, 32(6),* 751–759.

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association, 79(387),* 575–583.

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences, 2(1),* 13–43.

Reason, J. (1990). *Human Error*. Cambridge: Cambridge University Press.

Reason, J., Manstead, A., Stradling, S., Baxter, J., & Campbell, K. (1990). Errors and violations on the roads: a real distinction? *Ergonomics, 33(10),* 1315–1332.

Richardson, M., Dale, R., & Marsh, K. (2014). *Complex dynamical systems in social and personality psychology: Theory, modeling, and analysis*. In Reis, H. & Judd, C. (Eds.), Handbook of research methods in social and personality psychology (p. 253–282). Cambridge: Cambridge University Press.

Rimmö, P. (2002). Aberrant driving behaviour: homogeneity of a four-factor structure in samples differing in age and gender. *Ergonomics, 45(8),* 569–582.

Robinaugh, D. J., Hoekstra, R. H., Toner, E. R., & Borsboom, D. (2019). The network approach to psychopathology: a review of the literature 2008–2018 and an agenda for future research. *Psychological Medicine, 50(3),* 353–366.

Roman, G. D., Poulter, D., Barker, E., McKenna, F. P., & Rowe, R. (2015). Novice drivers' individual trajectories of driver behavior over the first three years of driving. *Accident Analysis & Prevention, 82,* 61–69.

Sakashita, C., Senserrick, T., Lo, S., Boufous, S., de Rome, L., & Ivers, R. (2014). The Motorcycle Rider Behavior Questionnaire: Psychometric properties and application amongst novice riders in Australia.

Transportation Research Part F: Traffic Psychology and Behaviour, 22, 126–139.

Saramäki, J., Kivelä, M., Onnela, J., Kaski, K., & Kertesz, J. (2007). Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E, 75(2),* 027105.

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology, 4,* 1–32.

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

Spano, G., Caffò, A. O., Lopez, A., Mallia, L., Gormley, M., Innamorati, M., … Bosco, A. (2019). Validating driver behavior and attitude measure for older Italian drivers and investigating their link to rare collision events. *Frontiers in Psychology, 10,* Article 368.

Stanojević, P., Lajunen, T., Jovanović, D., Sârbescu, P., & Kostadinov, S. (2018). The driver behaviour questionnaire in South-East Europe countries: Bulgaria, Romania and Serbia. *Transportation Research Part F: Traffic Psychology and Behaviour, 53,* 24–33.

Stanton, N. A., & Salmon, P. M. (2009). Human error taxonomies applied to driving: A generic driver error taxonomy and its implications for intelligent transport systems. *Safety Science, 47(2),* 227–237.

Stephens, A., & Fitzharris, M. (2016). Validation of the Driver Behaviour Questionnaire in a representative sample of drivers in Australia. *Accident Analysis & Prevention, 86,* 186–198.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103(2684),* 677–680.

Sullman, M. J., Meadows, M. L., & Pajo, K. B. (2002). Aberrant driving behaviours amongst New Zealand truck drivers. *Transportation Research Part F: Traffic Psychology and Behaviour, 5(3),* 217–232.

Sullman, M. J., Stephens, A. N., & Taylor, J. E. (2019). Dimensions of aberrant driving behaviour and their relation to crash involvement for drivers in New Zealand. *Transportation Research Part F: Traffic Psychology and Behaviour, 66,* 111–121.

Sümer, N. (2003). Personality and behavioral predictors of traffic accidents: testing a contextual mediated model. *Accident Analysis & Prevention, 35(6),* 949–964.

Summala, H. (2007). Towards Understanding Motivational and Emotional Factors in Driver Behaviour: Comfort Through Satisficing. In P. C. Cacciabue (Ed.), *Modelling Driver Behaviour in Automotive Environments* (pp. 189–207). London: Springer.

Tal, E. (2017). Measurement in Science. In Zalta, E.N. (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2017 ed.) Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/fall2017/entries/measurement-science/

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

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

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(2),* 339–356.

Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology, 60,* 58–71.

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In Goffin, R. D. & Helmes, E. (Eds.), *Problems and solutions in human assessment* (pp. 41–71). New York: Springer.

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some further observations. *Multivariate Behavioral Research, 25(1),* 97–114.

Wallén Warner, H., Ljung, M., Sandin, J., Johansson, E., & Björklund, G. (2008). *Manual for DREAM 3.0, Driving Reliability and Error Analysis Method. Deliverable D5.6 of the EU FP6 Project SafetyNet, TREN-04-FP6TRSI2.395465/506723.*

Wallén Warner, H. & Åberg L. (2006). Drivers' decision to speed: a study inspired by the theory of planned behavior. *Transportation Research Part F: Traffic Psychology and Behaviour, 9(6),* 427–433.

Wallén Warner, H., Özkan, T., Lajunen, T., & Tzamalouka, G. (2011). Cross-cultural comparison of drivers' tendency to commit different aberrant driving behaviours. *Transportation Research Part F: Traffic Psychology and Behaviour, 14(5),* 390–399.

Wells, P., Tong, S., Grayson, G., & Jones, E. (2008). *Cohort II – A study of learner and new drivers. Volume 1 – main report. Volume 2 – questionnaires and data tables.* London: Department for Transport.

Westerman, S., & Haigney, D. (2000). Individual differences in driver stress, error and violation. *Personality and Individual Differences, 29(5),* 981–998.

Williams, D. R., & Rast, P. (2020). Back to the basics: Rethinking partial correlation network methodology. *British Journal of Mathematical and Statistical Psychology, 73(2),* 187–212.

Xie, C., & Parker, D. (2002). A social psychological approach to driving violations in two Chinese cities. *Transportation Research Part F: Traffic Psychology and Behaviour, 5(4),* 293–308.

You, T., Fiedor, P., & Hołda, A. (2015). Network analysis of the Shanghai stock exchange based on partial mutual information. *Journal of Risk and Financial Management, 8(2),* 266–284.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2),* 301–320.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99(3),* 432–442.