KU LEUVEN

Hierarchical Models and Bayesian Structural Equation Models With Application in Health Services Research

Benedict Owino ORINDI

Promoter:	Prof. dr Emmanuel Lesaffre
Co-promoter:	Prof. dr Walter Sermeus
Co-promoter:	Dr Luk Bruyneel

KU Leuven
Biomedical Sciences Group
Faculty of Medicine
Department of Public Health and Primary Care
Leuven Biostatistics and Statistical Bioinformatics Centre

Hierarchical Models and Bayesian Structural Equation Models With Application in Health Services Research

Benedict Owino ORINDI

Jury:

Promoter:	Prof. dr Emmanuel Lesaffre	(KU Leuven & Universiteit Hasselt)
Co-promoter:	Prof. dr Walter Sermeus	(KU Leuven)
Co-promoter:	Dr Luk Bruyneel	(KU Leuven & Mutualités Libres)
Chair:	Prof. dr Geert Verbeke	(KU Leuven & Universiteit Hasselt)
Exam cttee:	Prof. dr Bart Meuleman	(KU Leuven)
Exam cttee:	Prof. dr Geert Molenberghs	(Universiteit Hasselt & KU Leuven)
Exam cttee:	Prof. dr Martin McKee	(London School of Hygiene & Tropical Medicine)
Exam cttee:	Prof. dr Sophie Vanbelle	(Maastricht University)

September 2022

Dedication

To Christian, Patricia, Syprose and Pauline

Acknowledgements

I would like to thank my promoter Prof. dr Emmanuel Lesaffre, and co-promoters Prof. dr Walter Sermeus and Dr Luk Bruyneel for first accepting me as their PhD student and for their supervision, advise and support during this research work. Professor Lesaffre has provided wonderful guidance, mentorship and extraordinary experiences throughout the doctoral period. Thanks Professor for introducing me to the Bayesian world, first during the masters course at *Universiteit Hasselt* (UHasselt), and later during this PhD. Professor Sermeus helped me appreciate and put in a better way how the work covered in this PhD fits into or impacts cross-cultural (cross-country) and multi-language situation(s). Dr Bruyneel helped me learn the nursing language and, therefore, contextualise the clinical research questions answered in this thesis. Together with Emmanuel and Luk, we had several meetings and calls. Indeed, this PhD study is a product of the continued multidisciplinary collaboration between the Centre for Health Services of KU Leuven and Nursing Research and the Leuven Biostatistics and Statistical Bioinformatics Centre.

May I also thank Prof. dr Geert Molenberghs and Prof. dr Bart Meuleman – my thesis advisory committee – for their helpful feedback during the research seminars and on the thesis text. Thanks to the external examination committee members, Prof. dr Martin McKee and Prof. dr Sophie Vanbelle for their helpful comments during the review of this thesis text. Much appreciation to Prof. dr Geert Verbeke for accepting to chair the public defense. To all of you, thanks for creating time for me amid your busy schedules.

I am grateful to Kirsten Verhaegen, Sophie Collart, Eamon Pease and Annemie Janssens for support with administrative matters at the department

and at the doctoral school.

Many thanks to colleagues at Erasmus MC, the Netherlands, for hosting me during the early stages of this PhD work. Dr Baoyue Li introduced me to some of the tricks of trade. Drs Eleni-Rosalina Andrinopoulou (aka Elrozy), Susan Bryan and Nicole Erler – thanks for the warm evenings and weekends in the cold weather in Rotterdam.

Special thanks to my parents and siblings for their support, encouragement and for their important roles in many ways over the years. Finally, I apologize to all other unnamed who helped me in various aspects to have this good training.

Benedict Owino Orindi
Leuven, Belgium

List of Publications

This thesis text is based on the following publications.

1. **Orindi, B.O.**, Lesaffre, E., Sermeus, W., and Bruyneel, L.(2017). Impact of cross-level measurement noninvariance on hospital rankings based on patient experiences with care in 7 European countries. *Medical Care*, **55**(12). p e150-e157.doi: 10.1097/MLR.0000000000000580.

2. **Orindi, B.O.**, Lesaffre, E., Quintero, L.A., Sermeus, W., and Bruyneel, L.(2019). Contribution of HCAHPS specific care experiences to global ratings varies across 7 countries: what can be learned for reporting these global ratings? *Medical Care*, **57**(11), e65-e72. doi: 10.1097/MLR.0000000 000001077.

3. **Orindi, B.O.**, Ziraba, A, Bruyneel, L., Floyd, S., and Lesaffre, E.(2021). Invariance of the WHO violence against women instrument among Kenyan adolescent girls and young women: Bayesian psychometric modeling. *PLoS ONE*, **16**(10), e0258651. doi: https: 10.1371/journal.pone.0258651.

4. **Orindi, B.O.**, E., Quintero, A., Bruyneel, L., Li, B., and Lesaffre, E. A combined multilevel factor analysis and covariance regression model with mixed effects in the mean and variance structure. *Statistics in Medicine* (Under review).

Contents

Dedication **iii**

Acknowledgements **v**

List of Publications **vii**

Summary - Samenvatting **xv**

List of Abbreviations **xix**

1 General Introduction **1**

1.1 Introduction 1

1.2 Thesis Contribution 4

1.3 Organization of the Thesis 5

2 Motivating Data Sets **7**

2.1 The Registered Nurse Forecasting Study 7

2.2 The DREAMS Impact Evaluation Study 8

2.3 Research Questions 9

3 Measurement Invariance Concepts and Inference **11**

3.1 Observed and Latent Variables 12

3.2 Factor Analytic Models 12

3.2.1 Exploratory and Confirmatory Factor Analysis 13

3.2.2 Multilevel Factor Analysis Model 15

3.2.3 Higher-Order Factor Analysis Model 16

3.3 Structural Equation Models 17
3.4 Identification of Latent Variable Models 18
3.5 Measurement Invariance 19
3.5.1 Definitions 19
3.5.2 Testing invariance 22
3.5.3 Causes of Measurement Noninvariance 24
3.6 The Frequentist and Bayesian Approaches 24
3.6.1 Markov chain Monte Carlo Sampling 27
3.6.2 Software 29

4 Contribution of HCAHPS Specific Care Experiences to Global Ratings Varies Across 7 Countries: What Can be Learned for Reporting These Global Ratings? 31
4.1 Introduction 32
4.2 Methods 35
4.2.1 Study design, setting and sample 35
4.2.2 Measures 35
4.2.3 Analysis 41
4.2.4 Model Evaluation 43
4.2.5 Computational Aspects 44
4.3 Results 45
4.3.1 Descriptive Findings 45
4.3.2 Association between Specific Patient Experiences and Global Ratings? 46
4.3.3 Generalizability across Countries of Contribution of Specific Patient Experiences to Global Ratings 49
4.3.4 Summary of model fit and predictive power 52
4.3.5 Do cut-off point(s) allow for an equitable comparison of patient hospital ratings across countries? 55
4.4 Discussion 57
4.5 Conclusions 60

5 Invariance of the WHO VAW Instrument Among Kenyan Adolescent Girls and Young Women: Bayesian Psychometric Modeling 61
5.1 Introduction . . . 62
5.2 Methods . . . 64
5.2.1 Study design, setting and sample . . . 64
5.2.2 Measures . . . 65
5.2.3 Statistical Methods . . . 67
5.3 Results . . . 75
5.3.1 Descriptive Findings . . . 75
5.3.2 Factorial structure of the WHO VAW instrument for the Kenyan AGYW population . . . 77
5.3.3 Measurement invariance of the WHO VAW instrument across groups of AGYW . . . 79
5.4 Discussion . . . 86
5.5 Concluding Remarks . . . 90

6 Impact of Cross-level Measurement Noninvariance on Hospital Rankings Based on Patient Experiences with Care 91
6.1 Introduction . . . 92
6.2 Methods . . . 95
6.2.1 Data sources . . . 95
6.2.2 Measures . . . 96
6.2.3 Statistical Analysis . . . 98
6.3 Software Implementation . . . 102
6.4 Results . . . 103
6.4.1 Multilevel Factor Structure and Cross-level Measurement Noninvariance of the HCAHPS . . . 103
6.4.2 Cross-group Measurement Invariance at Multiple Levels of the HCAHPS . . . 107
6.4.3 Modelling Cross-level Noninvariance Makes a Difference in Performance Assessment . . . 110
6.5 Discussion . . . 113

6.6 Concluding Remarks 117

7 A Combined Multilevel Factor Analysis and Covariance Regression Model with Mixed Effects in the Mean and Variance Structure 119

7.1 Introduction 120
7.2 The RN4CAST Study: Patient Satisfaction Data 121
7.2.1 Research Objectives 124
7.3 Proposed Model 125
7.3.1 Multilevel Factor Analysis (MFA) Model 125
7.3.2 Multilevel Covariance Regression (MCR) Model 126
7.3.3 Combined Multilevel Factor Analysis and Covariance Regression Model (CMFA-CRM) 127
7.3.4 Identification Issues 128
7.4 Computational Details 130
7.5 Simulation Study 131
7.5.1 Set up of the Simulation Study 131
7.5.2 Simulation Results 132
7.6 Application to the RN4CAST Patient Satisfaction Data 134
7.6.1 Transformation of the Response Variables and Covariates 134
7.6.2 A Multilevel Exploratory Factor Analysis 135
7.6.3 CMFA-CRM Analysis 136
7.6.4 Comparison with the Multi-stage Analysis 141
7.6.5 Comparison with the MCR model Analysis Using Sum Scores 141
7.6.6 Model Assessment 143
7.6.7 Clinical Interpretation 143
7.7 Concluding Remarks 145

8 General Conclusions and Further Research 147

8.1 What is Covered in this Thesis 147
8.2 Further Research 152

Appendices **157**

A Mathematical Details **159**

A.1 Default Priors in Mplus 159

Bibliography **161**

Summary - Samenvatting

English

Health Services Research applications often involve multivariate data which also have inherent complex hierarchical structures. Examples are patients clustered in medical centers, multiple clinical treatments in a patient , and so on but also in other research areas there are ample examples to find, such as children in schools, teeth clustered in mouths, teats clustered in udders, etc. Ignoring this correlation among observations in a cluster may lead to incorrect estimation of parameters of interest because of ignoring the correlation structure in the data. As a consequence the conclusions may be wrong and therefore misleading. The increasing complexity in study designs and research questions add to the need of novel statistical tools that fully exploit the nature of the data and adequately answer the research questions. In this thesis, we exploited the correlated structure of the data and made use of existing statistical procedures, and when necessary, have extended them.

We have two aims. First, to advance the knowledge of the relationship between organizational features of nursing care and patient outcomes by making use of Bayesian multilevel structural equation models (SEMs) that reflect the complex hierarchical structure of the motivating data. Second, we aim to critically evaluate the performance of existing Bayesian SEM methods. In particular, we aim to explore the performance of joint multilevel factor analytic models and multilevel covariance regression models.

We analyse the European Union's three-year (2009-2011) Registered Nurse Forecasting (RN4CAST) study data (Sermeus et al., 2011) and data from a

three-year (2017-2019) epidemiological study of adolescent girls and young women in Nairobi slums (Birdthistle et al., 2018). We also carry out a simulation study to assess the merit of the proposed method. We implement the models in Mplus, SAS and in JAGS running interactively from R. The relevant code can be found online as supplementary material of the published papers or from the first author upon request.

The approaches considered here are motivated in the context of Health Services Research, but they are relevant to a wide variety of research areas.

Flemish

Onderzoek in Health Services heeft vaak te maken met multivariate gegevens die een inherente complexe hiërarchische, ook genoemd multilevel, structuur hebben. We denken dan aan patiënten opgenomen in ziekenhuizen, meerdere klinische ingrepen bij een patiënt, enz. maar ook in andere onderzoeksdomeinen is het makkelijk om voorbeelden te vinden, bvb. kinderen in scholen, spenen in uiers, of tanden in monden. Het negeren van de correlatie tussen observaties kan leiden tot een foute schatting van de model parameters en daarom leiden tot foute en misleidende conclusies. De toenemende complexiteit in studieopzetten en onderzoeksvragen draagt bij aan de behoefte aan nieuwe statistische methoden die toelaten de aard van de gegevens ten volle te benutten en de onderzoeksvragen adequaat te beantwoorden. In deze thesis hebben we de correlaties tussen de gegevens benut en hebben gebruik gemaakt van bestaande statistische procedures en waar nodig uitbreidingen voorgesteld.

De doelstelling van dit onderzoek is tweeledig. Ten eerste beogen we een hernieuwd inzicht te bieden in de associatie tussen organisatorische kenmerken van verpleegkundige zorg en patiëntuitkomsten door gebruik te maken van Bayesiaanse multilevel structurele vergelijkingsmodellen (SEM) die de complexe hiërarchische structuur van de gehanteerde gegevens weerspiegelen. Ten tweede willen we de karakteristieken van de huidige Bayesiaanse SEM-methoden kritisch evalueren, met onder meer modellen die een multilevel model combineren met een factoranalyse model, maar ook van een covariantieregressiemodel dat

gecombineerd wordt met een multilevel model.

Binnen dit doctoraat analyseren we hoofdzakelijk gegevens die werden verzameld in het kader van de Registered Nurse Forecasting (RN4CAST) studie die liep van 2009 tot en met 2011 (Sermeus et al., 2011). Daarnaast maken we ook gebruik van gegevens die werden verzameld binnen een epidemiologische onderzoek naar adolescente meisjes en jonge vrouwen in de sloppenwijken van Nairobi (Birdthistle et al., 2018). We voeren een simulatiestudie uit om de waarde van de voorgestelde methodes te beoordelen. De modellen werden geïmplementeerd in Mplus, SAS en in JAGS vanuit het pakket R. De relevante syntax is online beschikbaar als aanvullend materiaal bij de gepubliceerde artikelen maar kan ook bekomen worden van de eerste auteur.

Tot slot is het belangrijk op te merken dat de beschouwde statistische technieken niet enkel relevant zijn voor Health Services Research, maar voor een breed scala aan onderzoeksdomeinen.

List of Abbreviations

AGYW	Adolescent Girls and Young Women
AIC	Akaike Information Criteria
BSEM	Bayesian Structural Equation Model
BOS	Bounded outcome scores
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CMFA-CRM	Combined Multilevel Factor Analysis and Covariance Regression Model
CTS	Conflict Tactics Scale
DIC	Deviance Information Criteria
DIF	Differential Item Functioning
DREAMS	Determined, Resilient, Empowered, AIDS-free, Mentored, and Safe women
EFA	Exploratory Factor Analysis
ERT	Extreme Response Tendency
ESEM	Exploratory Structural Equation Model
FA	Factor Analysis
FP7	Seventh Framework Programme
HCAHPS	Hospital Consumer Assessments of Healthcare Providers and Systems
HRS	Health Services Research
ICC	Intraclass Correlation Coefficient
JAGS	Just Another Gibbs Sampler

LRT	Likelihood Ratio Test
MCFA	Multilevel Confirmatory Factor Analysis
MCMC	Markov chain Monte Carlo
MCR	Multilevel Covariance Regression
MFA	Multilevel Factor Analysis
MGCFA	Multiple-Group Confirmatory Factor Analysis
MI	Measurement Invariance
MIMIC	Multiple Indicator Multiple Causes
ML-SEM	Multilevel Structural Equation Model
PCA	Principle Component Analysis
PEPFAR	US President's Emergency Plan for AIDS Relief
PPCs	Posterior Predictive Checks
PPP	Posterior Predictive P-value
PSRF	Potential Scale Reduction Factor
RMSEA	Root Mean Square Error of Approximation
RN4CAST	Registered Nurse Forecasting
SEMs	Structural Equation Models
TLI	Tusker-Lewis Index
US	United States
VAW	Violence Against Women
WHO	World Health Organization
WRMR	Weighted Root-Mean-Square Residual

Chapter 1

General Introduction

In this chapter, we present a general introduction to the subject of this thesis. We first take a look at health services research with a focus on patient experience with care, and highlight the need for statistical techniques used in this thesis. Then we describe the thesis objectives. We end the chapter by outlining the organization of this thesis text.

1.1 Introduction

Health services research is the multidisciplinary field of scientific investigation that studies how social factors, political factors, financial systems, organizational structures and processes, health technologies, personal behaviour and psychology affect access to health care, the quality and cost of health care and, ultimately, the health and wellbeing of citizens (see e.g., Robertson and Williams, 2009). Nursing outcomes research is part of health services research and has been defined by Lake (2002) as research on the association between the organization of nursing care (nurse staffing, nurse education and nurse work environments) and patient safety and nurse wellbeing. The European Union's Seventh Framework Programme (FP7) funded the registered nurse forecasting (RN4CAST) study (Sermeus et al., 2011) to raise research in health services research to a multi-country European level.

One of the challenges facing the health systems over the coming decades is

the provision of more patient-centered care. To this end, patient experience with care is receiving increased attention. To efficiently report information about the quality of care and service at hospitals, one needs to identify the important dimensions of hospital care and reliably evaluate hospital's performance in each dimension. For instance, Keller et al. (2005) identified a subset of items about hospital care quality from the US Hospital Consumer Assessments of Healthcare Providers and Systems (HCAHPS) survey.

In health services research, it is not uncommon to encounter latent constructs (or latent phenomena or latent variables) that in practice cannot be directly observed/measured by a single observed variable (e.g. patient satisfaction, individual preferences, feelings, behaviours, etc); rather, their values can be deduced from observed variables (measures). These measures are usually collected via surveys, which are often multilevel in nature in the sense that data are collected on nurses and/or patients who are clustered within nursing units which are in-turn nested within hospitals, possibly across multiple countries. In establishing a model to reflect reality, it is often necessary to assess various interrelationships among observed and latent variables. Structural equation models (SEMs) are recognized as the most useful statistical tool to serve this purpose (Kline, 2015). Although there has been a shift towards the application of statistical techniques that account for the inherent complex hierarchical structure of the data (Li et al., 2013; Diya et al., 2012) research questions are becoming too complex to be answered sufficiently by univariate multilevel analyses. There is a strong demand for a statistical method that simultaneously groups highly correlated observed variables into latent variables and that assesses interrelationships among latent variables through a regression model of latent variables, while also exploiting the complex hierarchical structure of the data (i.e., SEM and multilevel modeling need to be combined). A special kind of multivariate analysis which is gaining grounds in the field of health services research, and which can nicely combine all these features, is multilevel structural equation models (ML-SEM) or multilevel factor analytic models in general. Researchers have applied ML-SEM to nursing outcomes research (Bruyneel et al., 2017a; Quintero and Lesaffre, 2017; Diya et al., 2014; Li et al.,

2014; Gajewski et al., 2010; Bruyneel et al., 2009; Lake, 2002).
Usually, it is of interest to compare latent variable scores across groups or (sub)populations (e.g., gender, intervention arm, countries) or time points. For such comparisons to be valid, a measurement instrument (such as HCAHPS) needs to be measurement invariant across those groups (which might be at different hierarchical levels), or time points. An instrument is said to fulfill measurement invariance (MI) across populations when individuals who are from different populations but identical on the construct being measured, have the same probability of achieving any given score on that instrument (Vandenberg and Lance, 2000).

The Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) consists of 16 items measuring patient specific experiences with care in the hospital and two global ratings. The items that are influential on patient satisfaction may differ due to many factors. For example, the geographical regions in each country present cultural differences that possibly influence satisfaction standards of the inhabitants. Given the increased international interest in improvement strategies for patient experiences with care, it is important to understand whether the same specific care experiences affect global ratings across countries. In this project we investigate invariance across groups as well as across hierarchical levels, and assess their implications.

Models used in this thesis fully exploit the hierarchical structure and/or multivariate nature of the data. They are also anchored on the idea of heterogeneity in health services research; that is, patients are not alike. From a statistical point of view the total variability of the data comprises 1) the variability between individuals, and 2) the residual or error variability. Standard statistical models, such as linear regression, model the population average, which is the mean response of all individuals conditional on observed covariate values. These "one size fits all" approaches ignore the former part of variability, making their results to be vulnerable to aggregation bias and wrongly estimated standard errors.

The basic statistical prerequisite for the appropriate application of multilevel modeling includes clustered data with a positive intraclass correlation (ICC). A

positive ICC violates the independent observations assumption of ordinary least squares estimation, resulting in downwardly biased standard error estimates, overly large test statistics, and inflated type 1 error rates (Krull and MacKinnon, 2001).

It will also been shown that the constancy of variance assumption, typical in classical regression, is too strong in some cases.

1.2 Thesis Contribution

There are two objectives with this PhD study. First, we wish to advance the knowledge of the relationship between organizational features of nursing care and patient outcomes by making use of Bayesian multilevel SEMs that reflect the complex hierarchical structure of the RN4CAST data (see Section 3.6 for an explanation of what 'Bayesian' entails). Second, we aim to critically evaluate the performance of existing Bayesian SEM methods via simulations. In particular, we wish to explore the performance of joint multilevel factor analytic models and multilevel covariance regression models, with applications to the RN4CAST data. Thus, there are both health services research aims and statistical aims. More specifically our aims in this thesis are:

- To study the differential magnitude of associations between HCAHPS specific care experiences and the two overall measures of patient satisfaction (rating of the hospitals and recommending the hospital to friends and family members);
- To evaluate measurement invariance of a widely used World Health Organization (WHO) instrument for measuring violence against women among Kenyan adolescent girls and young women (AGYW) population;
- To verify whether the factor structure of the HCAHPS survey, which was developed in the US, can be reproduced in a European setting; and to examine invariance across groups and levels under study (i.e., patients, nursing units, hospitals, and countries) and illustrate the degree to which this method of analysis impacts hospital rankings;

- To propose and explore the performance of joint Bayesian multilevel factor analytic models and multilevel covariance regression models with applications to the RN4CAST data.

1.3 Organization of the Thesis

This thesis text is organized into eight chapters. After the general introduction we describe the different motivating data sets and the research questions which were posed and were used as guidance in the statistical analyses in Chapter 2. Then we introduce in Chapter 3 the concepts, terminologies and estimation techniques used in subsequent chapters. In Chapter 4, we consider a case of a single hierarchical outcome variable (leading naturally to mixed effect models). Here, we look at the differential magnitude of associations between specific patient experiences with care and global ratings across countries by means of interaction terms. In this chapter, the frequentist approach was used and we show that, while the statistical model may be quite complex, the frequentist approach will do the job, see the second paper in my publication list.

In Chapter 5, we consider multivariate data with a simple structure. The material in this chapter is described in the third publication of the main reference list. In Chapter 6, several but hierarchical outcome variables with covariates measured at hierarchical levels are considered, see the first paper in my list of publications. Chapter 7 considers a case where, for several hierarchical Gaussian outcomes, scores from a factor analytic model are used as input for a covariance regression model – leading to a joint factor analytic and covariance regression model (aka combined multilevel factor analysis and covariance regression model). In this chapter, we model both the mean and the covariance matrix of the outcomes with a mixed effects structure. The material presented in this chapter is described in the fourth publication in my list of publications. In Chapters 5 through 7, a Bayesian approach is invoked. Finally, Chapter 8 winds up the thesis with a general conclusion and suggestions for further research. Technical details and derivations that are excluded from the main text are presented in the Appendix.

Chapter 2

Motivating Data Sets

In this chapter, we introduce the data sets that will be used to address the research questions. The registered nurse forecasting (RN4CAST) study data are introduced in Section 2.1, and are analysed in Chapters 4, 6 and 7. This data set has been used repeatedly allowing to see the (dis)advantages of the different methods. In Section 2.2, data from a study assessing the impact of an intervention on reducing new HIV infections among the most vulnerable adolescent girls and young women (AGYW) are introduced. These data are analysed in Chapter 5.

2.1 The Registered Nurse Forecasting Study

The registered nurse forecasting (RN4CAST) project was a three-year (i.e., 2009–2011) nurse workforce study funded by the Seventh Framework Programme (FP7) of the European Union. The project involved researchers from 12 European countries (Belgium, England, Finland, Germany, Greece, Ireland, Poland, the Netherlands, Norway, Spain, Sweden and Switzerland), the USA, China, South Africa and Botswana.

With the Centre for Health Services and Nursing Research at the University of Leuven, Belgium, as the coordinating centre, a multi-country, multilevel cross-sectional design was used to collect data in the twelve European countries to determine how system level features in the organization of nursing care

(work environment, education and work load) impact individual measures of nurse well-being (burnout, job satisfaction and turnover) and patient safety outcomes and patient satisfaction. A four-level hierarchical structure sampling strategy was used, with nurses and patients clustered within nursing units within hospitals within countries. Details on the design of this study have been discussed elsewhere (Sermeus et al., 2011). Multi-country nurse and patient satisfaction surveys were conducted. The nurse survey was conducted in all twelve European countries, while the patient satisfaction survey was conducted in eight of these countries in either all hospitals (Belgium, Poland, Greece, Finland and Switzerland), or a selection of hospitals (Germany, Spain and Ireland). Uniform hospital discharge data were also collected to analyze the association between patient mortality and nursing system delivery strategies measured in the nurse survey. Finally, the general hospital management was also surveyed about hospital-wide characteristics like bed size, teaching status, and technology level. The nurse survey contained questions about the nursing working environment, burnout, job satisfaction, nurse-perceived quality of care, nurse staff levels (number and education), and demographics. The patient survey questionnaire had 16 questions evaluating six domains: care from nurses, care from doctors, hospital/physical environment, experiences in hospital, life after hospital, and overall rating of hospital. The instrument also included questions on the patient's overall health rating and education. This thesis will utilize data from the patient survey collected from 11289 patients in 824 nursing units from 186 hospitals.

2.2 The DREAMS Impact Evaluation Study

DREAMS (Determined, Resilient, Empowered, AIDS-free, Mentored, and Safe women) is an initiative of PEPFAR (the US President's Emergency Plan for AIDS Relief) to reduce HIV incidence among the most vulnerable adolescent girls and young women (AGYW) in areas identified as 'hot-spots' with high HIV burden. DREAMS is using an HIV prevention package, being delivered "at scale", to target the multiple sources of risks that AGYW face: behavioural,

biological and structural (see Saul et al., 2018). DREAMS impact evaluation is a three-year project (2017–2019) with the objectives to: 1) measure whether HIV-related outcomes change due to DREAMS interventions at a population level; 2) explore the pathways of protection by which DREAMS interventions influence the lives of young women and ultimately their risk for HIV; and 3) assess the extent to which the DREAMS interventions are delivered as intended. In Nairobi (Kenya), the African Population and Health Research Center is partnering with the London School of Hygiene and Tropical Medicine – with funding from Bill & Melinda Gates Foundation – to evaluate the impact of the DREAMS Initiative in two informal settlements of Korogocho and Viwandani. The full study protocol has been published elsewhere (Birdthistle et al., 2018). We focused on the AGYW survey component, in which a cohort of randomly selected girls/young women were enrolled during 2017 and followed prospectively, at ∼12 and ∼24 months. This thesis uses baseline data collected from Nairobi from 1,081 AGYW (Korogocho, $n = 617$; Viwandani, $n = 464$) aged 15–22 years in 2017. One of the secondary outcomes was experience of violence measured using the WHO's violence against women (VAW) survey instrument (World Health Organization, 2005) having 15 items.

The study protocol, including informed consent and study tools, were reviewed and approved by the Observational Research Ethics Committee of the London School of Hygiene and Tropical Medicine (Ref 21111835). Additionally, ethical approval was obtained from AMREF (ESRC P298/2016). Study participants also provided informed, written consent to participate in the study. For legal minors (i.e., those aged <18 years), assent was obtained from the minor after the parent or guardian gave consent.

2.3 Research Questions

The following research questions (which follow from the thesis objectives in Section 1.2) are principally addressed in relation to the registered nurse forecasting study (questions 1, 3 and 4) and the DREAMS impact evaluation study (question 2). They were used as guidance for the statistical analyses.

Research Question 1: What is the association between HCAHPS specific care experiences and the two overall measures of patient satisfaction (rating of the hospitals and recommending the hospital to friends and family members)? Are these associations consistent across countries? What cut-off point(s) for the hospital rating score is most pertinent to contrast patient overall ratings across countries?

Research Question 2: What is the configuration of the factor structure of the World Health Organization (WHO) violence against women instrument among Kenyan adolescent girls and young women population? Are there indications of measurement (non)invariance across sub-populations?

Research Question 3: What is the multilevel factorial structure of the HCAHPS? Are there indications of measurement (non)invariance across groups (i.e., for various extraneous variables at multiple levels such as overall self rated health at the patient level, unit specialty at the nursing unit level, and hospital size, teaching status and technology at the hospital level) and levels under study (i.e., patients, nursing units, hospitals, and countries), and how does this impact hospital rankings?

Research Question 4: Are the means of the latent patient satisfaction dimensions correlated with the patient-, nursing unit-, and hospital-level characteristics? Is the nature of the relationship among the latent satisfaction dimensions consistent across hierarchical levels (patients, nursing units and hospitals), after taking into account confounders across these levels?

Chapter 3

Measurement Invariance Concepts and Inference

In this chapter, we provide an overview of the concepts commonly used in the structural equation modeling literature in reference to measurement invariance assessment. We also provide some basic ideas of inference approaches. Many multivariate data sets consist of observations that can be classified into groups, such as the two gender classes, subjects speaking a different language, people belonging to different political parties, inhabitants of different countries, and so on. It is then of interest to determine whether the parameters of the structural equation model for the observed variables are invariant across the groups. Testing for invariance of measurement across groups (or time points) is essential before meaningful comparison across the groups can be made. Measurement invariance testing applications (or evaluations) can be found in a wide variety of research fields including psychology, cross-culture, organization, economics, political science, education, management and health. For the sake of clarity we will first introduce observed and latent variable terminologies and factor analytic models. We will then explain measurement invariance concepts. We end this chapter with a review of frequentist and the Bayesian approaches.

3.1 Observed and Latent Variables

A common strategy to obtain information about subjects regarding all kinds of behavior, feelings, memories, perceptions, knowledge, and attitude is the large-scale implementation of survey instruments that capture (multiple) dimensions of that behavior or feeling. An example of such an instrument is the US Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS; Darby et al., 2005), which was later adopted by the European Union's RN4CAST patient survey to measure patient experiences (or satisfaction) with care. HCAHPS measures patient experiences/satisfaction with care indirectly through a series/set of 22 questions that reflect the dimensions of patient experience/satisfaction like nurse communication, doctor communication, physical environment, pain control, medicine communication and discharge information. Such dimensions are usually referred to as latent constructs (or common factors or latent variables) and the (22) items as manifest (or observed) measures/variables. Thus, observed variables are those that can be measured directly, while latent constructs are those that cannot be measured directly (by a single observed variable). In this thesis an observed variable is sometimes referred to as a *manifest* variable, or as an *indicator* of the latent variable, or as an *item*.

One measure of the influence a latent variable has on manifest variables is the *factor loading*. These are akin to regression/path coefficients.

3.2 Factor Analytic Models

The relationships between latent constructs and manifest measures is often studied by factor analytic models, which constitute a specific type of multivariate analysis techniques. Factor analysis (FA) models are further classified into exploratory factor analysis (EFA) and confirmatory factor analysis (CFA; Nesselroade and Cattell, 2013). The origins of FA are generally credited to Sir Charles Spearman (1904) who introduced the idea in his seminal paper "General Intelligence: Objectively Determined and Measured," published in the *American Journal of Psychology* (see e.g., Vincent (1953) on the origin and

development of factor analysis).

3.2.1 Exploratory and Confirmatory Factor Analysis

The central aim of EFA is to determine the appropriate number of latent constructs, and to uncover which observed variables are reasonable indicators of the various latent dimensions (Gorsuch, 1988). EFA should not to be confused with principle component analysis (PCA), which is an exploratory variable reduction technique: Whereas EFA is used to identify underlying latent constructs that influence a set of observed variables, PCA is used to reduce the dimensionality of large multivariate data by replacing the p original variables by a smaller number, q, of derived variables, the principal components, which are linear combinations of the original observed variables (Jolliffe and Cadima, 2016). EFA is especially useful when we have no a priori idea about the pattern of relationships between the indicators and the latent constructs. Let $\boldsymbol{y}_i$ be a p-dimensional vector of observed measures for individual i, and let $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iq})'$ be a q-dimensional vector of underlying latent constructs (or dimensions/factors) for individual i with $q < p$. Then an EFA model specifies a set of linear relations between observed measures and underlying latent constructs as:

$$\boldsymbol{y}_i = \boldsymbol{\alpha} + \boldsymbol{\Lambda}\boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\xi}_i \sim N(\mathbf{0}, \boldsymbol{\Phi}), \quad i = 1, \ldots, N, \tag{3.1}$$

where $\boldsymbol{\alpha}$ is a vector of observable–variable intercepts (or thresholds) of length p, $\boldsymbol{\Lambda}$ is a factor loading matrix of size $p \times q$, $\boldsymbol{\Phi}$ is a $q \times q$ matrix, and $\boldsymbol{\varepsilon}_i$ is a p-dimensional vector of residuals which is distributed as $N(\mathbf{0}, \boldsymbol{\Psi})$. It is also usually assumed that all sources of covariation among the observed variables are expressed in the latent variables and therefore that $\boldsymbol{\Psi}$ is diagonal; it will be shown later that this assumption may easily be relaxed. It is further assumed that $\text{Cov}(\boldsymbol{\Lambda}, \boldsymbol{\varepsilon})$=0.

The aim of a CFA model is to test the underlying factor structure that we have a priori in mind. This structure may come from prior empirical analyses using EFA, as well as from theoretical reasoning. With a reasonable model fit,

the CFA model could provide evidence to confirm an assumed factor structure. While a typical CFA model has the same form as an EFA model in equation (3.1), in a CFA model the factor loading matrix $\mathbf{\Lambda}$ is different from that of an EFA model in that some elements are fixed. That is, the cross-loadings are fixed to zero as we have in mind a priori a particular factor structure as our testing hypothesis. As an example, suppose that for HCAHPS we are interested in studying the first three dimensions of patient satisfaction with care and it is known that nurse communication is measured through the first 4 items, doctor communication through the next 3 items, and physical environment by the next 2 items, then the conventional factor loading matrix for a CFA is given by:

$$\mathbf{\Lambda}^T = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \lambda_{13} & \lambda_{14} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_{25} & \lambda_{26} & \lambda_{27} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{38} & \lambda_{39} \end{bmatrix},$$

where next to the zeroes the λ_{jk} are unknown parameters. We will indicate that fixing cross-loadings to zero may be unrealistic. An alternative is to embed CFA in a structural equation modeling framework. This will be done in this thesis.

For either an EFA or a CFA model, the implied covariance matrix of the observed variable $\boldsymbol{y}$ takes the form:

$$\text{Cov}(\boldsymbol{y}) = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' + \mathbf{\Psi}.$$

That is, by implementing a factor model the covariance matrix of the observed responses could be reconstructed through the factor loadings $\mathbf{\Lambda}$, the covariance matrix of factors $\mathbf{\Phi}$ and the covariance matrix of residuals $\mathbf{\Psi}$.

In summary, whereas EFA and CFA are based on the common factor model, in a CFA the researcher controls every aspect of the model specification (e.g., to generate unstandardized solution; to specify correlated errors; and to place various constraints on the solution such as fixing cross-loadings to zero or holding model parameters to equality). These features allow CFA to address important research questions for which EFA is not well suited or incapable, such as the comparison of factor models across groups (e.g.,

measurement invariance of tests (Section 3.5)), the analysis of mean structures, scale reliability evaluation, and the inclusion of covariates in a factor model (e.g., multiple indicators multiple causes modeling, see Section 3.5.2).

3.2.2 Multilevel Factor Analysis Model

When data have a multilevel structure (e.g., patients within hospitals, children within schools), the correlated nature should be taken into account in the factor analysis model in order to obtain valid estimates (Longford and Muthén, 1992). This gives rise to the multilevel factor analysis (MFA) model. Suppose that we have two-level data from N individuals clustered within J clusters. It is classically assumed that the clusters are a simple random sample from a population of clusters, and that the individuals are a simple random sample within each cluster. With n_j the number of individuals in the jth cluster, the total sample size is $N = \Sigma_{j=1}^{J} n_j$. For a balanced data set, $n_j = n \; \forall j$. Let $\boldsymbol{y}_{ij}$ denote a $p-$dimensional response vector for individual i in cluster j. In a multilevel CFA, $\boldsymbol{y}_{ij}$ is decomposed into two latent random components reflecting two sources of random variation in multilevel data: between-cluster random components ($\boldsymbol{y}_{Bj}$) and within-cluster, between-individual random components ($\boldsymbol{y}_{Wij}$):

$$\boldsymbol{y}_{ij} = \boldsymbol{y}_{Bj} + \boldsymbol{y}_{Wij} \quad i = 1, 2, \ldots, n_j; \quad j = 1, 2, \ldots, J, \tag{3.2}$$

where $E\left(\boldsymbol{y}_{Wij}\right) = 0$ and $\text{Cov}\left(\boldsymbol{y}_{Bj}, \boldsymbol{y}_{Wij}\right) = 0$. The two-level CFA model specifies linear relations between $\boldsymbol{y}_{Bj}$ and underlying q_B-dimensional latent factors $\boldsymbol{\xi}_{Bj}$ at level 2, and linear relations between $\boldsymbol{y}_{Wij}$ and underlying q_W-dimensional latent factors $\boldsymbol{\xi}_{Wij}$ at level 1:

$$\begin{aligned} \boldsymbol{y}_{ij} = \boldsymbol{y}_{Bj} + \boldsymbol{y}_{Wij} &= \left[\boldsymbol{\alpha}_B + \boldsymbol{\Lambda}_B \boldsymbol{\xi}_{Bj} + \boldsymbol{\varepsilon}_{Bj}\right] + \left[\boldsymbol{\Lambda}_W \boldsymbol{\xi}_{Wij} + \boldsymbol{\varepsilon}_{Wij}\right], \\ \boldsymbol{\xi}_{Bj} \sim N(\mathbf{0}, \boldsymbol{\Phi}_B), &\quad \boldsymbol{\xi}_{Wij} \sim N(\mathbf{0}, \boldsymbol{\Phi}_W), \\ \text{Cov}(\boldsymbol{\xi}_{Bj}, \boldsymbol{\varepsilon}_{Bj}) = 0, &\quad \text{Cov}(\boldsymbol{\xi}_{Wij}, \boldsymbol{\varepsilon}_{Wij}) = 0, \\ i = 1, 2, \ldots, n_j; &\quad j = 1, 2, \ldots, J. \end{aligned} \tag{3.3}$$

where $\mathbf{\Lambda}_B$ is a $p \times q_B$ factor loading matrix associated with the between-level latent factor $\boldsymbol{\xi}_{Bj}$, $\mathbf{\Lambda}_W$ is a $p \times q_W$ factor loading matrix associated with the within-level latent factor $\boldsymbol{\xi}_{Wij}$, $\boldsymbol{\varepsilon}_{Bj}$ is a $p-$dimensional between-level vector of residuals with covariance matrix $\mathbf{\Psi}_B$, $\boldsymbol{\varepsilon}_{Wij}$ is a $p-$dimensional within-level vector of residuals with covariance matrix $\mathbf{\Psi}_W$ and all the latent factors and residual variances are assumed mutually independent with each other. The implied covariance matrix for the MFA model (3.3) is then:

$$\text{Cov}(\boldsymbol{y}_{ij}) = \mathbf{\Lambda}_B \mathbf{\Phi}_B \mathbf{\Lambda}_B' + \mathbf{\Psi}_B + \mathbf{\Lambda}_W \mathbf{\Phi}_W \mathbf{\Lambda}_W' + \mathbf{\Psi}_W \tag{3.4}$$

3.2.3 Higher-Order Factor Analysis Model

Higher-order factor analysis is a theory driven procedure in which the researcher imposes a more parsimonious structure to account for interrelationships among factors established by CFA (Maruyama, 1997; Brown, 2014). That is, one may want to find factors of the factors. In this context, the variables-by-factors loading matrix is a loading matrix of first order factors on second order factors. The second order factors, if they are themselves intercorrelated, may in turn be factored, leading to still higher order factors, and so on (Wind et al., 1973). This allows researchers to examine the pattern of constructs assumed to underlie the variables.

This analytic procedure can be used for several reasons. First, it is useful in situations where, for example, it is believed that more specialized facets of e.g. patient experience/satisfaction (i.e., doctor communication, nurse communication, and physical environment) are influenced by a broader dimension of general experience/satisfaction. Second, is to rescue a construct. Indeed, it could be that for a construct initially predicted to be unidimensional, the research evidence reveals that multiple factors are required to explain the covariation among the set of indicators of the construct. In these instances, a single higher-order factor can often be specified to account for the covariation among the multiple factors. As such, the construct consists of a single broader dimension, and several subdimensions. Third, in first-order multiple-factor CFA models the factors are presumed to be interrelated, but the researcher

makes no substantive claims about the nature of these relationships. In higher-order factor analyis, the focus is on the intercorrelations among the factors. We model intercorrelations among factors in Chapter 7.

3.3 Structural Equation Models

The basic goal of structural equation modeling (SEM) is to generalize the CFA model for assessing how latent variables affect each other in various ways. SEM is a combination of factor analysis and multiple regression. A typical SEM model consists of two major parts: a measurement model part which relates manifest variables to latent variables and a structural model part that relates latent variables to one another (i.e., regression model among the common factors; Kline, 2015). Let $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ be, respectively, p-dimensional and s-dimensional random vectors of manifest variables which are the respective indicators for latent variables $\boldsymbol{\xi}_i$ and $\boldsymbol{\eta}_i$ for individual i. The measurement part of SEM is then defined by the following CFA models:

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{\alpha}_y + \boldsymbol{\Lambda}_y \boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_{iy} \\ \boldsymbol{x}_i &= \boldsymbol{\alpha}_x + \boldsymbol{\Lambda}_x \boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_{ix}, \end{aligned} \tag{3.5}$$

with $\boldsymbol{\alpha}, \boldsymbol{\Lambda},$ and $\boldsymbol{\varepsilon}$ have the same meanings as before but having the y and x subscripts to distinguish between parameters associated with $\boldsymbol{y}$ and $\boldsymbol{x}$, respectively. With the two equations in model (3.5) modelled simultaneously, the structural part is then defined by the equation:

$$\boldsymbol{\xi}_i = \boldsymbol{\Gamma} \boldsymbol{\eta}_i + \boldsymbol{\delta}_i \tag{3.6}$$

where $\boldsymbol{\xi}$ a q_1-dimensional random vector of dependent latent factors and $\boldsymbol{\eta}$ a q_2-dimensional random vector of independent latent factors. $\boldsymbol{\Gamma}$ is an unknown $q_1 \times q_2$ matrix of regression coefficients representing the causal effects among $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, and $\boldsymbol{\delta}$ is a q_1-dimensional random vector of residuals.

3.4 Identification of Latent Variable Models

A key question in model identification is whether there is enough non-redundant information in the data to be able to estimate the required parameters uniquely. With regression, identification is never an issue because they are always *just-identified* models, meaning that the number of parameters to estimate exactly equals the amount of non-redundant information in the data. For now, think of the amount of non-redundant information in the data as being the number of non-redundant variances/covariances in the dataset. This can be easily calculated using the formula in Equation 3.7.

$$\text{Non-redundant information in a data set} = \frac{p(p+1)}{2} \tag{3.7}$$

where p is the number of manifest variables.

With a latent variable model, model identification is more complex than with regression, as the models can be just-identified, underidentified, or overidentified. If there are more parameters to estimate than there are pieces of non-redundant information, then the model is *underidentified*. Conversely, if there are more pieces of non-redundant information than there are parameters to estimate, then the model is usually *overidentified*. If a model is underidentified, the model's parameters cannot be estimated uniquely and latent variable model programs will return an error message. If a model is over- or just-identified, there should be unique estimates for each parameter. For overidentified models, not only do they provide unique parameter estimates, but they can also provide measures of model fit. This is because they have degrees of freedom (df). In latent variable models, df can be thought of as the number of non-redundant pieces of information in the data minus the number of parameters to estimate. Only overidentified models have $df > 0$. There are rules-of-thumb model identification conditions that generally work for most models with reflective latent variables. They mainly involve the number of indicator variables on each latent variable in the model (usually at least two indicators) and latent variable's scale (i.e., constraining the variance of the latent variables to 1.0, constraining a single factor loading for each latent variable to an arbitrary

value (usually 1.0), or effect's coding). See e.g., Beaujean (2014) for a detailed treatment of the topic.

3.5 Measurement Invariance

Measurement invariance (MI; also referred to“factorial invariance" or “measurement equivalence") occurs when the parameters of a measurement model are statistically equivalent across two or more groups, or time points. In the presence of MI, systematic differences between observed and test scores are attributable to true differences in the trait(s) that the test is measuring. A test is measurement invariant with respect to G if the following conditional independence holds:

$$f_1(Y \mid H = \xi, G = g) = f_2(Y \mid H = \xi), \tag{3.8}$$

where Y is a set of observed variable (or items/indicators, H is a set of attributes i.e., latent traits) measured by Y, and G is some grouping membership or socio-demographic variables. f_1 is the conditional distribution function of Y given ξ and g, while f_2 is the conditional distribution of Y given ξ. The measurement of H by Y is said to be biased with respect to G if the conditional independence does not hold; that is, if $f_1 \neq f_2$ (Jak et al, 2010).

The parameters of interest in MI testing are the factor loadings, indicator intercepts (when indicators are Gaussian) or indicator thresholds (when indicators are categorical) and residual variances. In a FA model, a latent mean is created from the factor loadings and intercepts (or thresholds) to ensure construct comparability (Meredith, 1993). For non-invariant items, the factor loadings and/or intercepts (or thresholds) contribute differently to the means, hence preventing valid and comparable factor score estimates.

3.5.1 Definitions

Cross-sectional factorial invariance concerns the invariance of corresponding parameters across population groups– whether in a single level or multilevel

setting. **Longitudinal factorial invariance** concerns the invariance of corresponding parameters across time within a population group. We explain these concepts in light of the former.

Consider the factor model (3.1) with the subscript i, representing subject, dropped for simplicity:

$$\boldsymbol{y} = \boldsymbol{\alpha} + \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\varepsilon} \tag{3.9}$$

Then for the multiple group case, the measurement model can be written as

$$\boldsymbol{y}_g = \boldsymbol{\alpha}_g + \boldsymbol{\Lambda}_g\boldsymbol{\xi}_g + \boldsymbol{\varepsilon}_g, \tag{3.10}$$

where the subscript g indicates group membership. Assuming that the latent factors are uncorrelated with the residuals, the mean and covariance structure of $\boldsymbol{y}_g$ are reproduced respectively by

$$E(\boldsymbol{y}_g) = \boldsymbol{\alpha}_g + \boldsymbol{\Lambda}_g\boldsymbol{\tau}_g, \tag{3.11}$$

$$\text{Cov}(\boldsymbol{y}_g) = \boldsymbol{\Lambda}_g\boldsymbol{\Phi}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g, \tag{3.12}$$

where $E(\boldsymbol{\xi}_g) = \boldsymbol{\tau}_g, \text{Cov}(\boldsymbol{\xi}_g) = \boldsymbol{\Phi}_g,$ and $\text{Cov}(\boldsymbol{\varepsilon}_g) = \boldsymbol{\Psi}_g$. The following factorial invariance types/levels can be defined:

(a) **Configural invariance** requires that the instrument represents the same dimension (i.e., number of factors) and pattern (i.e., each factor is associated with the same item sets) across groups and/or time points. That is, the dimension and the pattern of zero and non-zero loadings in $\boldsymbol{\Lambda}_g$ are the same across groups. Configural invariance must be established in order for subsequent tests to be meaningful. In itself, it is not sufficient to justify quantitative group comparisons.

(b) **Metric invariance** (also referred to as weak invariance) requires that, assuming configural invariance, corresponding factor loadings are equal across groups ($\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$ for all g). This level of invariance provides evidence that corresponding common factors have the same meaning across groups (Gregorich, 2006; Orindi et al., 2017). Metric invariance is also known as **pattern**

invariance stemming from the fact that the full set of factor loadings is known as the factor pattern matrix.

(c) **Scalar invariance** (also referred to as strong invariance) implies that the meaning of the factors/constructs (the factor loadings), and the levels of the underlying items (intercepts or thresholds) are equal across groups. That is, $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\alpha}_g$ are invariant across groups ($\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$ and $\boldsymbol{\alpha}_g = \boldsymbol{\alpha}$ for all g). Evidence that corresponding factor loadings and item intercepts/thresholds are invariant across groups suggests that (1) group differences in estimated factor means are unbiased and (2) group differences in observed means are directly related to group differences in factor means and are not contaminated by differential additive response bias (Gregorich, 2006). Thus scalar invariance is a necessary condition for meaningful comparison of group means (Meredith, 1993; Steenkamp and Baumgartner, 1998; Vandenberg and Lance, 2000; van de Schoot et al., 2012).

(d) **Strict factorial invariance** imposes cross-group equality constraints on corresponding factor loadings, item intercepts, and item residual variances. In other words, the strong factorial invariance model is further restricted so that corresponding item residual variances are invariant across groups ($\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}, \boldsymbol{\alpha}_g = \boldsymbol{\alpha}$, and $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$ for all g). Evidence for strict factorial invariance should be obtained for defensible comparisons of group means and variances. Defensible comparison requires not only that the measured constructs have the same meaning across groups, but also that group comparisons of sample estimates (e.g., means and variances) reflect true group differences and are not contaminated by group-specific attributes that are unrelated to the construct of interest (Gregorich, 2006). However, there is a consensus that it is not necessary to demonstrate invariance across residual variance parameters.

Cross-level measurement invariance can be checked in a multilevel setting. For example, cross-level configural invariance would require that the instrument represents the same factor structure (dimension and pattern) across the hierarchical levels (Zyphur et al., 2008; Orindi et al., 2017). In model 3.3 above, this implies $\boldsymbol{\Lambda}_B = \boldsymbol{\Lambda}_W$. Cross-level invariance (and its implications) is the topic of Chapter 6.

Structural invariance tests whether the arrows connecting the endogenous latent variables (i.e., $\boldsymbol{\xi}$) and exogenous latent variables (i.e., $\boldsymbol{\eta}$) to each other in expression 3.6 are properly drawn the same way for each group or level in the analysis. That is, one checks whether the structural relation (i.e., expression 3.6) is not the same across the groups or levels of analysis, and hence cross-group or cross-level structural non-invariance.

3.5.2 Testing invariance

Factorial invariance can be tested using three approaches. Firstly, with **multiple-groups confirmatory factor analysis** (MGCFA; Jöreskog, 1971; Bollen, 2014; Brown, 2014) verifying MI involves running a set of increasingly constrained structural equation models (SEM) to test the nested forms of factorial invariance in the following order: configural, metric, scalar and strict invariance.

Another possibility is to use **multiple indicators multiple causes** (MIMIC; Jöreskog and Goldberger, 1975; Abraham and Ledolter, 1979; Muthén, 1989). MIMIC is also referred to as CFA with covariates. MIMIC is a special case of SEM, which performs cross-group comparisons by including group membership as a covariate in the model. Both the factor(s) and indicators are then regressed on observable covariate(s)(which represent group membership). The model can be written as

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{c} + \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\varepsilon}; \quad \boldsymbol{\xi} = \boldsymbol{B}\boldsymbol{c} + \boldsymbol{\delta} \tag{3.13}$$

where $\boldsymbol{y}$ and $\boldsymbol{\xi}$ have the same meaning as before, $\boldsymbol{c}$ is a vector of observable group member covariates, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are the coefficient matrices. $\boldsymbol{\varepsilon}$ and $\boldsymbol{\delta}$ are vectors of the disturbances such that

$$E(\boldsymbol{\delta}\ \boldsymbol{\delta}') = I, \quad E(\boldsymbol{\delta}\ \boldsymbol{\varepsilon}') = 0, \quad E(\boldsymbol{\varepsilon}\ \boldsymbol{\varepsilon}') = \boldsymbol{\Psi}^2 \text{ (diagonal)}.$$

Then a significant direct effect of a covariate on a factor indicates population heterogeneity (i.e., the factor mean varies with the group). In model 3.13, this implies that the observed covariate in $\boldsymbol{c}$ has a significant impact on the latent

factor $\boldsymbol{\xi}$. A significant direct effect of an observed covariate on an indicator represents measurement noninvariance across groups. On its turn it implies that the observed covariate in $\boldsymbol{c}$ has a significant impact on $\boldsymbol{y}$ in model 3.13.

Alignment optimization (Asparouhov and Muthén, 2014; Muthén and Asparouhov, 2014) provides a third approach. In this approach, one replaces equality constraints with a procedure that is similar to looking for the 'best' solution by rotating the factors in EFA. It starts with a configural model with no invariance and finds as much variance as possible by letting the factor means and variances vary across groups. For each measurement parameter the biggest invariant set of groups is identified. A certain parameter in a given group is considered largely non-invariant when it differs significantly from the average value for that parameter across all groups. A strength of the method is the ability to conveniently estimate models for many groups. At the time of writing this text, this methodology was not yet developed to handle models with cross-loadings as well as models with covariates (Asparouhov and Muthén, 2014). Assessment of cross-group measurement invariance for single-level data is treated in Chapter 5.

For multilevel data, cross-group factorial invariance is established depending on whether the group membership is at level 1 or higher level, say level 2. The procedure for testing factorial invariance across groups when grouping variable is at a higher level is parallel to the standard procedure for testing single-level factorial invariance. Suppose that we have two-level data from N individuals clustered within J clusters. Then the two-level CFA model with level-2 group membership ($g = 1, \ldots, G$) can be written as

$$\begin{aligned} \boldsymbol{y}_{ij(g)} &= \boldsymbol{y}_{Bj(g)} + \boldsymbol{y}_{Wij(g)} \\ &= \left[\boldsymbol{\alpha}_{B(g)} + \boldsymbol{\Lambda}_{B(g)}\boldsymbol{\xi}_{Bj(g)} + \boldsymbol{\varepsilon}_{Bj(g)}\right] + \left[\boldsymbol{\Lambda}_{W(g)}\boldsymbol{\xi}_{Wij(g)} + \boldsymbol{\varepsilon}_{Wij(g)}\right] \end{aligned} \tag{3.14}$$

where the subscript (k) indicates the level-2 group membership, $E(\boldsymbol{\xi}_{Bj(g)}) = \boldsymbol{\tau}_{B(g)}$, $\mathrm{Cov}(\boldsymbol{\xi}_{Bj(g)}) = \boldsymbol{\Phi}_{B(g)}$, $\mathrm{Cov}(\boldsymbol{\varepsilon}_{Bj(g)}) = \boldsymbol{\Psi}_{B(g)}$, $\mathrm{Cov}(\boldsymbol{\xi}_{Wij(g)}) = \boldsymbol{\Phi}_{W(g)}$, and $\mathrm{Cov}(\boldsymbol{\varepsilon}_{Wij(g)}) = \boldsymbol{\Psi}_{W(g)}$. All four levels of cross-group invariance are testable in the level-2 model: configural, metric ($\boldsymbol{\Lambda}_{B(g)}$), scalar ($\boldsymbol{\Lambda}_{B(g)}, \boldsymbol{\alpha}_{B(g)}$) and strict ($\boldsymbol{\Lambda}_{B(g)}, \boldsymbol{\alpha}_{B(g)}$,

$\mathbf{\Psi}_{B(g)}$). Assessment of cross-group invariance at hierarchical levels is treated in Chapter 6.

3.5.3 Causes of Measurement Noninvariance

Factor loading noninvariance can occur when (i) the understanding (or conceptual meaning) of the construct differs across groups, (ii) certain items are more applicable for one group than for other groups, (iii) the scale was not properly translated, and/or (iv) certain groups respond to extreme items differently (Chen, 2008). Intercept (or threshold) noninvariance can occur due to one or more of the following reasons: social desirability reasons or social norms; particular group(s) displaying a propensity to respond more strongly to an item despite having the same latent or factor mean; and/or certain groups having different reference points when making statements about themselves (Chen, 2008). As other factors could also contribute to the lack of strong factorial invariance, the item content should be inspected carefully to ascertain the reasons for noninvariance. It should be noted that when all items are equally influenced by the factors listed above, measurement invariance would likely remain even though these measurement concerns or biases would still be existing.

3.6 The Frequentist and Bayesian Approaches

Statistical inference is a collection of activities aimed at extracting information from data and to generalize the observed results beyond the data at hand. Of course, inference depends on the collected data and on the assumed underlying probabilistic model that generated those data as well as the approach to generalize from the data to the population. The two main schools of drawing statistical inference are the frequentist and the Bayesian paradigms.

Frequentist approach is based on the idea that probability is a limiting frequency. This means that probability is assigned to a repeatable event in which the uncertainty is due to randomness. Because uncertainty about parameters is epistemic, frequentist inference does not allow probability statements about the

parameters of a statistical process (Wagenmakers et al., 2008). A popular way to estimate the model parameters is maximum likelihood. Maximum likelihood procedures aim to find the point in the parameter space that maximize the likelihood function, called the maximum likelihood estimate. For example, the *likelihood* of a regression model is defined as the probability of the data given the parameters and inputs. Thus, for n independent responses given a vector parameter $\boldsymbol{\theta}$ and predictors X, the likelihood of the sample is given by:

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta}, X) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}, X_i). \tag{3.15}$$

The maximum likelihood estimate is the vector $\boldsymbol{\theta}$ for which this expression is maximized, given data $X, \boldsymbol{y}$. It is common practice to use the notation $L(\boldsymbol{\theta}|\boldsymbol{y})\, p(\boldsymbol{y}|\boldsymbol{\theta})$ for the likelihood as a function of parameter vector $\boldsymbol{\theta}$, with the dependence on the predictors X implicit. For many classical models with no random effects, the likelihood is relatively simple and can be written analytically (with a closed-form expression). In situations where the model has random effects, say $\boldsymbol{b}_i\,(i = 1, ..., N)$, the marginal likelihood contribution is calculated by integrating out the random effects to obtain the maximum likelihood estimates. As an illustration, consider a model where $\boldsymbol{y}$ is the observed data vector and, conditional on the random effects, $\boldsymbol{b}_i$, assume that the elements of $\boldsymbol{y}$ are independent and drawn from a distribution in the exponential family, encompassing both normally distributed and non–normal outcomes. Also assume a distribution for $\boldsymbol{b}_i$ depending on parameter, $\boldsymbol{D}$. The marginal likelihood is given by,

$$L(\boldsymbol{\beta}, \boldsymbol{D}|\boldsymbol{y}) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{y_i}(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}) f_{b_i}(\boldsymbol{b}_i|\boldsymbol{D}) d\boldsymbol{b}_i. \tag{3.16}$$

Expression (3.16) cannot usually be evaluated in closed-form and has an integral with dimension equal to the number of random effects. Various methods have been proposed to solve this problem including numerical integration methods centered on the adaptative quadrature (McCulloch, 1997).

The main idea of the Bayesian approach is to combine the results of the present experiment or data (likelihood) with our past experience or prior knowledge (prior probability) to result in a revised probability (posterior probability). In the Bayesian paradigm, the parameter θ is given a probability distribution which expresses our prior knowledge about that parameter. There is still a true value for the parameter, but the parameter becomes stochastic because of our uncertainty of its value (Lesaffre and Lawson, 2012). Thus, Bayesian statistical conclusions are made in terms of probability statements conditionally on the observed data. That is, based on the posterior distribution

$$p(\theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\theta)p(\theta)}{p(\boldsymbol{y})} = \frac{L(\theta|\boldsymbol{y})p(\theta)}{p(\boldsymbol{y})}, \tag{3.17}$$

where $p(\theta)$ is the the prior distribution of θ, which can be specified on the basis of historical studies or expert knowledge. In the absence of such information, a non-informative prior could be used. In such a case, the likelihood dominates the prior and information extracted from the posterior is equivalent to the information extracted from the likelihood. Note that the denominator $p(\boldsymbol{y})$ does not depend on θ and can be regarded as a *normalizing constant*. It can be written as the integration of the likelihood, $L(\theta|\boldsymbol{y})$, over θ, as $p(\boldsymbol{y}) = \int L(\theta|\boldsymbol{y})p(\theta)d\theta$ and is therefore called the *averaged likelihood* or the *marginal likelihood*. It ensures that $p(\theta|\boldsymbol{y})$ integrates to one as required by the definition of a probability function.

The posterior distribution (3.17) combines the prior belief, $p(\theta)$, with the observed data, $\boldsymbol{y}$, through $p(\boldsymbol{y}|\theta)$ to perform inference. In some specific cases, expression (3.17) has a closed-form. For instance, when the data follow a binomial distribution and the success rate is assigned a beta prior, the posterior distribution of the success rate corresponds to a beta density. This facilitates the computation of posterior summary measures. However, in most cases, analytic determination of the posterior is not possible, mainly because the denominator may involve high-dimensional integration for the joint posterior distribution of high-dimensional parameters as the Bayesian method treats all parameters as random variables. This integration becomes even heavier in the presence of random effects or latent variables (Lesaffre and Lawson, 2012). As

such, alternative sampling algorithms are necessary to explore the posterior distribution and compute the summary measures for the parameters in the model. We present these in the next section.

3.6.1 Markov chain Monte Carlo Sampling

The Markov chain Monte Carlo (MCMC) technique (also called *Markov chain simulation*) is a general method based on drawing values of θ from approximate distributions and then correcting those draws to better approximate the target posterior distribution, $p(\theta|y)$. The values $\theta^1, \theta^2, \ldots,$ are drawn sequentially such that they form a Markov chain; that is, for any t, the distribution of $\theta^{(t)}$ given all previous θ's depends only on the most recent value, $\theta^{(t-1)}$. Key to the method's success is, however, not the Markov property, but rather that the approximate distributions are improved at each step in the simulation, in the sense of converging to the target distribution. The two main MCMC procedures are the Gibbs sampler and the Metropolis(-Hastings) algorithm.

The Gibbs sampling

The Gibbs sampler was introduced by Geman and Geman (1984) in the context of image processing, but it only became popular in the statistical community after Gelfand and Smith (1990) and Gelfand et al. (1990) demonstrated its elegance in tackling complex problems in a Bayesian framework. The basic idea of Gibbs sampling is to partition the set of unknown parameters and then estimate them one at a time (or one group at a time) with each parameter (or groups of parameters) estimated conditional on all the others. The algorithm is effective as, in most problems, estimating separate parts of a model is relatively easy, even if estimating all the parameters at once is difficult. The sampling approach is initiated using a starting value for parameters, say $\boldsymbol{\theta}^0$.

Formally, given a starting value $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0, \ldots, \theta_d^0)^T$, the procedure samples from the full conditional distributions (also called full conditionals) at iteration $(t+1)$ in the following d steps:

1. Sample $\theta_1^{(t+1)}$ from $p(\theta_1|\theta_2^t, \ldots, \theta_{d-1}^t, \theta_d^t, \boldsymbol{y})$

2. Sample $\theta_2^{(t+1)}$ from $p(\theta_2|\theta_1^{(t+1)}, \theta_3^t, \ldots, \theta_d^t, \boldsymbol{y})$

$\vdots$

d. Sample $\theta_d^{(t+1)}$ from $p(\theta_d|\theta_1^{(t+1)}, \ldots, \theta_{(d-1)}^{(t+1)}, \boldsymbol{y})$

Under mild regularity conditions (see Lesaffre and Lawson, 2012), the generated values $\boldsymbol{\theta}^t, \boldsymbol{\theta}^{(t+1)}, \ldots$ ultimately can be regarded as observations from the posterior distribution, called the *target distribution.* An initial portion of the chain, the *burn-in part*, is discarded.

There are several versions of the basic Gibbs sampler depending on how the sampler travels in the d dimensions.

In situations where the full conditional distribution is difficult to obtain (or difficult to sample from), one may use a different sampling algorithm, the Metropolis-Hastings algorithm. Of course, Gibbs sampling could be seen as a special case of Metropolis-Hastings sampling (see e.g., Gelman et al. (2013) for proof).

The Metropolis-Hastings algorithm

Unlike the Gibbs sampler, the Metropolis-Hastings algorithm does not require the full conditionals. Instead, the parameters $\boldsymbol{\theta}$ are first sampled from a proposal density q and in the second step, part of the sampled values are accepted to yield a sample from the posterior distribution $p(\theta|\boldsymbol{y})$. This involves the following two steps, at iteration t:

1. Sample a candidate $\widetilde{\boldsymbol{\theta}}$ from the proposal density $q(\widetilde{\boldsymbol{\theta}}|\boldsymbol{\theta})$, with $\boldsymbol{\theta} = \boldsymbol{\theta}^t$.

2. The next value $\boldsymbol{\theta}^{(t+1)} = \begin{cases} \widetilde{\boldsymbol{\theta}}, & \text{with probability } \alpha(\boldsymbol{\theta}^t, \widetilde{\boldsymbol{\theta}}) \quad \text{(accept proposal)} \\ \boldsymbol{\theta}^t, & \text{otherwise} \qquad\qquad\qquad\quad \text{(reject proposal)}, \end{cases}$

   with $\alpha(\boldsymbol{\theta}^t, \widetilde{\boldsymbol{\theta}}) = \min\left(r = \dfrac{p(\widetilde{\boldsymbol{\theta}}|\boldsymbol{y})q(\boldsymbol{\theta}^t|\widetilde{\boldsymbol{\theta}})}{p(\boldsymbol{\theta}^t|\boldsymbol{y})q(\widetilde{\boldsymbol{\theta}}|\boldsymbol{\theta}^t)}, 1\right)$.

A variety of Metropolis(-Hastings) algorithms are in use differing according to the choice of the proposal density. Common choices are the multivariate

normal density with mean equal to the current position $\boldsymbol{\theta}^t$ and covariance matrix Σ (which needs to be chosen by the user) and the multivariate $t-$distribution. See Gilks et al. (1995) for more on the Gibbs sampler and the Metropolis algorithm.

3.6.2 Software

It is impossible to give a comprehensive overview of frequentist and Bayesian software given the explosion of software developments. We shall therefore limit ourselves to a few software packages, which are also relevant for this thesis. Software such as R, SAS, Stata (from Version 15) and Mplus have the capabilities of fitting specialized models using both the frequentist and the Bayesian approaches. In this thesis, Stata version 14.2 (StataCorp, College Station,TX) is used for data management only (see e.g. Chapter 5).

For the frequentist approach, Gaussian and adaptive Gaussian quadrature is implemented in SAS procedures GLIMMIX and NLMIXED (generalized linear mixed models); Stata procedure GLLAMM; and the package lme4 (Bates et al., 2015) in R. We have used SAS version 9.4 (SAS Institute, Cary NC) to perform (univariate) multilevel ordered logistic (also probit and complementary log-log) regression analyses in Chapter 4. Exploratory single- and multi-level factor analyses were performed using Mplus Version 7.4 (Muthén and Muthén, 2015), which is a specialised software for fitting latent variable models.

For the Bayesian approach, Gibbs sampling, together with many other sampling methods, are nowadays implemented in many statistical programs and packages, such as the **B**ayesian inference **U**sing **G**ibbs **S**ampling (BUGS) software comprising both WinBUGS and OpenBUGS (Spiegelhalter et al., 2003), **J**ust **A**nother **G**ibbs **S**ampler (JAGS; Plummer, 2003), Mplus (Muthén and Muthén, 2015), and many more. The BUGS software has several advantages including flexibility to accommodate more complex models – a point we illustrate in Chapter 7 – and ability to be called from other software packages (see e.g., `http://www.mrc-bsu.cam.ac.uk/software/bugs/calling-winbugs-1-4-from-other-programs/`), enabling users with varied preferences for statistical software to integrate them into their toolkit of statisti-

cal software. We note that as these programs may differ in the default sampling method for specific types of parameters, they may behave differently. JAGS, as opposed to BUGS, is platform independent. In this thesis we have mainly used Mplus and JAGS for Bayesian analyses. **RStudio** (`http://rstudio.org/`) is a useful Open source software tool to access R packages, allowing for interaction between R and other softwares.

Chapter 4

Contribution of HCAHPS Specific Care Experiences to Global Ratings Varies Across 7 Countries: What Can be Learned for Reporting These Global Ratings?

In this chapter, we begin by analysing two hospital global ratings. Cross country invariance is investigated using a model with country-by-items interraction terms. Similarity in regression coefficients suggest patients' opinion about the hospital as expressed by the response were driven by the same arguments, hence invariance across the countries.

Given increased international interest in improvement strategies for patient experiences with care, it is important to understand if the same specific care experiences affect global ratings across countries. Also, reporting of these global ratings currently substantially varies in both research and public reporting. We examine differential magnitude of associations between Hospital

Consumer Assessment of Healthcare Providers and Systems (HCAHPS) specific care experiences and global ratings, and explore implications for reporting these global ratings, using multilevel ordered logistic regression analysis. By means of interaction terms, we examine consistency of these associations across countries. Assuming homogeneous use and interpretations of response categories of these specific experiences across patients within and between countries, we investigate what the associations between specific experiences and global ratings imply for reporting global ratings across countries. It will be shown that all specific experiences were associated with both global ratings, and that there were relatively little differential effects across countries. A reporting strategy with different cut-off points across countries might provide a more equitable comparison of global ratings.
Thus, research and public reporting may consider examining in more detail, within and across populations, global rating cut-off points that represent the same true level of positive patient experiences.

4.1 Introduction

Patients' reports of their experiences with care are increasingly part of a validated set of metrics to measure the quality of care. The most widely used instrument is the US HCAHPS – Hospital Consumer Assessment of Healthcare Providers and Systems (Giordano et al., 2010). In addition to survey items on experiences with specific components of care, it includes two items that assess patients' willingness to recommend the hospital (on a scale from 1 to 4) and their global rating of the hospital on a scale from 0 to 10. HCAHPS or HCAHPS-derived patient experience surveys are used in public reporting practices and value based purchasing around the globe (Zhu et al., 2015; Elliott et al., 2009a; Weech-Maldonado et al., 2008; Damiano et al., 2004). Many substantive-methodological researchers have however created awareness of measurement challenges and limitations of the HCAHPS. Much scholarly attention has gone into examining if better experiences across groups present superior care or can be explained by differences in response style (Zhu

et al., 2015; Elliott et al., 2009a; Weech-Maldonado et al., 2008; Damiano et al., 2004). Others have shown that the correlation structure of the items measuring specific components of care varies by type of hospital (O'Malley et al., 2005b) and across countries (Orindi et al., 2017). Both streams of research imply that caution is warranted when comparing findings across groups. It has also been investigated how the specific experiences of care are related to the global ratings, and how this association varies across groups. Elliott et al. (2009b) showed that communication with nurses was the most important dimension overall to explain the hospital rating of 0 to 10, but also found that the association between these and other specific experiences and the hospital rating substantially varied by type of hospitalization. Hays et al. (2018) recently showed for the CAHPS Medicare survey that the association between experiences with specific care components and the global doctor rating were stronger for beneficiaries with worse self-rated health. The authors concluded that it is important to understand whether specific experiences have similar relevance across subgroups, as different drivers of overall experiences may require different targeted approaches to focus on. With a growing number of studies that are evaluating and comparing patient experiences across countries, including one large study that used the HCAHPS in 8 countries (Aiken et al., 2012), we sought to understand whether the previously established differential magnitude of associations between components of care and global ratings across groups can also be seen across countries.

In addition, we deemed it important to explore from the association between specific experiences and the global ratings what can be learned for reporting these global ratings in international research. Originally, it was thought that these global ratings had the potential to be used as summative measures of patient experiences with care (Elliott et al., 2009a). Later it was shown that these global ratings might be more subjective ratings prone to response style, whereas differences in the more specific experiences are more likely to reflect true differences rather than differences in expectation or scale use (Weinick et al., 2011). Either way, many academics have used in particular the 0 to 10 hospital rating as the major study outcome. Associations have been found

between better hospital ratings and several system and individual patient or provider aspects (Aiken et al., 2012, 2017; Boulding et al., 2011; Bruyneel et al., 2015; Jha et al., 2008). Patient hospital global ratings are almost always converted into a dichotomous score. Scores are mostly collapsed as 0-8 versus 9-10. Jha and colleagues argued that such dichotomization correlated highly with patients' willingness to recommend the hospital (Jha et al., 2008). Other authors have indicated that dichotomization was said to best reflect the negatively skewed responses (Bruyneel et al., 2015; Aiken et al., 2012). Hawkins et al. (2016) in their study on linkages between patient ratings and social media primarily focused on the percentage of patients rating the hospital at least 9 out of 10, but also considered ratings of 7 and higher. In a recent Canadian study, (Kemp et al., 2015) used the percentage of patients responding with a rating of 10 only, with the idea that ratings of 0-9 present a potential opportunity for improvement. This may be problematic since 10-only CAHPS ratings have shown increased sensitivity to response tendency (Weech-Maldonado et al., 2008; Damiano et al., 2004). Different methods are also observed in public reporting practices. The Alberta Health Services' public reporting of strategic performance measures uses the percentage of patients who rated the hospital 8, 9 or 10. The US Centers for Medicare & Medicaid Services (Centers for Medicare and Medicaid Services, 2016) on the other hand consider the percentage of patients who gave their hospital a rating of 9 or 10 for purposes of public reporting, an approach that is also used in Belgium.

This study extends the existing literature by (1) examining the association between specific patient experiences and global ratings, and the consistency of this association across countries, and (2) exploring what can be learned from possible differential magnitudes of these associations across countries for reporting overall ratings that allow fair comparisons across countries.

4.2 Methods

4.2.1 Study design, setting and sample

We used data from the 12-country cross-sectional Registered Nurse Forecasting (RN4CAST) study (Sermeus et al., 2011). We focussed on the patient survey component of the study conducted in the year 2009 in eight European countries, using a one-day census. Data for Germany were excluded from the analysis as they did not have nursing unit identifiers, which was an important aspect in our anlysis, as described later. Data for this study were from 11289 patients from 824 nursing units in 186 hospitals in 7 countries (Belgium, $n = 2623$; Finland, $n = 1947$; Ireland, $n = 285$; Greece, $n = 847$; Poland, $n = 4136$; Spain, $n = 470$; Switzerland, $n = 981$). The full details of the data collection protocol have been published (Sermeus et al., 2011). The survey instrument was subject to rigorous translation and a piloting and validation process prior to data collection to reduce threats to data validity and to improve instrument reliability (Squires et al., 2012). The overall patient response rate was 73% and the hospital response rate averaged 63%, but these varied across countries (see Table 4.1). To cope with hospital non-response, a check for representativeness (i.e., hospital type, size) was performed at the planning stage. When necessary, some corrective measures (like extra motivating hospitals to participate) were taken. For patient non-response, no corrective measures could be taken at the analysis stage as no information is available on the non-respondents. If this information was available, statistical weighting could be used to adjust for the non-responses. We address item non-response (or item missing data) in Section 4.2.2.

4.2.2 Measures

The outcome of interest was overall patient satisfaction measured using two items: (1) patient overall rating of hospital on a scale of 0 (worst) to 10 (best); and (2) patient willingness to recommend the hospital to friends and family, rated on a 4-point scale of “definitely no”(=1), “probably no”(=2), “probably yes”(=3), “definitely yes”(=4).

Table 4.1: RN4CAST Patient Survey Data: Hospital and patient response rates for the eight European countries

Country	Eligible hospitals	Target number of hospitals	Number of Hospitals invited	Number of Hospitals in general study	Number of Hospitals in patient survey	Response rate
Belgium	104	30	104	56	51	53.8
Finland	49	34	34	32	31	94.1
Germany*	1104	31	100	49	12	49.0
Greece	110	29	57	24	17	42.1
Ireland	33	33	33	30	10	90.9
Poland	195	30	32	30	30	93.8
Spain	185	34	37	33	15	89.2
Switzerland	88	35	41	35	25	85.4

Country	Patient survey		
	Distributed	Retrieved	Response Rate
Belgium	3870	2623	67.8
Finland	3752	1947	51.9
Germany*	390	262	67.2
Greece	1334	616	46.2
Ireland	307	285	92.8
Poland	4416	4136	93.7
Spain	N/A	470	
Switzerland	1458	997	68.4
	15527	11336	73.0**

Notes: *Excluded from the analyses; **Calculated without Spain.

Main explanatory variables were indicators for patient experience with care. Sixteen specific items were used in the HCAHPS, five of which were conditional on screeners. Only the 11 questions that applied to all patients were used in the present analyses (Table 4.2). They were measured on a 4-point Likert scale with categories "never"(=1), "sometimes"(=2), "usually"(=3), "always"(=4), except for the last two items which had "yes"(=1) or "no"(=0) responses. The instrument also had a question on patient's self-rated health measured on a 5-point Likert scale ["excellent"(=1), "very good"(=2), "good"(=3), "fair"(=4), "poor"(=5)]. Individual patient characteristics like gender and age were not collected. These covariates were selected on the basis of previous research in which they have been found to relate to patient experience domains (Keller et al., 2005; Orindi et al., 2017).

Data on nursing unit speciality (surgical= 0; medical or mixed = 1) and hospital characteristics (teaching status: 1=teaching hospital, 0=non-teaching; technological level: 1=advanced, 0=not advanced; and size, that is, number of beds) were also provided by hospital management from their own institutional records. These unit and hospital characteristics were adjusted for in the analyses. Frequency distribution showed that the data were negatively skewed (see Figure 4.1) and with some missing observations. For all 11 items and both outcomes, missing data ranged from a low of 0.4% for *Nurses treat with courtesy and respect* (Item 1) to a high of 9% for *Have you gotten information in writing about what symptoms or health problems to look out for after you leave the hospital?* (Item 11; Table 4.3). Reasons for patients not responding were unknown. Missing data values were multiply imputed (Molenberghs and Kenward, 2007; Little and Rubin, 2014). Details on how this was implemented is provided in Section 4.2.5.

Table 4.2: RN4CAST Patient Survey Data: Wording for the 11 Items Used to Measure Patient Experience With Care

	Item
1:	How often did nurses treat you with courtesy and respect?
2:	How often did nurses listen carefully to you?
3:	How often did nurses explain things in a way you could understand?
4:	After you called for assistance, how often did you get help as soon as you wanted it?
5:	How often did doctors treat you with courtesy and respect?
6:	How often did doctors listen carefully to you?
7:	How often did doctors explain things in a way you could understand?
8:	How often were your room and bathroom kept clean?
9:	How often was the area around your room quiet at night?
10:	Have doctors, nurses or other hospital staff talked with you about your care after you leave the hospital?
11:	Have you gotten information in writing about what symptoms or health problems to look out for after you leave the hospital?

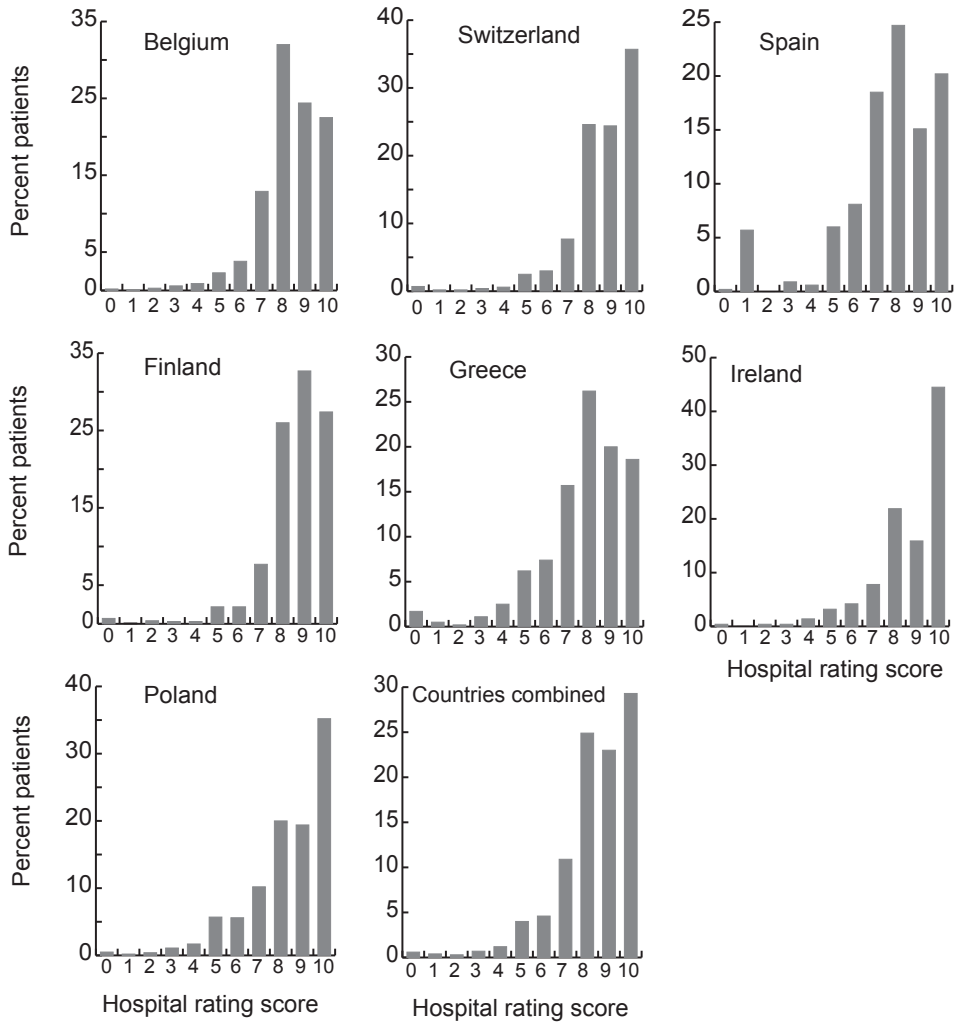

Figure 4.1: RN4CAST Patient Survey Data: Distribution of answers for patients' hospital ratings for each country separately and for all seven countries combined.

Table 4.3: RN4CAST Patient Survey Data: Frequencies of observed and missing observations for all 11 experience items, self-rated health status, hospital rating and hospital recommendation.

Item	# observed	# missing	% missing
Nurses courtesy and respect	11 243	46	0.41
Nurses listen carefully	11 232	57	0.50
Nurses explain things well	11 215	74	0.66
Got help soon	11 076	213	1.89
Doctors courtesy and respect	11 205	84	0.74
Doctors listen carefully	11 187	102	0.90
Doctors explain things well	11 115	174	1.54
Clean room and bathroom	11 072	217	1.92
Room quiet at night	11 052	237	2.10
Care after hospital	10 543	746	6.61
Symptoms to look out for	10 278	1 011	8.96
Health status	10 854	435	3.90
Hospital rating	10 954	335	3.00
Recommending hospital	10 916	373	3.30

4.2.3 Analysis

We studied the distribution of observed scores across the 11-point answer scale for hospital rating in each country. We also examined the Spearman correlation between patients' hospital rating and their willingness to recommend the hospital. We evaluated which hierarchical level in the design of our study (country, hospital, nursing unit, and patient) contributed to overall satisfaction most by estimating the intraclass correlation coefficient (ICC; Killip et al., 2004), and the portion of the total variance explained for each of the two global ratings.

We then took a three-step analytic approach for evaluating our research objectives.

First, to examine the associations between HCAHPS items measuring specific experiences and the overall rating and recommendation, we fitted a three-level model including a hospital random effect, a nursing unit random effect, a country fixed effect taking Belgium as the reference group, the eleven specific patient experience items as main explanatory variables, and the two global ratings as outcome variables. We also adjusted for the effect of self-rated health status, unit specialty and hospital characteristics. See Molenberghs and Verbeke (2005) on multilevel models for discrete longitudinal data focusing on biostatistical applications. A random effects proportional odds model was chosen (McCullagh, 1980), as the response categories are ordered. As both outcomes were highly (negatively) skewed (Figure 4.1), we specified three candidate models, namely probit, logit, and complementary log-log (clog-log). We used latent variable formulation. The latent variable model assumes that for each ordered categorical satisfaction indicator, there is an underlying continuous, unobserved variable y^* such that the exact continuous measurements of y^* are not available, but are related to the observed ordered categorical variable y as follows. Let y^*_{iuh} be the continuous measurement underlying the observed ordinal hospital rating for patient i from nursing unit u from hospital h, then for $k = 0, 1, 2, \ldots 10$,

$$y_{iuh} = k \quad \text{if } \alpha_k \leq y^*_{iuh} < \alpha_{k+1},$$

where $-\infty = \alpha_0 < \alpha_1 < \alpha_2 < \ldots < \alpha_{11} = \infty$, and $\alpha\{s\}$ are the thresholds which are also estimated from the model. Similarly, for hospital recommendation, but with $k = 1, 2, 3, 4$.

Then we specify the 3-level main effects model as,

$$\begin{aligned} y^*_{iuh} &= \alpha + \beta_0^{(c)} + b_h + b_{uh} + \beta_1 Item1_{iuh} + \beta_2 Item2_{iuh} + \ldots + \beta_{11} Item11_{iuh} \\ &\quad + \beta_{12} Health_{iuh} + \beta_{13} Specialty_{uh} + \beta_{14} Beds_h + \beta_{15} Tech_h + \beta_{16} Teach_h + \varepsilon_{iuh}, \\ &\quad \varepsilon_{iuh} \sim N(0, \sigma^2), \quad i = 1, \ldots, N_{uh}, \quad u = 1, \ldots, N_h, \quad h = 1, \ldots, H, \end{aligned} \tag{4.1}$$

such that N_{uh} is the number of patients in nursing unit u of hospital h, N_h is the number of nursing units in hospital h, and H is the number of hospitals, $\beta_0^{(c)}$ is the country fixed effect, βs are the coefficients associated with the covariates, and ε_{iuh} are error measurements with standard deviation, σ. b_h and b_{uh} are, respectively, the hospital and nursing unit random effects. Further, ε_{iuh} are assumed to be independent of the random intercepts.

Second, to investigate whether the association between the 11 specific items and the two global ratings was similar across countries, we extended the model (4.1) to allow for a differential effect of each item across seven countries by including country-items interactions. With Belgium as the reference, the three-level model with country-items interactions can be expressed as,

$$\begin{aligned} y^*_{iuh} &= \alpha + \beta_0 + b_h + b_{uh} + \beta_1 Item1_{iuh} + \beta_2 Item2_{iuh} + \ldots + \beta_{11} Item11_{iuh} \\ &\quad + \beta_0^{(c)} + \beta_1^{(c)} Item1_{iuh}^{(c)} + \beta_2^{(c)} Item2_{iuh}^{(c)} + \ldots + \beta_{11}^{(c)} Item11_{iuh}^{(c)} \\ &\quad + \beta_{12} Health_{iuh} + \beta_{13} Specialty_{uh} + \beta_{14} Beds_h + \beta_{15} Tech_h + \beta_{16} Teach_h + \varepsilon_{iuh} \\ &\quad \varepsilon_{iuh} \sim N(0, \sigma^2), \quad i = 1, \ldots, N_{uh}, \quad u = 1, \ldots, N_h, \quad h = 1, \ldots, H, \end{aligned} \tag{4.2}$$

where $Item1_{iuh}^{(c)}$ equals $Item1_{iuh}$ if the observation iuh belongs to country c (CH= Switzerland, ES=Spain, FI=Finland, GR=Greece, IE=Ireland, PL=Poland) and it is zero otherwise. Therefore, $Itemj_{iuh}^{(c)}$ is the product of a dummy variable for each country and response score for item $j (j = 1, 2, \ldots, 11)$. For the reference group (i.e., Belgium), $Itemj_{iuh}^{(c)}$ is always zero.

Model (4.2) is especially advantageous as we can study the differential effect of each item j across countries. This can be determined by studying the significance of β_j^c; that is, by testing the hypothesis $H_0 : \beta_j^{(CH)} = \beta_j^{(ES)} = \beta_j^{(FI)} = \beta_j^{(GR)} = \beta_j^{(IE)} = \beta_j^{(PL)} = 0$. For instance, if the parameter $\beta_2^{(ES)}$ is significant, then $Item2$ (i.e., *nurse listened carefully to you?*) is different in Spain compared to Belgium. The likelihood ratio test (LRT) was then used to test the interaction terms. Next, we estimated the effect of each item on overall satisfaction for each country. If regression coefficients are equal, then patients' opinion about the hospital, as expressed by the response, were driven by the same arguments, that is, it implies invariance of these associations across the countries.

Third, to investigate what the associations between specific experiences and the global ratings within countries imply for determining cut-off points that reflect a high probability of positive hospital ratings, we plotted, from the model without interaction terms (i.e., Model 4.1), the regression lines of the country-specific scores against the overall hospital rating. We performed this analysis under an assumption of homogeneous use and interpretations of response categories of these specific experiences across patients within and between countries. Implications are discussed in Section 4.4.

4.2.4 Model Evaluation

We compared the performance of the three links by comparing the value of the log-likelihood of the main effects model (i.e., Model 4.1) with probit, logit, and complementary log-log (Clog-log) link functions. For each link function, models were evaluated using deviance test; we first compared an intercept only model to model with main effects (Model 4.1), then main effects model (Model 4.1) to model with interactions (Model 4.2). To check the predictive ability of the models, we first compared the observed and predicted proportions of patients rating their hospitals high or recommending the hospital at each score in the outcome scale. Next, following Agresti (2010), we estimated the Spearman correlation between the observed responses and the estimated means of the conditional distributions from the model. This mimics the multiple

correlation in multiple regression modeling (pp. 65–66). We also obtained the concordance index for the models with and without interactions (Harrell et al., 1996). The concordance index estimates the probability that the predictions and the outcomes are concordant, that is, the observation with the larger $y-$value also has a stochastically higher set of model-estimated probabilities.

4.2.5 Computational Aspects

Multiple imputation (Little and Rubin, 2014) was done in R 3.3.2 (R Core Team, 2014) using the `aregImpute` function from the Hmisc package (Harrell Jr et al., 2016). `aregImpute` uses flexible parametric additive regression spline models to predict target variable. Fisher's optimal scoring algorithm is used to transform the categorical variables. Both global outcomes, all 11 specific experience items, health status, unit specialty and hospital characteristics variables were included in the imputation model. All analyses were performed in SAS software Version 9.4 (SAS Institute, Cary NC) using the GLIMMIX and NLMIXED procedures. Following Rubin's rules (Little and Rubin, 2014), the MIANALYZE procedure was then used to combine the inferences into a single one. That is, by averaging the parameter estimates over the multiple imputations and obtaining their variances. The NLMIXED procedure was used to fit the complementary log-log model as we experienced computational problems with the GLIMMIX procedure. Since convergence can be challenging, it was necessary to try several NLOPTIONS arguments in GLIMMIX. For NLMIXED, we made use of the QPOINTS option and also the double-dogleg optimization technique as it converges properly and takes less time compared to the quasi-Newton technique. The global outcomes were defined such that we modelled the probabilities of patients rating their hospitals high and recommending the hospital to friends and family. Probabilities modelled were cumulated over the higher ordered values.

4.3 Results

4.3.1 Descriptive Findings

For the hospital rating, the frequency distribution of answers across the 11-point scale showed highly left skewed distributions with hospital ratings concentrated in the top three scores (i.e., ≥ 8), except for Spain and Greece where they concentrated in the top 4 scores (i.e., ≥ 7). The proportion of patients rating their hospital 10 ranged from 19% in Greece to 45% in Ireland. For all countries except Finland, 8 was selected more often than 9 (Figure 4.1). For all seven countries combined about 29%, 52%, 77%, 88% and 93% of patients rated their hospital care high using a cut-off point of 10, 9, 8, 7 and 6, respectively. Using cut-off point 9, the proportion of patients rating their hospitals high ranged from 35% in Spain to 61% in Ireland and Finland (Figure 4.2).

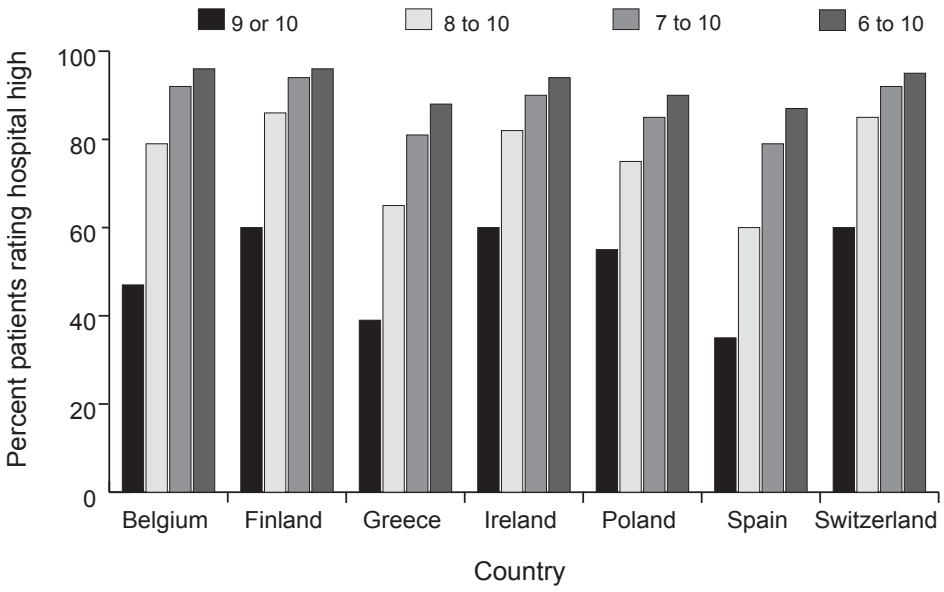

Figure 4.2: RN4CAST Patient Survey Data: Percent of Patients Rating the Hospital High at 9, 8, 7 and 6 Cut-off Points for each Country Separately.

Overall the two global ratings had a high positive correlation (Spearman

rho=0.62, 95%CI 0.61-0.63), but this varied from 0.47 in Spain to 0.66 and 0.68 in Greece and Poland, respectively (Figure 4.3). The variation in hospital global rating (hospital recommendation) explained at the country, hospital, nursing unit, and individual levels was 5% (4.5%), 4.4% (5.8%), 6.6% (6.3%), and 84% (83%), respectively. Thus, the individual level by far explained most variation in both outcomes while the hospital level explained least variation in hospital rating. For hospital recommendation, the country level explained least variation.

About one in ten (12.6%) patients rated their health as being very good or excellent. 44% nursing units were surgical, and approximately 20% and 29% of the hospitals were university and technologically advanced, respectively. The mean (median) number of beds in a hospital was 467 (397).

4.3.2 Association between Specific Patient Experiences and Global Ratings?

Table 4.4 presents the effect of experience items on the two satisfaction measures for probit, logit, and Clog-log models using a 3-level main effects model (i.e., equation 4.1). It shows that all experience items were significantly positively associated with patients' propensity to rate their hospitals higher scores. Items of *nurses courtesy and respect, nurses listen carefully, got help soon, clean room and bathroom, room quiet at night, doctors/nurses talked about care after hospital and symptoms to look out for* contributed most to patient's rating of the hospital. This was very similar to patients' propensity to recommend the hospital to friends and family. Better health status was also positively associated with the two global ratings, as was unit specialty with the overall rating.

Whereas the probit, logit, and Clog-log models resulted in similar inference, they showed some small differences. For instance, unlike in the probit model, the logit model indicated a significant effect of unit specialty for hospital rating. As expected the parameter estimates for the logit model are slightly larger. Further, based on the log-likelihood the logit model provided a better fit (rating= -16853.22; recommend= -7885.24) compared to a probit model

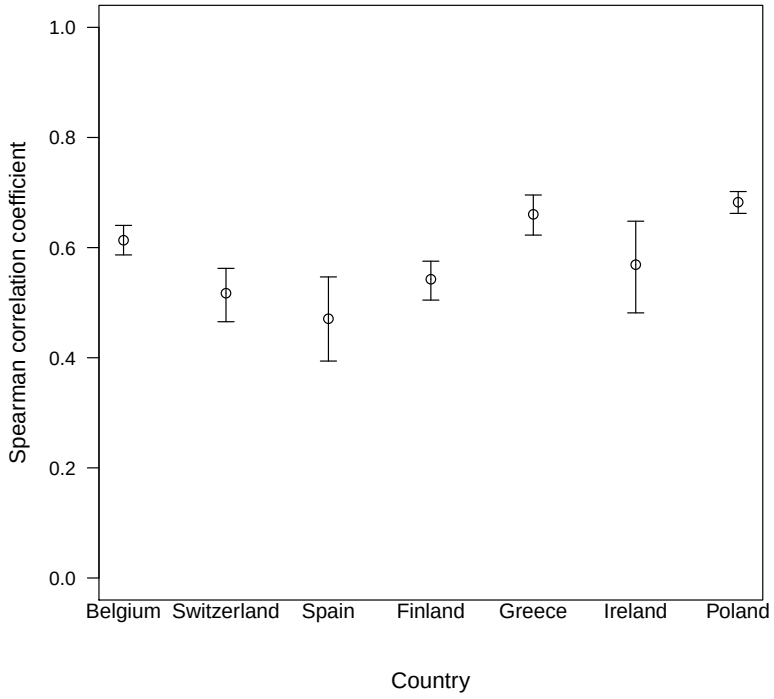

Figure 4.3: RN4CAST Patient Survey Data: Spearman correlation coefficient between the two outcomes and the associated 95% confidence interval for each country separately.

(rating = −16985.14; recommend = −7934.20) and a complementary log-log model (rating=convergence problem; recommend = −8132.01; Table 4.4). Thus, complentary log-log provided the worst fit. From now on we base our inferences on the logit model.

The model fit and predictive power is summarized in Section 4.3.4. The observed and predicted proportion of patients that score their hospitals more than 9, 8, 7, and so on, using a logit link, are presented in Figure 4.5 for each country. It can be seen that the model describes the data reasonably well. Similar observations were made for hospital recommendation as well.

Table 4.4: RN4CAST Patient Survey Data: Association between eleven experience items and the outcomes using main effects model (Model 4.1) for probit, logit and Clog-log links. Data are adjusted standardized parameter estimates (standard error). Values with asteric are statistically significant.

Variable	Hospital rating		Recommend hospital		
	Probit	Logit	Probit	Logit	clog-log
	Est (se)	Est (se)	Est (se)	Est (se)	Est (se)
Nurses courtesy and respect	0.119 (0.014)*	0.227 (0.024)*	0.131 (0.017)*	0.252 (0.029)*	0.099 (0.016)*
Nurses listen carefully	0.167 (0.017)*	0.295 (0.028)*	0.119 (0.017)*	0.203 (0.030)*	0.114 (0.016)*
Nurses explain things well	0.078 (0.014)*	0.155 (0.024)*	0.083 (0.016)*	0.152 (0.029)*	0.072 (0.015)*
Got help soon	0.124 (0.014)*	0.237 (0.024)*	0.100 (0.016)*	0.182 (0.029)*	0.093 (0.015)*
Doctors courtesy and respect	0.085 (0.015)*	0.162 (0.026)*	0.109 (0.019)*	0.199 (0.033)*	0.093 (0.018)*
Doctors listen carefully	0.096 (0.016)*	0.162 (0.028)*	0.059 (0.019)*	0.106 (0.034)*	0.055 (0.018)*
Doctors explain things well	0.088 (0.014)*	0.166 (0.025)*	0.102 (0.017)*	0.179 (0.029)*	0.096 (0.016)*
Clean room and bathroom	0.117 (0.012)*	0.211 (0.020)*	0.120 (0.013)*	0.218 (0.024)*	0.105 (0.013)*
Room quiet at night	0.143 (0.013)*	0.253 (0.022)*	0.120 (0.015)*	0.215 (0.027)*	0.112 (0.014)*
Care after hospital	0.224 (0.026)*	0.373 (0.044)*	0.204 (0.036)*	0.343 (0.061)*	0.202 (0.034)*
Symptoms to look out for	0.159 (0.028)*	0.275 (0.046)*	0.122 (0.035)*	0.229 (0.061)*	0.086 (0.033)*
Health status	0.056 (0.012)*	0.090 (0.021)*	0.040 (0.015)*	0.065 (0.026)*	0.038 (0.015)*
Unit specialty	0.048 (0.026)	0.092 (0.045)*	-0.007 (0.032)	-0.011 (0.056)	-0.005 (0.034)
Beds	0.021 (0.024)	0.023 (0.041)	0.045 (0.035)	0.084 (0.062)	0.052 (0.027)
Tech	0.068 (0.048)	0.120 (0.083)	0.077 (0.069)	0.146 (0.120)	0.053 (0.051)
Teach	-0.015 (0.061)	-0.009 (0.105)	0.082 (0.083)	0.151 (0.147)	0.039 (0.064)
Country					
Switzerland	0.125 (0.060)*	0.246 (0.101)*	0.379 (0.081)*	0.679 (0.144)*	0.342 (0.059)*
Spain	-0.487 (0.080)*	-0.718 (0.14)*	-0.123 (0.106)	-0.226 (0.189)	-0.102 (0.082)
Finland	0.283 (0.053)*	0.509 (0.089)*	0.207 (0.074)*	0.440 (0.131)*	0.091 (0.057)
Greece	0.052 (0.065)	0.129 (0.110)	-0.06 (0.083)	0.007 (0.149)	-0.220 (0.067)*
Ireland	0.542 (0.095)*	0.937 (0.164)*	0.489 (0.128)*	0.893 (0.228)*	0.440 (0.104)*
Poland	0.013 (0.053)	0.070 (0.090)	-0.168 (0.071)*	-0.308 (0.126)*	-0.145 (0.051)*
log-likelihood	−16985.14	−16853.22	−7934.20	−7885.24	−8132.01

4.3.3 Generalizability across Countries of Contribution of Specific Patient Experiences to Global Ratings

Using the model with country-items interactions (Model 4.2), a differential effect across countries was detected for hospital rating for two items: *clean room and bathroom* ($p = 0.0152$) and *care after hospital* ($p = 0.0035$; Table 4.5). For recommendation of the hospital to friends and family, differential effects were detected for items of *nurses courtesy and respect* (p=0.0463), and *clean room and bathroom* (p=0.0097; Table 4.5).

Figure 4.4 presents the regression coefficients and their 95% confidence intervals related to each item for hospital rating by country. The figure shows that the regression coefficients were similar across countries for all items except *clean room and bathroom* (Greece and Poland are significantly higher than Belgium) and *care after hospital* (Spain higher than Belgium, Switzerland and Finland). The equal coefficients suggest patients' opinion about the hospital as expressed by the response were driven by the same arguments, hence invariance across the countries.

Table 4.5: RN4CAST Patient Survey Data: Differential effect of experience items across countries: Minus twice loglikelihood, likelihood ratio test statistic comparing the full country-items interactions model predicting hospital global outcome ratings with that without an item's corresponding interaction term for each of the eleven experience items (G^2) and associated $p-$values using 3-level multiple logit model*

	Hospital rating			Recommending hospital		
Model	$-2ll$	G^2	$p-$value	$-2ll$	G^2	$p-$value
Full interaction	33464.1			15650.7		
Model without interaction for:						
Nurse courtesy	33472.85	8.77	0.1869	15663.47	12.82	0.0463
Nurse listen	33470.78	6.70	0.3495	15661.70	11.05	0.0853
Nurse explain	33473.29	9.21	0.1621	15656.22	5.57	0.4695
Nurse help soon	33469.17	5.09	0.5323	15660.79	10.14	0.1205
Doctor courtesy	33467.51	3.43	0.7533	15656.93	6.28	0.3904
Doctor listen	33470.10	6.02	0.4210	15658.58	7.93	0.2455
Doctor explain	33476.67	12.59	0.0500	15655.14	4.49	0.6093
Clean room	33479.83	15.75	0.0152	15667.57	16.92	0.0097
Room quiet	33475.02	10.94	0.0902	15661.25	10.60	0.1016
Careafter hospital	33483.50	19.42	0.0035	15660.30	9.65	0.1425
Symptoms to look for	33476.12	12.04	0.0611	15654.19	3.54	0.7440

*Model predicting hospital outcome ratings with interaction terms for country and items (Model 4.2).

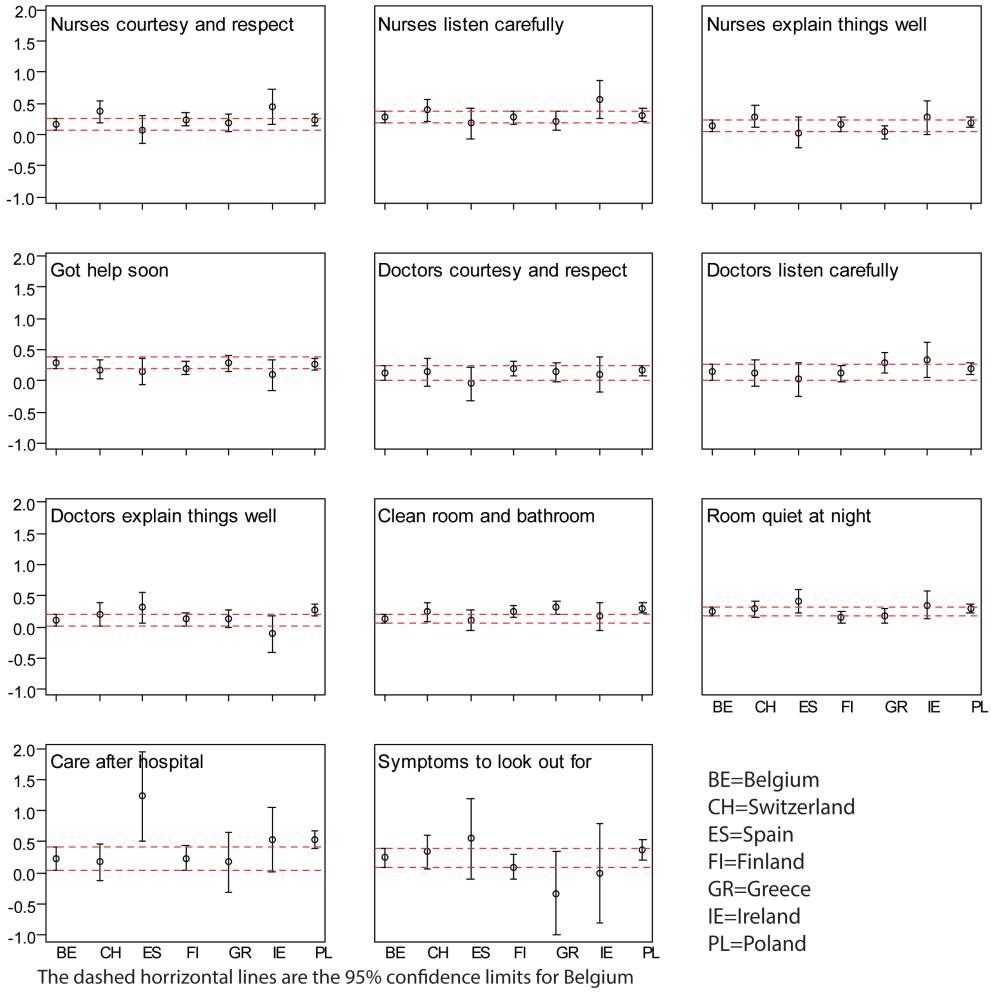

Figure 4.4: RN4CAST Patient Survey Data: 95% Confidence interval for regression coefficients for hospital rating in each country (BE=Belgium, CH=Switzerland, ES=Spain, FI=Finland, GR=Greece, IE=Ireland, PL=Poland) from the logit model. Complete item listing and wordings are presented in Table 6.2.

4.3.4 Summary of model fit and predictive power

The deviance test indicated that Model 4.1 was a better fit than an intercept only model for both hospital rating ($\chi^2 = 4369.6, df = 22; p < 0.0001$) and hospital recommendation ($\chi^2 = 2926.9, df = 22; p < 0.0001$); and that adding interaction terms to Model 4.1 improved its fit for both hospital rating ($\chi^2 = 242.4, df = 66, p < 0.0001$) and recommending the hospital ($\chi^2 = 119.8, df = 66, p < 0.0001$). However, Figure 4.5 presenting the observed and predicted proportions rating the hospital high indicate that whereas there is statistical significance of the interaction terms, the predicted models with and without interactions seem to be close, suggesting low impact of the significant interaction terms. Table 4.6 shows correlations between the observed scores and the estimated means of the conditional distributions from the models with and without interactions for each outcome as well as the concordance indices. The table shows moderate to high correlations between observed and estimated scores, and that adding interaction terms provided no substantive improvement. The overall concordance index for the model without interactions show that, for 80.4% (88.7%) of the untied pairs on hospital rating (recommendation to friends and family), the observation with a higher outcome score also had a stochastically higher estimated distribution.

Table 4.6: RN4CAST Patient Survey Data: Predictive Power for Multilevel Cummulative Logit Models fitted to for each Country Separately and for all Seven Countries Combined.

Country	Spearman correlation			
	Hospital rating		Hospital recommendation	
	No-interaction	With interaction	No-interaction	With interaction
Belgium	0.587	0.593	0.550	0.548
Switzerland	0.553	0.554	0.485	0.487
Spain	0.563	0.611	0.493	0.433
Finland	0.579	0.586	0.567	0.562
Greece	0.609	0.628	0.582	0.587
Ireland	0.569	0.584	0.531	0.529
Poland	0.643	0.640	0.622	0.622
Overall	0.621	0.624	0.572	0.572
	Concordance index			
Belgium	0.795	0.842	0.911	0.931
Switzerland	0.872	0.838	0.939	0.868
Spain	0.802	0.849	0.898	0.929
Finland	0.793	0.833	0.913	0.915
Greece	0.792	0.813	0.856	0.855
Ireland	0.829	0.773	0.897	0.826
Poland	0.821	0.770	0.895	0.886
Overall	0.804	0.797	0.887	0.882

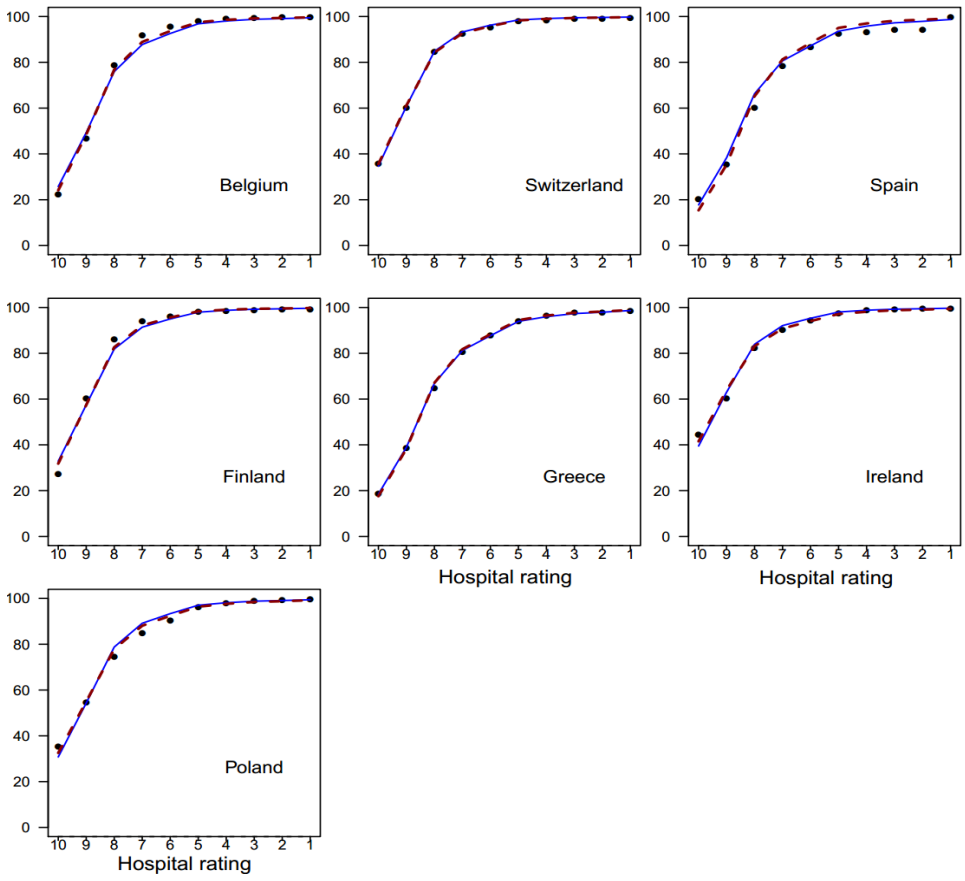

Figure 4.5: RN4CAST Patient Survey Data: Observed (points) and predicted percent of patients rating their hospitals high at different levels in the hospital rating scale in each country based on Model without (solid line) and with interactions (dotted line) using a logit link function.

4.3.5 Do cut-off point(s) allow for an equitable comparison of patient hospital ratings across countries?

Figure 4.6 presents the country-specific and overall curves estimated from our model assessing the association between the 11 specific experience items and the overall hospital rating. These are cumulative probabilities from the ordered logit model. This means that, at score 10, we plotted the probability of giving the hospital a rating of 10. At score 9, we plotted the probability of rating the hospital 9 or 10. At score 8, we plotted the probability of rating the hospital 8, 9 or 10, and so on. Findings showed that, for example, at a cut-off point of 10 the predicted probability of positive experiences, excluding Spain and Greece, varied from 0.257 in Belgium to 0.395 in Ireland. Spain and Greece have much lower probabilities at this cut-off point, 0.177 and 0.185 respectively. Looking at the various cut-off points, these results suggest that (1) the commonly used cut-off point of 9 represents a much lower probability of positive specific experiences than a cut-off point of 8, after which the increase in probability of positive experiences from lowering the cut-off point becomes less steep, (2) excluding Spain and Greece, the range in the probability of positive experiences is lower at cut-off point 9 (0.493 to 0.630) compared to cut-off point 8 (0.760 to 0.848), and (3) for Spain and Greece to have a comparable probability of high hospital rating to the other countries, a cut-off point of 7 would be a good choice.

However, because this analysis assumes homogeneity in using and interpreting response categories, and because it is based on 11 specific experience items only, these results should be interpreted with caution, as discussed below.

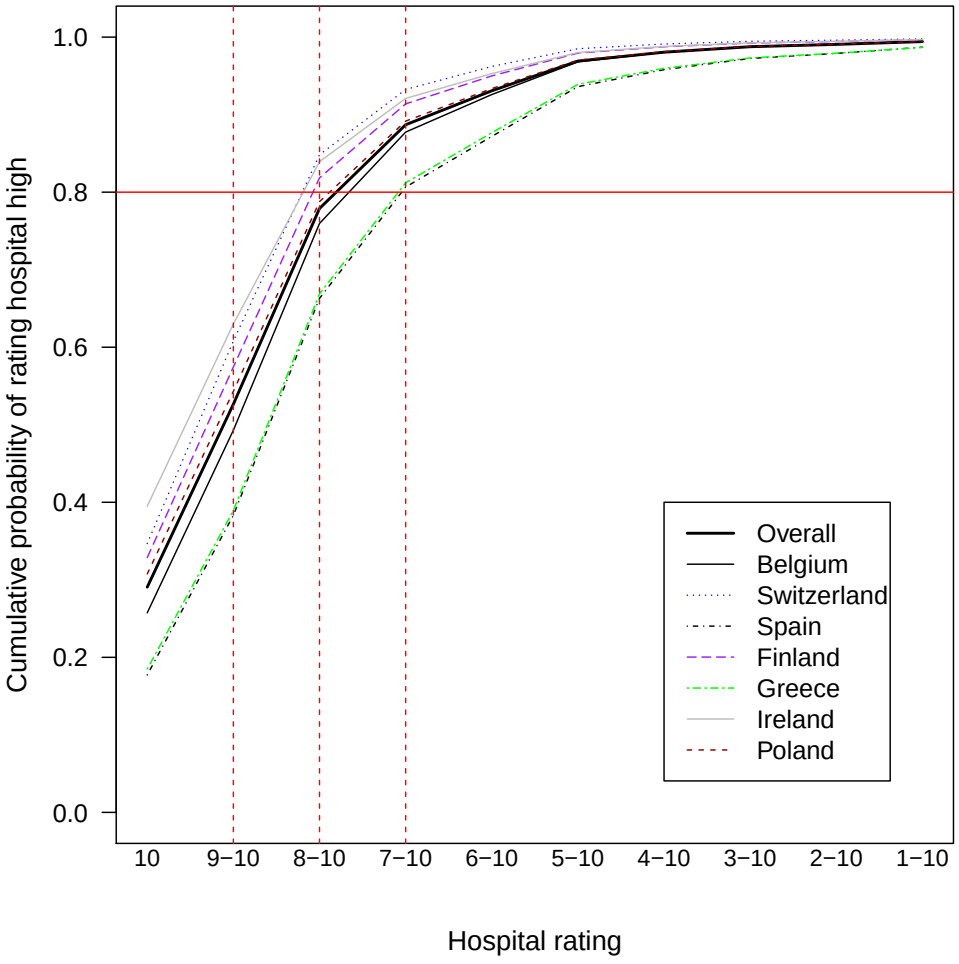

Figure 4.6: RN4CAST Patient Survey Data: Estimated Overall and Country Specific Probability of Rating the Hospital Higher at different Levels of the Outcome Profile from Model 4.1 Using Logit Link Function.

4.4 Discussion

We have illustrated from HCAHPS data collected within a large European study that patients' responses to items measuring specific experiences are reflected in their global ratings of the hospital. Although some differences in the importance of items were noticed across countries, a helpful finding in our study is that overall all items were strongly positively associated with higher global ratings and their predictive power was high. In the development of the HCAHPS a range of exploratory and confirmatory factor analytic techniques were used to assess construct validity of the specific survey items but with the explicit objective to make the survey as short as possible to reduce administration costs and to allow room for users to add customized content. This meant that over half of the items identified in an early phase to reflect patient experiences were dropped from the questionnaire (O'Malley et al., 2005b; Keller et al., 2005). A recent study found that feedback shared on Yelp strongly correlates with the overall hospital ratings and covers additional information not included in the HCAHPS. The domains such as amenities, cost of hospital visit, billing, ancillary testing, facilities, family member care among others, though reported on Yelp, were omitted from the CAHPS surveys as they do not reflect quality of care (Ranard et al., 2016). The authors put forward that it is unclear whether patients base their overall rating on the HCAHPS items. Our findings however provide support for the limited number of items used to measure specific patient experiences. Previous research examining the contribution of specific experiences to global ratings repeatedly found that items related to nurse communication contributed strongest to these global ratings (Bruyneel et al., 2015; Boulding et al., 2011; Elliott et al., 2009b). Here, these items also contributed significantly, but communication about discharge was by far the largest contributor. Clearly, for these European countries we cannot corroborate a conclusion from previous US research that there is little potential for discharge communication to contribute to variability in overall hospital ratings because it may be perceived by patients as obvious, routine or important for many conditions (Elliott et al., 2009b).

We studied the comparison across countries of a uniformly dichotomized

0-10 patient rating of the hospital. Dichotomizing continuous variables may reduce power to detect associations, conceal possible non-linearity, increase risk of false positives, and underestimate variation (Altman and Royston, 2006). Our findings illustrate the importance of a conceptual and statistical motivation for determining cut-off points on the continuum from 0 to 10. Traditional methods of straightforward comparison of dichotomized HCAHPS ratings may not reflect true positive patient experiences and are likely to result in unfair comparisons because respondents use the range of values between 0 and 10 differently. These findings raise concerns over previous between-nation analyses (Aiken et al., 2012; Bruyneel et al., 2015; Aiken et al., 2017).

Previous studies have repeatedly shown that there are systematic differences in the way patients use and interpret response categories of specific experience items as well as global ratings. As such, we cannot say for our data that all patients have the same true patient experience irrespective of any individual characteristics or cultural aspects, i.e. our assumption of homogeneity does not hold. Therefore, the analysis performed in the third step should be viewed as exploratory and hypothesis generating, and needs further study. In addition, in our study, the derivation of cut-off points was based on 11 items measuring specific patient experiences. This means that the respective cut-off points represent a high probability of positive specific experiences, but for these 11 items only. The specific experiences do not represent a gold standard of patient-centred care. The issues presented here may apply to any other self-reported outcome measure. We therefore encourage researchers interested in patient reported experience measures to search for cut-off points that, as precisely as possible, reflect positive experiences. Also of importance is to consider the trade-off between the probability of positive experiences and remaining variability at those cut-off points within countries.

A promising strategy to detect and adjust for differences in response category cut-off points is available. Anchoring vignettes provide a method to directly measure response category (in)comparability and to correct for it if necessary. Respondents are asked to assess hypothetical scenarios (anchoring vignettes) with the same questions and response scales used for self-assessments

(King et al., 2004; Salomon et al., 2004). Although the technique is often cited as a possible solution to tackle incomparability, applications in the health sciences are still rare and many existing vignettes failed tests of cross-national equivalence (Grol-Prokopczyk et al., 2015).

This study has some limitations. First, whereas the overall response rate to the HCAHPS survey was high (73%), it varied largely across countries (46.2% in Greece to 93.7% in Poland) as well as within countries. A recent study in England showed that high response rates for a patient experience survey may be a marker of higher quality, but also reiterated that low or high response rates alone are not an indicator that findings from a particular hospital are more or less likely to be biased (Saunders et al., 2016). In this multi-country study, it is unclear how non-response may have biased our findings. Second, while we devised multilevel models that included hospital and nursing unit random effects and controlled for patient self-reported health and other potential confounders, information on several patient level characteristics that have previously been linked to patient global ratings, such as age and gender, was not available for our analysis. Relatedly, our data were highly negatively skewed with – other than Greece and Spain – most of 0-10 ratings falling in top 3 categories (i.e., 8, 9 or 10). Whereas such distribution may be due to positive response tendency (PRT; a tendency for some respondents to evaluate care more positively than others given the same underlying experiences), Elliott et al. (2009a) argue it could be a case of extreme response tendency (ERT; a systematic tendency for some respondents to prefer the endpoints of a response scale more than other respondents given the same underlying experiences). Directly measuring ERT for individual subjects often requires both a very large number of items (i.e., ≥ 100) and data that are not skewed (Greenleaf, 1992). Strategies for dealing with ERT when data are highly skewed have been developed (Elliott et al., 2009a). Finally, our analyses could have benefitted from a sensitivity analysis using a standard differential item functioning (DIF) approach such as the one proposed by (Zumbo, 1999), and Gelin and Zumbo (2003) to get insight into whether the "preferred" use of items and their sum scores performs in an equivalent fashion across countries, but also to assess the magnitude of

DIF. However, to the best of our knowledge, these methods have only been developed for a marginal OLR model and no extensions to a random effects OLR model for hierarchical data (such as ours) have been developed. Thus, we are not certain if the conclusions based on these pseudo R^2 methods would still be valid.

4.5 Conclusions

HCAHPS global patient ratings are good reflections of specific aspects of patient experiences. Strategies to improve these specific experiences in order to improve global ratings can be exchanged across countries. In terms of reporting the overall hospital rating, whereas the common practice of collapsing 0-10 hospital ratings has the benefit of simplicity, we caution against straightforward comparisons across multiple populations without understanding what these dichotomized ratings reflect. One uniform cut-off point may not be the best solution for drawing fair comparisons as individuals across populations appear to use the hospital rating scale differently.

Chapter 5

Invariance of the WHO VAW Instrument Among Kenyan Adolescent Girls and Young Women: Bayesian Psychometric Modeling

In the previous chapter, we analysed a single (categorical) outcome variable at a time. In the present chapter and chapters that follow, we shift gears to situations where several outcome variables are measured and are studied simultaneously. We begin with a case of multivariate data having a non-hierarchical structure in this chapter. We will analyse the DREAMS impact evaluation data set introduced in Chapter 2. Interest is in assessing whether valid comparison(s) can be performed across subpopulations (or groups). To make valid comparisons across groups, a measurement instrument needs to be measurement invariant across those groups. We will evaluate measurement invariance for experience of violence among adolescent girls and young women (AGYW) in two informal settlements in Nairobi, Kenya, for a widely used WHO's violence against women instrument for measuring exposure to violence.

Factorial (non)invariance (see Chapter 3 for definitions of various forms or levels of invariance) will be assessed using exploratory structural equation modeling (ESEM) approach. Cross-group measurement invariance will be assessed using a Multiple Indicator Multiple Causes (MIMIC) model within a Bayesian framework. It is shown that the Bayesian approach offers a flexible way to modeling by relaxing some of the conditions and assumptions often required in the frequentist framework such as fixing some parameters to certain values (e.g., fixing cross-loadings to zeroes).

5.1 Introduction

Violence against women was put high on the agenda as an important public health problem and human rights abuse at key international conferences in the 1990s. The Fourth World Conference on Women held in Beijing in 1995 (United Nations, 1995), in particular, established a strategic objective to study the causes and consequences of violence against women and the efficacy of preventive measures; and to encourage promotion of research on this subject. Prevalence studies on violence against women perpetrated by (intimate) male partners have since grown considerably, but with variability in figures reported across settings (Orindi et al., 2020; World Health Organization, 2013; Stöckl et al., 2013; Alhabib et al., 2010; Ellsberg et al., 2008; Garcia-Moreno et al., 2006; Kury et al., 2004; World Health Organization, 2005; Watts and Zimmerman, 2002; Krug et al., 2002). Whereas these differences may correspond to the actual differences in the magnitude of the problem in different settings, they may be due to study design and/or methodological differences (such as questionnaire content and questionnaire administration methods), or could be a reflection of cultural differences. Such discrepancies would limit meaningful comparisons among diverse settings.

To fill this gap, the "multi-country study on women's health and domestic violence against women" was developed by the World Health Organization (WHO; World Health Organization, 2005). The study tool included the violence against women (VAW) survey instrument to measure primarily violence by

intimate partners experienced by women, especially physical, psychological and sexual violence. The WHO VAW instrument was crafted from the Conflict Tactics Scale (CTS; Straus, 2017; Straus et al., 1996), and followed a long process of discussion and consultation with technical experts in specific areas, largely because of the special ethical care the topic demands (World Health Organization, 2005). Several studies have since reported good internal consistency/reliability and validity of the instrument (Garcia-Moreno et al., 2006; Schraiber et al., 2010; Nybergh et al., 2012, 2013; Ribeiro et al., 2014).

Of critical importance is whether the instrument can be used among individuals with different characteristics or at different time points. If the instrument behaves differently across subgroups of the population, such as adolescent girls and young women, measurement biases could occur, resulting into invalid (or inappropriate) comparisons and interpretations that are not meaningful. In practice, these assumptions can be assessed through a statistical assessment of measurement invariance – also referred to as measurement equivalence (Meredith, 1993; Vandenberg and Lance, 2000). Different types of invariance can be distinguished. Configural invariance requires equality in terms of dimension (i.e., number of factors) and pattern (i.e., items within each factor) across groups. This type of invariance indicates similarity of concepts measured across groups. Metric invariance implies corresponding factors have the same meaning across groups; that is, equal factor loadings across groups. Scalar invariance implies equality in the meaning of the dimensions (i.e., equality of the factor loadings), and the levels of the underlying items (i.e., intercepts or thresholds) across groups. Scalar invariance is a necessary condition for meaningful comparison of group means (Meredith, 1993; Vandenberg and Lance, 2000; Gregorich, 2006; Orindi et al., 2017). Only a limited number of studies have rigorously evaluated this aspect of the WHO VAW instrument. Ribeiro et al. (2017) assessed configural invariance and invariance of the factor loadings, but did not evaluate scalar invariance. In another study in two Brazilian municipalities, Ribeiro et al. (2014) used confirmatory factor analysis (CFA) to investigate whether violence is a uni-or-multidimensional construct. They did not evaluate any cross-group measurement invariance. Other studies used either exploratory

factor analysis (EFA; Schraiber et al., 2010), or principal component analysis (PCA), which do not permit assessment of invariance (Nybergh et al., 2012, 2013).

In this study we investigate whether (1) the factorial structure of the WHO VAW instrument for the adolescent girls and young women (AGYW) in Nairobi slums is configurally invariant (Steenkamp and Baumgartner, 1998) with the WHO's classification (World Health Organization, 2005), and (2) there is any evidence of cross-group invariance in experience of violence for important covariates; that is, whether the sub-populations of AGYW attribute the same meaning to the dimensions and the levels of the underlying items.

5.2 Methods

5.2.1 Study design, setting and sample

We analyzed data from the DREAMS IE (Determined, Resilient, Empowered, AIDS-free, Mentored, and Safe women Impact Evaluation) study. DREAMS is an initiative of PEPFAR (the US President's Emergency Plan for AIDS Relief) to reduce new HIV infections among the most vulnerable AGYW in areas identified as 'hot-spots' with high HIV burden. DREAMS is using an HIV prevention package, being delivered "at scale", to target the multiple sources of risks that AGYW face: behavioural, biological and structural (see Saul et al., 2018). DREAMS IE is a three-year project (2017–2019) with the objectives to: 1) measure whether HIV-related outcomes change due to DREAMS interventions at a population level; 2) explore the pathways of protection by which DREAMS interventions influence the lives of young women and ultimately their risk for HIV; and 3) assess the extent to which the DREAMS interventions are delivered as intended. In Nairobi (Kenya), the African Population and Health Research Center is partnering with the London School of Hygiene & Tropical Medicine to evaluate the impact of the DREAMS Initiative in two informal settlements of Korogocho and Viwandani. The full study protocol has been published elsewhere (Birdthistle et al., 2018). We focused on the AGYW survey component, in which a cohort of randomly selected girls/young women

were enrolled during 2017 and followed prospectively, at $\sim$ 12 and $\sim$ 24 months. The present analysis uses baseline data collected from Nairobi on 1,081 AGYW (Korogocho, $n = 617$; Viwandani, $n = 464$) aged 15–22 years in 2017.

5.2.2 Measures

Experience of violence was measured using the WHO's VAW survey instrument (World Health Organization, 2005). The 15 items are listed in Table 5.1. The questions were like "Has any male ever done any of the following things to you in the past 12 months?". The questions had two response options of "yes"(=1) or "no"(=0). The World Health Organization classifies the questions into three dimensions, i.e., psychological/emotional violence (items 1–3), physical violence (items 4–11), and sexual violence (items 12–15) (World Health Organization, 2005).

Data were collected electronically using face-to-face interviews by forty-two carefully selected, properly trained and experienced field interviewers who were also well conversant with the study area. The tool was piloted and adjustments were made where needed.

Table 5.1: DREAMS Survey Data: The 15 items of the World Health Organization Violence Against Women questionnaire used to measure experience of violence among adolescent girls and young women in the DREAMS study.*

	Item
1	Say or do something to humiliate you in front of others?
2	Threaten to hurt or harm you or someone close to you?
3	Insult you or make you feel bad about yourself?
4	Push you, shake you, or throw something at you?
5	Slap you?
6	Twist your arm or pull your hair?
7	Punch you with his fist or something that could hurt you?
8	Kick you, drag you, or beat you up?
9	Try to choke you or burn you on purpose?
10	Threatened to attack you with a knife or other weapon?
11	Attacked you with a weapon?
12	Touched you in a sexual way (e.g. kissing, grabbing, or fondling), when you did not want them to?
13	Try to have sexual intercourse with you when you did not want to but did not succeed?
14	Physically forced you to have sexual intercourse even when you did not want to?
15	Forced you to perform sexual acts when you did not want to?

*The questions were like "Has any male ever done any of the following things to you in the past 12 months?". The questions had two response options of "yes"(=1) or "no"(=0)

5.2.3 Statistical Methods

Factorial Structure Assessment

To assess whether the structure of the VAW instrument proposed by the WHO, with three dimensions, can be replicated in the Kenyan AGYW population (i.e., configurally invariant), we adopted the exploratory structural equation modeling (ESEM) framework of Asparouhov and Muthen (Asparouhov and Muthén, 2009). ESEM is a more recent technique which, in addition to or instead of a CFA measurement model, allows an EFA measurement model with factor loading matrix rotations to be used in a structural equation model (Asparouhov and Muthén, 2009; Marsh et al., 2009). The ESEM was likelihood-based using a limited-information weighted least squares estimation method with a Geomin rotation criterion equal to 0.001 (the 0.001 value is to improve the shape of the rotation function, so that it is easier to minimize and to reduce the number of local solutions, with larger values being used for models with more factors). To assess the fit of the ESEM model, we considered two indices which are functions of the likelihood ratio statistic, i.e., the comparative fit index (CFI; Bentler, 1990) and the Tucker-Lewis index (TLI; Tucker and Lewis, 1973) and two indices based on how well a given model approximates the true model, i.e., the root mean square error of approximation (RMSEA; Browne and Cudeck, 1992) and the weighted root-mean-square residual (WRMR). CFI- and TLI-values of at least 0.95, RMSEA values less than 0.06, and WRMR less than 1.0 are indicative of a good fit (DiStefano et al., 2018; Hu and Bentler, 1999).

Cross-group Measurement Invariance Evaluation

To compare the latent means and to evaluate measurement invariance across groups, the MIMIC model (Abraham and Ledolter, 1979; Jöreskog and Goldberger, 1975) was established by regressing the latent factors obtained from the ESEM model as well as the 15 items on covariates. If configurally invariant, the latent factors obtained from the ESEM model correspond with the 3 factors proposed by the WHO. In our MIMIC model, a significant effect of a covariate

on any of the violence latent variables (i.e., factors) indicates population heterogeneity (i.e., group differences on latent means). A significant direct effect of a covariate on any of the 15 items, over and above the indirect effect via the factors, indicates that the item is not invariant across the levels of that covariate (i.e., group differences on the indicator's intercept or scalar noninvariance) Brown (2014). That is, that particular item does not behave similarly across the levels of that covariate. Absence of such direct effect does not necessarily provide evidence of noninvariance. Potential covariates for which invariance was assessed included self-reported invitation to participate in DREAMS (not-invited=0, invited=1), slum of residence (site: Korogocho=0, Viwandani=1), age at survey (15–17 years = 0, 18–22 years= 1), marital status (never married=1, previously married/lived with partner=2, currently married/living with partner=3), currently in school (no=0, yes=1), educational level (none/some primary=1, complete primary=2, some secondary=3, complete secondary=4, tertiary=5), religion (Muslim=1, Christian=2, and other=3), ethnicity (Somali=1, Kamba=2, Kikuyu=3, Kisii=4, Luhya=5, Luo=6, Other=7), ever had sex (no=0, yes=1), slept hungry at night in past 4 weeks (no=0, yes=1), and wealth index. Wealth index was constructed using principle component analysis (PCA) with input as indicator variables on ownership of household and individual assets/items (such as television, electricity, fridge, radio, bicycle, motorcycle, shoes, blanket, clothes, etc), household structure (i.e., floor, roof and wall material), and on household's water supply and sanitation (Filmer and Pritchett, 2001; Vyas and Kumaranayake, 2006). It is common to split wealth index into quantiles. For our case, it was grouped into three categories of "poor"(=1), "medium"(=2), and "wealthy"(=3).

We specified the MIMIC model as follows using the latent variable parameterization via a probit link (see e.g. Gelman and Hill (2006) for more on latent variable parameterization):

$$\begin{aligned} \boldsymbol{y}_i^* &= \mathbf{A}\boldsymbol{x}_i + \boldsymbol{\Lambda}\boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_i, \\ \boldsymbol{\xi}_i &= \mathbf{B}\boldsymbol{x}_i + \boldsymbol{\delta}_i, \\ \boldsymbol{\delta}_i &\sim N(0, \Sigma), \quad \boldsymbol{\varepsilon}_i \sim N(0, I_p), \quad i = 1, \ldots, n, \end{aligned} \tag{5.1}$$

where $\boldsymbol{y}_i^*$ is a $p \times 1$ vector of items for girl/woman i. It is the underlying continuous measure of the observed dichotomous experience of violence item $\boldsymbol{y}_i$ such that for $\boldsymbol{y}_i^* \geq 0$, violence experience is expressed on a manifest scale indicated by $\boldsymbol{y}_i = 1$, otherwise zero. The probit model assumes that the underlying latent variable $\boldsymbol{y}_i^*$ follows a normal distribution with standard deviation unity. Thus a linear regression for $\boldsymbol{y}_i^*$ is equivalent to a probit regression for $\boldsymbol{y}_i$. The $p \times r$ matrix $\mathbf{A}$ of the direct effects is associated with the $r \times 1$ vector of covariates $\boldsymbol{x}_i$ specified above. Further, Λ is a $p \times q$ matrix of factor loadings (i.e., regression coefficients relating the common factor to the factor items, λs), $\boldsymbol{\xi}_i$ is a $q \times 1$ vector of latent factors distributed as $N(0, \boldsymbol{\Phi})$, $\boldsymbol{\varepsilon}_i$ is a $p \times 1$ random vector of error measurements and is independent of $\boldsymbol{\xi}_i$, $\mathbf{B}$ is a $q \times r$ unknown parameter matrix associated with the vector of covariates $\boldsymbol{x}_i$, δ_i is a $q \times 1$ vector of residual errors. Cross-group (non)invariance is then assessed by studying the significance of the A-coefficients, i.e. a significant A-coefficient indicates that the item is non-invariant across the levels of that covariate. The effect of the covariates on the violence factors is assessed by studying the significance of B-coefficients in Model 5.1 above.

Selection of covariates for the assessment of direct effect on the items in the A-coefficient

We assume a probit model again. First, invitation to DREAMS, site and age adjusted model was run with one covariate at a time for each of the 15 items. Next, using a likelihood ratio test (LRT) all covariates significant at $p \leq 0.10$ in the first step above were included in a multivariable model. Finally, covariates found to be significant in the multivariable model at $p \leq 0.05$ (adjusted for invitation to DREAMS, site and age) were included in the direct effect of covariates part of the MIMIC model (i.e., the A part of the MIMIC model in equation 5.1) for assessment of cross-group measurement (non)invariance. Invitation to DREAMS, site, and age were retained even if they were not significant as they were of interest to the research, but we also wished to correct for their impact. Table 5.2 summarizes the results of this exercise. All eleven covariates were included in the B part of equation 5.1 for assessment of

group mean differences.

Table 5.2: DREAMS Survey Data. Covariates For the Direct Effect Assessment in the A-coefficient. Data are LRT p-values.

Covariate	Items														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Simple univariate model (invited to DREAMS, site and age adjusted)															
Invited to DREAMS	.994	.205	.472	.029	.964	.265	.759	.816	.342	.623	.139	.542	.684	.030	.769
Site	.549	.298	.221	.001	.008	.034	.022	.003	.104	<.001	.035	.899	.174	.050	.272
Age	.032	.014	.337	.865	.001	.089	.427	.354	.553	.694	.270	.530	.643	.046	.228
Marital status	.032	.066	<.001	.027	<.001	.345	.010	.011	.239	.0177	.756	.941	.987	.598	.613
Currently in school	.681	.008	.204	.108	<.001	.217	.018	.006	.002	.007	.234	.775	.013	<.001	.120
Education level	.286	.456	.244	.889	<.001	.262	.240	.006	.374	.615	.705	.439	.213	.780	.848
Religion	<.001	.003	<.001	.074	.035	.024	.304	.068	.073	.056	.796	.001	<.001	.812	.110
Ethnicity	.013	.031	<.001	.449	.080	.133	.469	.257	.898	.269	.610	.031	.007	.541	.412
Ever had sex	.013	.004	.004	.007	<.001	.015	.004	.001	.011	.003	.395	<.001	<.001	<.001	<.001
Slept hungry at night past 4 weeks	<.001	<.001	.046	.021	.012	.021	.040	.027	.925	.009	.300	.101	<.001	.001	<.001
Wealth quantile	.548	.205	.601	.175	.176	.181	.598	.440	.501	.012	.768	.369	.060	.013	.012
Multiple univariate model (fully adjusted)															
Invited to DREAMS	.589	.518	.420	.061	.424	.288	.936	.631	.554	.473	.139	.764	.610	.056	.671
Site	.591	.804	.569	.007	.020	.038	.087	.033	.108	<.001	.035	.849	.397	.990	.400
Age	.334	.524	.931	.146	.841	.010	.170	.885	.024	.072	.270	.068	.024	.080	.182
Marital status	.034	.028	<.001	.105	.018		.182	.474		.064					
Currently in school		.080			.871		.453	.512	.035	.218			.745	.532	
Education level					.019			.073							
Religion	.168	.653	.529	.806	.522	.041		.131	.179	.229		.076	.192		
Ethnicity	.405	.416	.559	.886	.755							.516	.764		
Ever had sex	.055	.102	.059	.057	.023	.055	.138	.271	.407	.165		<.001	<.001	<.001	<.001
Slept hungry at night past 4 weeks	<.001	<.001	.130	.024	.096	.031	.068	.092		.028		.155	.001	.011	<.001
Wealth quantile										.016			.122	.083	.164

Computational Details

Data management was performed using Stata version 14.2 (StataCorp, College Station,TX) and all analyses were performed using M*plus* version 7.4 (Muthén and Muthén, 2015). We used a combination of frequentist and Bayesian approaches. For the variable selection described in the preceding paragraph, frequentist methods were used as they are considerably faster than the Bayesian methods. The MIMIC model was fit in a Bayesian framework (see e.g. Lesaffre and Lawson (2012) for a full, pedagogical introduction to Bayesian inference) using the Bayesian structural equation modeling (BSEM) approach proposed by Muthén and Asparouhov (2012a). The choice of a Bayesian analysis needs some motivation. First, in a conventional frequentist approach freely estimating the cross-loadings and all direct effects of all the covariates on the 15 items would render the model non-identified because the q^2 restrictions (where q is the number of factors) necessary to eliminate indeterminacies are absent (Hayashi and Marcoulides, 2006). The indeterminacies are resolved by fixing cross-loadings to zero, but this is—in many cases—overly stringent. Substantive theory does not dictate that the cross-loadings are exactly 0, just that they are low in magnitude. Second, we report here that we experienced computational challenges with either A- or B-coefficients estimated in the MIMIC model 5.1 using frequentist ESEM approach due to the low frequencies for items 9 and 11 (see Figure 1). The Bayesian approach overcomes these challenges by allowing easy imposition of "hard" constraints (e.g. a cross-loading equals 0) and "soft" constraints (e.g. a cross-loading is likely near 0) while preserving the core of the substantive theory by varying the prior variances. Replacing cross-loadings specifications of exact zeros with approximate zeros based on small-variance priors brings information into the analysis that avoids the non-identification problem (Muthén and Asparouhov, 2012a). It is not immediately clear how small the prior variances need to be to solve the indeterminacy. Increasing them too much can also lead to identification problem. So we vary them empirically. In our analysis, all item cross-loadings were given normal priors with zero mean and variance 0.02 (allow for small cross-loadings which provide a more flexible solution than fixing these cross-loadings to zero), while the direct effects

of covariates on the 15 violence experience items were given normal priors with zero mean and variance 1. The $N(0, 1)$ prior variance was selected here because of three reasons: 1) Smaller prior variances produced worse model fit; 2) whereas same results were obtained when the variance was gradually increased up to 20, time to get convergence increased considerably; and 3) very large variances rendered the model non-identified. This is also in line with Gelman et al. (2008) in which weakly informative priors, such as $N(0, 1)$, $N(0, 5)$, and $N(0, 20)$, are recommended for probit and logistic regression although preference is given there to priors based on the T-distribution and the Cauchy distribution. In addition, Asparouhov and Muthén (2010) warned against the use of generic non-informative priors in factor analytic models with binary items. They argued that specifying priors that have unlimited uniform range for parameters that are on probit scale is not ideal as such priors induce skewed priors on probability scale. They encourage researchers to choose priors with a reasonable and finite range. Other than the cross-loadings and the direct effects, the default non-informative prior settings in Mplus were not overridden (Muthén and Asparouhov, 2012a). The coefficients for the MIMIC model were standardized by setting the variances of the continuous latent variables to unity.

Two chains each having 200,000 Markov Chain Monte Carlo (MCMC) iterations were initiated. Thinning of one in 20 iterations was implemented. Convergence of all parameters was assessed using trace plots and the Brooks and Gelman (1998). The mixing of MCMC chains (see Figure 5.1) together with values of potential scale reduction factor (PSRF)$<$ 1.1 were considered indicative of no convergence problems.

The model fit was assessed using the posterior predictive checks (PPCs). PPCs are obtained by generating replicated data sets from the posterior predictive distribution of the fitted model. These data sets are then compared with the observed data set with respect to the feature of interest by using discrepancy measures (Gelman et al., 2013). From these PPCs, posterior predictive p-values (PPP-values) can be computed. In this chapter the chi-square discrepancy measure (Muthén and Asparouhov, 2012a; Gelman et al.,

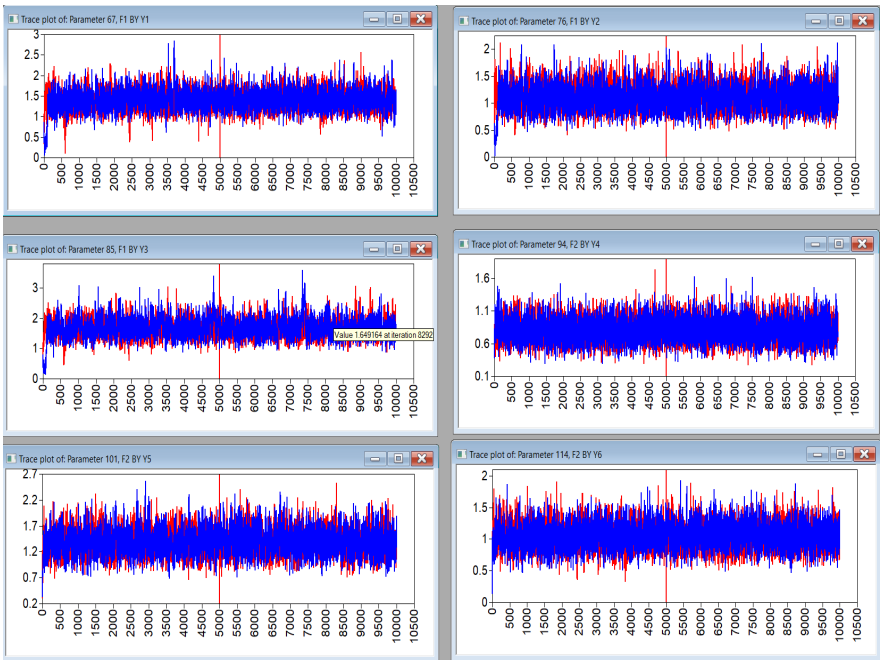

Figure 5.1: DREAMS Survey Data. Trace plots for the first six hypothesized factor loadings from the Bayesian MIMIC model. They all show "caterpillar"-like behaviour, indicating MCMC sampler is mixing well; efficiently sampling from a maximum in the underlying distribution.

1996) was used. Consider data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ consisting of n random vectors of random variables, given the parameter vector $\boldsymbol{\theta}$. The chi-square discrepancy is then

$$\chi^2(\boldsymbol{y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{(\boldsymbol{y}_i - E(\boldsymbol{y}_i|\boldsymbol{\theta}))^2}{\text{var}(\boldsymbol{y}_i|\boldsymbol{\theta})}.$$

The PPP-value is the proportion of the discrepancy measures obtained from the replicated data that are greater than the discrepancy measure of the observed data. In the Mplus implementation the PPP-value is computed using every 10th iteration among the iterations used to describe the posterior distribution

of parameters. A 95% confidence interval is produced for the difference in the discrepancy measure for the observed and replicated data (Muthén and Asparouhov, 2012a). If the model is true or close to true, the PPP-value will tend to 0.5 and the discrepancy measure difference will be close to zero falling close to the middle of the confidence interval (Muthén and Asparouhov, 2012a). We present standardized estimates.

5.3 Results

5.3.1 Descriptive Findings

The mean and median ages of the AGYW were 17.9 years and 17 years respectively. The majority had never been married (843/1081), were in school (625/1081), were Christians (917/1081), and ever had sex (642/1081). The AGYW were from different ethnic groups. Table 5.3 shows the distribution of AGYW by socio-demographic characteristics. Figure 5.2 shows, for each of the 15 items, the percentage of AGYW who reported to have experienced violence in the past 12 months. It shows that the proportions ranged from a high of 26.5% for "insult you or make you feel bad about yourself", to a low of 1.6% for "attacked you with a weapon".

About 44% ($n = 474$) of the AGYW experienced at least one act of violence, 2.7% ($n = 29$) experienced at least half of the acts of violence, and 0.3%($n = 3$) experienced up to 14 acts (0.3%, $n = 3$). No respondent experienced all 15 acts. Figure 5.3 shows the percentage of AGYW who experienced one act or more, two or more acts, three or more acts, and so on up to 8 or more acts, stratified by invitation to participate in DREAMS, age, site and religion. The numbers were similar between invited and non-invited, and between 15-17 year- and 18-22 year-olds experiencing at least 4 acts. The numbers were consistently higher for Korogocho than Viwandani; about 5% ($n = 23$) and 1% ($n = 6$) reported at least 8 items in Korogocho and Viwandani, respectively. These proportions also varied by religion.

Table 5.3: DREAMS Survey data. Distribution of AGYW by socio-demographics.

Characteristics	Number of AGYW	Percentage of AGYW
Overall (all AGYW 15-22 years)	1,081	100.0
Invited to DREAMS		
Not-invited	545	50.4
Invited	536	49.6
Site/slum		
Korogocho	617	57.1
Viwandani	464	42.9
Age		
15-17 years	547	50.6
18-22 years	534	49.4
Marital/co-habitation status		
Never married	843	78.0
Previously married/lived with partner	33	3.1
Currently married/living with partner	205	19.0
Currently in school		
No	455	42.1
Yes	626	57.9
Educational level		
None/some primary	124	11.5
Complete primary	217	20.1
Some secondary	491	45.4
Complete secondary	198	18.3
Tertiary: university/college/vocational	51	4.7
Religion		
Christian	917	84.8
Muslim	142	13.1
Other	22	2.0
Ethnicity		
Somali	90	8.3
Kamba	208	19.2
Kikuyu	319	29.5
Kisii	51	4.7
Luhya	176	16.3
Luo	176	16.3
Other	61	5.6
Ever had sex		
No	439	40.6
Yes	642	59.4
Sleep hungry at night past 4 weeks		
No	729	67.4
Yes	352	32.6
Wealth quantile		
Poor	361	33.4
Medium	360	33.3
Wealthy	360	33.3

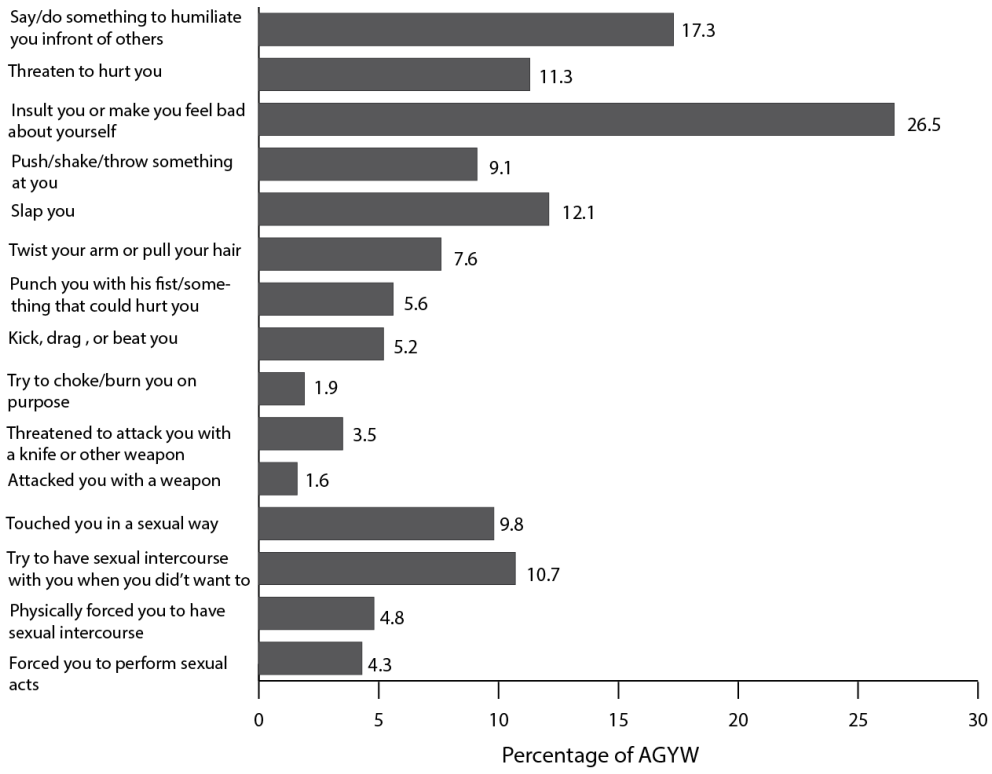

Figure 5.2: DREAMS Survey Data: Percent of AGYW reporting each of the 15 acts of violence.

5.3.2 Factorial structure of the WHO VAW instrument for the Kenyan AGYW population

Goodness-of-fit indices for this study indicated the ESEM solution summarized the 15 items well (CFI= 0.996, TLI= 0.993, RMSEA= 0.026 (90% CI: 0.017 − 0.034), WRMR= 0.634).

Table 5.4 shows the standardized factor loadings and factor correlations obtained from the ESEM model. The results verified the hypothesized factor structure. It can be seen that all the hypothesized major loadings were high ($|\lambda| = 0.543 - 0.943$). The factor loadings gave a clear and parsimonious interpretation of the factors in terms of psychological (the first 3 items), physical (items 4 through 11), and sexual violence (items 12 through 15).

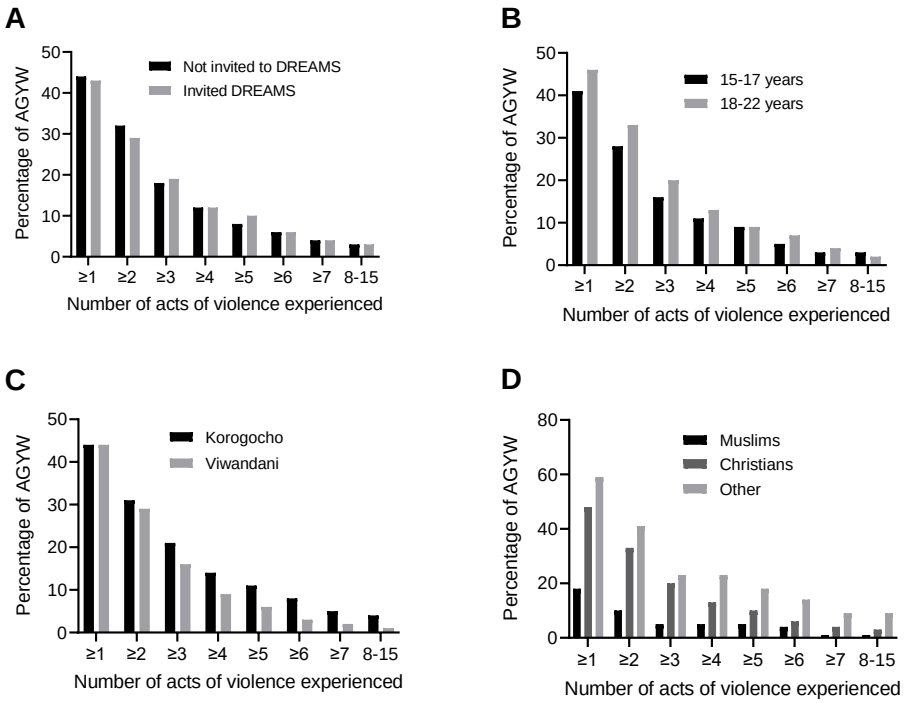

Figure 5.3: DREAMS Survey Data: Percentage of AGYW who experienced one or more violence acts, two or more acts, three or more acts, and so on up to 8 or more acts of violence, stratified by invitation to participate in DREAMS (A), age (B), site (C), and religion (D).

There were some cross-loadings ($|\lambda| = 0.001 - 0.352$)—suggesting the cross-loadings need not be restricted to zero (as is usually done in CFA using the frequentist approach to identify the model). As explained earlier, we took care of this in the Bayesian MIMIC model by allowing for cross-loadings using informative, small-variance priors. The factor correlations were medium to high and were positive, pointing to the fact that these three factors are measuring an underlying domain of experience of violence. The highest factor correlation was between psychological and physical violence factors. The smallest correlation was between psychological and sexual violence factors. In summary, the obtained factor structure was configurally similar to the one proposed by the World Health Organization (2005).

5.3.3 Measurement invariance of the WHO VAW instrument across groups of AGYW

First, we observe that the Bayesian MIMIC model provided a good fit to the data (PPP-value= 0.468). Including the effects of the 11 covariates (i.e., specified in the $\boldsymbol{x}_i$ vector in equation 5.1 for B-coefficients) on the three violence factors as well as the effects of the covariates selected on the basis of a series of univariate models—as described in Section 5.2.3—on the 15 items (i.e., specified in the $\boldsymbol{x}_i$ vector in equation 5.1 for A coefficients) did not undermine the hypothesized factor solution obtained from the ESEM model presented in Table 5.4 (see Table 5.5). Factor correlations were, however, relatively higher than those from the ESEM model. We will come back to factor correlations in Section 6.5.

Table 5.4: DREAMS Survey Data: Exploratory Structural Equation Modeling factor solution.

Item	Factor		
	Psychological	Physical	Sexual
1. Say or do something to humiliate you in front of others?	**0.886***	0.001	-0.048
2. Threaten to hurt or harm you or someone close to you?	**0.543***	0.288*	0.092
3. Insult you or make you feel bad about yourself?	**0.755***	0.126	0.041
4. Push you, shake you, or throw something at you?	0.352*	**0.516***	-0.021
5. Slap you?	0.086	**0.756***	-0.038
6. Twist your arm or pull your hair?	0.044	**0.664***	0.133
7. Punch you with his fist or something that could hurt you?	-0.007	**0.943***	-0.061
8. Kick you, drag you, or beat you up?	0.040	**0.934***	-0.047
9. Try to choke you or burn you on purpose?	-0.047	**0.873***	0.004
10. Threatened to attack you with a knife or other weapon?	0.183	**0.663***	0.019
11. Attacked you with a weapon?	-0.247	**0.816***	0.143
12. Touched you in a sexual way (e.g. kissing, grabbing, or fondling), when you did not want them to?	0.099	0.228*	**0.644***
13. Try to have sexual intercourse with you when you did not want to but did not succeed?	-0.037	0.278*	**0.752***
14. Physically forced you to have sexual intercourse even when you did not want to?	0.020	0.052	**0.917***
15. Forced you to perform sexual acts when you did not want to?	0.167	-0.017	**0.896***
Factor Correlations			
Psychological	1.000		
Physical	0.575*	1.000	
Sexual	0.446*	0.467*	1.000

Bold values indicate hypothesized major loadings; * indicates values are significant at 5% level;

Table 5.6 shows the effect of covariates on the latent mean scores for the three factors (i.e., B-coefficients in equation 5.1) from the MIMIC model. The B-coefficients show the means for psychological violence factor were significantly higher among those who had ever had sex. For physical violence factor, the means were significantly lower among AGYW with complete secondary education, compared to those who never attended school or had some primary level of education. For sexual violence factor, the means were significantly lower among those with complete primary education than those with no education or some primary education, and among those in the medium wealth quantile. On average, sexual violence was less in non-Somali ethnic groups. The means of sexual violence factor were significantly higher among non-Muslims and girls/women who had ever had sex (compared to those who have never have sex).

Table 5.7 shows the effect of covariates on the 15 items (A-coefficients in equation 5.1) from the MIMIC model. The A-coefficients show evidence of measurement noninvariance for five items: 1) Three items "*Say or do something to humiliate you in front of others*", "*Threaten to hurt or harm you or someone close to you*", and "*Forced you to perform sexual acts when you did not want to*" each had one significant direct effect, in the sense of the 95% Bayesian credibility interval not containing zero, from the covariate slept hungry at night past 4 weeks; 2) "*Slap you*" had two significant direct effects from the covariates marital/co-habitation status and education level; and 3) "*Threatened to attack you with a knife or other weapo*n" had two significant direct effects from the covariates site and wealth quantile. The significant direct effects can be interpreted as follows, in the case of site as an example. For a given factor value, AGYW in Viwandani slum had a lower propensity of giving a yes answer to the question "*Has any male ever threatened to attack you with a knife or other weapon in the past 12 months*?" than those in Korogocho slum. We note that whereas other direct effects were not significant, small nonzero estimates were obtained.

In general, for most items, we found no evidence of deviation from measurement invariance across groups of Kenyan AGYW. Thus, the instrument is

invariant in practical terms and the differences in factor means could therefore be interpreted as actual differences.

Table 5.5: DREAMS Survey Data. Bayesian MIMIC Model: Confirmatory Factor Analysis Solution

Item	Factor loading		
	Psychological	Physical	Sexual
1. Say or do something to humiliate you in front of others?	**0.915***	-0.046	-0.038
2. Threaten to hurt or harm you or someone close to you?	**0.729***	0.081	0.079
3. Insult you or make you feel bad about yourself?	**0.903***	0.026	0.006
4. Push you, shake you, or throw something at you?	0.201*	**0.575***	0.033
5. Slap you?	0.002	**0.820***	-0.013
6. Twist your arm or pull your hair?	0.017	**0.710***	0.097
7. Punch you with his fist or something that could hurt you?	-0.027	**0.962***	-0.030
8. Kick you, drag you, or beat you up?	-0.025	**0.967***	-0.007
9. Try to choke you or burn you on purpose?	0.002	**0.841***	-0.020
10. Threatened to attack you with a knife or other weapon?	0.098	**0.626***	0.018
11. Attacked you with a weapon?	-0.057	**0.716***	0.056
12. Touched you in a sexual way (e.g. kissing, grabbing, or fondling), when you did not want them to?	0.066	0.033	**0.828***
13. Try to have sexual intercourse with you when you did not want to but did not succeed?	-0.010	0.027	**0.922***
14. Physically forced you to have sexual intercourse even when you did not want to?	-0.027	-0.018	**0.936***
15. Forced you to perform sexual acts when you did not want to?	0.011	-0.009	**0.943***
Factor Correlations			
Psychological	1.000		
Physical	0.782*	1.000	
Sexual	0.617*	0.632*	1.000

Bolden values indicate hypothesized major loadings; * indicates 95% credibility interval does not contain zero;

Table 5.6: DREAMS Survey Data. Bayesian MIMIC model: Effect of Covariates on Violence Factors (B-coefficients)

	B-coefficients		
Covariates	Psychological	Physical	Sexual
Invited to DREAMS (Ref: not invited)	-.005	-.021	-.088
Site/slum (Ref: Korogocho)	-.006	-.207	.011
Age (Ref: 15-17yrs)	.011	-.097	-.133
Marital/co-habitation status (Ref: never married)			
Previously married/lived with partner	.090	.071	-.067
Currently married/living with partner	-.184	-.020	-.096
Currently in school (no/yes. Ref: no)	-.059	-.090	-.017
Educational level (Ref: None/some primary)			
Complete primary	-.103	-.103	-.133*
Some secondary	.022	-.134	-.076
Complete secondary	-.044	-.167*	-.036
Tertiary: university/college/vocational	-.045	-.109	.032
Religion (Ref: Muslim)			
Christian	.208	.196	.578*
Other	.096	.101	.250*
Ethnicity (Ref: Somali)			
Kamba	.079	-.137	-.535*
Kikuyu	.050	-.153	-.590*
Kisii	.052	-.029	-.305*
Luhya	.021	-.159	-.481*
Luo	.062	-.081	-.458*
Other	.002	-.049	-.266*
Ever had sex (no/yes. Ref: no)	.133*	.090*	.372*
Slept hungry at night past 4 weeks (no/yes. Ref: no)	.066	.052	.052
Wealth quantile (Ref: Poor)			
Medium	-.010	-.077	-.103*
Wealthy	.011	.002	-.060)

* Indicates 95% credibility interval does not contain zero.

Table 5.7: Bayesian MIMIC model: Effect of Covariates on the 15 Experience of Violence Items (A-coefficients)

	Effect of covariates on the 15 violence items (A-coeffients)														
Covariates	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Invited to DREAMS (Ref: not invited)	-.039	.024	.012	.095	.040	.056	-.019	.012	-.055	-.047	-.134	.036	.041	-.066	.015
Site/slum (Ref: Korogocho)	.020	-.008	.035	-.046	.024	.023	.050	-.002	.044	-.244*	-.074	.026	-.013	-.039	.030
Age (Ref: 15-17yrs)	.033	.066	-.031	-.038	.057	-.108	.032	.039	-.159	-.028	.093	-.023	-.036	-.001	-.007
Marital/co-habitation status (Ref: never married)															
Previously married/lived with partner	-.038	-.032	.016		.030										
Currently married/living with partner	.047	.006	.050		.116*										
Currently in school (no/yes. Ref: no)									-.139						
Educational level (Ref: None/some primary)															
Complete primary					-.068										
Some secondary					-.135*										
Complete secondary					-.062										
Tertiary					-.036										
Religion (Ref: Muslim)															
Christian						.058									
Other						-.057									
Ethnicity (Ref: Somali)															
Kamba															
Kikuyu															
Kisii															
Luhya															
Luo															
Other															
Ever had sex (no/yes. Ref: no)					.071							-.020	-.034	.098	.001
Slept hungry at night past 4 weeks (no/yes. Ref: no)	.090*	.100*		.032		.029				.068			.072	.073	.139*
Wealth quantile (Ref: Poor)															
Medium										-.215*					
Wealthy										.033					

* Indicates 95% credibility interval does not contain zero; Full definitions of items 1 to 15 are listed in Table 5.1

5.4 Discussion

The purpose of the present chapter was to investigate the factorial structure and measurement invariance of the WHO VAW instrument among vulnerable Kenyan AGYW population. We rigorously evaluated measurement invariance of the WHO VAW instrument among AGYW Kenyan population in an urban slum setting. Violence against AGYW in the two Nairobi slums is common, with about four in ten AGYW reporting to have experienced at least one act of violence in the past 12 months. Using state-of-the-art statistical techniques, we obtained a 3-factor solution comprising psychological, physical, and sexual domains, configurally similar to the one proposed by World Health Organization (2005). Our findings of no indications of noninvariance for the WHO VAW instrument items with respect to marital status, education level, religion, ethnicity, and whether one had ever had sex implies the significant differences in their latent factor means need be interpreted as actual differences in the magnitude of exposure to violence. That is, the instrument is equally reflective of the constructs of interest in all instances considered and the differences in the latent mean scores correspond to actual differences and may not be attributed to artifacts resulting from study design, methodological differences, or cultural differences. There was, however, evidence of noninvariance for items 1, 2, and 15 (with respect to slept hungry at night past 4 weeks), item 5 (for single levels of marital/cohabitation status and education level), and item 10 (for slum of residence and wealth quantile). Sass (2011) provides guidelines on how to handle noninvariant items. That is, either 1) use invariant items only; or 2) apply a partial measurement invariance model; 3) assume that for the items with measurement noninvariance, the differences are too small to influence the results and proceed using all the items; or 4) simply avoid using the scale. He argues the third option is feasible when the degree of measurement noninvariance is minimal and the majority of items are invariant. For our case, we assumed this option as noninvariance was detected for five items only – spread across the three domains – for single levels of the covariates, with three of them being attributed to a single covariate of slept hungry at night past 4 weeks. We, however, note that in another study in Brazil using the

same tool – although it did not evaluate measurement invariance – the authors reported a Heywood situation (i.e., a negative value for the residual variance) in a CFA model with respect to item 10 (Ribeiro et al., 2014). They resolved it by excluding the item from the CFA analysis. There are several causes of Heywood cases including "indefiniteness" (underidentification) of the model (Van Driel, 1978), empirical underidentification (Rindskopf, 1984), outliers (Bollen, 1987), structural misspecification of the model (Van Driel, 1978; Sato, 1987; Kolenikov and Bollen, 2012), sampling fluctuations (Van Driel, 1978; Anderson and Gerbing, 1984), or missing data (Wothke, 1993). Thus, even as we assume the third option for the present study, the contents of item 10 need to be given careful attention.

Whereas our findings may permit comparison of Kenyan results across groups within the country, caution needs be exercised as there are other factors such as rural/urban residence, entire age continuum, etc, which we did not look at but may impact on how the tool performs. Consider residence as an example, the slum population is generally unique with different social challenges compared to their rural and urban non-slum counterparts.

Our finding of configural invariance of the factor structure in the Kenyan AGYW population alone does not, however, permit valid international comparisons (Byrne and Campbell, 1999; Meredith and Teresi, 2006). When interest is to make comparisons across countries, then it is important to ascertain that these measurements are invariant across countries. Such international studies often result into data that have a multilevel structure (e.g., data of girls/women clustered in countries). Multilevel factor analytic models can be applied to evaluate measurement invariance across the hierarchical levels of the study and across groups at those specific levels. Several authors have underscored the importance of assessment of cross-level invariance, including the fact that the meanings of the factors may differ across those hierarchical levels, that it easily allows for deeper understanding of differences across countries by allowing for inclusion of country-level variables (such as general income level or literacy level) in the model to explain potential country bias in survey items, and that one can evaluate how much of the common factor variance exist between

countries and how much exist within countries (Orindi et al., 2017; Bruyneel et al., 2017a; Van de Vijver and Poortinga, 2002; Zyphur et al., 2008; Jak and Jorgensen, 2017; Davidov et al., 2012a). Heise and Kotsadam (2015) recently used a multilevel model to investigate how macro-level factors impact women's risk of intimate partner violence among 44 countries, but without evaluating cross-country invariance. Studies with univariate data can assess invariance across settings by including score-by-country interaction terms in the model as in standard differential functioning approaches.

The factor correlations from the likelihood-based ESEM were relatively smaller than the Bayesian factor correlations. The cross-loadings in the ESEM may contribute to the lower factor correlations as less correlations among the items need to go through the factors. However, the Bayesian factor correlations are not excessively high since the factors are expected to correlate to a substantial degree according to theory. These high correlations may also suggest a need to model a second-order factor(s) (Wind et al., 1973). We advocate for further research to explore this.

We evaluated invariance across groups using MIMIC modeling within the BSEM framework. The MIMIC modeling approach is more parsimonious, allow smaller sample sizes, can accommodate several covariates simultaneously, and can allow for continuous covariates (e.g., age) as well as interactions. However, it only studies higher level invariance (i.e., intercept/threshold/scalar invariance). As we have demonstrated, a Bayesian framework offers a flexible approach that allows for comprehensive evaluation of measurement invariance, overcoming some challenges often experienced in a frequentist approach such as convergence/nonidentification. A discussion on this, and other related issues, can be found in Garrett and Zeger (2000), Muthén and Asparouhov (2012b) and Levy and Mislevy (2016). Other methods for studying measurement invariance have been developed. First, is multiple-group confirmatory factor analysis (MGCFA; Jöreskog, 1971; Sörbom, 1974), involving running a set of increasingly constrained structural equation models (SEM) to test the nested forms of invariance. It is used when measurement invariance is tested with respect to a grouping variable (e.g., invited to DREAMS vs not invited, girls vs.

boys). The second is alignment optimization (Asparouhov and Muthén, 2014), which replaces setting equality constraints with a procedure that is similar to rotation in exploratory factor analysis which looks for the 'best' solution. Third, is restricted factor analysis (RFA; Oort, 1992, 1998) which is similar to MIMIC analysis except that in MIMIC models, the covariates have causal effects on the latent factors, whereas in the RFA approach the covariates and latent factors are merely associated.

Two important methodological recommendations emerge from the present study. First, concerns model evaluation tools. In the present analyses we evaluated models based on posterior predictive checks (PPCs), as the only available tool in Mplus for single-level SEM with categorical variables. Despite its importance to structural equation modeling (SEM), model evaluation remains underdeveloped for the BSEM. The PPP-value is a Bayesian tool for assessing goodness of fit available in popular software (Gelman et al., 1996; Spiegelhalter et al., 2002). Deviance information criteria (DIC; Spiegelhalter et al., 2014) is a generalization of frequentist Akaike information criteria (AIC) to choose between models, in which the model complexity penalty is determined using the deviance of the hypothesized model (Spiegelhalter et al., 2002). Recently, Cain and Zhang (2019) evaluated PPP-value and DIC in a series of Monte Carlo simulation studies, but for continuous variables only. In another study, Hoofs and colleagues evaluated the Bayesian variant of the root mean square error of approximation (RMSEA; Hoofs et al., 2018). Zhang et al. (2019) also introduced five variants of DIC as a model selection index for multilevel IRT models with dichotomous outcomes in WinBUGS. Second, concerns Bayesian variable selection (BVS) in SEMs. We selected covariates for which to study direct effects in the MIMIC model by first applying classical forward selection and backward elimination techniques on each of the 15 violence items. A number of approaches for BVS have been developed (see e.g. Lesaffre and Lawson, 2012; Miller, 2002). Extending these approaches to Bayesian SEMs will be the focus of our future research.

In summary, meaningful comparisons across groups can be performed with confidence if measurement invariance is evaluated. While the present analysis

has not detected noninvariance for a large number of covariates in this population, we hope researchers of violence against women/girls will find value in assessing measurement invariance as a way of addressing methodological issues in the study of violence against women/girls. This is especially important because the WHO violence against women instrument was built on the tradition of CTS, in the sense that respondents are asked questions about their experiences of specific acts of psychological, physical or sexual violence by a male (partner). Whereas asking such behaviourally specific questions encourages greater disclosure than requiring respondents to identify themselves as abused (Ellsberg et al., 2001), the interpretations are dependent on subjective perceptions.

5.5 Concluding Remarks

Our findings support comparisons across groups, which is important when comparing violence against women/girls prevalence rates between groups to reduce abuse or violence experienced by girls/women perpetrated by male partners. Further studies examining invariance and other psychometric properties of the instrument—especially among AGYW—need to be conducted in other countries before comparing prevalence at the international level. This chapter follows from a call by Sharpe (2013) for an increase in papers that bridge knowledge from the statistical and psychometric community to researchers who apply these methods to their empirical data. We believe researchers concerned with instrumentation in other fields will also find merit in our work and apply it in their respective areas.

Chapter 6

Impact of Cross-level Measurement Noninvariance on Hospital Rankings Based on Patient Experiences with Care

Hospital-level findings on patient experiences with care are increasingly reported publicly. A critical aspect left unexamined is the commonality of composite measures of patient experiences across different groups of patients, nursing units, hospitals and countries. Absence of commonality is termed measurement noninvariance and is hypothesized to have a strong impact on performance assessment. In Chapter 5, we examined measurement invariance across groups using cross-sectional data of a simple design structure. In this chapter, we examine measurement invariance both across groups and across levels under study (patients, nursing units, hospitals and countries) and illustrate the degree to which this method of analysis impacts hospital rankings. We show that leaving unexamined possible noninvariance across groups and hierarchical levels may have far reaching consequences for how the public perceives hospitals'

position relative to other hospitals.

6.1 Introduction

Patient experiences with care are widely recognized as an important indicator to evaluate and improve the quality of services. Associations have been shown with several hospital organizational structures and processes, such as nurse staffing and various indicators of the quality of clinical care, as well as with the well-being of health care providers (Jha et al., 2008; Aiken et al., 2012; Bruyneel et al., 2015). In many countries, hospital performance scores are used for ranking, reporting to the public, and even incentive payments (Delnoij, 2009; Giordano et al., 2010). These applications require extensive validation of survey instruments and developing an adequate adjustment methodology to allow comparability. Most attention has so far been paid to adjustments for patient case-mix and other confounders. Extant research shows that patient experiences are associated with survey mode (Elliott et al., 2009d; Vries et al., 2005), hospital-level response rates (Saunders et al., 2016), structural characteristics such as hospital size and hospital service (Young et al., 2000; O'Malley et al., 2005a; McFarland et al., 2015), physician gender (Bertakis et al., 2002) and patient race/ethnicity, primary language, education, age, gender, self-perceived health and proxy status (Elliott et al., 2001; Zaslavsky et al., 2001; O'Malley et al., 2005a; Goldstein et al., 2009; Elliott et al., 2009d,c, 2012). Several studies have subsequently shown that adjusting for case-mix alters inter-hospital variance in scores and hospital rankings (Hargraves et al., 2001; Paddison et al., 2012; Abel et al., 2014; Raleigh et al., 2015).

A first imperative condition for any measurement instrument is, however, to provide empirical evidence of its ability to distinguish true differences from measurement differences. Analysis of measurement invariance must be applied to evaluate the extent to which survey items have equal meaning across groups (Meredith, 1993; Steenkamp and Baumgartner, 1998; Vandenberg and Lance, 2000; van de Schoot et al., 2012). In the case of patient experience surveys, measurement noninvariance across groups of patients is present when these

groups have an equal standing on an instrument's underlying dimensions (e.g., doctor communication) but have different probabilities of giving particular responses to the questions (e.g., how often did doctors listen carefully to you?) which comprise those dimensions. Some explorations of this issue are available in research on patient experiences with care. This field is dominated by Consumer Assessment of Healthcare Providers and Systems (CAHPS) surveys which aim to standardize the assessment of patient experiences with US health care (Crofton et al., 1999). An early study found evidence of metric invariance between the Spanish and English versions of the Hospital CAHPS (HCAHPS; Hurtado et al., 2005). This means that Spanish and English speaking patients attributed the same meaning to the dimensions under study (i.e., equality of factor loadings in statistical terminology). However, scalar equivalence was left unevaluated. Scalar invariance would imply that patients attribute the same meaning to the dimensions, and in addition the levels of the underlying items are equal across groups (i.e., equality of factor loadings and intercepts in statistical terminology). This assumption must be empirically evaluated, as scalar invariance is a condition for meaningful comparison of mean scores across groups (Meredith, 1993; Steenkamp and Baumgartner, 1998; Vandenberg and Lance, 2000; van de Schoot et al., 2012). Other studies evaluating measurement equivalence were limited to a selection of CAHPS dimensions, thereby neglecting survey item cross-loadings (Carle et al., 2012; Carle and Weech-Maldonado, 2012; Zhu et al., 2015).

Second, of equal importance is the evaluation of cross-level measurement invariance. This pertains to equivalence of dimensions across aggregation levels (Van de Vijver and Poortinga, 2002; Zyphur et al., 2008), and applies directly to comparisons of patient experiences across hospitals or other system levels, which involves an inherently multilevel design. For example, in a study with patients clustered in nursing units, hospitals and countries, the question is whether patient experience dimensions are equivalent across these levels. It is not unlikely that by using patient-level dimensions as aggregate variables, one does not contrast sufficiently between groups of patients clustered within units, hospitals or countries, whereas dimensions at the unit, hospital or country level

(i.e., between-group dimensions) would provide this contrast. While similar factor structures may appear at different levels, this cannot be presupposed, which would indicate configural noninvariance (i.e., factor loadings patterns are not invariant across levels of analysis (or across groups when studying cross-group measurement noninvariance)) (Van de Vijver and Watkins, 2006; Bruyneel et al., 2017a; Schweig, 2014; Diya et al., 2014; Huang and Cornell, 2015). Further, factor correlations in single-level analyses may differ from those in multilevel analysis (Julian, 2001; Schweig, 2014). Hence, when drawing on patient experience data to compare hospitals or countries, dimensions derived at those specific levels are to be used. To this end multilevel factor analytic techniques are applied. Factor models that overlook the clustering effect may also lead to underestimated standard errors and incorrect parameter estimates and model fit indices (Julian, 2001; Pornprasertmanit et al., 2014; Huang and Cornell, 2015). A recent study on school quality is most pertinent in examining the consequences of cross-level noninvariance (Schweig, 2014). Findings showed different dimensions across the teacher and school level, with a different number of items and different definitions for those dimensions. A logic conclusion with important policy relevance is that the type of factor analysis used for clustered data would result in different appraisal of school quality and different exploration of relationships using these factors. This issue may equally apply to current practices in analysing and reporting patient experience surveys and many other measures of quality of care, in which correspondence in dimensions between the patient, nursing unit, hospital and country levels is often readily assumed. An early pilot study on the development of the HCAHPS showed that dimensions are not similar at the patient and hospital level, and concluded that hospital-level analysis may be more appropriate to compare hospitals (O'Malley et al., 2005b).

This chapter applies multilevel factor analytic techniques to inherently multilevel patient experience data and provides follow-up analysis of hospital rankings. The following research questions are addressed:

*Q*1. What is the multilevel factorial structure of the HCAHPS? Are there indications of cross-level measurement (non)invariance?

*Q*2. Is there evidence of measurement invariance of patient experiences with care across countries and for various extraneous variables at multiple levels such as patients' self-rated health at the individual level or nursing unit speciality at the nursing unit level?

*Q*3. How do inferences about cross-hospital comparison differ depending on whether single-level or between-level factors are used? How do these inferences differ depending on how hospital-level scores are calculated?

6.2 Methods

6.2.1 Data sources

Data are available from the three year (2009–2011) registered nurse forecasting (RN4CAST) study for 11289 patients from 824 nursing units in 186 hospitals from 7 European countries using a multilevel cross-sectional design (Table 6.1). All patient survey data were collected in 2009 only using a one-day census approach. The full study protocol and previous findings on patient experiences with care in relation to hospital structures and processes have previously been extensively described (Sermeus et al., 2011; Aiken et al., 2012; Bruyneel et al., 2015).

Table 6.1: RN4CAST Patient Survey Data. Hospitals Sampled in 7 European Countries with Patient Survey Data

Country	No. of hospitals	No. of nursing units	No. of patients
Belgium	51	244	2623
Finland	31	128	1947
Greece	24	65	847
Ireland	10	38	285
Poland	30	119	4136
Spain	15	104	470
Switzerland	25	126	981
Total	186	824	11289

6.2.2 Measures

Translated versions of the HCAHPS were used to collect information on patient experiences with care (Squires et al., 2012). The standard single-level factor analytic solution for the HCAHPS proposes seven domains: nurse communication, doctor communication, physical environment, responsiveness, pain control, medicine communication, and discharge information (Keller et al., 2005). The 16 items used in this chapter and proposed associated factors are presented in Table 6.2. Questions were of the form "During this hospital stay, . . . ," followed by a statement. Except for the questions on discharge information which were coded yes/no, answers were measured on a 4-point Likert scale with categories "never"(=1), "sometimes"(=2), "usually"(=3), and "always"(=4).

Other information was also collected for analytic purposes. First, the survey contained a question on patient's self-rated health. Other individual patient characteristics such as age and gender were not collected. Second, nursing unit speciality (general surgery, internal medicine, mixed) was accounted for. Hospital characteristics such as the number of beds, teaching status and size were also available, but were not included in the model for computational reasons.

Table 6.2: RN4CAST Patient Survey Data. The 16 Items Used to Measure Patient Satisfaction With Care and the Single-level Factor Solution Proposed by Keller et al. (2005).

Item
Nurse Communication
1: How often did nurses treat you with courtesy and respect?
2: How often did nurses listen carefully to you?
3: How often did nurses explain things in a way you could understand?
Doctor Communication
5: How often did doctors treat you with courtesy and respect?
6: How often did doctors listen carefully to you?
7: How often did doctors explain things in a way you could understand?
Responsiveness
4: After you called for assistance, how often did you get help as soon as you wanted it?
10: How often did you get help in getting to the bathroom or in using a bedpan as soon as you wanted?
Physical Environment
8: How often were your room and bathroom kept clean?
9: How often was the area around your room quiet at night?
Pain Control
11: How often was your pain well controlled?
12: How often did the hospital staff do everything they could to help you with your pain?
Medicine Communication
13: How often did hospital staff tell you what the medicine was for?
14: How often did hospital staff describe possible side effects in a way you could understand?
Discharge Information
15: Have doctors, nurses or other hospital staff talked with you about your care after you leave the hospital?
16: Have you gotten information in writing about what symptoms or health problems to look out for after you leave the hospital?

6.2.3 Statistical Analysis

Step 1. Multilevel Factor Structure and Evaluation of Cross-level Measurement Invariance

A probit model was fit using latent variable strategy. That is, we considered a latent variable regression model with normal errors where the exact continuous measurements of vector $\boldsymbol{y}^*_{iuh}$ are not available, but are related to the observed ordered categorical vector $\boldsymbol{y}_{iuh}$ as follows for patient i in nursing unit u in hospital h with items $j(\ j = 1, 2, \ldots, 14)$,

$$y_{iuhj} = \begin{cases} 1 & \text{if } -\infty \leq y^*_{iuhj} < \alpha_{1j} \\ 2 & \text{if } \alpha_{1j} \leq y^*_{iuhj} < \alpha_{2j} \\ 3 & \text{if } \alpha_{2j} \leq y^*_{iuhj} < \alpha_{3j} \\ 4 & \text{if } \alpha_{3j} \leq y^*_{iuhj} < \infty \end{cases},$$

where αs are unknown thresholds which must also be estimated; and for the dichotomous items ($j = 15, 16$),

$$y_{iuhj} = \begin{cases} 1 & \text{if } y^*_{iuhj} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

As a preliminary analysis, the intraclass correlation coefficient (ICC) was computed for each of the 16 items to assess overall variation explained by clustering. The ICC measures the relatedness of the subjects within a group and ranges from 0 (individuals within a group are as heterogeneous as individuals between groups) to 1 (members within a group show identical items) (Killip et al., 2004). The ICCs were computed for each of the three clustering levels (i.e., country, hospital and nursing unit) in the following formulas whereby all variances are replaced by their estimates:

$$ICC_c = \sigma^2_c/\sigma^2_{all}, \quad ICC_h = (\sigma^2_c + \sigma^2_h)/\sigma^2_{all}, \quad ICC_u = (\sigma^2_c + \sigma^2_h + \sigma^2_u)/\sigma^2_{all},$$

with $\sigma^2_{all} = \sigma^2_c + \sigma^2_h + \sigma^2_u + \sigma^2_i$, where $c, h, u,$ and i represent, respectively,

country, hospital, nursing unit and patient levels. In a probit model, the lowest level variance, σ_i^2 , is set to unity using the latent variable strategy. Table 6.3 presents the ICCs and the percent of variance explained at the different hierarchical levels. Higher level ICC was subtracted from lower level ICC to obtain the partitioned proportion of the total variance into each level. Nursing unit and hospital levels were considered random, with ICC values consistently >0.1 at the nursing unit level and >0.05 at the hospital level. As most of the ICC values at the country level were not more than trivial (i.e., <0.05), and as countries were not selected at random and their number is small, the country level was considered fixed.

Table 6.3: RN4CAST Patient Survey Data: Intraclass correlation coefficients (ICCs) and proportion of total variance explained at country, hospital, nursing unit and patient level.

Items	ICC			Percent of variance explained			
	Country	Hospital	Nursing unit	Country	Hospital	Nursing unit	Patient
1	0.020	0.038	0.115	2.0	1.8	7.7	88.5
2	0.012	0.039	0.103	1.2	2.7	6.4	89.7
3	0.065	0.085	0.154	6.5	2.0	6.9	84.6
4	0.019	0.068	0.157	1.9	4.9	8.9	84.3
5	0.033	0.049	0.165	3.3	1.6	11.6	83.5
6	0.021	0.038	0.144	2.1	1.7	10.6	85.6
7	0.018	0.027	0.135	1.8	0.9	10.8	86.5
8	0.092	0.155	0.221	9.2	6.3	6.6	77.9
9	0.011	0.051	0.134	1.1	4.0	8.3	86.6
10	0.069	0.129	0.181	6.9	6.0	5.2	81.9
11	0.078	0.083	0.148	7.8	0.5	6.5	85.2
12	0.035	0.052	0.125	3.5	1.7	7.3	87.5
13	0.029	0.058	0.115	2.9	2.9	5.7	88.5
14	0.097	0.113	0.169	9.7	1.6	5.6	83.1
15	0.307	0.332	0.380	30.7	2.5	4.8	62.0
16	0.375	0.390	0.446	37.5	1.5	5.6	55.4

First, a multilevel exploratory factor analysis (MEFA) was performed to understand the factor structure underlying the items, especially at the system level (i.e., at nursing unit and hospital levels). Since at present MEFA with categorical variables in M*plus* has been developed only for two-level data (Muthén and Muthén, 2015), we adopted a strategy of fitting 2-level models

of patients within nursing units (treating nursing units as random effects and ignoring hospitals), and patients within hospitals (treating hospitals as random effects and ignoring nursing units). The MEFA was likelihood-based using a robust weighted least squares estimation method with Geomin rotation. The comparative fit index (CFI), Tucker-Lewis Index (TLI), root mean squared error of approximation (RMSEA), and standardized root mean square residual (SRMR) were used to evaluate model fit. A reasonably good fitting model is obtained when SRMR is close to 0.08 or below, RMSEA is close to 0.06 or below, and CFI and TLI are close to 0.95 or greater (Hu and Bentler, 1999).

Second, a three-level confirmatory factor analysis (CFA) was performed to validate the factor structures proposed by the MEFA. The factor loadings were standardized by setting factor variances equal to 1. Thus, at the patient level a factor loading represents the relationship between a factor and the indicators after controlling for the clusters. A factor loading at the between-group levels (nursing units, hospitals) represents the relationship between factors and indicators at those levels (Pornprasertmanit et al., 2014). Unlike EFA, three-level CFA with categorical indicators can be performed in M*plus* in a Bayesian framework. Software implementation details are provided in Section 6.3.

Step 2. Evaluation of Cross-group Measurement Invariance

Cross-group MI was evaluated using multiple indicators multiple causes (MIMIC) modeling (Jöreskog and Goldberger, 1975; Abraham and Ledolter, 1979; Brown, 2014). That is, the survey items and factors from the three-level CFA (resulting from the second stage in step 1) were regressed on covariates. The covariates were measured at the patient level (self rated health: health = 1 if health status was rated poor, fair or good; 0 if rated very good or excellent), and at the nursing unit level (unit speciality: 1 = medical or mixed, 0 = surgical). We also compared countries by including, at the hospital level, 6 dummies. A significant direct effect of a covariate on a factor represents population heterogeneity (i.e., the factor means vary with the covariate). A significant direct effect of a covariate on an item represents cross-group measurement

noninvariance (i.e., when the factor is held constant, the means of the indicator vary with the covariate). Both the CFA and MIMIC models were fit within the Bayesian structural equation modeling (BSEM) framework (Muthén and Asparouhov, 2012a).

The multilevel MIMIC model is given by the equation:

$$\boldsymbol{y}^*_{iuh} = \boldsymbol{A}\boldsymbol{c}_{iuh} + \boldsymbol{\Lambda}_3\boldsymbol{\xi}_{3h} + \boldsymbol{\epsilon}_{3h} + \boldsymbol{\Lambda}_2\boldsymbol{\xi}_{2uh} + \boldsymbol{\epsilon}_{2uh} + \boldsymbol{\Lambda}_1\boldsymbol{\xi}_{1iuh} + \boldsymbol{\epsilon}_{1iuh}, \tag{6.1}$$

$$\boldsymbol{\xi}_{3h} = \boldsymbol{B}_3\boldsymbol{x}_h + \boldsymbol{\delta}_h, \quad \boldsymbol{\xi}_{2uh} = \boldsymbol{B}_2\boldsymbol{x}_{uh} + \boldsymbol{\delta}_{uh}, \quad \boldsymbol{\xi}_{1iuh} = \boldsymbol{B}_1\boldsymbol{x}_{iuh} + \boldsymbol{\delta}_{iuh},$$

where $\boldsymbol{y}^*_{iuh}$ is a $p\times 1$ vector of items for patient i in nursing unit u in hospital h and the subscripts $1, 2,$ and 3 refer to levels $1, 2,$ and 3, respectively. $\boldsymbol{A}$ is a $p \times r_1$ matrix of unknown coefficients, $\boldsymbol{c}_{iuh}$ is an $r_1 \times 1$ vector of covariates, $\boldsymbol{\Lambda}_3, \boldsymbol{\Lambda}_2$ and $\boldsymbol{\Lambda}_1$ are respectively, $p \times q_3, p \times q_2$ and $p \times q_1$ matrices of factor loadings, $\boldsymbol{\xi}_{3h}, \boldsymbol{\xi}_{2uh}$ and $\boldsymbol{\xi}_{1iuh}$ are, respectively, the $q_3 \times 1, q_2 \times 1$ and $q_1 \times 1$ vectors of latent factors distributed as $N(\boldsymbol{0}, \boldsymbol{\Phi}_3), N(\boldsymbol{0}, \boldsymbol{\Phi}_{2h})$, and $N(\boldsymbol{0}, \boldsymbol{\Phi}_{1uh})$. $\boldsymbol{\epsilon}_{3h}$ is a $p\times 1$ random vector of error measurements which is independent of $\boldsymbol{\xi}_{3h}$ and is distributed as $N[\boldsymbol{0}, \boldsymbol{\Psi}_3], \boldsymbol{\epsilon}_{2uh}$ is a $p \times 1$ random vector of error measurements which is independent of $\boldsymbol{\xi}_{2uh}$ and is distributed as $N[\boldsymbol{0}, \boldsymbol{\Psi}_2], \boldsymbol{\epsilon}_{1iuh}$ is a $p\times 1$ random vector of error measurements which is independent of $\boldsymbol{\xi}_{1iuh}$ and is distributed as $N[\boldsymbol{0}, \boldsymbol{\Psi}_1]$, where $\boldsymbol{\Psi}_3, \boldsymbol{\Psi}_2$ and $\boldsymbol{\Psi}_1$ are diagonal matrices. We further assume that $\boldsymbol{\epsilon}_{3h}, \boldsymbol{\epsilon}_{2uh}$ and $\boldsymbol{\epsilon}_{1iuh}$ are independent. $\boldsymbol{B}_3, \boldsymbol{B}_2$ and $\boldsymbol{B}_1$ are unkown parameter matrices associated with the vector of covariates $\boldsymbol{x}_h, \boldsymbol{x}_{uh}$, and $\boldsymbol{x}_{iuh}$, respectively with $\boldsymbol{c}_{iuh}$ and $\boldsymbol{x}_{iuh}$ having the same covariates; $\boldsymbol{\delta}_h, \boldsymbol{\delta}_{uh}$ and $\boldsymbol{\delta}_{iuh}$ are respectively $q_3 \times 1, q_2 \times 1$ and $q_1 \times 1$ vectors of residual errors.

Population heterogeneity then implies that the covariate has a significant impact on the terms $\boldsymbol{\xi}_{3h}$, $\boldsymbol{\xi}_{2uh}$ or $\boldsymbol{\xi}_{1iuh}$. Cross-group measurement noninvariance implies that, in model 6.1 above, the covariate in $\boldsymbol{c}_{iuh}$ has a significant impact on $\boldsymbol{y}^*_{iuh}$.

Step 3. Assessing the Policy Relevance of Assuming Cross-level Invariance

Findings for the final multilevel CFA model with covariates (resulting from step 2) were compared with a 6-factor-single-level CFA model. The latter model, which is the current standard practice, erroneously assumes independence of observations as well as cross-level measurement invariance, and does not include covariates for assessing cross-group measurement invariance. We then studied the implications of these assumptions on the inferences. More specifically, we verified how the hospitals are ranked when based on the most important (i.e., explaining most variance) hospital-level factor scores and compared these ranks with those obtained by averaging over the most important individual factor scores from the single-level analysis. A Spearman's rank correlation was run to assess the strength of agreement between the two rankings. Software implementation details are in Section 6.3.

6.3 Software Implementation

All analyses were performed using M*plus* version 7.31 (Muthén and Muthén, 2015). For the Bayesian analysis, the following priors were used. In the confirmatory factor analyses we used the default noninformative prior settings in M*plus* except for the indicator cross-loadings which were given normal priors with zero mean and variance .01 in line with Muthén and Asparouhov (2012a). These priors allow for a small cross-loadings which provide a more flexible solution than fixing these cross-loadings to zero, see e.g., Muthén and Muthén (2015) and Muthén and Asparouhov (2012a). In the MIMIC model, we further used normal priors with zero mean and variance .04 for the direct effects. Smaller variances may produce worse model fit whereas larger variances may render the model non-identified. The IW(Σ, d) residual covariance prior was chosen with Σ as the residual variances of the CFA model above and d, the degrees of freedom, set to 2000 following Asparouhov et al. (2015). Like with cross-loadings, this prior allows the covariances to be neither completely fixed to to zero nor completely free, but are instead approximately fixed to

zero. Larger d values may produce a worse fit wile smaller values may lead to non-identification. When Σ is not set as in the CFA model and d is large, there can be a misfit due to mismatch in the residual variance rather than the residual covariance. Two chains each having 100000, 180000 and 300000 MCMC iterations for the single-level CFA, MCFA and MIMIC models, respectively, were initiated. Convergence of all parameters was assessed using trace plots and the Gelman, Brooks and Rubin diagnostic (Brooks and Gelman, 1998). The quick mixing of Markov chain Monte Carlo (MCMC) chains together with values of potential scale reduction factor (PSRF)<1.1 were considered indicative of no convergence problems. Because the choice of informative priors can strongly affect Bayesian analysis, a sensitivity analysis was conducted by reconducting the analyses with cross-loading and direct effect prior variances of 0.001, 0.02, 0.04, 0.06, and 0.08. For the Inverse-Wishart residual covariance prior, d was varied from 500 to 1000 to 2000. The reported estimates are the posterior median values.

6.4 Results

6.4.1 Multilevel Factor Structure and Cross-level Measurement Noninvariance of the HCAHPS

Beside the ICCs, the multilevel variances decomposition indicated need for multilevel analysis. First, the two MEFA models resulted in a six-factor within and three-factor between nursing unit solution [fit statistics: CFI=0.985, TLI=0.969, RMSEA=0.027, SRMR=0.014 (patient level) and 0.053 (nursing unit level)], and a six-factor within and three-factor between hospital solution [CFI=0.992, TLI=0.983, RMSEA=0.024, SRMRs=0.013 (patient level) and 0.055 (hospital level)]. For both two-level models, the patient level factors were configurally similar to the solution proposed by Keller et al. (2005). Only the factor of responsiveness did not appear in our data. Items originally loading significantly on this factor loaded on other factors. At the nursing unit level, 11 items loaded on a single factor comprising nurse communication, physical environment, pain control, and medicine communication. The factors of doctor

communication and discharge information appeared similarly as at the patient level. At the hospital level, 8 items loaded on a first factor, but items 1, 11 and 12 now loaded on the second factor together with items 5 to 7. Discharge information again appeared similarly as at the patient level. There were also some significant cross-loadings at all three levels, but in general these were low.

Second, a three-level CFA (MCFA) model results verified the hypothesized factor structure that clearly demonstrated cross-level measurement noninvariance (Figure 6.1). The patient level reproduced the factor structure obtained for the US HCAHPS albeit with the slight modifications regarding the responsiveness factors as evidenced above. Table 6.4 shows the standardized factor loading and factor correlation estimates from the Bayesian multilevel CFA model. It can be seen that at all three levels, all the hypothesized major loadings were statistically important (i.e., 95% credibility interval doesn't contain zero), and there were also some important cross-loadings. The patient level shows the effect of idiosyncratic influences while the nursing and hospital levels show the effect of shared functional/structural and environmental influence. Most of the inter-factor correlations were high and statistically important, pointing to these factors measuring an underlying domain of patient experiences.

In summary, the 16 items reflected different factors across the patient, nursing unit, and hospital level (i.e., configural cross-level noninvariance).

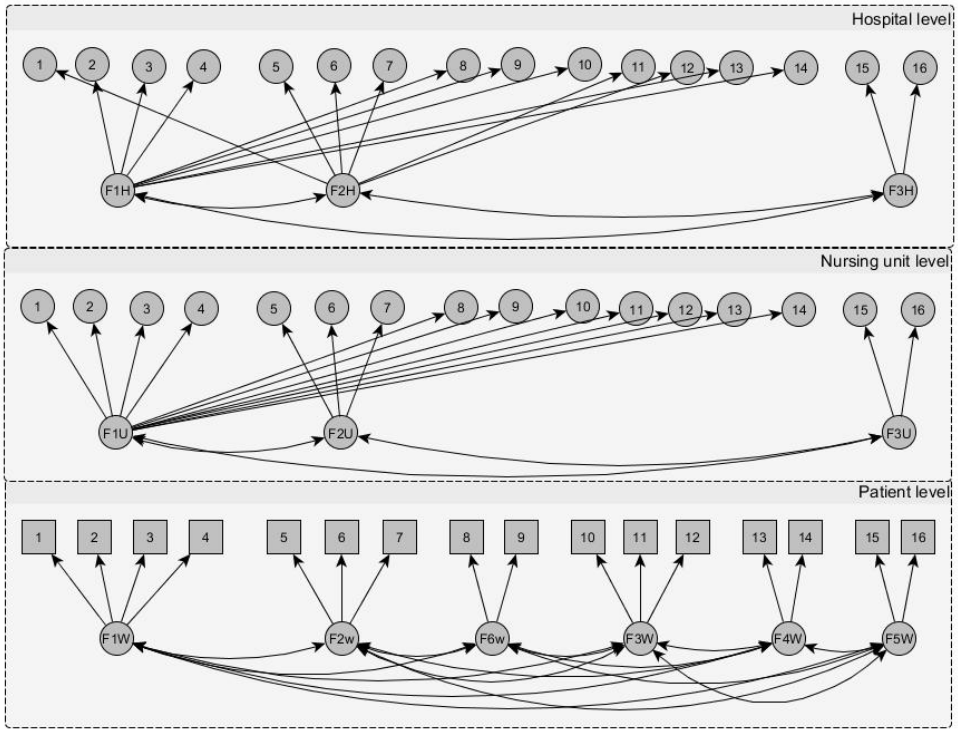

Figure 6.1: RN4CAST Patient Survey: $6-3-3$ factorial structure for the three-level CFA analysis. Cross-loadings and other details have been suppressed to make the figure easily readable. F1W to F6W are factors 1 to 6 at the within level (i.e, patient level); F1U, F2U, F3U are factors 1,2,3 at the nursing unit level; F1H, F2H, F3H are factors 1,2,3 at the hospital level.

Table 6.4: RN4CAST Patient Survey: Estimates from the Bayesian multilevel CFA with cross-loadings having normal priors with zero mean and variance= .01

Item	Within level						Nursing unit level			Hospital level		
	NC	DC	PC	MC	DI	PE	Gu	DC	DI	Gh	CPC	DI
1	**0.929***	-0.060	-0.048	-0.034	0.020	0.030	**0.938***	0.015	0.026	0.186	**0.852***	-0.015
2	**0.921***	0.018	0.000	-0.016	-0.020	-0.068	**0.924***	0.010	0.102	**0.711***	0.334	-0.246
3	**0.585***	0.130*	0.007	0.133*	0.010	-0.011	**0.912***	0.026	0.024	**0.850***	0.039	0.158
4	**0.553***	-0.011	0.160*	0.013	-0.007	0.102*	**0.923***	-0.038	0.099	**1.106***	-0.295*	-0.274*
5	0.060	**0.853***	-0.008	-0.046	0.007	0.027	-0.022	**1.005***	-0.047	-0.339*	**1.080***	0.116
6	-0.014	**0.977***	-0.009	-0.011	-0.006	-0.013	0.008	**0.984***	0.013	0.015	**0.949***	-0.190
7	-0.035	**0.715***	0.057*	0.116*	0.031	0.018	0.077	**0.949***	-0.055	0.454*	**0.627***	0.068
8	0.016	-0.019	0.006	0.001	-0.011	**0.642***	**0.569***	0.049	-0.232	**0.769***	0.160	-0.065
9	-0.040	0.078	-0.018	0.009	-0.025	**0.598***	**0.709***	0.111	-0.300*	**0.928***	-0.147	-0.680*
10	0.127*	-0.060	**0.376***	0.039	-0.009	0.178*	**1.011***	-0.183	-0.079	**0.703***	0.304*	-0.127
11	-0.090*	0.007	**0.880***	-0.010	0.013	-0.064	**1.037***	-0.151	-0.106	0.241	**0.525***	-0.084
12	0.026	0.001	**0.902***	-0.007	-0.004	0.000	**0.903***	0.053	0.014	0.260*	**0.775***	0.167
13	0.000	-0.002	0.000	**0.986***	-0.011	-0.005	**0.852***	0.013	0.013	**0.832***	0.015	0.056
14	0.021	0.008	-0.002	**0.660***	0.155*	0.069*	**0.675***	0.035	0.273	**0.645***	-0.169	0.448*
15	0.016	-0.014	0.017	0.042	**0.810***	-0.045	0.149	-0.113	**0.893***	-0.081	0.151*	**0.996***
16	-0.022	0.002	-0.023	-0.030	**0.886***	0.016	-0.018	0.078	**0.927***	0.091	-0.110	**0.911***

Factor correlations

NC	1.000						Gu	1.000			Gh	1.000		
DC	0.724*	1.000					DC	0.702*	1.000		DC	0.462*	1.000	
PC	0.691*	0.589*	1.000				DI	0.540*	0.386	1.000	DI	0.587*	0.105	1.000
MC	0.542*	0.476*	0.530*	1.000										
DI	0.313*	0.292*	0.284*	0.381*	1.000									
PE	0.666*	0.547*	0.586*	0.427*	0.272*	1.000								

NC=Nurse communication, DC=Doctor communication, PE=Physical environment, PC=Pain control, MC=medicine communication, DI= Discharge information, Gu=Unit level general factor, Gh=Hospital level general factor, CPC=Communication and pain control. * Indicate 95% credibility interval excludes zero. Bolded values indicate hypothesized major loadings.

6.4.2 Cross-group Measurement Invariance at Multiple Levels of the HCAHPS

The results of the multilevel MIMIC model are summarized in Tables 6.5 (factor solution) and 6.6 (effects of covariates on the factors and on the 16 items). It can be seen from Table 6.5 that adding covariates to the model did not disturb the factor solution discussed in the previous section. The upper part of Table 6.6 shows that at the patient level, the factor means were significantly lower among patients who rated their overall health status as poor, fair or good than those who rated their overall health as being very good or excellent. There were no significant differences in the factor means with respect to nursing unit specialization. With Belgium as the reference category, findings also indicated that the factor means varied significantly across the countries. The lower part of the table shows that the 16 items were invariant across the different levels of self-rated health, after controlling for the 6 factors and the clusters. This means that the significant differences in factor means can be interpreted as such. At the nursing unit level, there was evidence of measurement noninvariance for two items with respect to nursing unit specialization. The results further indicated measurement noninvariance for 6 items in relation to country membership, with 5 of them being attributed to Poland.

In summary, there was a strong indication of measurement invariance across patients with different perceptions of self-rated health and across nursing units with different specialities. There was however strong indication of measurement noninvariance across patients from different countries.

Table 6.5: RN4CAST Patient Survey: Bayesian three-level MIMIC model with covariates at the within and between levels: Factor solution.

Item	Within level						Nursing unit level			Hospital level		
	NC	DC	PC	MC	DI	PE	Gu	DC	DI	Gh	CPC	DI
1	**0.932***	-0.060	-0.041	-0.038	0.023	0.025	**1.004***	-0.014	-0.061	0.450	**0.429**	-0.092
2	**0.913***	0.025	0.000	-0.014	-0.024	-0.064	**1.010***	-0.063	0.046	**0.398***	0.336*	-0.054
3	**0.583***	0.132*	0.012	0.131*	0.001	-0.010	**0.886***	0.040	0.040	**0.809***	0.079	0.031
4	**0.563***	-0.013	0.143*	0.015	-0.006	0.104*	**0.956***	-0.100	0.115	**0.755***	-0.015	-0.044
5	0.062*	**0.856***	-0.011	-0.050	0.005	0.027	0.017	**0.980***	-0.053	-0.416	**0.775***	0.083
6	-0.018	**0.982***	-0.010	-0.011	-0.007	-0.014	0.014	**0.976***	0.027	-0.065	**0.828***	-0.011
7	-0.029	**0.713***	0.058*	0.118*	0.027	0.017	-0.035	**0.993***	0.032	0.191	**0.645***	0.107
8	-0.014	-0.017	0.004	-0.005	-0.004	**0.689***	**0.647***	0.056	-0.263*	**0.450***	0.072	-0.009
9	-0.037	0.083	-0.013	0.012	-0.015	**0.570***	**0.709***	0.103	-0.298*	**0.437***	0.111	-0.148
10	0.145*	-0.056	**0.350***	0.025	-0.011	0.208*	**0.907***	-0.054	-0.053	**0.589***	0.137	-0.055
11	-0.103*	-0.007	**0.961***	-0.030	0.004	-0.072	**0.959***	-0.111	-0.052	0.118	**0.473***	-0.024
12	0.073	0.008	**0.812***	0.006	-0.004	0.023	**0.838***	0.103	0.010	0.208	**0.670***	0.139
13	-0.001	-0.001	0.000	**0.992***	-0.007	-0.003	**0.891***	0.018	0.023	**0.765***	-0.012	0.043
14	0.038	0.011	-0.002	**0.652***	0.161*	0.064*	**0.709***	-0.053	0.294*	**0.457***	0.057	0.311*
15	0.033	-0.007	0.031	0.090*	**0.720***	-0.048	0.145	-0.100	**0.867***	0.022	-0.032	**0.631***
16	-0.012	0.000	-0.010	-0.021	**0.947***	0.006	-0.007	0.022	**0.970***	0.022	0.016	**0.784***
Factor correlations												
NC	1.000						Gu 1.000			Gh 1.000		
DC	0.720*	1.000					DC 0.721*	1.000		CPC 0.561	1.000	
PC	0.673*	0.575*	1.000				DI 0.512*	0.344*	1.000	DI 0.379	0.221	1.000
MC	0.528*	0.467*	0.513*	1.000								
DI	0.270*	0.258*	0.242*	0.319*	1.000							
PE	0.667*	0.536*	0.567*	0.405*	0.232*	1.000						

* Indicate 95% credibility interval excludes zero. Bolded values indicate hypothesized major loadings.

Table 6.6: RN4CAST Patient Survey: Bayesian 3-level Multiple Inidcators Multiple Causes Model With Covariates at the Within and Between Levels: Effects of Covariates

Parameter	Covariate							
	Health	Speciality	Country					
			Switzerland	Spain	Finland	Greece	Ireland	Poland
Factors (within level)								
Nurse communication	-0.419*	_	_	_	_	_	_	_
Doctor communication	-0.327*	_	_	_	_	_	_	_
Pain control	-0.442*	_	_	_	_	_	_	_
Medicine communication	-0.290*	_	_	_	_	_	_	_
Discharge information	-0.252*	_	_	_	_	_	_	_
Physical environment	-0.251	_	_	_	_	_	_	_
Factors (Level 2: units)								
General factor	_	-0.311	_	_	_	_	_	_
Doctor communication	_	0.197	_	_	_	_	_	_
Discharge information	_	-0.058	_	_	_	_	_	_
Factors (Level 3: hospital)								
General factor	_	_	2.050*	0.990	1.875*	-2.763	1.083	2.120*
Communication and pain control	_	_	6.060*	2.165	-0.734	1.749	1.094	3.122*
Discharge information	_	_	-1.146	5.42*	-0.853	-9.971*	-6.092*	1.050
Items								
1	0.028	0.143	0.040	-0.120	-1.314	-0.855	1.682	-0.464
2	-0.013	0.260	-0.632	0.401	-0.181	0.167	0.583	0.970*
3	0.002	-0.007	0.534	0.114	0.011	-0.669	0.621	0.269
4	-0.047	0.303*	-0.722	-0.346	-0.102	0.842	-0.920	0.946*
5	0.020	-0.005	0.981	-0.242	-1.018	-0.853	0.372	-1.701*
6	-0.006	0.052	-0.021	-0.392	-0.387	0.832	-1.367	-0.160
7	-0.030	-0.131	0.752	0.013	0.375	-0.079	-0.135	0.304
8	-0.029	0.291	0.108	-0.993*	-0.262	-0.788	0.246	0.526
9	-0.030	0.193	-0.897	-0.269	0.371	0.340	-0.963	0.735
10	0.018	-0.534*	0.731	0.064	0.189	-0.023	-0.306	-0.594
11	-0.033	-0.302	0.217	0.719	0.073	0.197	0.399	0.467
12	0.029	-0.272	0.430	-0.742	0.167	-0.072	-0.848	-0.409
13	0.001	-0.107	0.015	0.074	0.100	-0.025	0.091	-0.259
14	0.021	0.265	-0.272	-0.204	0.011	0.232	-0.480	0.990*
15	-0.057	-0.243	0.804	-0.283	-0.479	-0.869	0.733	-1.361*
16	0.015	0.040	-0.274	0.193	0.078	0.189	-0.269	0.450

* Indicate 95% credibility interval excludes zero. Bolded values indicate hypothesized major loadings.

6.4.3 Modelling Cross-level Noninvariance Makes a Difference in Performance Assessment

We now explore whether accounting for cross-level invariance makes a difference in hospital rankings. We extract from the multilevel analysis hospital-level factor scores and compare hospital rankings for this method with current practice of averaging up individual factor scores from a single-level analysis. A single-level factor solution for patients (see Table 6.7) was similar to the patient-level factor solution from the multilevel CFA and MIMIC model (Table 6.4), although the factor correlations were generally higher in the former model. We have discussed more about factor correlations in Section 5.4 of Chapter 5. Rankings are compared for the most important (i.e., explaining most variance) hospital-level factor, which is the hospital-level general factor from the multilevel CFA analysis (see Table 6.4), which explains 55.3% of the variance and the most important patient-level factor from the single-level analysis, which is nurse communication (see Table 6.7), which explains 45.1% of the variance.

The inter-hospital variance in scores consistently increased from the single-level to the hospital-level analyses scores for all seven countries: from 0.04 to 0.25 in Belgium, 0.08 to 0.56 in Switzerland, 0.11 to 0.25 in Spain, 0.06 to 0.19 in Finland, 0.06 to 0.64 in Greece, 0.10 to 0.51 in Ireland, and 0.03 to 0.10 in Poland. Figure 6.2, presenting for each country a scatterplot of hospital rankings for the two methods, envinces discordance in rankings . Spearman's rank correlation coefficients are shown. Although these seem high, the percentage of hospitals with discordant ranks by more than 10 percentile points (Paddison et al., 2012) equalled 26.7% in Spain, 28.0% in Switzerland, 41.9% in Finland, 50% in Ireland, 54.9% in Belgium, 66.7% in Greece, and 70% in Poland.

Table 6.7: RN4CAST Patient Survey. Estimates From the Bayesian Single-level CFA with Cross-loadings Having Normal Priors with Zero Mean and Variance= .01

Items	Factors					
	NC	DC	PC	MC	DI	PE
1	**0.933***	-0.019	-0.041	-0.051	0.007	0.000
2	**0.955***	0.005	-0.003	-0.022	-0.025	-0.074*
3	**0.569***	0.082*	-0.008	0.166*	0.034	0.048
4	**0.559***	-0.040	0.129*	0.032	0.009	0.151*
5	0.043	**0.896***	-0.007	-0.063*	0.008	0.013
6	-0.003	**0.972***	-0.006	-0.006	-0.008	-0.016
7	-0.059*	**0.738***	0.044	0.134*	0.022	0.049
8	0.041	-0.031	0.018	-0.026	0.040	**0.631***
9	-0.023	0.099	-0.008	-0.002	-0.060	**0.597***
10	0.122*	-0.069*	**0.386***	0.047	-0.030	0.215*
11	-0.079	-0.020	**0.907***	-0.013	-0.008	-0.054
12	0.025	0.017	**0.898***	-0.010	0.005	-0.009
13	0.009	0.007	0.009	**0.939***	-0.048	-0.026
14	0.006	-0.015	-0.015	**0.748***	0.135*	0.040
15	0.013	0.001	0.021	0.062	**0.796***	-0.053
16	-0.007	-0.001	-0.010	-0.036	**0.920***	0.019
Factor correlations						
NC	1.000					
DC	0.715*	1.000				
PC	0.730*	0.596*	1.000			
MC	0.606*	0.498*	0.586*	1.000		
DI	0.332*	0.254*	0.304*	0.448*	1.000	
PE	0.673*	0.524*	0.603*	0.514*	0.275*	1.000

* Indicate 95% credibility interval excludes zero. Bolded values indicate hypothesized major loadings.

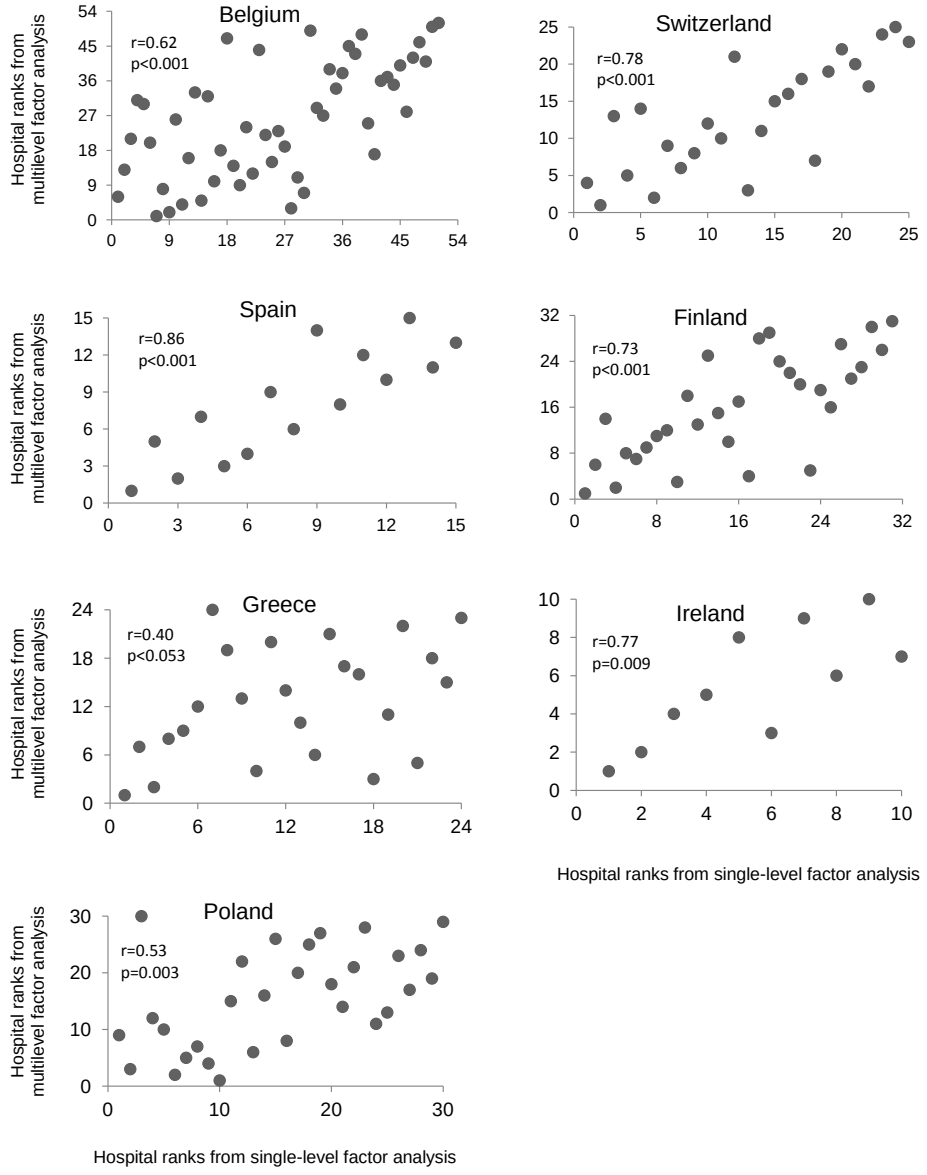

Figure 6.2: RN4CAST Patient Survey Data: Scatterplot of hospital ranks by country: comparison of the most important hospital-level factor scores from the multilevel analysis versus the most important patient level factor scores from the single-level analysis.

6.5 Discussion

This study creates awareness for the analysis and use of system-level factors when comparisons across these systems are to be made. This study is in response to a call by the HCAHPS development team made a decade ago (O'Malley et al., 2005b). In our final analysis we have shown that using different statistical methods for analysis may have important consequences on hospital rankings. Although the strength and direction of the correlation may look comforting, the issue is not necessarily the extent to which hospital rankings on the two methods correlate, it is the extent to which the ranking of individual hospitals change. This ranking markedly changed for many hospitals in each of the 7 countries included in this study. It was also shown that inter-hospital variance was much larger for the scores extracted from the between-level model compared to those from the single-level model. This means that the multilevel approach is better able to contrast patient experience scores between hospitals. This is of relevance, for example, to hospital league tables that draw on a hospital's position relative to other hospitals. As we have demonstrated, the underlying issue is that the single-level factor structure does not always match the results found at the nursing unit or hospital level. This should lead to theoretical perspectives on the system-level factor structure (Zyphur et al., 2008). It may also have various ramifications in further regression analyses (Schweig, 2014).

Cross-group invariance was studied within our analytic models that simultaneously assessed cross-level measurement invariance. With no indication of noninvariance for any of the HCAHPS items for various levels of patients' self-perceived health, it was concluded that the significant differences in factor means, which are consistent with the international literature on case-mix , can be interpreted as such. More problematic are the findings on noninvariance at the country level. For Poland, all three hospital-level factors had at least one item which showed measurement non-invariance. This means that these findings cannot be instantly compared to other countries.

A major strength of the present study is its ability to explore the impact of cross-level invariance on hospital rankings across a range of European countries with different health systems. This multi-country approach however also brings

with it a potential limitation of translation issues that for example would explain noninvariance for Poland. We have gone to great lengths to ensure conceptual, contextual, content, semantic and technical equivalence of the instrument in relationship to the patient care experience across countries (Squires et al., 2012). Nevertheless, findings presented here may imply a re-examination of the content validity index ratings from the RN4CAST qualitative examination of translated HCAHPS items and assessing whether these ratings may accurately predict measurement invariance. This would present an opportunity to help researchers understand how to limit, evaluate and explain measurement noninvariance. A second potential challenge for interpreting the differences in hospital rankings is that we opted for a comparison of the factors that explained most variance in each method of analysis. Sensitivity analyses showed similar conclusions when using the sole factor that configurally appeared similar in both methods, i.e., discharge information (see Figure 6.3). Also, using responses in the most positive category (i.e., top box scores; Elliott et al., 2009d), a commonly used method to publicly report findings on patient experiences, would not alter our conclusions, as it would not circumvent the issue of configural noninvariance of the factor structures across levels (see Figure 6.4).

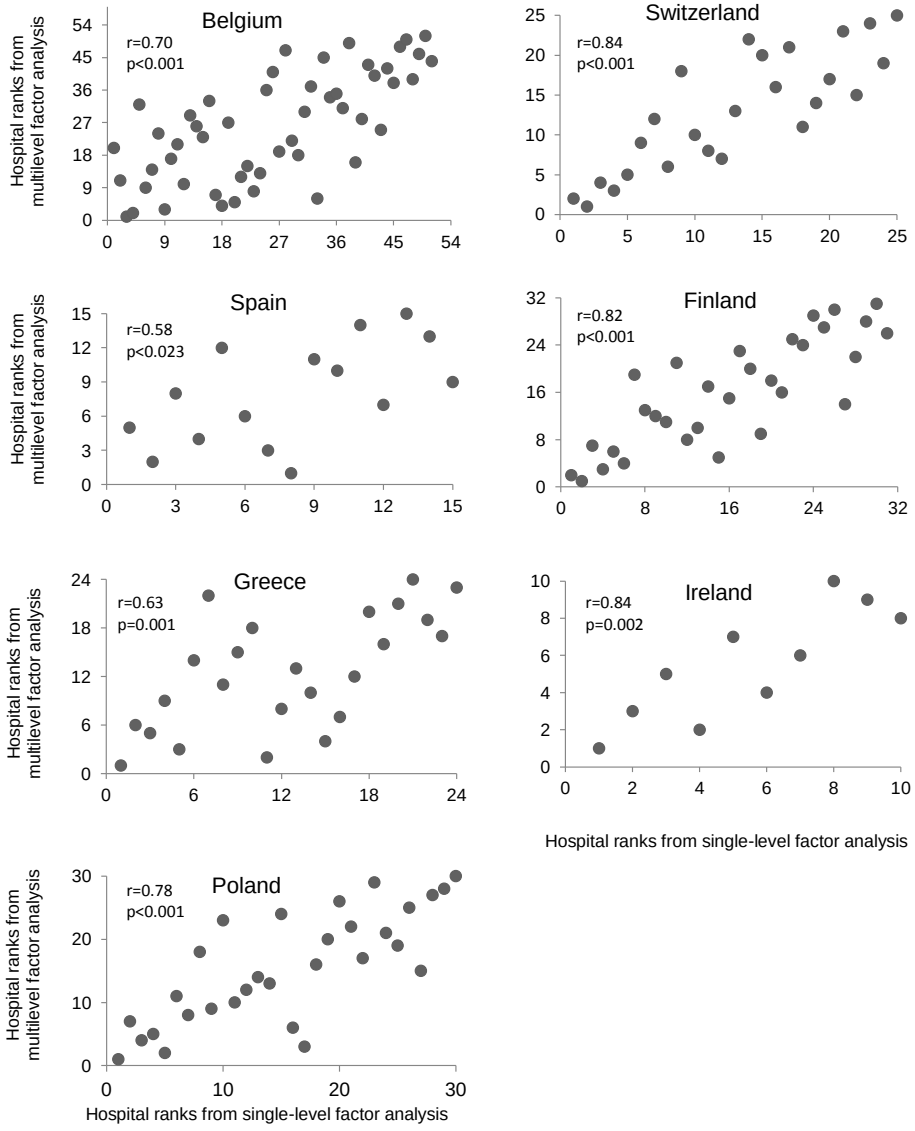

Figure 6.3: RN4CAST Patient Survey Data: Scatterplot of hospital ranks by country: comparison of discharge information hospital-level factor scores from the multilevel analysis versus discharge information individual factor scores from the single-level analysis.

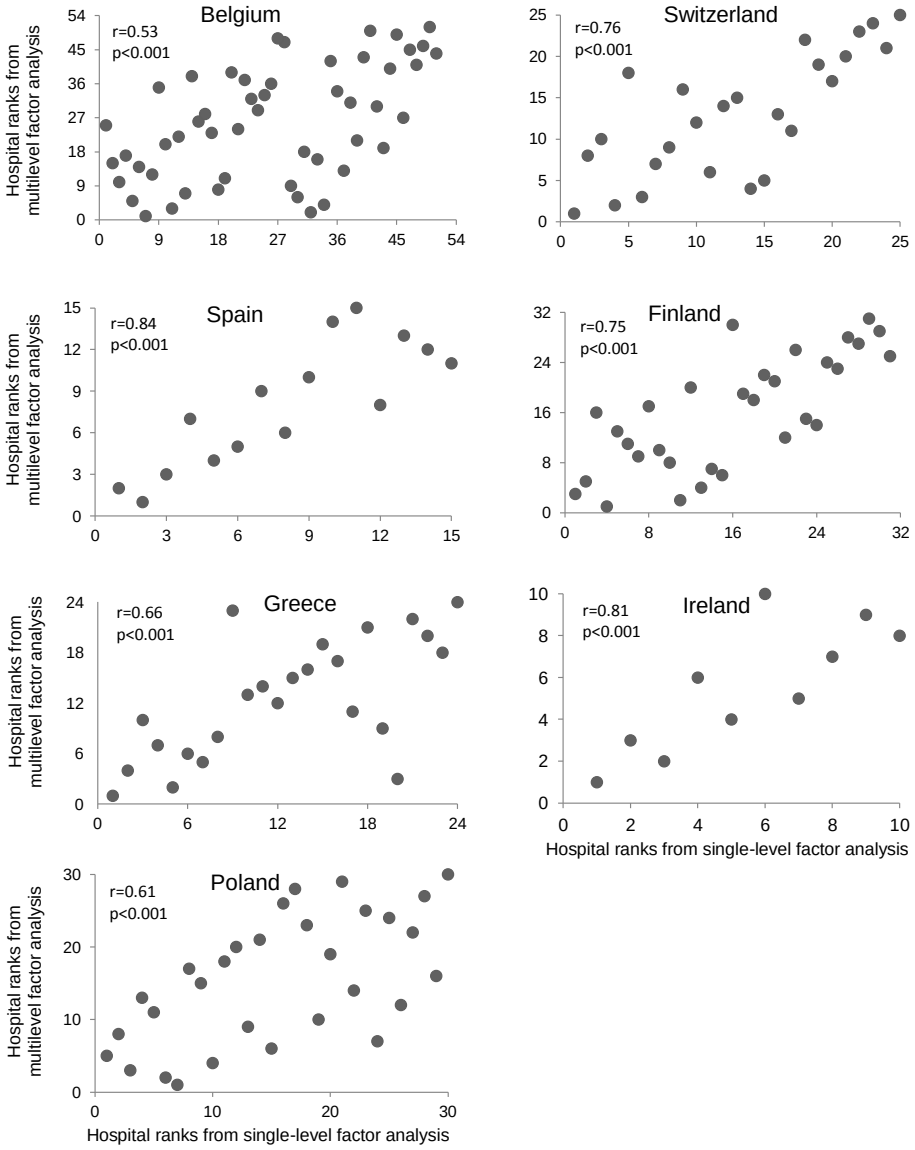

Figure 6.4: RN4CAST Patient Survey Data: Scatterplot of hospital ranks by country: comparison of most important hospital-level factor scores from the multilevel analysis versus the most important individual factor scores from the single-level analysis using top-box scoring.

6.6 Concluding Remarks

Whereas single-level factor analytic models have long gained grounds in health care sciences, the utility of multilevel factor analytic techniques is insufficiently recognized. As demonstrated in this study, multilevel analysis is not a statistical nicety but a necessity. In a climate of public reporting, not accounting for the inherent multilevel aspects in comparing patient experiences across hospitals presents an important source of bias. We believe that research that readily assumes cross-level invariance is largely caused by unawareness of the serious implications that the inability to demonstrate invariance may have such as demonstrated in this study. We note that the evaluation of measurement invariance across groups has gained increased attention in health care sciences. It is thus mostly a matter of expanding the scope on measurement invariance from a cross-group to a cross-level perspective, without which hospital managers and policy-makers may implement misinformed policies and programmes.

Chapter 7

A Combined Multilevel Factor Analysis and Covariance Regression Model with Mixed Effects in the Mean and Variance Structure

Previous chapters dealt with models for categorical outcomes and without modeling the covariance structure of the latent factors. In this chapter, we consider hierarchical Gaussian outcomes and model the covariance matrix of the responses with a mixed effects structure via a factor analytic model with structured factor loadings. This novel analysis technique combines features of the multilevel covariance regression (MCR) model of Li et al. (2014) with a multilevel factor analytic (MFA) framework. It will be shown that this technique, which replaces the responses in the MCR part with the factor scores coming from an MFA model, is quite efficient when the responses of the MCR are not measured directly but are latent variables, which is often the case in various fields of research.

7.1 Introduction

Traditional multivariate regression analysis models the mean response vector as a linear combination of explanatory variables, assuming a constant covariance matrix. This easily admits maximum likelihood estimation. Most often homoscedasticity is assumed, i.e. the covariance matrix remains the same for the different clusters. However, ignoring heteroscedasticity when present can have significant impact on inference. For instance, it can lead to inefficient parameter estimates resulting into inefficient regression predictions and inconsistent covariance matrix estimates and, consequently, invalid conclusions. Spurred by the increasing interest in capturing heteroscedasticity in multivariate measurements, a large collection of extensions have been proposed that allow the covariance matrix to depend on explanatory variables (Chiu et al., 1996; Pourahmadi, 1999; Barnard et al., 2000; Yin et al., 2010; Hoff and Niu, 2012). Hoff and Niu (2012), in particular, proposed the rank-1 regression model $\boldsymbol{y}_i = A\boldsymbol{x}_i + \gamma_i B\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ where $\boldsymbol{y}_i \in \mathbb{R}^p$ is a random multivariate response vector for subject i, $\boldsymbol{x}_i \in \mathbb{R}^q$ is a vector of explanatory variables for subject i, A is a matrix of unknown regression coefficients, $\gamma_i \sim N(0,1)$ is a latent variable allowing for extra-variation in the response, B is a $p \times q$ real matrix of factor loadings, and $\boldsymbol{\epsilon}_i$ is a random vector of error measurements. The implied marginal covariance matrix for the response is a quadratic function of the covariates $\Sigma_x = \Psi + B\boldsymbol{x}_i\boldsymbol{x}_i^T B^T$, where Ψ is a $p \times p$ "baseline" positive-definite matrix. This approach is analogous to the mean regression model, and is similar to a factor analysis model in which the factor loadings depend on the explanatory variables. Of critical importance is to ensure positive definiteness of Σ_x for all $\boldsymbol{x}_i$.

Li and colleagues (Li et al., 2014) extended the model of Hoff and Niu (2012) to accommodate multilevel data. The novelty of the paper by Li and colleagues was to simultaneously model the mean and the covariance matrix of the multidimensional response with a mixed effects model. The authors called this modeling approach the multilevel covariance regression (MCR) model. They applied their procedure on data from the RN4CAST study (Sermeus et al., 2011), a European study examining the impact of hospital organization

aspects on the well-being of nurses and patient outcomes. The RN4CAST data have a hierarchical structure with nurses and patients nested in nursing units within hospitals across 12 countries. The authors fitted a MCR model to examine if the mean and covariance matrix of 3 nurse burnout subscales depends on recorded system-level features in the organization of nursing care, but also on not-recorded factors that vary with countries, hospitals, and nursing units. They used, as responses, three sum scores obtained by summing the respective items comprising each of those subscales.

Despite the elegance of the above extension proposed by Li *et al.*, the use of sum scores may not be preferable in general. An alternative approach is to combine a multilevel factor analysis (MFA) model with the MCR approach. Therefore, we bring together in a single modeling framework the attractive features of both MFA model and MCR model: the *combined multilevel factor analysis and covariance regression model* (CMFA-CRM).

Section 7.2 describes the motivating dataset and the research objectives. In Section 7.3, the MFA and MCR models are reviewed, and the proposed CMFA-CRM is presented. Computational details on fitting the considered models to data is the subject of Section 7.4. In Section 7.5, a simulation study to compare the parameter estimates between the multi-stage approach and the simultaneous approach (CMFA-CRM) is carried out. The motivating dataset is analysed in Section 7.6. The significant advancements in statistical analyses by introducing the CMFA-CRM are discussed in Section 7.7.

7.2 The RN4CAST Study: Patient Satisfaction Data

The registered nurse forecasting (RN4CAST) study was conducted between 2009 and 2011 in 12 European countries, see Sermeus et al. (2011) for details. This multilevel cross-sectional study aimed to determine how system level features in the organization of nursing care impact individual measures of nurse well-being and patient safety outcomes and patient satisfaction. A multi-country patient survey was conducted to obtain patient level data. Here, we focus on the Belgian data only to reduce the computing time. Another

reason is that the CMFA-CRM may vary between countries thus necessitating an even more complex statistical approach. The outcome of interest here is the multidimensional patient experiences with care measured using 16 items in the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS; Darby et al., 2005). Eleven of these items applied to all patients and 9 of them were about the process of care, which we will focus on in this chapter. All 9 items were measured on a 4-point Likert scale with categories from 1 “Never" to 4 “Always". Using a multilevel factor analysis, Orindi et al. (2017) have previously examined the factor structure of the HCAHPS. The 9 items we analyse here comprised three domains of nurse communication, doctor communication and physical environment at the patient level. The items are listed in Table 7.1. All items have a non-standard, J-shaped distribution (Figure 7.1). In Section 7.6.1 we address the implication of this deviation from the standard Gaussian assumption in the factor analysis approach.

We included one covariate at the patient level, i.e. patient's self-rated health measured on a 5-point Likert scale (from 5 “Excellent” to 1 “Poor”). Note that data on demographics such as sex and age were not collected. At the nursing unit level we considered unit speciality (non-surgical = 1, surgical = 0) as only covariate. Finally, the covariates considered at the hospital level include teaching status (teaching hospital = 1, non-teaching = 0), technological level (advanced if heart surgery and/or transplant surgery available = 1, else not advanced = 0), and hospital size (defined as the number of beds). Data were available on 2623 patients from 244 nursing units in 51 hospitals in Belgium, with 15 technologically advanced hospitals and four university hospitals. Of the 244 nursing units, 108 were surgical and 136 were either medical or mixed surgical and medical units. The number of patients who participated per nursing unit varied between 1 and 24, with an average size of 11 patients. The median number of nursing units per hospital was 4, ranging from 3 to 9. About one in five patients rated their health as either excellent or very good. Note that there were only a few missing values for all the responses, which ranged from a low of 0.41% for *nurse courtesy and respect* to a high of 2.1% for *room quiet at night*.

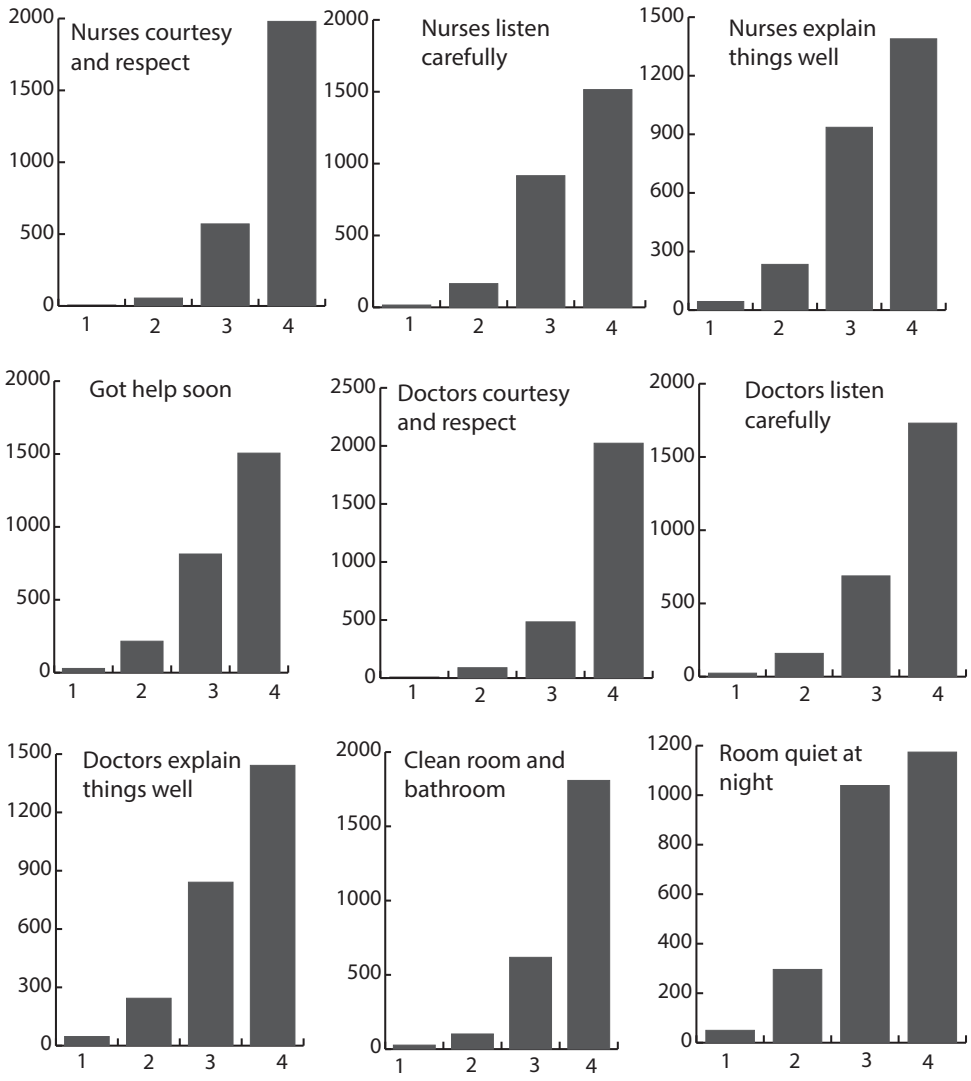

Figure 7.1: RN4CAST Patient Survey Data: Bar plot for each of the 9 HCAHPS items used to measure patient satisfaction with care.

Table 7.1: RN4CAST Patient Satisfaction Survey Data. Wording for the 9 HCAHPS items used to measure patient satisfaction with care in the RN4CAST data and the asscociated patient level domains obtained by Orindi et al. (2017).†

Items
Nurse Communication
1. How often did nurses treat you with courtesy and respect?
2. How often did nurses listen carefully to you?
3. How often did nurses explain things in a way you could understand?
4. After you called for assistance, how often did you get help as soon as you wanted it?
Doctor Communication
5. How often did doctors treat you with courtesy and respect?
6. How often did doctors listen carefully to you?
7. How often did doctors explain things in a way you could understand?
Physical Environment
8. How often were your room and bathroom kept clean?
9. How often was the area around your room quiet at night?

† The questions were like "During this hospital stay, . . . , " followed by a statement. They had responses measured on a 4-point Likert scale of 1 "Never", 2 "Sometimes", 3 "Usually", 4 "Always".

7.2.1 Research Objectives

The following research objectives were addressed here in relation to the patient experiences measurements:

> **Research Objective 1:** To examine the associations between the means and (co)variances of the latent patient experience dimensions with the patient-, nursing unit-, and hospital-level characteristics.

> **Research Objective 2:** To determine the impact of the scoring method (simultaneous approach (CMFA-CRM) vs. multi-stage approach vs. MCR using subscale sum scores) on the conclusions.

7.3 Proposed Model

For simplicity, we assume, in this section, that (for all models) the responses have a Gaussian distribution. We handle the non-normality of the patient experiences items in Section 7.6.

7.3.1 Multilevel Factor Analysis (MFA) Model

The purpose of factor analysis is to summarize a (large) number of measurements in terms of a limited number of unobservable, latent variables (also called common factors) thereby preserving most of the original information (Mulaik, 2009). The MFA model is a multilevel version of the classical confirmatory factor analysis model, see Hox (2010) and Goldstein (2010) for a general and detailed description of the MFA model. Our motivating dataset has a three-level structure (with patients within nursing units within hospitals), implying a three-level factor model. As we are interested in further modeling of the patient level common factors, we estimate an unstructured covariance matrix at the nursing unit and hospital levels, and keep only the patient level factor structure. The considered three-level factor model is expressed as

$$\begin{gathered} \boldsymbol{y}_{iuh} = \boldsymbol{\mu} + \boldsymbol{b}_{uh} + \boldsymbol{b}_h + \boldsymbol{\Lambda}\boldsymbol{\xi}_{iuh} + \boldsymbol{\epsilon}_{iuh}, \\ \boldsymbol{b}_{uh} \sim N(\mathbf{0}, \Sigma_{bu}), \quad \boldsymbol{b}_h \sim N(\mathbf{0}, \Sigma_{bh}), \quad \boldsymbol{\xi}_{iuh} \sim N(\mathbf{0}, \Sigma_z), \\ \boldsymbol{\epsilon}_{iuh} \sim N(\mathbf{0}, \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_P^2)), \end{gathered} \tag{7.1}$$

where $\boldsymbol{y}_{iuh}$ represents the p-dimensional response for patient i in nursing unit u in hospital h; $\boldsymbol{\xi}_{iuh}$ is the q-dimensional patient-level common factor following a multivariate normal distribution with an unstructured covariance matrix Σ_z, and $\boldsymbol{\Lambda}$ is its $p \times q$-dimensional loading matrix; $\boldsymbol{b}_{uh}$ and $\boldsymbol{b}_h$ represent the higher-level random effects with an unstructured covariance matrices Σ_{bu} and Σ_{bh}, respectively. $\boldsymbol{\epsilon}_{iuh}$ represents the patient level p-dimensional residual having a normal distribution with mean zero and variance σ_p^2 for each of its p elements. The unique factors are assumed to be independent with each other at each level, respectively, for the sake of identification. This will be elaborated upon in Section 7.3.4.

7.3.2 Multilevel Covariance Regression (MCR) Model

The covariance regression model was first proposed by Hoff and Niu (2012) with a multivariate regression model that has a built-in factor analysis model with factor loadings having a linear model. The covariance matrix of the multivariate responses, determined by the factor part, depends quadratically on covariates (see Equation 7.4). The MCR model extends Hoff and Niu's covariance regression model to multilevel data. In a MCR model, both the mean and the covariance parts are allowed to have a mixed effects model and can be modeled simultaneously. The MCR model for the Belgian patient satisfaction data is given by:

$$\begin{aligned}
\boldsymbol{\xi}_{iuh} &= \boldsymbol{B}\boldsymbol{x}_{iuh} + \boldsymbol{\nu}_{uh} + \boldsymbol{\nu}_h + \boldsymbol{\delta}_{iuh}, \\
\boldsymbol{\delta}_{iuh} &= \boldsymbol{L}_{iuh}\boldsymbol{\gamma}_{iuh} + \boldsymbol{\epsilon}_{iuh}, \quad \boldsymbol{L}_{iuh} = \boldsymbol{B}^*\boldsymbol{x}^*_{iuh} + \boldsymbol{\nu}^*_{uh} + \boldsymbol{\nu}^*_h, \\
\boldsymbol{\nu}_{uh} &\sim N(\mathbf{0}, \Sigma_u), \quad \boldsymbol{\nu}_h \sim N(\mathbf{0}, \Sigma_h), \quad \boldsymbol{\gamma}_{iuh} \sim N(0,1), \\
\boldsymbol{\epsilon}_{iuh} &\sim N(\mathbf{0}, \Sigma_\varepsilon), \quad \boldsymbol{\nu}^*_{uh} \sim N(\mathbf{0}, \Sigma^*_u), \quad \boldsymbol{\nu}^*_h \sim N(\mathbf{0}, \Sigma^*_h),
\end{aligned} \tag{7.2}$$

where $\boldsymbol{\xi}_{iuh}$ is the q-dimensional latent patient satisfaction variables for patient i in nursing unit u in hospital h, in place of which sum scores were used in the MCR model of Li et al. (2014); $\boldsymbol{B}$ is a $q \times m$ matrix of fixed effects associated with the m-dimensional vector $\boldsymbol{x}_{iuh}$ gathering information from all levels. $\boldsymbol{\nu}_{uh}$ and $\boldsymbol{\nu}_h$ represent the q-dimensional random intercepts in the mean part at the nursing unit and hospital levels with general covariance matrices Σ_u and Σ_h, respectively. The residuals $\boldsymbol{\delta}_{iuh}$ are decomposed into a fixed part assumed constant across patients with general covariance matrix Σ_ε and a part that varies with m^* characteristics $\boldsymbol{x}^*_{iuh}$ possibly different from $\boldsymbol{x}_{iuh}$ and are allowed to depend on all levels through a factor part. $\gamma_{iuh} \sim N(0,1)$ is a latent variable allowing for extra-variation and is assumed common for each dimension of the response. $\boldsymbol{B}^*$ is a $q \times m^*$ matrix of fixed effects associated with $\boldsymbol{x}^*_{iuh}$, while $\boldsymbol{\nu}^*_{uh}$ and $\boldsymbol{\nu}^*_h$ represent the q-dimensional random intercepts in the variance part at the nursing unit and hospital levels with general covariance matrices Σ^*_u and Σ^*_h, respectively.

7.3.3 Combined Multilevel Factor Analysis and Covariance Regression Model (CMFA-CRM)

The CMFA-CRM is a combination of a MFA and a MCR model, realized by letting the common factors serve both as the latent variables in the MFA part, as well as dependent variables in the MCR part. It is formally defined as,

$$\begin{gathered}
\textbf{The MFA part:} \quad \boldsymbol{y}_{iuh} = \boldsymbol{\mu} + \boldsymbol{b}_{uh} + \boldsymbol{b}_h + \boldsymbol{\Lambda}\boldsymbol{\xi}_{iuh} + \boldsymbol{\epsilon}^{FA}_{iuh}, \\
\boldsymbol{b}_{uh} \sim N(\boldsymbol{0}, \Sigma_{bu}), \quad \boldsymbol{b}_h \sim N(\boldsymbol{0}, \Sigma_{bh}), \quad \boldsymbol{\epsilon}^{FA}_{iuh} \sim N(\boldsymbol{0}, \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_P^2)). \\
\textbf{and the MCR part:} \quad \boldsymbol{\xi}_{iuh} = \boldsymbol{B}\boldsymbol{x}_{iuh} + \boldsymbol{\delta}_{iuh}, \\
\boldsymbol{\delta}_{iuh} = \boldsymbol{L}_{iuh}\boldsymbol{\gamma}_{iuh} + \boldsymbol{\epsilon}^{CR}_{iuh}, \quad \boldsymbol{L}_{iuh} = \boldsymbol{B}^*\boldsymbol{x}^*_{iuh} + \boldsymbol{\nu}^*_{uh} + \boldsymbol{\nu}^*_h, \\
\boldsymbol{\gamma}_{iuh} \sim N(0,1), \quad \boldsymbol{\epsilon}^{CR}_{iuh} \sim N(\boldsymbol{0}, \Sigma_\varepsilon), \quad \boldsymbol{\nu}^*_{uh} \sim N(\boldsymbol{0}, \Sigma^*_u), \quad \boldsymbol{\nu}^*_h \sim N(\boldsymbol{0}, \Sigma^*_h).
\end{gathered} \tag{7.3}$$

The MFA part is the same as in Model 7.1 except for the distribution of $\boldsymbol{\xi}_{iuh}$, which is incorporated in the MCR model later on. The $\boldsymbol{\xi}_{iuh}$ in the MCR part are output from the MFA model. All parameters have the same meaning as in Model 7.2 except that now, the random effects at the nursing unit and hospital levels in the mean, that is, $\boldsymbol{\nu}_{uh}$ and $\boldsymbol{\nu}_h$, have been dropped for identification purposes discussed in Section 7.3.4. The implied marginal covariance matrix for the latent common factors, $\boldsymbol{\xi}_{iuh}$, is given by,

$$\Psi_{iuh} = \boldsymbol{B}^*\boldsymbol{x}^*_{iuh}\boldsymbol{x}^{*T}_{iuh}\boldsymbol{B}^{*T} + \Sigma^*_u + \Sigma^*_h + \Sigma_\varepsilon. \tag{7.4}$$

From Equation (7.4) it can clearly be seen that the covariance parameters depend on covariates quadratically. It can be proved that the marginal distribution for each of the $\boldsymbol{\xi}$'s is not normal. Whereas the skewness for each marginal density is still zero when all random effects are mutually independent (assumed here), the kurtosis for the qth marginal density of $\boldsymbol{\xi}_{iuh}$ becomes,

$$\text{kurtosis}_q = \frac{6a^{*2}_q + 12a^*_q b_q}{(a^*_q + b_q + c_q)^2}, \tag{7.5}$$

where, a^*_q, b_q, c_q are the qth diagonal elements of $(\Sigma^*_u + \Sigma^*_h)$, $\boldsymbol{B}^*\boldsymbol{x}^*_{iuh}\boldsymbol{x}^{*T}_{iuh}\boldsymbol{B}^{*T}$ and Σ_ε, respectively.

Model 7.3 can be re-written in a compact form as:

$$\begin{aligned}
&\boldsymbol{y}_{iuh} = \boldsymbol{\mu} + \boldsymbol{b}_{uh} + \boldsymbol{b}_h + \boldsymbol{\Lambda}\boldsymbol{B}\boldsymbol{x}_{iuh} + \boldsymbol{\Lambda}(\boldsymbol{B}^*\boldsymbol{x}^*_{iuh} + \boldsymbol{\nu}^*_{uh} + \boldsymbol{\nu}^*_h)\boldsymbol{\gamma}_{iuh} + \boldsymbol{\Lambda}\boldsymbol{\epsilon}^{CR}_{iuh} + \boldsymbol{\epsilon}^{FA}_{iuh}, \\
&\quad \boldsymbol{b}_{uh} \sim N(\boldsymbol{0}, \Sigma_{bu}), \quad \boldsymbol{b}_h \sim N(\boldsymbol{0}, \Sigma_{bh}), \quad \boldsymbol{\epsilon}^{FA}_{iuh} \sim N(\boldsymbol{0}, \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_P^2)). \qquad (7.6) \\
&\quad \boldsymbol{\nu}^*_{uh} \sim N(\boldsymbol{0}, \Sigma^*_u), \quad \boldsymbol{\nu}^*_h \sim N(\boldsymbol{0}, \Sigma^*_h), \quad \boldsymbol{\gamma}_{iuh} \sim N(0, 1), \quad \boldsymbol{\epsilon}^{CR}_{iuh} \sim N(\boldsymbol{0}, \Sigma_\varepsilon).
\end{aligned}$$

It can be seen from this expression that the nursing unit and hospital level-specific random intercepts, that is, $\boldsymbol{b}_{uh}$ and $\boldsymbol{b}_h$, have a multivariate normal distribution with a general covariance matrix, respectively. This is already a saturated model for the covariance matrix at each of the two levels, which prevents the inclusion of any other kinds of random effects at these levels to the mean of the observed p-dimensional responses, and thus to the mean of $\boldsymbol{\xi}_{iuh}$. We re-visit this in Section 7.3.4 on identification issues.

Generally, there are two approaches to combine the MFA and MCR models: multi-stage and simultaneous. In a multi-stage approach, factor scores are created for each dimension separately using the MFA model and saved in one step. In a second step, the factor scores are then treated as observed data for the MCR model. The simultaneous approach combines the two estimation processes in a single step, resulting into a CMFA-CRM.

7.3.4 Identification Issues

There are three main identification concerns for the CMFA-CRM, emanating from the MFA part, the MCR part, and the simultaneous modeling of these two models. We describe them in turn.

First, for the MFA part, it is well-known that scaling constraints are necessary, for either the loading parameters or the variances of the common factors. Here we scaled the loading parameters of the common factors, $\boldsymbol{\xi}_{iuh}$. A common strategy is to constrain a single loading for each common factor to a constant value, usually 1, see e.g., Kline (2015) or Brown (2014). We estimated cross-loadings as well as they are, in practice, most often small but non-zero. This, however, violates the rule of having at least q^2 (where q is the number of factors) constraints in a frequentist factor analytic model (see e.g.

Asparouhov and Muthén, 2009). For the Bayesian approach, informative priors for the cross-loadings can restrict them stochastically thereby overcoming the identification problem. Muthén and Asparouhov (2012a) called this Bayesian structural equation modeling (BSEM). An informative normal prior with zero mean and standard deviation around .01 was chosen for all cross-loadings. In addition, we assumed that the p unique factors (residuals) are mutually independent.

Second, for the MCR part, $\boldsymbol{L}_{iuh}$ is known upto the sign as $\boldsymbol{\gamma}_{iuh} \sim N(0,1)$. This implies that there are two possible solutions for the model as $(\boldsymbol{B}^*, \boldsymbol{\gamma}_{iuh}, \boldsymbol{\nu}^*_{uh}, \boldsymbol{\nu}^*_h)$ and $-(\boldsymbol{B}^*, \boldsymbol{\gamma}_{iuh}, \boldsymbol{\nu}^*_{uh}, \boldsymbol{\nu}^*_h)$. To avoid this "flipping states" situation in the MCMC iterations (Lesaffre and Lawson, 2012) and ensure the model is identified we proceeded as follows: first, sample the factor loadings $\boldsymbol{B}^*$, the latent variable $\boldsymbol{\gamma}_{iuh}$ and the random effects $\boldsymbol{\nu}^*_{uh}$ and $\boldsymbol{\nu}^*_h$. Next, sample a variable $\boldsymbol{S}_i \sim$ Bernoulli$(\frac{1}{2})$. Finally, change the sign of $\boldsymbol{\gamma}_{iuh}, \boldsymbol{\nu}^*_{uh}$ and $\boldsymbol{\nu}^*_h$ if $\boldsymbol{S}_i = 1$ else keep the original sign. Thus, the sampled $\boldsymbol{\gamma}_{iuh}, \boldsymbol{\nu}^*_{uh}$ and $\boldsymbol{\nu}^*_h$ are allowed to change the sign to follow the solution sampled for $\boldsymbol{B}^*$ which is not allowed to change sign through the MCMC iterations. This is a departure from the strategy of assigning a mixture prior to the coefficient $\boldsymbol{B}^*$ and the random effects, $\boldsymbol{\nu}^*_{uh}$ and $\boldsymbol{\nu}^*_h$, proposed by Li et al. (2014). Given a reasonable variation of $\boldsymbol{x}^*_{iuh}$, all parameters in the MCR part, including the whole covariance matrix of the residual $\boldsymbol{\epsilon}^{CR}_{iuh}$, could be identified.

The third identification issue stems from the joint modeling of the MFA and the MCR parts. As described above, the key part of this joint modeling are the commom factors, $\boldsymbol{\xi}_{iuh}$, that connect the MFA part and the MCR part. From the CMFA-CRM expressed both in models 7.3 and 7.6, the nursing unit and hospital level-specific covariance matrices formulated by $\boldsymbol{b}_{uh}$ and $\boldsymbol{b}_h$, respectively, are saturated, therefore adding random effects to the factor scores ($\boldsymbol{\xi}_{iuh}$), which will further contribute to the saturated covariance matrices at higher levels, cannot be identified. For this reason, we did not include the random effects ($\boldsymbol{\nu}_{uh}$ and $\boldsymbol{\nu}_h$) when modeling the mean (i.e., the common factor scores, $\boldsymbol{\xi}_{iuh}$). There is no such constraint for modeling of the variance/covariance of the factor scores, i.e. modeling the heteroscedasticity. It is valid to assume that

the variance of the outcome within each nursing unit is different and may also depend on some covariates. That is also to say that the modeling of the (co)variance of $\boldsymbol{\xi}_{iuh}$ could be multilevel structured, even when the modeling of the mean structure does not have a hierarchical structure.

7.4 Computational Details

One of the reasons for choosing a Bayesian approach is that the considered model has a relatively large number of random effects, which is known to cause problems in maximum likelihood algorithms, although Chib and Carlin (1999) have noted that Bayesian approach may also have difficulties when applied to high-dimensional statistical models. For the three-level CMFA-CRM, there are $2p$ random intercepts for the p indicators, $q(= 3)$ common factors from the MFA part (i.e., $\boldsymbol{\xi}_{iuh}$), one common factor in the variance structure (i.e., γ_i), and two extra random intercepts for each of the q factor loadings for γ_i. Just another Gibbs sampler (JAGS; Plummer, 2003) Markov chain Monte Carlo (MCMC) program was used. In particular, JAGS was run from R (R Core Team, 2017) through the *rjags* (Plummer, 2013) package. We initiated three chains each having 30000 iterations for the burn-in part and another 30000 iterations for model estimation with thinning factor 10. The Brooks-Gelman-Rubin plots with the potential scale reduction factor (PSRF, which should be smaller than 1.1) was used for convergence checking. In addition, we ran the Markov chain until the Monte Carlo (MC) error was around or smaller than 5% of the posterior standard deviation (Brooks and Gelman, 1998).

A Bayesian analysis needs prior distributions for all its parameters. Reasonable informative priors can lead to more accurate parameter estimates, which is one of the benefits of using the Bayesian approach. Here, we used informative priors for the cross-loadings of the common factors in the MFA part to overcome the identification issue described in Section 7.3.4. For the other loading parameters and the intercepts, a vague normal distribution was used for each with mean zero and a large variance (10^6). For the variance/covariance parameters, we used an inverse gamma distribution with small shape and

rate parameters (0.001) for the univariate case. For the multivariate case, we used an inverse Wishart distribution with a small scale matrix ($0.01I$, with I the identity matrix) and degrees of freedom equal to the dimension of the matrix. For the MCR part, all parameters (including the variance/covariance parameters and the rest) have the same vague normal distribution or inverse gamma/Wishart distribution as for the MFA part.

Models were compared using the deviance information criteria (DIC; Spiegelhalter et al., 2002). Models with smaller DIC values are preferred. As JAGS cannot provide DIC due to the Bounded outcome scores (BOS) strategy employed (see Section 7.6.1), we calculated DIC outside JAGS. The likelihood required for this is calculated from the density of the latent multivariate LN distribution as its integral between the limits of the interval censored outcomes. Celeux et al. (2006) proposed a large number of DIC possibilities for missing data models such as mixture models and random effect models, which they classified as "observed" (DICs 1 to 3), "complete" (DICs 4 to 6) and "conditional" (DICs 7 & 8). Here, we used the 'conditional DIC_7' as it is (1) a natural choice that applies the original definition of DIC to the conditional distributuion given the random effects according to the authors, and (2) the default approach for calculating DIC for hierarchical models in WinBUGS (window-based Bayesian inference Using Gibbs Sampling; Spiegelhalter et al., 2003). We revisit the topic of model selection using DIC in Section 7.7.

7.5 Simulation Study

We conduct a limited simulation study to compare the performance of the multi-stage model and the proposed CMFA-CRM that fosters a simultaneous approach.

7.5.1 Set up of the Simulation Study

The data were first simulated from the CMFA-CRM. The multi-stage model and the CMFA-CRM were then applied to the simulated data. We defined the population parameters in this simulation study based on the primary results

from the analysis of the RN4CAST dataset. The simulated data had a 3-level structure as the motivating dataset (i.e., 1000 patients from 200 nursing units in 50 hospitals), with 9 continuous responses. We defined three common factors for the MFA (i.e., the $\boldsymbol{\xi}_s$), which are further modelled with a MCR model with a single common factor (i.e., the $\boldsymbol{\gamma}$). One covariate was specified in both the mean and the covariance parts of the MCR model. Except for cross-loadings for which we have chosen informative normal priors with zero mean and standard deviation of 0.22, vague priors were chosen to estimate the models as described in Section 7.4. 500 simulations were performed.

Following Burton et al. (2006), the performance was assessed by the standardized bias (stbias) and the 95% coverage for each parameter. The standardized bias is calculated as $100|\bar{\hat{\beta}} - \beta|/SE(\hat{\beta})$, with β the true value for each parameter, $\bar{\hat{\beta}}$ the average parameter estimate over all simulations, and $SE(\hat{\beta})$ the empirical standard error calculated as standard deviation of the parameter estimate from all simulations. When the standardized bias is less than 40%, the parameter is considered to be reasonably well estimated (Collins et al., 2001). In addition, we require that the coverage of the 95% credible interval should be close to 95%.

7.5.2 Simulation Results

Table 7.2 presents the main results. The two-stage model appears to give proper (unbiased with good coverage) estimates for the intercepts of each item and the factor loadings for the lower-order factors. However, it underestimated the regression coefficients, i.e. both $\boldsymbol{B}$ and $\boldsymbol{B}^*$, and some of the variance/covariance parameters for the random effects in the MCR part. The two-stage model also seriously overestimated the covariance matrix of the higher-order common factors.

Table 7.2: Simulation study: Standard bias and coverage for the two-stage and CMFA-CRM models.

Parameter	Standardized bias		Coverage	
	Two-stage	CMFA-CRM	Two-stage	CMFA-CRM
$\mu 0[1]$	-2.57	-1.32	89.3	91.3
$\mu 0[2]$	0.30	1.92	92.3	91.3
$\mu 0[3]$	2.44	3.46	92.7	93.3
$\mu 0[4]$	13.89	15.76	93.0	92.7
$\mu 0[5]$	0.84	2.33	90.0	92.0
$\mu 0[6]$	1.83	3.19	93.0	91.3
$\mu 0[7]$	0.29	0.10	92.3	91.0
$\mu 0[8]$	1.91	2.82	92.3	92.3
$\mu 0[9]$	2.74	3.44	92.0	91.0
$\beta[1]$	-67.01	2.15	80.0	94.0
$\beta[2]$	-64.82	5.93	81.0	93.7
$\beta[3]$	-59.81	9.54	79.7	93.7
loading[1]	-6.77	-7.62	95.3	95.0
loading[2]	3.68	2.60	94.7	94.3
loading[3]	5.65	2.78	96.7	96.0
loading[4]	1.37	-1.71	95.0	95.0
loading[5]	3.67	1.39	95.0	95.7
loading[6]	11.64	9.50	95.0	96.0
$\beta_0^*[1]$	-11.55	2.59	95.0	94.7
$\beta_0^*[2]$	-14.82	-0.91	95.0	95.7
$\beta_0^*[3]$	-4.29	10.06	95.7	96.0
$\beta_1^*[1]$	-55.79	2.70	86.7	92.7
$\beta_1^*[2]$	-58.18	1.01	86.0	92.7
$\beta_1^*[3]$	-57.74	-1.75	84.7	94.0
$\Sigma_\varepsilon[1,1]$	286.41	35.44	3.0	95.7
$\Sigma_\varepsilon[2,1]$	264.37	15.45	6.3	91.0
$\Sigma_\varepsilon[3,1]$	276.81	17.24	4.7	94.7
$\Sigma_\varepsilon[1,2]$	264.37	15.45	6.3	91.0
$\Sigma_\varepsilon[2,2]$	291.19	29.81	3.7	92.3
$\Sigma_\varepsilon[3,2]$	286.25	16.75	3.0	92.3
$\Sigma_\varepsilon[1,3]$	276.81	17.24	4.7	94.7
$\Sigma_\varepsilon[2,3]$	286.25	16.75	3.0	92.3
$\Sigma_\varepsilon[3,3]$	297.53	30.22	4.0	92.7
$\sigma^2[1]$	-5.43	-3.80	92.0	92.3
$\sigma^2[2]$	7.82	9.91	95.3	95.0
$\sigma^2[3]$	16.95	15.78	92.7	93.7
$\sigma^2[4]$	7.62	6.57	92.3	94.0
$\sigma^2[5]$	6.72	9.13	96.0	97.0
$\sigma^2[6]$	12.89	15.07	95.3	95.7
$\sigma^2[7]$	9.85	9.76	94.3	93.7

Continued on next page

Table 7.2 – *Continued from previous page*

Parameter	Standardized bias		Coverage	
	Two-stage	CMFA-CRM	Two-stage	CMFA-CRM
$\sigma^2[8]$	10.56	11.67	95.7	94.0
$\sigma^2[9]$	2.29	4.99	96.0	95.7
$\Sigma_h^*[1,1]$	-18.36	16.32	91.7	93.7
$\Sigma_h^*[2,1]$	-7.30	-2.96	91.7	92.0
$\Sigma_h^*[3,1]$	-13.94	-10.50	93.0	94.7
$\Sigma_h^*[1,2]$	-7.30	-2.96	91.7	92.0
$\Sigma_h^*[2,2]$	-13.12	20.87	92.0	92.3
$\Sigma_h^*[3,2]$	-14.09	-9.87	91.7	92.3
$\Sigma_h^*[1,3]$	-13.94	-10.50	93.0	94.7
$\Sigma_h^*[2,3]$	-14.09	-9.87	91.7	92.3
$\Sigma_h^*[3,3]$	-25.05	10.93	91.3	95.3
$\Sigma_u^*[1,1]$	-42.08	30.30	91.7	93.7
$\Sigma_u^*[2,1]$	7.09	13.94	93.7	90.7
$\Sigma_u^*[3,1]$	5.26	13.82	94.0	94.0
$\Sigma_u^*[1,2]$	7.09	13.94	93.7	90.7
$\Sigma_u^*[2,2]$	-49.02	18.52	89.7	92.3
$\Sigma_u^*[3,2]$	5.60	12.51	93.3	94.3
$\Sigma_u^*[1,3]$	5.26	13.82	94.0	94.0
$\Sigma_u^*[2,3]$	5.60	12.51	93.3	94.3
$\Sigma_u^*[3,3]$	-41.00	27.42	89.0	93.3

7.6 Application to the RN4CAST Patient Satisfaction Data

We now apply the proposed CMFA-CRM analysis to the RN4CAST patient satisfaction data. The multi-stage approach and the MCR model using subscale sum scores as responses (proposed in Li et al., 2014) are also applied to assess the impact of scoring approaches on estimates and conclusions. We begin by addressing the nonstandard distribution of the scale items.

7.6.1 Transformation of the Response Variables and Covariates

The J-shaped patient satisfaction items in Figure 7.1 were transformed to normality using the BOS technique proposed by Lesaffre et al. (2007). This method assumes that a standardized version of the satisfaction measurement is

a coarsened latent continuous variable which has a Gaussian distribution after a logit transformation. Specifically, the observed response is first transformed to a (discrete) response y on the unit interval (by a change of scale). Then, a latent random variable z on (0,1) is assumed, with the property that $\log[z/(1-z)] \sim N(\mu, \sigma^2)$ such that y is obtained by coarsening z.

The candidate covariates involved in the mean and/or variance structures are listed in Section 7.2. For computational reasons and to make the regression coefficients of these variables comparable, covariates were standardized (mean=0, SD=1).

7.6.2 A Multilevel Exploratory Factor Analysis

To obtain an appropriate factor structure for the MFA part of the CMFA-CRM, and for completeness, we first conducted an exploratory MFA using the 9 HCAHPS items listed in Table 7.1 using Belgian sample only. We estimated an unstructured covariance matrix at the nursing unit and hospital levels. This task was performed using *Mplus* version 7.4 (Muthén and Muthén, 2015) in a frequentist fashion, and the number of factors determined on the basis of parallel analysis, fit statistics and comprehensibility of the factors (Hoyle and Duvall, 2004). We found a patient level factor structure with configuration similar to that obtained by Orindi et al. (2017) for the HCAHPS using the full multi-country RN4CAST data. That is, with the 9 items comprising the 3 factors of nurse communication (NC: 'nurses courtesy and respect', 'nurses listen carefully', 'nurses explain things well', 'got helpsoon'), doctor communication (DC: 'doctors courtesy and respect', 'doctors listen carefully', 'doctors explain things well'), and physical environment (PE: 'clean room and bathroom', 'room quiet at night'). The 3-factor patient level solution summarised the 9 items reasonably well [fit statistics: Comparative fit index (CFI)=0.984, Tucker-Lewis index (TLI)=0.970, root mean squared error of approximation (RMSEA)=0.058, and standardized root mean square residual (SRMR)=0.034]. A reasonably good fitting model is obtained when SRMR is $\leq$0.08, RMSEA is $\leq$0.06, and CFI and TLI are $\geq$0.95 (Hu and Bentler, 1999).

7.6.3 CMFA-CRM Analysis

Next, we applied our proposed CMFA-CRM to the RN4CAST patient satisfaction data. For the multilevel factor analysis part of the CMFA-CRM, we used the factor structure obtained from the exploratory analysis in Section 7.6.2.

For the MCR part, we first included covariates at the patient level (self-rated health), unit level (unit speciality), and hospital level (teaching status, technological level, hospital size expressed in terms of number of beds) as x_{ijk} and x^*_{ijk}. However unit speciality, teaching status and technological level were dropped from x^*_{ijk} as we experienced computational challenges with them. Whether to include random effects in the factor loadings part of the covariance matrix (i.e., $\boldsymbol{\nu}^*_{uh}$ and $\boldsymbol{\nu}^*_h$ in Model 7.3) was assessed by comparing the DIC of the model without and with random effects. For the model with no random effects, the DIC was 28600 with effective number of parameters $p_D = 3285$; for the model with nursing unit random effects only the DIC was 28319 with $p_D = 3110$; for the model with hospital random effects only the DIC was 27877 with $p_D = 2906$; and for the model with both nursing unit and hospital random effects the DIC was 28011 with $p_D = 2926$. Thus the model with only hospital random effects was preferred.

Full results are included in the supplementary material. Table 7.3 shows the posterior Mean estimates and the 95% credible intervals (CI) for the coefficients in the mean part (upper panel) and covariance part (lower panel) of modeling patient experiences with care. The upper panel of Table 7.3 shows that at the patient level, self-perceived health status was significantly related to all three patient experience dimensions in the sense of the 95% CIs not containing zero. Patients who reported being in better health had, on average, better nurse and doctor communication and physical environment than those who reported being in worse health. At the nursing unit level, unit speciality was significantly related to nurse communication domain only. Patients in medical or mixed nursing units experienced worse nurse communication than those in surgical units.

The lower panel of Table 7.3 presents the covariance part of modeling patient experience with care. It shows that at the patient level, health sta-

tus was significantly associated with variations in nurse communication and physical environment domains. At the hospital level, variations in doctor communication and physical environment rating were significantly associated with the number of beds in a hospital.

Figure 7.2 shows the associations between the (co)variances/correlations among the patient experience dimensions and the covariate *number of beds/hospital size* at the hospital level. It shows that the variations in the variances for doctor communication and physical environment generally increased with number of beds. Whereas the correlation between nurse and doctor communication was constant, that between nurse communication and physical environment decreased with increasing hospital size. The correlation between doctor communication and physical environment generally increased with hospital size. Figure 7.3 shows the relationships between the (co)variances/correlations among the patient experience dimensions and the covariate *health status* at the patient level. It can be seen that the correlations between the dimensions of nurse and doctor communication, and nurse communication and physical environment were optimum for average health then decreased when the health status was either worse or better. The correlation between doctor communication and physical environment increased with better self-rated health status. A clinical interpretation of these findings is given in Section 7.6.7.

Table 7.3: RN4CAST Patient Survey Data: Posterior mean and 95% credible interval for the parameters in the mean part and loadings in the covariance part of the CMFA-CRM.

Parameter	Patient Experience Domain†	Mean (95%CI)
Mean part of modeling patient satisfaction		
At the patient level		
Health status	NC	$0.251(0.190, 0.315)*$
	DC	$0.255(0.179, 0.332)*$
	PE	$0.163(0.096, 0.230)*$
At the nursing unit level		
Speciality	NC	$-0.076(-0.153, -0.0004)*$
	DC	$-0.077(-0.170, 0.018)$
	PE	$0.065(-0.009, 0.137)$
At the hospital level		
Teaching status	NC	$0.004(-0.088, 0.105)$
	DC	$-0.044(-0.160, 0.072)$
	PE	$0.031(-0.082, 0.145)$
Technology	NC	$-0.009(-0.117, 0.094)$
	DC	$0.018(-0.106, 0.137)$
	PE	$0.040(-0.074, 0.155)$
Number of beds	NC	$0.048(-0.074, 0.177)$
	DC	$0.074(-0.076, 0.231)$
	PE	$-0.124(-0.276, 0.019)$
Covariance part of modeling patient satisfaction		
At the patient level		
Health status	NC	$0.172(0.007, 0.304)*$
	DC	$-0.129(-0.326, 0.045)$
	PE	$-0.184(-0.343, -0.006)*$
At the hospital level		
Number of beds	NC	$0.091(-0.053, 0.224)$
	DC	$0.172(0.006, 0.332)*$
	PE	$0.236(0.073, 0.390)*$

†NC=Nurse communication, DC=Doctor communication, PE=Physical environment; * indicates the 95%CI does not contain zero.

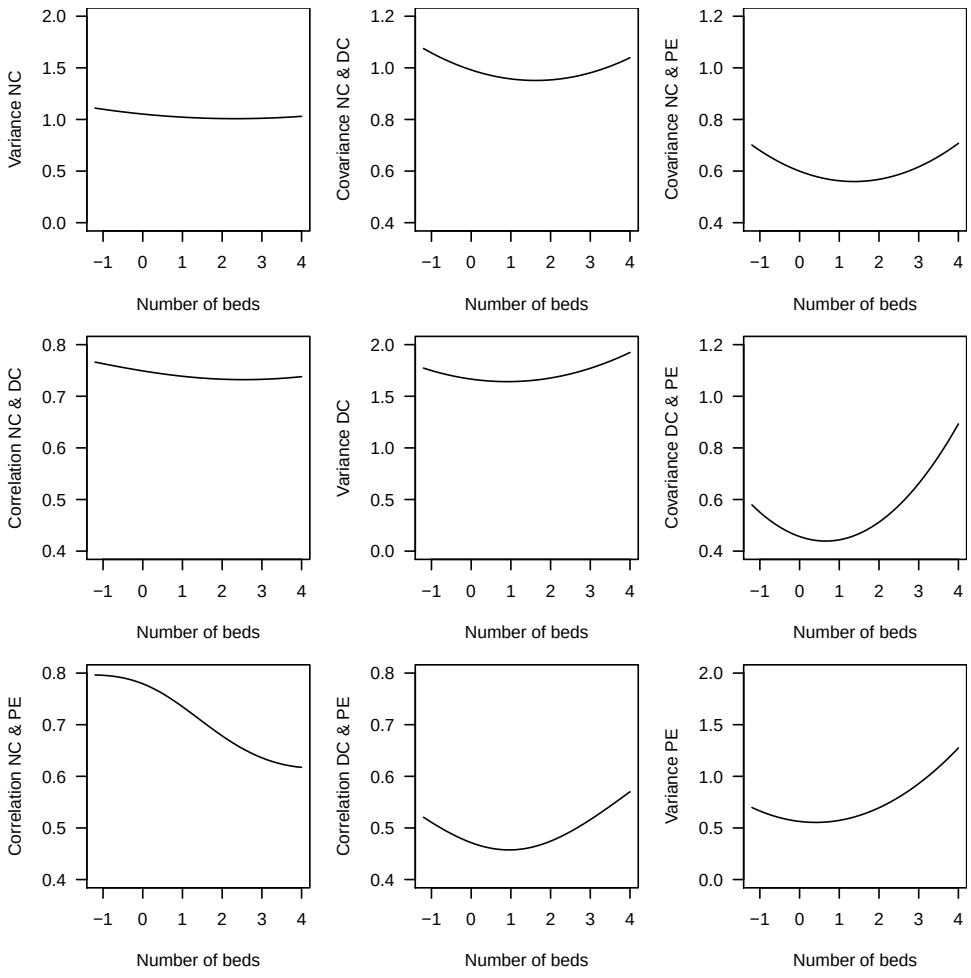

Figure 7.2: RN4CAST Patient Survey Data: (Co)variances (upper triangle) and correlations (lower triangle) with hospital size at the hospital level. NC=Nurse communication, DC=Doctor communication, and PE=Physical environment.

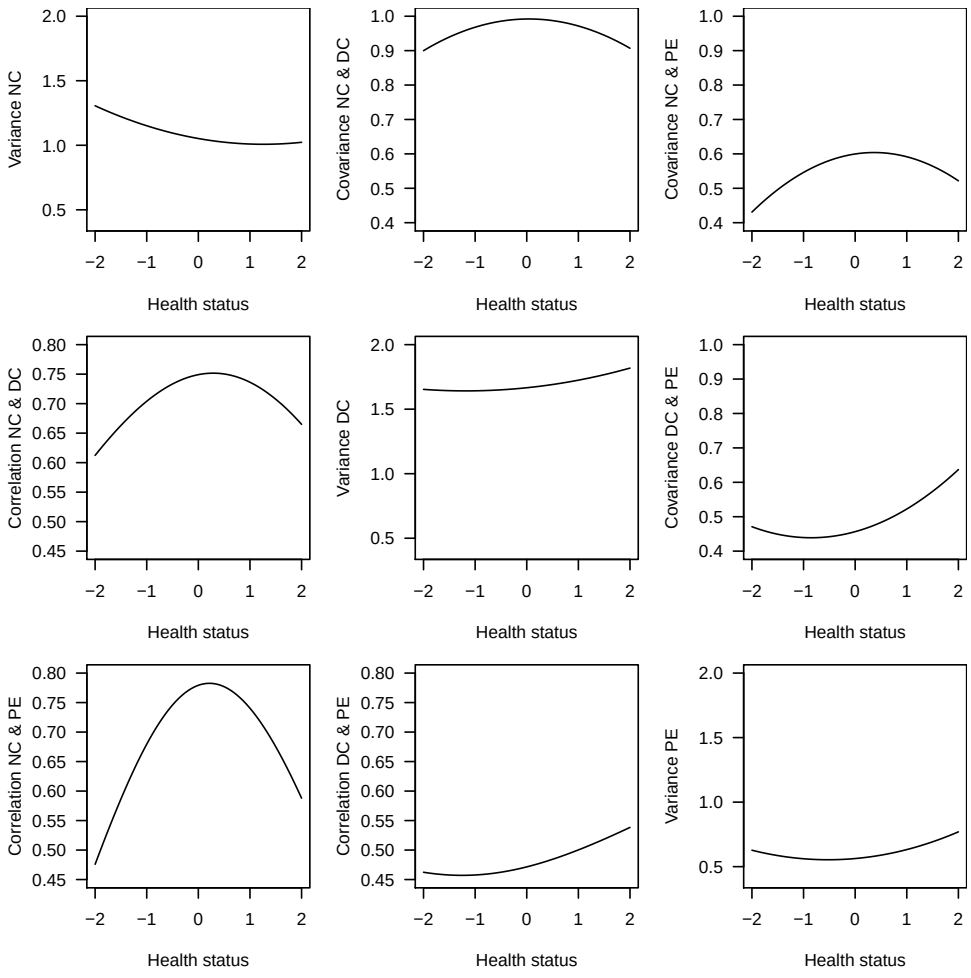

Figure 7.3: RN4CAST Patient Survey Data: (Co)variances (upper triangle) and correlations (lower triangle) with perceived health status at the patient level. NC=Nurse communication, DC=Doctor communication, and PE=Physical environment

7.6.4 Comparison with the Multi-stage Analysis

We compared the parameter estimates obtained from the final CMFA-CRM analysis with those obtained from a multi-stage analysis. The results indicated that for both the CMFA-CRM and the multi-stage model, the covariates *health status* and *unit speciality* showed significant effects on the mean part of modeling patient satisfaction at the patient and nursing unit levels, respectively. At the hospital level, however, all three covariates showed significant association at the mean part for the multi-stage model only. In the covariance part, both models found a significant effect *health status* and *number of beds (hospital size)* at the patient and hospital levels, respectively.
With regard to the magnitude of the parameter estimates, we found that the multi-stage model generally provided smaller estimates than the CMFA-CRM, especially in the covariance part, which was in line with what we observed in the simulation study. A more detailed listing of the parameter estimates from the two models was given in the Supplementary Materials. In conclusion, for our motivating data set, the CMFA-CRM seemed to show more power and efficiency for the parameter estimates.

7.6.5 Comparison with the MCR model Analysis Using Sum Scores

As the responses used in the CMFA-CRM (i.e., 9 items) are not exactly the same as those used in the MCR model (i.e., 3 sum scores from the 9 items), it is not straightforward to compare the parameter estimates from the two analyses. As such, we assessed the consistency of the two approaches by 1) studying the correlations between the sum scores and the factor scores from the CMFA-CRM; and 2) checking whether the same covariates were found to be significant in both models in the sense of the 95% credibility interval not containing zero.

Figure 7.4 shows for each dimension a scatter plot of the factor scores obtained from the congeneric model and the sum scores. It can be seen that the model yields scores that are not perfectly linearly related to the sum scores. The estimated Spearman correlation coefficients are Nurse communication

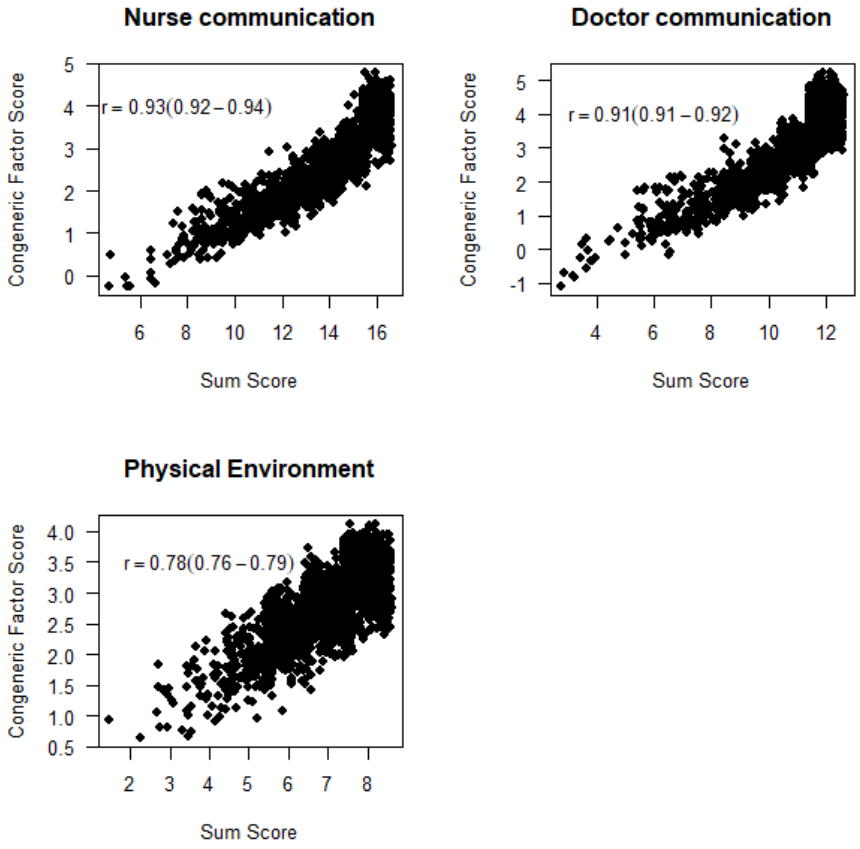

Figure 7.4: RN4CAST Patient Survey Data: Scatter plot of patient level (jittered) sum scores with congeneric factor scores from the MFA model for Nurse Communication, Doctor Communication, and Physical Environment constructs.

($r = 0.93$), Doctor communication ($r = 0.91$), and Physical environment ($r = 0.78$). Whereas these correlations are high, they are not one. In terms of significance of covariates, both the CMFA-CRM and the MCR model detected significant associations between patient experience dimensions and *health status*, both in the mean and covariance parts. Unlike the CMFA-CRM, the MCR model did not find associations with *unit specialty* (in the mean part) and *hospital size* (in the covariance part).

7.6.6 Model Assessment

The model adequacy was assessed using the posterior predictive check (PPC; Gelman et al., 1996). For the CMFA-CRM, we used the Gelman χ^2 statistic (Gelman et al., 2013) as the discrepancy function to calculate the PPC for each of the 9 items, which is defined as:

$$\chi^2(\boldsymbol{y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{[y_i - E(y_i|\boldsymbol{\theta})]^2}{var(y_i|\boldsymbol{\theta})},$$

with $\boldsymbol{y}$, in our case, representing the 9 latent continuous responses instead of the observed experience items. $\boldsymbol{\theta}$ represents all parameters including the random effects. The Bayesian P-values of PPC for the 9 experience items were 0.52, 0.51, 0.52, 0.51, 0.52, 0.51, 0.50, 0.52 and 0.51 with a mean value 0.51, indicating a good fit of the CMFA-CRM.

7.6.7 Clinical Interpretation

Three important findings emerged for patient health status and its association with experiences with care. First, patients who reported being in better health had, on average, better nurse and doctor communication and physical environment than those who reported being in worse health. Second, worse health status was associated with increased variation in nurse communication, while worse or better health status, compared to average health status, was associated with increased variation in physical environment scores. Third, correlations between the dimensions of nurse and doctor communication, and between nurse communication and physical environment were highest for average health, and decreased when the health status was either worse or better. The correlation between doctor communication and physical environment increased with better self-rated health status. The first finding is a long-standing finding in the international literature, as, for example, shown in a narrative review of patient- and hospital-level predictors of patient satisfaction (Mazurenko et al., 2017). Patient health status and other factors endogenous to the patient explain

a significant portion of the variation in patient experiences with care, and hence are consistently used in patient-mix adjustment in countries all over the world (Bruyneel et al., 2017b; Health Services Advisory Group, 2022). Xiao and Barber (2008) hypothesized that patients in better health require less medical care and interactions with healthcare providers, and as such have less risk of experiencing problems. Hall et al. (1998) argued that (1) poor health produces dissatisfaction directly in the sense that sicker patients may blame the physician for not alleviating their problems and/or may respond aggressively to frustrations; and (2) poor health produces dissatisfaction through the mediating effect of physicians' behaviour. The second and third findings are novel, and strongly suggest that predicting the mean part of the model only sheds light on part of the impact of patients' health status on their experiences with care. The findings may suggest that, among others, various conditions and comorbidities underlying the health status may affect experiences with care, but such information was not collected in this research. The type of condition has in previous research been shown to be related to patient experience scores (O'Malley et al., 2005a).

Hospital size has previously been reported to be associated with patient experiences. Some studies reported worse experiences in larger hospitals (McFarland et al., 2015, 2017). Another study observed better experiences in larger hospitals, although this was specifically for the child HCAHPS, a satisfaction survey among parents of hospitalized children (Toomey et al., 2017). For our data, no association for hospital size could be observed for the mean part of patient experiences with care. However, variation in doctor communication as well as physical environment generally increased with a larger number of beds. Not only does the variation in these two domains increases with hospital size, but also does their correlation. On the other hand, the correlation between nurse communication and physical environment decreased with increasing hospital size. The reasons for these associations are not entirely understood, and will hopefully generate new hypotheses in the field of health services research. We theorise that in larger hospitals there exist subclusters that drive patient experiences with care differently. In larger hospitals, among

others, there may be more interactions between a larger number of doctors and patients, and the physical environment may vary from one nursing unit to another, hence the larger variation.

The narrative review of patient- and hospital-level predictors of patient satisfaction reported that the association between hospital teaching status and patient satisfaction is ambiguous, with four studies concluding that patients treated in teaching hospitals were more satisfied, and one study observing no significant association. In our study also, no association was found for hospital teaching status with the mean or covariance part of patient experiences with care.

Last, we found that patients in medical or mixed nursing units experienced worse nurse communication than those in surgical units, but experiences with doctor communication and physical environment were similar. In the US HCAHPS, type of unit is expressed as surgical, medical or maternity service line, and, next to patient characteristics and survey mode, is used for case-mix adjustment. In early findings, compared to surgical patients, medical patients gave lower overall doctor ratings but relatively similar nurse and hospital ratings, which could not be confirmed in our data (O'Malley et al., 2005a). Nowadays, the patient-mix adjustment of HCAHPS scores includes only a service line by gender adjustment (Health Services Advisory Group, 2022).

7.7 Concluding Remarks

We have proposed a combined multilevel factor analysis and covariance regression model which provides a way to directly assess the heteroscedasticity of the multidimensional factor scores from a multilevel factor analysis model in a complex situation. Our results show that the proposed CMFA-CRM is quite efficient when the responses of the MCR model are latent variables which are not measured directly. Further extensions could include correlated residuals for the patient satisfaction indicators at the hierarchical levels such that $\boldsymbol{\epsilon}^{FA}$ is nolonger diagonal, but identifiability of the parameters must be ensured.

The multi-stage model is clearly less computationally intensive, and there-

fore a possibly attractive alternative to the CMFA-CRM. However, it may not only lead to biased parameter estimates if there is heteroscedasticity in the covariance part of the model, but it may also result in less efficient estimation because of the separation of the correlated information in the MFA model and the MCR model. This latter observation has also been made by e.g., Molenberghs et al. (2011) in a different frequentist context. Our results show that when using scores in subsequent analyses, the choice of scoring method can affect the conclusions of the analysis, even when the correlation between sum scores and factor scores is very high. A detailed treatment of this topic has been presented by McNeish and Wolf (2020). While sum scoring can be expressed as a parallel model (i.e., via linear transformation), optimal weighting of items with a congeneric model is a more general approach as every item is differentially related to the construct of interest and every item has a unique error variance (Graham, 2006).

We compared models using 'conditional DIC'. Whereas there is little agreement on appropriate criterion for a Bayesian model selection, Ariyo et al. (2020) recently showed that the 'marginal' criteria (averaged over random effects) are superior to 'conditional' criteria (given random effects). Therefore, the 'marginal' criteria will be implemented in the future.

Chapter 8

General Conclusions and Further Research

In this chapter, we reflect on the findings observed in Chapters 4 to 7. We also suggest potential areas for further research.

8.1 What is Covered in this Thesis

Statistical methods for analysing multilevel and multivariate data are in dire demand, not only in the health services research but also in other areas. The increasing interest in large studies, multi-country comparisons and complexity in study designs and research questions add to the need for these tools.

In this thesis, we have applied and developed hierarchical models and Bayesian SEMs to data of different types and structures. We analysed data from health services research and epidemiological studies from diverse settings. Methodological developments were inspired by the research questions in the application areas, and the merit was assessed through a simulation study. Chapter 4 handles univariate multilevel categorical (actually, ordinal) outcomes. Subsequent chapters handle multivariate single-level (Chapter 5) and multilevel (Chapters 6 and 7) outcomes which are categorical (Chapters 5 and 6) and continuous (non-standard distributed) (Chapter 7) in nature.

In Chapter 4, we handled heterogeneity in patient global ratings by means

of multilevel ordered logit models. We studied the differential magnitude of associations between HCAHPS specific patient experiences with care and global ratings across countries by means of country-by-item interactions; and explored implications for reporting these global ratings. We found that all specific experiences were associated with both global ratings, and that there were relatively little differential effects across countries. Our findings cast doubts on previous ambitious cross-country comparisons on the basis of single level analyses which have not only dichotomized the global ratings, but have also used a common cut-off point for all countries. A reporting strategy with different cut-off points across countries might provide a more equitable comparison of global ratings. Thus, research and public reporting may consider examining in more detail, within and across populations, global rating cut-off points that represent the same true level of positive patient experiences. At best we encouraged researchers to use methods that exploit fully the nature of the global ratings.

Individuals differ in the way they use rating scales to describe themselves. In Chapter 4, we mentioned in passing that one promising remedy is to correct (or "anchor") an individual's responses according to the way they use the scale when they rate an anchoring vignette (Chevalier and Fielding, 2011; King et al., 2004; Salomon et al., 2004). An anchoring vignette is a brief, hypothetical description of a fictional character who exemplifies the trait of interest (such as pain) to a lesser or greater degree (Grol-Prokopczyk et al., 2015). For example, "Laura has a headache once a month that is relieved one hour after taking a pill. During the headache she can carry on with her day to day affairs." Respondents are asked to rate their own level of the trait and, using the same set of response categories, to also rate the fictional character's level. Respondents are given multiple vignettes per domain, each representing different points along the health spectrum.

By design, vignettes are fixed across respondents (i.e., identical vignettes are given to all respondents). As such, any differences in ratings of a given vignette are considered indicative of reporting heterogeneity. That is, vignette ratings can be used to determine what different groups mean by terms such

as “mild” or “moderate”, and to statistically estimate the locations of each group’s intercategory thresholds – thereby overcoming the identification problem. Group differences in rating style can then be statistically accounted for, allowing for intergroup comparisons unbiased by reporting heterogeneity. Key identifying assumptions of the approach are 1) vignette equivalence – respondents perceiving the vignettes as representing the same absolute position on the latent health spectrum; and 2) response consistency – that respondents rate themselves using the same underlying response scale that they rate the vignettes (Grol-Prokopczyk et al., 2015).

We studied cross-group and cross-level measurement invariance for the WHO violence against women instrument and the HCAHPS in different studies. For the WHO violence against women instrument our findings support comparisons across groups. However, we reiterate that for meaningful international comparisons using this instrument a multilevel analysis assessing both cross-level and cross-country measurement invariance is needed (Chapter 5). For the HCAHPS survey, our findings showed that leaving unexamined possible noninvariance across groups and hierarchical levels may have far reaching consequences for how the public perceives their hospital’s position relative to other hospitals (Chapter 6). The sixteen HCAHPS items reflected different factor structure configuration across the hierarchical levels, with fewer factors at the nursing unit and hospital levels. This finding was consistent with multilevel factor analytic literature where researchers often found fewer number of factors at the contextual (“between”; e.g., hospital) levels than at the individual (“within”; e.g., patient) levels. This is common practice because a good fit for the within level variance-covariance matrix usually requires several factors while a good fit for the between level variance-covariance matrix can be achieved with only one, two or three factors. The within level sample size (i.e., the total sample size in the data) would be larger and misfits in the factor model are likely to be significant. The between level sample (i.e., the number of clusters) would be small and misfits in the factor model are likely to be insignificant. Thus more factors will be added on the within level and few on the between level to a chieve a good fitting model.

There was a strong indication of measurement invariance across patients with different perceptions of self-rated health and across nursing units with different specialities. However, there was strong indication of measurement noninvariance across patients from different countries, especially Poland. In Section 3.5.3, it was indicated that measurement noninvariance may be due to differences in cultures and concepts between countries (organizations, groups, etc), translation, and/or relevance of an item and overall instrument. RN4CAST was a complex twelve-country, eleven-language (Dutch, English, Finnish, French, German, Greek, Italian, Norwegian, Polish, Spanish, and Swedish) study across differently organized and financed health systems, different sociopolitical histories, and nursing professions in varying stages of professional development (Sermeus et al., 2011). To improve the chances that the instrument measures consistently and accurately the same concepts and constructs across countries (i.e., 'culture-free'), the RN4CAST research team employed a rigorous instrument development and cross-cultural adaptation processes, including translation (Squires et al., 2013) and piloting (Bruyneel et al., 2009; Rafferty et al., 2007) pre-data collection. The team quantified cross-cultural applicability of the instrument in terms of relevance score (evaluating the content, construct and conceptual relevance of the items) and translation score (evaluating if items were semantically and technically equivalent). Flaherty et al. (1988) provide a taxonomy of issues for adapting instruments for use in cross-cultural research.

We have considered a combined multilevel factor analysis model and multilevel covariance regression model in which we not only modelled the mean but also the covariance matrix of the $p-$dimensional response with a mixed effect model simultaneously, assuming a Gaussian distribution. Whereas sum scoring can be expressed as a parallel model, optimal weighting of items with a congeneric model (i.e., different factor loadings of items) is a more general approach. The basic idea of a congeneric model is that every item is differentially related to the construct of interest and every item has a unique error variance (Graham, 2006). In sum, our findings showed that this novel technique is quite efficient, especially when the responses of the MCR model are latent variables.

The Bayesian models considered in Chapters 5, 6 and 7 were fitted within a BSEM framework (Muthén and Asparouhov, 2012a). Two key factor analysis application areas for BSEM are cross-loadings and residual covariances. Whereas cross-loadings model meaningful associations between indicators and nontarget factors, residual covariances model shared sources of influence on the indicators that are unrelated to the factors, such as wording effects (e.g., negatively worded items) and context (e.g., items referring to similar contexts). It has been shown in the statistical literature that constraining cross-loadings to zero like in traditional factor analysis results in upwardly biased factor correlations (Marsh et al., 2009; Asparouhov and Muthén, 2009; Schmitt and Sass, 2011; Marsh et al., 2013). The key to a successful use of cross-loadings in a BSEM application is determining the small variance of cross-loadings prior that will allow the non-zero cross-loadings to be estimated while keeping the rest near zero. Asparouhov et al. (2015) recommend that this should be done by starting with a small prior variance that makes the BSEM and the CFA model identical and slowly increasing the variance until improvements in model fit diminish and/or quality of model identification diminishes as measured by the rate of convergence in the Bayesian iterations. The same strategy applies to BSEM residual covariances as well, although priors for the BSEM residual covariance analysis are more advanced due to its multivariate nature.

Finally, the development and dissemination of new statistical techniques is only a first step towards the actual application of these techniques. Many hurdles are to be overcome to ensure that substantive researchers apply the proposed new statistical techniques: a lack of awareness of statistical developments (Wilcox, 1998), publish or perish pressures (Fanelli, 2009; Gardenier and Resnik, 2002), unavailability and/or inaccessibility of (user-friendly) software (Aiken et al., 2008; Keselman et al., 1998), inadequate training (or capacity) in statistics (Muthén, 2017; Henson et al., 2010; Aiken et al., 2008), the failure by journal editors to mandate change (Finch et al., 2001; Thompson, 1999), mindset (i.e., atavism or a fear of deviating from normative practices; Thompson (1999)), among others. The type of data, research questions and proposed statistical approaches in this thesis are also relevant to various other research

areas, such as preventive medicine (Motl et al., 2000), education (Marsh et al., 2009), social and behavioral (Davidov et al., 2012b), environmental (Liu et al., 2003), political sciences (Ariely and Davidov, 2011), industrial research, organizational psychology and management (Lievens and Anseel, 2004) and marketing (Mavondo et al., 2003). Like ours, these research areas often collect multivariate and/or hierarchical data and, in one way or another, involve aspects of cross-cultural comparison or generalizability of measures, cross-group comparison of e.g. organizational levels, ethnic groups, gender, age-groups, etc, or scale development/validation. Therefore, we hope that our innovative approaches will allow applied researchers working in these areas to probe more deeply into mechanistic questions and bridge knowledge gaps in respective fields. After all, awareness and application of newly proposed statistical tools allow substantive researchers to generate new, previously unimaged hypotheses and expand the literature with incremental or revelatory knowledge.

8.2 Further Research

Further research could focus on some of the topics considered in this thesis. We also suggest further refinements in modeling multivariate multilevel data. Where necessary, we provide key references for interested readers.

In health services research, patients do react differently to interventions. Often the underlying mechanism that causes variability of intervention effects is not known; thus, the modest benefit observed in many interventions can be misleading as moderate average effects may be a mixture of substantial benefits for some, little, or no benefit for many and harm for a few. This is a case of unobserved heterogeneity of patients or populations. A mixture model could be an important tool for analyzing this type of heterogeneous data. In a frequentist context, Böhning (2003) provides a good treatment of mixture modeling while Schlattmann (2009) presents medical applications. In a Bayesian paradigm, Denison et al. (2002) wrote a text on classification. The Bayesian paradigm has been shown to be particularly suitable for analyzing mixture models with an unkown number of components (Richardson and Green,

1997). Due to the heterogeneous character of the RN4CAST data, it would be helpful to explore various Bayesian mixture SEM models, see e.g., Lee and Song (2012). Chapter 7 of Lee and Song (2012) introduces a general finite mixture SEM (in which the probabilities of component memberships are unkown) and a modified mixture SEM (in which the probabilities of component memberships are modelled through a multinomial logit model).

In Chapter 5, we evaluated models based on posterior predictive checks, as the only available tool in Mplus for single-level Bayesian SEM with categorical variables – underscoring underdevelopment of model evaluation tools for Bayesian SEMs. In a frequentist framework, a number of indices for assessing fit of SEMs (or factor analytic models) exist: the comparative fit index, Tucker-Lewis index, root mean square error of approximation, and standardized root-mean-square residual (Hu and Bentler, 1999). Development and evaluation of Bayesian variants of these indices, through simulation studies and analytically, would be a great addition to both the statistics and non-statistics communities. Related to this is Bayesian variable selection for the direct effects in MIMIC models. As has been demonstrated in this thesis, assessment of cross-group invariance is a necessary step before latent variable means can be compared across those groups. Often, the number of candidate grouping variables can be large and this can be problematic if they intervene on both the factors and on each of the items, which are also usually many. Whereas a number of approaches for Bayesian variable selection have been developed (see e.g., Lesaffre and Lawson, 2012), their application in the context of Bayesian SEMs is conspicuously missing. This can be an active area of research in the future. An increasingly popular Bayesian variable selection approach is the "spike and slab priors" technique which combines variable selection with regression paramters estimation.

Patient experience items measured in the RN4CAST study have non-standard, J–shaped distributions. Such types of data are often collected in a wide variety of fields. In Chapter 7, we used the BOS strategy (Lesaffre et al., 2007) to transform the data to normality to allow us apply the combined factor analytic and covariance regression model. Extending the model to

accommodate non-Gaussian outcomes will be appealing.

Scientific Acknowledgement

The data presented and analysed in this thesis were collected by the Registered Nurse Forecasting (RN4CAST) Consortium and the African Population and Health Research Center in collaboration with the London School of Hygiene and Tropical Medicine. We thank the anonymous editors and referees of the different articles for their comments that helped improve the manuscripts. We thank the jury members for their helpful comments.

Conflict of Interest

There is no conflict of interest.

Personal Contribution

Personal contribution for each of the research articles are as follows:

Chapter 4: Analysis of the RN4CAST patient survey data using multilevel models in SAS; interpretation of the results of the analysis; preparation of the manuscript.

Chapter 5: Analysis of the DREAMS IE data using Bayesian MIMIC models; interpretation of the results of the analysis using Mplus; preparation of the manuscript.

Chapter 6: Analysis of the RN4CAST patient survey data using multilevel SEM models using Mplus; the interpretation of the results of the analysis; preparation of the manuscript.

Chapter 7: Proposal of the combined multilevel factor analysis and multilevel covariance regression model; simulation study; analysis of the RN4CAST patient survey data using JAGS, interpretation of the analysis results; preparation of the manuscript.

Appendices

Appendix A

Mathematical Details

A.1 Default Priors in Mplus

Within the Bayesian framework, prior distributions need to be specified for all parameters in the model. To make Bayesian SEM accessible to a broad audience, popular software programs for analyzing Bayesian SEMs, such as Mplus, offer default prior distributions. We review the default prior settings used in Mplus and were also used for some types of models in this thesis. For a good understanding of distributions, we refer readers to standard probability and/or statistics texts.

The default prior for intercepts, regression slopes, and loading parameters in $N(0, \infty)$, unless these parameters are directly in a probit regression and in that case the default is $N(0, 5)$. The default priors for the threshold parameters in $N(0, \infty)$. The fault for variance parameters, i.e., variance-covariance blocks of size 1 is $IG(0, -1)$. The default for variance-covariance blocks of size bigger than 1 is $IW(0, -p, -1)$, where p is the size of the matrix, unless the variance-covariance matrix also includes parameters from a probit regression, and in that case the default is $IW(I, p + 1)$. For unrestricted model imputation even when all variables are continuous the default prior is $IW(I, p + 1)$. The default prior for all class proportions is the Dirichlet prior $D(10, 10, \ldots, 10)$.

Bibliography

Abel, G. A., Saunders, C. L., and Lyratzopoulos, G. (2014). Cancer patient experience, hospital performance and case mix: evidence from England. *Future Oncology*, 10(9):1589–1598.

Abraham, B. and Ledolter, J. (1979). A note on the multiple indicator-multiple cause model with several latent variables. *Metrika*, 26(1):215–217.

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. John Wiley & Sons.

Aiken, L. H., Sermeus, W., Van den Heede, K., Sloane, D. M., Busse, R., McKee, M., Bruyneel, L., Rafferty, A. M., Griffiths, P., Moreno-Casbas, M. T., et al. (2012). Patient safety, satisfaction, and quality of hospital care: cross sectional surveys of nurses and patients in 12 countries in Europe and the United States. *BMJ*, 344:e1717.

Aiken, L. H., Sloane, D., Griffiths, P., Rafferty, A. M., Bruyneel, L., McHugh, M., Maier, C. B., Moreno-Casbas, T., Ball, J. E., Ausserhofer, D., et al. (2017). Nursing skill mix in european hospitals: cross-sectional study of the association with mortality, patient ratings, and quality of care. *BMJ Quality & Safety*, 26(7):559–568.

Aiken, L. S., West, S. G., and Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of aiken, west, sechrest, and reno's (1990) survey of phd programs in north america. *American Psychologist*, 63(1):32–50.

Alhabib, S., Nur, U., and Jones, R. (2010). Domestic violence against

women: Systematic review of prevalence studies. *Journal of Family Violence*, 25(4):369–382.

Altman, D. G. and Royston, P. (2006). The cost of dichotomising continuous variables. *BMJ*, 332:1080.

Anderson, J. C. and Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, 49(2):155–173.

Ariely, G. and Davidov, E. (2011). Can we rate public support for democracy in a comparable way? cross-national equivalence of democratic attitudes in the world value survey. *Social Indicators Research*, 104(2):271–286.

Ariyo, O., Quintero, A., Muñoz, J., Verbeke, G., and Lesaffre, E. (2020). Bayesian model selection in linear mixed models for longitudinal data. *Journal of Applied Statistics*, 47(5):890–913.

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

Asparouhov, T. and Muthén, B. (2010). Bayesian analysis of latent variable models using Mplus. *Retrieved June*, 17:2014.

Asparouhov, T. and Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(4):495–508.

Asparouhov, T., Muthén, B., and Morin, A. J. (2015). Bayesian Structural Equation Modeling with Cross-Loadings and Residual Covariances: Comments on Stromeyer et al. *Journal of Management*.

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4):1281–1311.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*. Routledge.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107(2):238.

Bertakis, K. D., Franks, P., and Azari, R. (2002). Effects of physician gender on patient satisfaction. *Journal of the American Medical Women's Association (1972)*, 58(2):69–75.

Birdthistle, I., Schaffnit, S. B., Kwaro, D., Shahmanesh, M., Ziraba, A., Kabiru, C. W., Phillips-Howard, P., Chimbindi, N., Ondeng'e, K., and Gourlay, A. (2018). Evaluating the impact of the DREAMS partnership to reduce HIV incidence among adolescent girls and young women in four settings: a study protocol. *BMC Public Health*, 18(1):912.

Böhning, D. (2003). *Computer-Assisted Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping and Others*. Chapman and Hall/CRC press, Boca Raton.

Bollen, K. A. (1987). Outliers and improper solutions: A confirmatory factor analysis example. *Sociological Methods & Research*, 15(4):375–384.

Bollen, K. A. (2014). *Structural Equations with Latent Variables*. John Wiley & Sons.

Boulding, W., Glickman, S. W., Manary, M. P., Schulman, K. A., and Staelin, R. (2011). Relationship between patient satisfaction with inpatient care and hospital readmission within 30 days. *The American Journal of Managed Care*, 17(1):41–48.

Brooks, S. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Brown, T. A. (2014). *Confirmatory Factor Analysis for Applied Research*. Guilford Publications.

Browne, M. W. and Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21(2):230–258.

Bruyneel, L., Heede, K. V. d., Diya, L., Aiken, L., and Sermeus, W. (2009). Predictive validity of the International Hospital Outcomes Study questionnaire: an RN4CAST pilot study. *Journal of Nursing Scholarship*, 41(2):202–210.

Bruyneel, L., Li, B., Ausserhofer, D., Lesaffre, E., Dumitrescu, I., Smith, H. L., Sloane, D. M., Aiken, L. H., and Sermeus, W. (2015). Organization of hospital nursing, provision of nursing care, and patient experiences with care in Europe. *Medical Care Research and Review*, 72(6):643–664.

Bruyneel, L., Li, B., Squires, A., Spotbeen, S., Meuleman, B., Lesaffre, E., and Sermeus, W. (2017a). Bayesian multilevel MIMIC modeling for studying measurement invariance in cross-group comparisons. *Medical Care*, 55(4):e25–e35.

Bruyneel, L., Tambuyzer, E., Coeckelberghs, E., De Wachter, D., Sermeus, W., De Ridder, D., Ramaekers, D., Weeghmans, I., and Vanhaecht, K. (2017b). New instrument to measure hospital patient experiences in flanders. *International Journal of Environmental Research and Public Health*, 14(11):1319.

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292.

Byrne, B. M. and Campbell, T. L. (1999). Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure: A look beneath the surface. *Journal of Cross-Cultural Psychology*, 30(5):555–574.

Cain, M. K. and Zhang, Z. (2019). Fit for a Bayesian: An Evaluation of PPP and DIC for Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 26(1):1–12.

Carle, A. C. and Weech-Maldonado, R. (2012). Does the Consumer Assessment of Healthcare Providers and Systems Cultural Competence Survey provide equivalent measurement across English and Spanish versions? *Medical Care*, 50(9 Suppl 2):S37.

Carle, A. C., Weech-Maldonado, R., Ngo-Metzger, Q., and Hays, R. D. (2012). Evaluating measurement equivalence across race and ethnicity on the CAHPS® Cultural Competence Survey. *Medical Care*, 50(9 Suppl 2):S32–S36.

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, 1(4):651–673.

Centers for Medicare and Medicaid Services (2016). Hospital Compare Webpage. `https://www.medicare.gov/hospitalcompare/search.html`. [Online; accessed October 5,2016].

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, 95(5):1005.

Chevalier, A. and Fielding, A. (2011). An introduction to anchoring vignettes. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(3):569–574.

Chib, S. and Carlin, B. P. (1999). On MCMC sampling in hierarchical longitudinal models. *Statistics and Computing*, 9(1):17–26.

Chiu, T. Y., Leonard, T., and Tsui, K.-W. (1996). The matrix-logarithmic covariance model. *Journal of the American Statistical Association*, 91(433):198–210.

Collins, L., Schafer, J., and Kam, C. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4):330–351.

Crofton, C., Lubalin, J. S., and Darby, C. (1999). Consumer Assessment of Health Plans Study (CAHPS). Foreword. *Medical Care*, 37(3 Suppl):MS1–9.

Damiano, P. C., Elliott, M., Tyler, M. C., and Hays, R. D. (2004). Differential use of the cahps® 0–10 global rating scale by medicaid and commercial populations. *Health Services and Outcomes Research Methodology*, 5(3):193–205.

Darby, C., Hays, R. D., and Kletke, P. (2005). Development and evaluation of the CAHPS® Hospital Survey. *Health Services Research*, 40(6):1973–1976.

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., and Meuleman, B. (2012a). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology*, 43(4):558–575.

Davidov, E., D́ulmer, H., Schĺuter, E., Schmidt, P., and Meuleman, B. (2012b). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology*, 43(4):558–575.

Delnoij, D. M. (2009). Measuring patient experiences in europe: what can we learn from the experiences in the usa and england? *The European Journal of Public Health*, 19(4):354–356.

Denison, D. G., Holmes, C. C., Mallick, B. K., and Smith, A. F. (2002). *Bayesian Methods for Nonlinear Classification and Regression*, volume 386. John Wiley & Sons.

DiStefano, C., Liu, J., Jiang, N., and Shi, D. (2018). Examination of the weighted root mean square residual: Evidence for trustworthiness? *Structural Equation Modeling: A Multidisciplinary Journal*, 25(3):453–466.

Diya, L., Li, B., Heede, K., Sermeus, W., and Lesaffre, E. (2014). Multilevel factor analytic models for assessing the relationship between nurse-reported adverse events and patient safety. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 177(1):237–257.

Diya, L., Van den Heede, K., Sermeus, W., and Lesaffre, E. (2012). The relationship between in-hospital mortality, readmission into the intensive care nursing unit and/or operating theatre and nurse staffing levels. *Journal of Advanced Nursing*, 68(5):1073–1081.

Elliott, M. N., Haviland, A. M., Kanouse, D. E., Hambarsoomian, K., and Hays, R. D. (2009a). Adjusting for subgroup differences in extreme response tendency in ratings of health care: impact on disparity estimates. *Health Services Research*, 44(2p1):542–561.

Elliott, M. N., Kanouse, D. E., Edwards, C. A., and Hilborne, L. H. (2009b). Components of care vary in importance for overall patient-reported experience by type of hospitalization. *Medical Care*, 47(8):842–849.

Elliott, M. N., Lehrman, W. G., Beckett, M. K., Goldstein, E., Hambarsoomian, K., and Giordano, L. A. (2012). Gender differences in patients' perceptions of inpatient care. *Health Services Research*, 47(4):1482–1501.

Elliott, M. N., Lehrman, W. G., Goldstein, E., Hambarsoomian, K., Beckett, M. K., and Giordano, L. A. (2009c). Do hospitals rank differently on HCAHPS for different patient subgroups? *Medical Care Research and Review*, 67(1):56–73.

Elliott, M. N., Swartz, R., Adams, J., Spritzer, K. L., and Hays, R. D. (2001). Case-mix adjustment of the National CAHPS benchmarking data 1.0: a violation of model assumptions? *Health Services Research*, 36(3):555.

Elliott, M. N., Zaslavsky, A. M., Goldstein, E., Lehrman, W., Hambarsoomians, K., Beckett, M. K., and Giordano, L. (2009d). Effects of survey mode, patient mix, and nonresponse on CAHPS® hospital survey scores. *Health Services Research*, 44(2):501–518.

Ellsberg, M., Heise, L., Pena, R., Agurto, S., and Winkvist, A. (2001). Researching domestic violence against women: methodological and ethical considerations. *Studies in Family Planning*, 32(1):1–16.

Ellsberg, M., Jansen, H. A., Heise, L., Watts, C. H., and Garcia-Moreno, C. (2008). Intimate partner violence and women's physical and mental health in the who multi-country study on women's health and domestic violence: an observational study. *The Lancet*, 371(9619):1165–1172.

Fanelli, D. (2009). How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PLoS ONE*, 4(5):e5738.

Filmer, D. and Pritchett, L. H. (2001). Estimating wealth effects without expenditure data–or tears: an application to educational enrollments in states of india. *Demography*, 38(1):115–132.

Finch, S., Cumming, G., and Thomason, N. (2001). Reporting of statistical inference in the journal of applied psychology: Little evidence of reform. *Educational and Psychological Measurement*, 61:181–210.

Flaherty, J. A., Gaviria, F. M., Pathak, D., Mitchell, T., Wintrob, R., Richman, J. A., and Birz, S. (1988). Developing instruments for cross-cultural psychiatric research. *Journal of Nervous and Mental Disease*, 176(5):260–263.

Gajewski, B. J., Boyle, D. K., Miller, P. A., Oberhelman, F., and Dunton, N. (2010). A multilevel confirmatory factor analysis of the practice environment scale: a case study. *Nursing Research*, 59(2):147–153.

Garcia-Moreno, C., Jansen, H. A., Ellsberg, M., Heise, L., and Watts, C. H. (2006). Prevalence of intimate partner violence: findings from the who multi-country study on women's health and domestic violence. *The Lancet*, 368(9543):1260–1269.

Gardenier, J. and Resnik, D. (2002). The misuse of statistics: concepts, tools, and a research agenda. *Accountability in Research: Policies and Quality Assurance*, 9(2):65–74.

Garrett, E. S. and Zeger, S. L. (2000). Latent class model diagnosis. *Biometrics*, 56(4):1055–1067.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, 85(412):972–985.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelin, M. N. and Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the center for epidemiologic studies depression scale. *Educational and Psychological Measurement*, 63(1):65–74.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge university press.

Gelman, A., Jakulin, A., Pittau, M. G., Su, Y.-S., et al. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–760.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov chain Monte Carlo in Practice*. Chapman and Hall/CRC.

Giordano, L. A., Elliott, M. N., Goldstein, E., Lehrman, W. G., and Spencer, P. A. (2010). Development, implementation, and public reporting of the HCAHPS survey. *Medical Care Research and Review*, 67(1):27–37.

Goldstein, E., Elliott, M. N., Lehrman, W. G., Hambarsoomian, K., and Giordano, L. A. (2009). Racial/ethnic differences in patients' perceptions of inpatient care using the HCAHPS survey. *Medical Care Research and Review*, 67(1):74–92.

Goldstein, H. (2010). *Multilevel Statistical Models*. Wiley, 4th edition.

Gorsuch, R. L. (1988). Exploratory factor analysis. In *Handbook of multivariate experimental psychology*, pages 231–258. Springer.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement*, 66(6):930–944.

Greenleaf, E. A. (1992). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research*, 29(2):176–188.

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.

Grol-Prokopczyk, H., Verdes-Tennant, E., McEniry, M., and Ispány, M. (2015). Promises and pitfalls of anchoring vignettes in health survey research. *Demography*, 52(5):1703–1728.

Hall, J. A., Milburn, M. A., Roter, D. L., and Daltroy, L. H. (1998). Why are sicker patients less satisfied with their medical care? tests of two explanatory models. *Health Psychology*, 17(1):70.

Hargraves, J. L., Wilson, I. B., Zaslavsky, A., James, C., Walker, J. D., Rogers, G., and Cleary, P. D. (2001). Adjusting for patient characteristics when analyzing reports from patients about hospital care. *Medical Care*, 39(6):635–641.

Harrell, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387.

Harrell Jr, F. E., with contributions from Charles Dupont, and many others. (2016). *Hmisc: Harrell Miscellaneous*. R package version 4.0-2.

Hawkins, J. B., Brownstein, J. S., Tuli, G., Runels, T., Broecker, K., Nsoesie, E. O., McIver, D. J., Rozenblum, R., Wright, A., Bourgeois, F. T., et al. (2016). Measuring patient-perceived quality of care in us hospitals using twitter. *BMJ Quality & Safety*, 25(6):404–413.

Hayashi, K. and Marcoulides, G. A. (2006). Teacher's corner: Examining identification issues in factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(4):631–645.

Hays, R. D., Mallett, J. S., Haas, A., Kahn, K. L., Martino, S. C., Gaillot, S., and Elliott, M. N. (2018). Associations of cahps composites with global ratings of the doctor vary by medicare beneficiaries' health status. *Medical Care*, 56(8):736–739.

Health Services Advisory Group (2022). Patient-Mix Adjustments and National Means for January 2022 Publicly Reported HCAHPS Results. `https://hcahpsonline.org/en/mode--patient-mix-adj/`.

Heise, L. L. and Kotsadam, A. (2015). Cross-national and multilevel correlates of partner violence: an analysis of data from population-based surveys. *The Lancet Global Health*, 3(6):e332–e340.

Henson, R. K., Hull, D. M., and Williams, C. S. (2010). Methodology in our education research culture: Toward a stronger collective quantitative proficiency. *Educational Researcher*, 39(3):229–240.

Hoff, P. D. and Niu, X. (2012). A covariance regression model. *Statistica Sinica*, 22(2):729–753.

Hoofs, H., van de Schoot, R., Jansen, N. W., and Kant, I. (2018). Evaluating model fit in Bayesian confirmatory factor analysis with large samples: Simulation study introducing the BRMSEA. *Educational and Psychological Measurement*, 78(4):537–568.

Hox, J. (2010). *Multilevel Analysis: Techniques and Applications*. Routledge, 2nd edition.

Hoyle, R. H. and Duvall, J. L. (2004). Determining the number of factors in exploratory and confirmatory factor analysis. In Kaplan, D., editor, *The Sage handbook of quantitative methodology for the social sciences*, pages 301–315. Sage, Thousand Oaks, CA.

Hu, L.-t. and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1):1–55.

Huang, F. L. and Cornell, D. G. (2015). Using Multilevel Factor Analysis With Clustered Data Investigating the Factor Structure of the Positive Values Scale. *Journal of Psychoeducational Assessment*, 34(1):3–14.

Hurtado, M. P., Angeles, J., Blahut, S. A., and Hays, R. D. (2005). Assessment of the equivalence of the spanish and english versions of the cahps® hospital survey on the quality of inpatient care. *Health Services Research*, 40(6 Part II):2140–2161.

Jak, S. and Jorgensen, T. D. (2017). Relating measurement invariance, cross-level invariance, and multilevel reliability. *Frontiers in Psychology*, 8:1640.

Jha, A. K., Orav, E. J., Zheng, J., and Epstein, A. M. (2008). Patients' perception of hospital care in the united states. *New England Journal of Medicine*, 359(18):1921–1931.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 36(4):409–426.

Jöreskog, K. G. and Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 70(351a):631–639.

Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 8(3):325–352.

Keller, S., O'Malley, A. J., Hays, R. D., Matthew, R. A., Zaslavsky, A. M., Hepner, K. A., and Cleary, P. D. (2005). Methods used to streamline the CAHPS® hospital survey. *Health Services Research*, 40(6 Part II):2057–2077.

Kemp, K. A., Chan, N., McCormack, B., and Douglas-England, K. (2015). Drivers of inpatient hospital experience using the HCAHPS survey in a canadian setting. *Health Services Research*, 50(4):982–997.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., et al. (1998). Statistical practices of educational researchers: An analysis of their anova, manova, and ancova analyses. *Review of Educational Research*, 68(3):350–386.

Killip, S., Mahfoud, Z., and Pearce, K. (2004). What is an intracluster correlation coefficient? crucial concepts for primary care researchers. *The Annals of Family Medicine*, 2(3):204–208.

King, G., Murray, C. J., Salomon, J. A., and Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, 98(1):191–207.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Kolenikov, S. and Bollen, K. A. (2012). Testing negative error variances: Is a heywood case a symptom of misspecification? *Sociological Methods & Research*, 41(1):124–167.

Krug, E. G., Mercy, J. A., Dahlberg, L. L., and Zwi, A. B. (2002). The world report on violence and health. *The Lancet*, 360(9339):1083–1088.

Krull, J. L. and MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research*, 36(2):249–277.

Kury, H., Obergfell-Fuchs, J., and Woessner, G. (2004). The extent of family violence in europe: A comparison of national surveys. *Violence Against Women*, 10(7):749–769.

Lake, E. T. (2002). Development of the practice environment scale of the Nursing Work Index. *Research in Nursing & Health*, 25(3):176–188.

Lee, S.-Y. and Song, X.-Y. (2012). *Basic and Advanced Bayesian Structural Equation Modeling: With Applications in the Medical and Behavioral Sciences*. John Wiley & Sons.

Lesaffre, E. and Lawson, A. B. (2012). *Bayesian Biostatistics*. John Wiley & Sons.

Lesaffre, E., Rizopoulos, D., and Tsonaka, R. (2007). The logistic transform for bounded outcome scores. *Biostatistics*, 8(1):72–85.

Levy, R. and Mislevy, R. J. (2016). *Bayesian psychometric modeling*. CRC Press.

Li, B., Bruyneel, L., and Lesaffre, E. (2014). A multivariate multilevel Gaussian model with a mixed effects structure in the mean and covariance part. *Statistics in Medicine*, 33(11):1877–1899.

Li, B., Bruyneel, L., Sermeus, W., Van den Heede, K., Matawie, K., Aiken, L., and Lesaffre, E. (2013). Group-level impact of work environment dimensions

on burnout experiences among nurses: a multivariate multilevel probit model. *International Journal of Nursing Studies*, 50(2):281–291.

Lievens, F. and Anseel, F. (2004). Confirmatory factor analysis and invariance of an organizational citizenship behaviour measure across samples in a dutch-speaking context. *Journal of Occupational and Organizational Psychology*, 77(3):299–306.

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Liu, C.-W., Lin, K.-H., and Kuo, Y.-M. (2003). Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in taiwan. *Science of the Total Environment*, 313(1-3):77–89.

Longford, N. and Muthén, B. (1992). Factor analysis for clustered observations. *Psychometrika*, 57(4):581–597.

Marsh, H. W., L′udtke, O., Nagengast, B., Morin, A. J., and Von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right–camouflaging misspecification with item parcels in cfa models. *Psychological Methods*, 18(3):257.

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J., and Trautwein, U. (2009). Exploratory structural equation modeling, integrating cfa and efa: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3):439–476.

Maruyama, G. M. (1997). *Basics of Structural Equation Modeling*. SAGE publications.

Mavondo, F., Gabbott, M., and Tsarenko, Y. (2003). Measurement invariance of marketing instruments: An implication across countries. *Journal of Marketing Management*, 19(5-6):523–540.

Mazurenko, O., Collum, T., Ferdinand, A., and Menachemi, N. (2017). Predictors of Hospital Patient Satisfaction as Measured by HCAHPS: A Systematic Review. *Journal of Healthcare Management*, 62(4):272–283.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B*, 42(2):109–127.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Atatistical Association*, 92(437):162–170.

McFarland, D. C., Ornstein, K. A., and Holcombe, R. F. (2015). Demographic factors and hospital size predict patient satisfaction variance–implications for hospital value-based purchasing. *Journal of Hospital Medicine*, 10:503–9.

McFarland, D. C., Shen, M. J., Parker, P., Meyerson, S., and Holcombe, R. F. (2017). Does hospital size affect patient satisfaction? *Quality Management in Health Care*, 26(4):205–209.

McNeish, D. and Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, 52(6):2287–2305.

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

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

Miller, A. (2002). *Subset Selection in Regression*. Chapman and Hall/CRC.

Molenberghs, G. and Kenward, M. (2007). *Missing Data in Clinical Studies*, volume 61. John Wiley & Sons.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Spriner.

Molenberghs, G., Verbeke, G., and Iddi, S. (2011). Pseudo-likelihood methodology for partitioned large and complex samples. *Statistics & Probability Letters*, 81(7):892–901.

Motl, R. W., Dishman, R. K., Trost, S. G., Saunders, R. P., Dowda, M., Felton, G., Ward, D. S., and Pate, R. R. (2000). Factorial validity and invariance of questionnaires measuring social-cognitive determinants of physical activity among adolescent girls. *Preventive Medicine*, 31(5):584–594.

Mulaik, S. A. (2009). *Foundations of Factor Analysis*. Chapman and Hall/CRC.

Muthén, B. (2017). Teaching students of educational psychology new sophisticated statistical techniques. In *The future of educational psychology*, pages 181–189. Routledge.

Muthén, B. and Asparouhov, T. (2012a). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychological Methods*, 17(3):313.

Muthén, B. and Asparouhov, T. (2012b). Rejoinder to MacCallum, Edwards, and Cai (2012) and Rindskopf (2012): Mastering a new method. *Psychological Methods*, 17(3):346–353.

Muthén, B. and Asparouhov, T. (2014). IRT studies of many groups: the alignment method. *Frontiers in Psychology*, 5(978):1–7.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, 54(4):557–585.

Muthén, L. K. and Muthén, B. O. (1998–2015). *Mplus User's Guide. 7th Edition*. Los Angeles, CA: Muthén & Muthén.

Nesselroade, J. R. and Cattell, R. B. (2013). *Handbook of multivariate experimental psychology*. Springer Science & Business Media.

Nybergh, L., Taft, C., and Krantz, G. (2012). Psychometric properties of the WHO Violence Against Women instrument in a male population-based sample in Sweden. *BMJ Open*, 2(6):e002055.

Nybergh, L., Taft, C., and Krantz, G. (2013). Psychometric properties of the WHO Violence Against Women instrument in a female population-based sample in Sweden: a cross-sectional survey. *BMJ Open*, 3(5):e002053.

O'Malley, A. J., Zaslavsky, A. M., Elliott, M. N., Zaborski, L., and Cleary, P. D. (2005a). Case-mix adjustment of the CAHPS® Hospital Survey. *Health Services Research*, 40(6 Part II):2162–2181.

O'Malley, A. J., Zaslavsky, A. M., Hays, R. D., Hepner, K. A., Keller, S., and Cleary, P. D. (2005b). Exploratory factor analyses of the CAHPS® hospital pilot survey responses across and within medical, surgical, and obstetric services. *Health Services Research*, 40(6 Part II):2078–2095.

Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika*, 6(2):150–166.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 5(2):107–124.

Orindi, B. O., Lesaffre, E., Sermeus, W., and Bruyneel, L. (2017). Impact of cross-level measurement noninvariance on hospital rankings based on patient experiences with care in 7 European countries. *Medical Care*, 55(12):e150–e157.

Orindi, B. O., Maina, B. W., Muuo, S. W., Birdthistle, I., Carter, D. J., Floyd, S., and Ziraba, A. (2020). Experiences of violence among adolescent girls and young women in Nairobi's informal settlements prior to scale-up of the DREAMS Partnership: Prevalence, severity and predictors. *PLoS ONE*, 15(4):e0231737.

Paddison, C., Elliott, M., Parker, R., Staetsky, L., Lyratzopoulos, G., Campbell, J. L., and Roland, M. (2012). Should measures of patient experience in primary care be adjusted for case mix? Evidence from the English General Practice Patient Survey. *BMJ Quality Safety*, 21(8):634–640.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *The 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March.*

Plummer, M. (2013). *Package rjags version 3-10.*

Pornprasertmanit, S., Lee, J., and Preacher, K. J. (2014). Ignoring clustering in confirmatory factor analysis: Some consequences for model fit and standardized parameter estimates. *Multivariate Behavioral Research*, 49(6):518–543.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690.

Quintero, A. and Lesaffre, E. (2017). Multilevel covariance regression with correlated random effects in the mean and variance structure. *Biometrical Journal*, 59(5):1047–1066.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rafferty, A. M., Clarke, S. P., Coles, J., Ball, J., James, P., McKee, M., and Aiken, L. H. (2007). Outcomes of variation in hospital nurse staffing in english hospitals: cross-sectional analysis of survey data and discharge records. *International Journal of Nursing Studies*, 44(2):175–182.

Raleigh, V., Sizmur, S., Tian, Y., and Thompson, J. (2015). Impact of case-mix on comparisons of patient-reported experience in NHS acute hospital trusts in England. *Journal of Health Services Research & Policy*, 20(2):92–99.

Ranard, B. L., Werner, R. M., Antanavicius, T., Schwartz, H. A., Smith, R. J., Meisel, Z. F., Asch, D. A., Ungar, L. H., and Merchant, R. M. (2016). Yelp reviews of hospital care can supplement and inform traditional surveys of the patient experience of care. *Health Affairs*, 35(4):697–705.

Ribeiro, M. R. C., Alves, M. T. S. S. d. B. E., Batista, R. F. L., Ribeiro, C. C. C., Schraiber, L. B., Barbieri, M. A., Bettiol, H., and da Silva, A. A. M. (2014). Confirmatory factor analysis of the who violence against women

instrument in pregnant women: results from the brisa prenatal cohort. *PLoS ONE*, 9(12):e115382.

Ribeiro, M. R. C., Silva, A. A. M. d., Alves, M. T. S. S. d. B. e., Batista, R. F. L., Ribeiro, C. C. C., Schraiber, L. B., Bettiol, H., and Barbieri, M. A. (2017). Effects of socioeconomic status and social support on violence against pregnant women: a structural equation modeling analysis. *PLoS ONE*, 12(1):e0170469.

Richardson, S. and Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):731–792.

Rindskopf, D. (1984). Structural equation models: Empirical identification, Heywood cases, and related problems. *Sociological Methods & Research*, 13(1):109–119.

Robertson, D. and Williams, G. H. (2009). *Clinical and Translational Science: Principles of Human Research*. Academic Press.

Salomon, J. A., Tandon, A., and Murray, C. J. (2004). Comparability of self rated health: cross sectional multi-country survey using anchoring vignettes. *BMJ*, 328(7434):258.

Sass, D. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, page 0734282911406661.

Sato, M. (1987). Pragmatic treatment of improper solutions in factor analysis. *Annals of the Institute of Statistical Mathematics*, 39(2):443–455.

Saul, J., Bachman, G., Allen, S., Toiv, N. F., and Cooney, C. (2018). The DREAMS core package of interventions: A comprehensive approach to preventing HIV among adolescent girls and young women. *PloS ONES*, 13(12):e0208167.

Saunders, C. L., Elliott, M. N., Lyratzopoulos, G., and Abel, G. A. (2016). Do differential response rates to patient surveys between organizations lead to unfair performance comparisons?: evidence from the english cancer patient experience survey. *Medical Care*, 54(1):45.

Schlattmann, P. (2009). *Medical Applications of Finite Mixture Models.* Springer.

Schmitt, T. A. and Sass, D. A. (2011). Rotation criteria and hypothesis testing for exploratory factor analysis: Implications for factor pattern loadings and interfactor correlations. *Educational and Psychological Measurement*, 71(1):95–113.

Schraiber, L. B., Latorre, M. d. R. D. O., França Jr, I., Segri, N. J., and d'Oliveira, A. F. P. L. (2010). Validity of the WHO VAW study instrument for estimating gender-based violence against women. *Revista de Saude Publica*, 44(4):658–666.

Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys implications for policy and practice. *Educational Evaluation and Policy Analysis*, 36(3):259–280.

Sermeus, W., Aiken, L., Van den Heede, K., Rafferty, A., Griffiths, P., Moreno-Casbas, M., Busse, R., Lindqvist, R., Scott, A., Bruyneel, L., et al. (2011). Nurse forecasting in Europe (RN4CAST): Rationale, design and methodology. *BMC Nursing*, 10(1):1–9.

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods*, 18(4):572.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, 27(2):229–239.

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

Spiegelhalter, D., Best, N., Carlin, B., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

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

Spiegelhalter, D. J., Thomas, A., Best, N., and Lunn, D. (2003). Winbugs version 1.4 user manual. *MRC Biostatistics Unit, Cambridge. URL http://www. mrc-bsu. cam. ac. uk/bugs*.

Squires, A., Aiken, L. H., van den Heede, K., Sermeus, W., Bruyneel, L., Lindqvist, R., Schoonhoven, L., Stromseng, I., Busse, R., Brzostek, T., et al. (2013). A systematic survey instrument translation process for multi-country, comparative health workforce studies. *International Journal of Nursing Studies*, 50(2):264–273.

Squires, A., Bruyneel, L., Aiken, L. H., Van den Heede, K., Brzostek, T., Busse, R., Ensio, A., Schubert, M., Zikos, D., and Sermeus, W. (2012). Cross-cultural evaluation of the relevance of the HCAHPS survey in five European countries. *International Journal for Quality in Health Care*, 24(5):470–475.

Steenkamp, J.-B. E. and Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25(1):78–107.

Stöckl, H., Devries, K., Rotstein, A., Abrahams, N., Campbell, J., Watts, C., and Moreno, C. G. (2013). The global prevalence of intimate partner homicide: a systematic review. *The Lancet*, 382(9895):859–865.

Straus, M. A. (2017). Measuring intrafamily conflict and violence: The conflict tactics (ct) scales. In *Physical violence in American families*, pages 29–48. Routledge.

Straus, M. A., Hamby, S. L., Boney-McCoy, S., and Sugarman, D. B. (1996).

The revised conflict tactics scales (CTS2) development and preliminary psychometric data. *Journal of Family Issues*, 17(3):283–316.

Thompson, B. (1999). Why "encouraging" effect size reporting is not working: The etiology of researcher resistance to changing practices. *The Journal of Psychology*, 133(2):133–140.

Toomey, S. L., Elliott, M. N., Zaslavsky, A. M., Klein, D. J., Ndon, S., Hardy, S., Wu, M., and Schuster, M. A. (2017). Variation in family experience of pediatric inpatient care as measured by child HCAHPS. *Pediatrics*, 139(4):e20163372.

Tucker, L. R. and Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1):1–10.

United Nations (1995). The fourth world conference on women, beijing, china, sept 4-15, 1995. Report, World Health Organization.

van de Schoot, R., Lugtig, P., and Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, 9(4):486–492.

Van de Vijver, F. J. and Poortinga, Y. H. (2002). Structural equivalence in multilevel research. *Journal of Cross-Cultural Psychology*, 33(2):141–156.

Van de Vijver, F. J. and Watkins, D. (2006). Assessing similarity of meaning at the individual and country level. *European Journal of Psychological Assessment*, 22(2):69–77.

Van Driel, O. P. (1978). On various causes of improper solutions in maximum likelihood factor analysis. *Psychometrika*, 43(2):225–243.

Vandenberg, R. J. and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1):4–70.

Vincent, D. F. (1953). The origin and development of factor analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2(2):107–117.

Vries, H., Elliott, M. N., Hepner, K. A., Keller, S. D., and Hays, R. D. (2005). Equivalence of mail and telephone responses to the CAHPS® Hospital Survey. *Health Services Research*, 40(6 Part II):2120–2139.

Vyas, S. and Kumaranayake, L. (2006). Constructing socio-economic status indices: How to use principal components analysis. *Health Policy and Planning*, 21(6):459–468.

Wagenmakers, E.-J., Lee, M., Lodewyckx, T., and Iverson, G. J. (2008). Bayesian versus frequentist inference. In *Bayesian evaluation of informative hypotheses*, pages 181–207. Springer.

Watts, C. and Zimmerman, C. (2002). Violence against women: global scope and magnitude. *The Lancet*, 359(9313):1232–1237.

Weech-Maldonado, R., Elliott, M. N., Oluwole, A., Schiller, K. C., and Hays, R. D. (2008). Survey response style and differential use of cahps rating scales by Hispanics. *Medical Care*, 46(9):963.

Weinick, R. M., Elliott, M. N., Volandes, A. E., Lopez, L., Burkhart, Q., and Schlesinger, M. (2011). Using standardized encounters to understand reported racial/ethnic disparities in patient experiences with care. *Health Services Research*, 46(2):491–509.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, 53(3):300–314.

Wind, Y., Green, P. E., and Jain, A. K. (1973). Higher order factor analysis in the classification of psychographic variables. *Journal of the Market Research Society*, 15(4):224–232.

World Health Organization (2005). WHO multi-country study on women's health and domestic violence against women: initial results on prevalence, health outcomes and women's responses. Report, World Health Organization.

World Health Organization (2013). Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and non-partner sexual violence. Report 9241564628, World Health Organization.

Wothke, W. (1993). Nonpositive definite matrices in structural modeling. InK. A. Bollen, & JS Long (Eds.), Testing structural equation models (pp. 256-293).

Xiao, H. and Barber, J. P. (2008). The effect of perceived health status on patient satisfaction. *Value in Health*, 11(4):719–725.

Yin, J., Geng, Z., Li, R., and Wang, H. (2010). Nonparametric covariance model. *Statistica Sinica*, 20:469–479.

Young, G. J., Meterko, M., and Desai, K. R. (2000). Patient satisfaction with hospital care: effects of demographic and institutional characteristics. *Medical Care*, 38(3):325–334.

Zaslavsky, A. M., Zaborski, L. B., Ding, L., Shaul, J. A., Cioffi, M. J., and Cleary, P. D. (2001). Adjusting performance measures to ensure equitable plan comparisons. *Health Care Financing Review*, 22(3):109.

Zhang, X., Tao, J., Wang, C., and Shi, N.-Z. (2019). Bayesian model selection methods for multilevel irt models: A comparison of five dic-based indices. *Journal of Educational Measurement*, 56(1):3–27.

Zhu, J., Weingart, S. N., Ritter, G. A., Tompkins, C. P., and Garnick, D. W. (2015). Racial/Ethnic disparities in patient experience with communication in hospitals: real differences or measurement errors? *Medical Care*, 53(5):446–454.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zyphur, M. J., Kaplan, S. A., and Christian, M. S. (2008). Assumptions of cross-level measurement and structural invariance in the analysis of multilevel data: Problems and solutions. *Group Dynamics: Theory, Research, and Practice*, 12(2):127–140.

L-BioStat
Kapucijnenvoer 35
Block D, bus 7001
3000 LEUVEN, BELGIË
tel. + 32 16 37 33 89
fax + 32 16 33 70 15
www.kuleuven.be