Multilevel Regression Models for Mean and (Co)variance

with Applications in Nursing Research

Baoyue Li

Multilevel Regression Models for Mean and (Co)variance

with applications in nursing research

Multilevel regressie modellen voor mean en (co)variantie

met toepassingen in verplegingswetenschap

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus

Prof.dr. H.A.P. Pols

en volgens besluit van het College voor Promoties

De openbare verdediging zal plaatsvinden op
dinsdag 17 juni 2014 om 13:30 uur

door

Baoyue Li

geboren te LiaoNing, China

Promotiecommissie

Promotor

Prof.dr. E.M.E.H. Lesaffre

Overige leden

Prof.dr. E.W. Steyerberg
Prof.dr. L.R. Arends
Prof.dr. W. Sermeus

To my wife, my parents and my sisters

CONTENTS

1

GENERAL INTRODUCTION

1.1 Introduction

In this chapter, a concise overview is provided for the statistical techniques that are applied in this thesis. This includes two classes of statistical modeling approaches which have been commonly applied in plenty of research areas for many decades. Namely, we will describe the fundamental ideas about mixed effects models and factor analytic (FA) models. To be specific, this chapter covers several types of these two classes of modeling approaches. For the mixed effects models, we briefly describe the linear, generalized and multivariate mixed effects models, while for the FA models, exploratory FA (EFA), confirmatory FA (CFA) models and multilevel FA (MFA) models are covered. As an extension of FA models, structural equation modeling (SEM) and multilevel SEM are also briefly described. we also discuss the two classical estimating methods, i.e. the frequentist and the Bayesian approach, with the latter chosen as the analytic algorithm for our proposed models.

1.2 Hierarchical data

Hierarchical (also called multilevel or clustered) data are abundantly present in empirical research. For example, a quality of life survey collects information of residents from each household; a clinical trial recruits patients from multiple medical centers; an evaluation of the teaching quality samples students from different schools; a rehabilitation test records daily patients' performance for a certain period; etc. An important feature of all these kinds of multilevel structured data is "non-independence", e.g. residents from the same household tend to act more similar than those from different households as they share the same household environment and are genetically related. The data set used in most chapters of this thesis is taken from a multi-country European nurse survey, the RN4CAST (registered nurse forecasting) project. This project involved a large number of nurses within nursing units within hospitals across countries, implying a four-level hierarchical structure. Feelings of work-related burnout, measured with the multidimensional 22-item Maslach Burnout Inventory (MBI, (Maslach and Jackson, 1981)), are examined in this thesis with relation to work environment variables and personal characteristics. Three dimensions were extracted by Maslach and Jackson (1981) using a factor analytic model.

1.3 Mixed effects models

1.3.1 Linear mixed effects models

Linear mixed effects models (LMM) are linear models with both fixed and random effects. A specific case of a LMM is a longitudinal growth study, where the baseline responses for the individuals differ but their linear growth is the same. This yields the random intercepts model, given by:

$$\begin{aligned} y_{ij} &= \boldsymbol{\beta}^T \boldsymbol{x}_{ij} + u_j + \varepsilon_{ij}, \quad i = 1, 2, ..., n_j;\ j = 1, 2, ..., k, \\ u_j &\sim N(0, \sigma_u^2), \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \quad u_j \perp \varepsilon_{ij}, \end{aligned} \tag{1.1}$$

where y_{ij} is the response measurement for individual j at the ith time, k is the number of individuals, n_j is the number of responses for individual j, $\boldsymbol{x}_{ij}$ represents the q_x-dimensional covariates vector with fixed effects vector $\boldsymbol{\beta}$ having length q_x, u_j represents the random intercept that follows a normal distribution with mean zero and variance σ_u^2, ε_{ij} is the residual part following a normal distribution with mean zero and variance σ_ε^2 with u_j and ε_{ij} mutually independent with each other.

The correlation of the repeated measurements for the same individual j is:

$$cor(y_{ij}, y_{i'j}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2}, \qquad (i \neq i'),$$

and is known as the intra-class correlation coefficient (ICC), expressing the degree of the dependency of the observations in the multilevel data set. ICC ranges from 0 to 1, with higher values indicating more dependency. Another way to quantify the dependency is to calculate the design effect (DE), as well as the effective sample size (ESS), which are respectively:

$$\begin{aligned} \text{DE} &= 1 + \text{ICC} * (\bar{n} - 1), \\ \text{ESS} &= N/\text{DE} = \bar{n}k/\text{DE}, \end{aligned}$$

where $\bar{n}$ represents the average repeated times across the k individuals and N is the total number of observations with $N = \bar{n}k$. We see from the two expressions that DE increases with ICC and $\bar{n}$, while ESS increases when N increases or DE decreases. For ICC$= 0$, we obtain the smallest DE, i.e. 1, and the largest ESS, i.e. N, indicating completely independent measurements within an individual. When ICC increases to 1, we obtain the largest DE, i.e. $\bar{n}$, and the smallest ESS, i.e. k. This also means that with ICC equal to 1, the design effect is the average number of the repeated times and the effective sample size is actually equal to the number of individuals. This is quite important for the sample size calculation in a multilevel design. For a sound analysis one must take ICC (or DE and ESS) into account.

There are two strategies towards using mixed effects models for hierarchical data. One strategy suggests using a rule of thumb to apply a mixed effects model only if ICC is greater than 0.05 (Raudenbush and Liu, 2000). The second strategy recommends applying a mixed effects model for hierarchical data irrespective of the value of ICC, since ignoring ICC often results in too small estimates of the standard errors, leading to inflated type I errors (Krull and MacKinnon, 2001). In this thesis, we adhere to the second strategy but our analyses also satisfy the rule of thumb that ICC> 0.05.

The LMM in model (1.1) can be extended by adding other random effects on top of the random intercept. For example in the previous longitudinal growth study, not only the baseline measurement, but also the true linear growth trend for each individual can be assumed to be different. This results in a linear random intercept and slope model. In general, we can write a LMM as follows:

$$\begin{aligned} y_{ij} = \boldsymbol{\beta}^T \boldsymbol{x}_{ij} + \boldsymbol{u}_j^T \boldsymbol{z}_{ij} + \varepsilon_{ij}, \quad i = 1, 2, ..., n_j;\ j = 1, 2, ..., k, \\ \boldsymbol{u}_j \sim N(\boldsymbol{0}, \Sigma_u), \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \quad u_j \perp \varepsilon_{ij}, \end{aligned} \tag{1.2}$$

where $\boldsymbol{u}_j$ represents the q_z-dimensional random effect with a multivariate normal distribution having mean zero and covariance matrix Σ_u, $\boldsymbol{z}_{ij}$ is the corresponding q_z covariates vector. The other terms are the same as in model (1.1). Note that $\boldsymbol{z}_{ij}$ may be different from $\boldsymbol{x}_{ij}$. Model (1.1) is obtained when $\boldsymbol{z}_{ij}$ has a single value 1. A specific case of model (1.2) is the cross-classified mixed effects model. This model arises when there exist more than one hierarchical structure for the data, e.g. students come from different schools and different districts where both school and district are clusters but are not nested within each other.

1.3.2 Generalized linear mixed models

The LMM is a special case of a generalized LMM (GLMM) whereby the response has a normal distribution with an identity link function. In general, the GLMM can handle a large amount of probability distributions coming from the exponential family such as the normal, binomial, Poisson and gamma distributions up to random effects. For a GLMM, it is assumed that the expected value of the response y_{ij} can be modeled as a linear function of fixed and random effects up to a link function $g()$, i.e.:

$$\begin{gathered} g(E(y_{ij}|\boldsymbol{u}_j)) = \boldsymbol{\beta}^T\boldsymbol{x}_{ij} + \boldsymbol{u}_j^T\boldsymbol{z}_{ij}, \quad i = 1, 2, ..., n_j;\ j = 1, 2, ..., k, \\ \boldsymbol{u}_j \sim N(\boldsymbol{0}, \Sigma_u), \end{gathered} \tag{1.3}$$

whereby u_j is usually assigned a multivariate normal distribution, but other distributions such as a multivariate t distribution are possible. GLMMs have been suggested to address the overdispersion of e.g. count responses (Breslow, 1984).

When the response has a binomial distribution with a *logit* link function, we obtain a logistic mixed effects model, where:

$$E(y_{ij}|\boldsymbol{u}_j) = p_{ij} = \frac{e^{\boldsymbol{\beta}^T\boldsymbol{x}_{ij} + \boldsymbol{u}_j^T\boldsymbol{z}_{ij}}}{1 + e^{\boldsymbol{\beta}^T\boldsymbol{x}_{ij} + \boldsymbol{u}_j^T\boldsymbol{z}_{ij}}}, \tag{1.4}$$

where p_{ij} represents the conditional expected probability for the observed binomial data. The right-hand side part of model (1.4) is actually known as the logistic function of $\boldsymbol{\beta}^T\boldsymbol{x}_{ij} + \boldsymbol{u}_j^T\boldsymbol{z}_{ij}$. An alternative to the logistic model for the binomial data is the *probit* model, with the *probit* link function having the form:

$$p_{ij} = \Phi(\boldsymbol{\beta}^T\boldsymbol{x}_{ij} + \boldsymbol{u}_j^T\boldsymbol{z}_{ij}), \tag{1.5}$$

where Φ is the cumulative distribution function (CDF) of a standard normal distribution. It has been shown that the parameter estimates from the logistic and *probit* models are similar up to a constant value, i.e. the coefficients from a logistic model are 1.6 times of the corresponding ones under a *probit* link model (Gelman and Hill, 2006).

1.3.3 Multivariate mixed effects models

The multivariate mixed effects model is the generalization of the mixed effects model to multiple responses at the same time. The multivariate LMM has the following form:

$$\begin{aligned} \boldsymbol{y}_{ij} &= \boldsymbol{B}\boldsymbol{x}_{ij} + \boldsymbol{U}_j\boldsymbol{z}_{ij} + \boldsymbol{\varepsilon}_{ij}, \quad i = 1, 2, ..., n_j;\ j = 1, 2, ..., k, \\ \boldsymbol{U}_j &\sim N(\boldsymbol{0}, \Sigma_u), \quad \boldsymbol{\varepsilon}_{ij} \sim N(\boldsymbol{0}, \Sigma_\varepsilon), \quad \boldsymbol{U}_j \perp \boldsymbol{\varepsilon}_{ij}, \end{aligned} \tag{1.6}$$

where $\boldsymbol{y}_{ij}$ represents the p-dimensional response vector for individual i from group j. The covariates $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ have the same meaning as in model (1.2), and now with the fixed and random effects being a $p \times q_x$ matrix $\boldsymbol{B}$ and a $p \times q_z$ matrix $\boldsymbol{U}_j$, respectively. The p-dimensional residual vector $\boldsymbol{\varepsilon}_{ij}$ is usually assumed to have a multivariate normal distribution with mean zero and the covariance matrix Σ_ε.

We see from model (1.6) that the correlated nature of the responses is reflected by correlated residuals and correlated random effects. The latter means that not only the random effects within each response are correlated, but also the random effects across the responses are correlated. This may increase the power for estimation because the parameter estimates for each of the p responses can borrow information from each other through the correlations. In addition, tests for the equality of the parameter estimates across multiple responses and global tests based on all responses can be constructed. Take the three-dimensional burnout measurements as an example. Through a multivariate linear random effects model with the covariate *work environment*, we can test whether the effects of work environment are the same for all the three burnout dimensions taking into account the multilevel structure. We can also check whether the *work environment* variable has a significant effect on all of the three burnout dimensions simultaneously, thereby dealing with the multiple testing problem.

1.4 Estimation methods

Generally speaking, there are two main classes of estimating methods: the frequentist approach and the Bayesian approach. In this section, we describe some basic features of each approach and the performance of these two approaches for handling some of the models mentioned earlier.

1.4.1 Frequentist approach

In the frequentist approach, probability is defined as a limiting relative frequency. That is, the probability of an event is the limit of the relative frequency of that event in a large number of studies. Further, in frequentist statistics one estimates the unknown but fixed model parameter θ. Prediction is done given the estimated θ and the uncertainty of the prediction is based on the sampling property of the estimated value of θ (Feller, 1968).

Maximum likelihood (ML) is a popular way of estimating the model parameters. It finds the parameter estimates that maximize the likelihood function, $L(\boldsymbol{\theta}|\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta})$, therefore are called the maximum likelihood estimates (MLEs). For many models without random

effects, the likelihood function is relatively simple and can be written analytically (with a closed form). When the model contains random effects, the marginal likelihood is calculated by integrating over the random effects to obtain the MLEs. For example for a LMM expressed in model (1.2), let $\boldsymbol{\theta}$ denote all parameters except the random effects, i.e. $(\boldsymbol{\beta}, \Sigma_u, \sigma_\varepsilon^2)$, the marginal likelihood is:

$$\begin{aligned} L_m(\boldsymbol{\theta}, \boldsymbol{y}) &= p(\boldsymbol{y}|\boldsymbol{\beta}, \Sigma_u, \sigma_\varepsilon^2) \\ &= \prod_{j=1}^{k} \int \prod_{i=1}^{n_j} p(y_{ij}|\boldsymbol{\beta}, \sigma_\varepsilon^2, \boldsymbol{u}_j) p(\boldsymbol{u}_j|\Sigma_u) d\boldsymbol{u}_j. \end{aligned} \tag{1.7}$$

The integral in this expression can be solved analytically, and the likelihood function is written as:

$$L_m(\boldsymbol{\theta}, \boldsymbol{y}) = \prod_{j=1}^{k} \{(2\pi)^{-n_j/2} |V_j|^{-1/2} \times exp(-\frac{1}{2}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta})^T V_j^{-1}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta}))\}, \tag{1.8}$$

where $\boldsymbol{y}_j$ represents the response vector for group j with length n_j, N is the total number of individuals ($N = \sum_{j=1}^{k} n_j$), $\boldsymbol{\beta}$ is the q_x-dimensional fixed effects vector with $\boldsymbol{X}_j$ its corresponding covariate matrix of dimension $n_j \times q_x$, V_j is the $n_j \times n_j$ marginal covariance matrix of $\boldsymbol{y}_j$, which has the form:

$$V_j = \boldsymbol{Z}_j \Sigma_u \boldsymbol{Z}_j^T + \sigma_\varepsilon^2 I.$$

In this form $\boldsymbol{Z}_j$ is the corresponding $n_j \times q_z$ covariate matrix for the random effects having a $q_z \times q_z$ covariance matrix Σ_u, and I is the identity matrix of size n_j.

Unfortunately, for most of the GLMMs such as the logistic random effects models, there exists no closed form for the likelihood function (1.7). To solve this, numerical approximations have been developed, e.g. the non-adaptive Gaussian quadrature method, the adaptive Gaussian quadrature method with the Laplacian approximation as the simplest case, etc. Further, based on the approximated marginal likelihood function, the maximization algorithms such as the Newton-Raphson and the iterative generalized least square (IGLS) algorithms, are required to find the MLEs.

1.4.2 Bayesian approach

In the Bayesian approach 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 (Lesaffre and Lawson, 2012), but the parameter becomes stochastic because of our uncertainty of its value. We denote $p(\theta)$ as the prior distribution of θ obtained from expert knowledge, historical information, etc., but without observing the current data $\boldsymbol{y}$. $L(\theta|\boldsymbol{y})$ is the likelihood defined by the model specification. The probability distribution of θ obtained from combining the information from the prior and the data is given by Bayes' Theorem

and is called the posterior distribution given by:

$$p(\theta|\boldsymbol{y}) = \frac{L(\theta|\boldsymbol{y})p(\theta)}{p(\boldsymbol{y})} = \frac{L(\theta|\boldsymbol{y})p(\theta)}{\int L(\theta|\boldsymbol{y})p(\theta)d\theta}. \tag{1.9}$$

The denominator $p(\boldsymbol{y})$ can be written as the integration of the likelihood $L(\theta|\boldsymbol{y})$ over the variable θ, therefore is called the averaged likelihood. Bayes' Theorem shows one of the advantages of the Bayesian approach, namely that it can utilize the informative prior which may increase the power for estimation. For example, previous similar studies could be used to represent our prior belief when analyzing the data from the current study, which may result in a more precise conclusion. However informative priors have also caused a lot of controversy between Bayesians and frequentists, since frequentists accused the Bayesian approach to be subjective. When no prior information is available, a non-informative prior could be used. Then, the likelihood dominates the prior and information from the posterior is actually equivalent to the information extracted from the likelihood.

Bayesian estimation involves integration as shown in Bayes' Theorem. 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. For the LMM of model (1.1), the joint posterior distribution of all parameters, including the random effects $\boldsymbol{u}$, is then:

$$\begin{aligned} p(\boldsymbol{\beta}, \Sigma_u, \sigma^2_\varepsilon, \boldsymbol{u}|\boldsymbol{y}) &= \frac{L(\boldsymbol{\beta}, \Sigma_u, \sigma^2_\varepsilon, \boldsymbol{u}|\boldsymbol{y})p(\boldsymbol{u}|\Sigma_u)p(\boldsymbol{\beta})p(\Sigma_u)p(\sigma^2_\varepsilon)}{p(\boldsymbol{y})} \\ &= \frac{p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2_\varepsilon, \boldsymbol{u})p(\boldsymbol{u}|\Sigma_u)p(\boldsymbol{\beta})p(\Sigma_u)p(\sigma^2_\varepsilon)}{\int_{\boldsymbol{\beta}} \int_{\Sigma_u} \int_{\sigma^2_\varepsilon} \int_{\boldsymbol{u}} L(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2_\varepsilon, \boldsymbol{u})p(\boldsymbol{u}|\Sigma_u)p(\boldsymbol{\beta})p(\Sigma_u)p(\sigma^2_\varepsilon)d\boldsymbol{\beta}d\Sigma_u d\sigma^2_\varepsilon d\boldsymbol{u}}. \end{aligned}$$

Note that in the denominator, each integral may involve multiple integrations depending on their respective dimensions. Because of the high-dimensional integration, the Bayesian approach was for about two centuries impossible to use for real-life problems (Lesaffre and Lawson, 2012).

1.4.2.1 Bayesian computational techniques

In 1990, a powerful class of numerical procedures, called Markov Chain Monte Carlo (MCMC) techniques (Gelfand and Smith, 1990), was launched which revolutionized the Bayesian approach. The MCMC technique is based on a sampling approach, i.e. the integral is approximated by Monte Carlo sampling (Ripley, 1987). There are two major classes of MCMC techniques: Gibbs sampling and Metropolis-Hastings (MH) sampling. We describe here both methods but focus on Gibbs sampling which is the most popular approach used for the considered models in this thesis.

Gibbs sampling Gibbs sampling was first introduced by Geman and Geman (1984) and is commonly used nowadays for Bayesian inference. It explores the M-dimensional joint posterior distributions of $\boldsymbol{\theta}$, and therefore of each parameter θ_m. This is done by sampling

from the full conditional distributions $p(\theta_m|\boldsymbol{\theta}_{(-m)}, \boldsymbol{y})$, where $\boldsymbol{\theta}_{(-m)}$ represents all parameters except θ_m ($m = 1, 2, ..., M$). To initialize the updating phase of all parameters in Gibbs sampling, a set of starting values are first given to each parameter in $\boldsymbol{\theta}$, denoted as $\boldsymbol{\theta}^0$. For the first iteration, the Gibbs sampling proceeds as follows:

- Sample $\theta_1^{(1)}$ from the full conditional distribution $p(\theta_1|\boldsymbol{\theta}_{(-1)}^0, \boldsymbol{y})$
- Sample $\theta_2^{(1)}$ from the full conditional distribution $p(\theta_2|\theta_1^{(1)}, \theta_3^0, ..., \theta_M^0, \boldsymbol{y})$
- ...
- Sample $\theta_M^{(1)}$ from the full conditional distribution $p(\theta_M|\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_{M-1}^{(1)}, \boldsymbol{y})$

Thus a new set of values $\boldsymbol{\theta}^{(1)} = (\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_M^{(1)})$ is obtained from the starting values $\boldsymbol{\theta}^0$ and the observed data $\boldsymbol{y}$. The second iteration is conducted based on the new set of values $\boldsymbol{\theta}^{(1)}$ and the data, and so on so forth. In general, the Gibbs sampling for the mth parameter in the lth iteration is conducted from the following full conditional distribution:

$$\theta_m^{(l)} \sim p(\theta_m|\theta_1^{(l)}, ...\theta_{m-1}^{(l)}, \theta_{m+1}^{(l-1)}, ..., \theta_M^{(l-1)}, \boldsymbol{y}), \qquad m = 1, 2, ..., M, \quad l = 1, 2, ..., L$$

where M is the total number of parameters and L is the total number of iterations. The iteration continues till the Markov chain(s) for each parameter have converged. When multiple chains are launched, convergence can be tested using, e.g. the BGR diagnostic (Brooks and Gelman, 1998) which compares the between- and within-chain variability. At the end, we obtain for each parameter L samples. After removing the non-converged part of the chain, called the "burn-in" part, the remaining samples represent well the marginal posterior distribution for each parameter.

We note that the samples obtained using Gibbs sampling are not independent. Each set of samples $\boldsymbol{\theta}^{(l)}$ depends on the previous samples $\boldsymbol{\theta}^{(l-1)}$ but is conditionally independent with all other previous samples given $\boldsymbol{\theta}^{(l-1)}$, i.e.,

$$p(\boldsymbol{\theta}^{(l)}|\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(l-1)}, \boldsymbol{y}) = p(\boldsymbol{\theta}^{(l)}|\boldsymbol{\theta}^{(l-1)}, \boldsymbol{y}).$$

This is known as the Markov property. Gibbs sampling assumes that the full conditional distributions $p(\theta_m|\boldsymbol{\theta}_{(-m)}, \boldsymbol{y})$ should be relatively easy to sample from. In cases that the full conditional distribution is hard to obtain or hard to sample from, we may use a different sampling algorithm, called the Metropolis-Hastings (MH) algorithm which is described below.

Metropolis-Hastings sampling Another class of MCMC sampling methods is the Metropolis-Hastings (MH) algorithm. Here the parameters $\boldsymbol{\theta}$ are first sampled from a proposal distribution q and in a second step part of the sampled values are accepted to yield a sample from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$. This involves the following two steps:

1. For iteration l, a candidate sample $\tilde{\boldsymbol{\theta}}$ is sampled from $q(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}^{(l)})$.

2. Calculate the acceptance ratio $\alpha = \frac{p(\tilde{\boldsymbol{\theta}}|\boldsymbol{y})q(\boldsymbol{\theta}^{(l)}|\tilde{\boldsymbol{\theta}})}{p(\boldsymbol{\theta}^{(l)}|\boldsymbol{y})q(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}^{(l)})}$ and choose the next sample on:

$$\boldsymbol{\theta}^{(l+1)} = \begin{cases} \tilde{\boldsymbol{\theta}}, & \text{with the probability } min(1, \alpha); \\ \boldsymbol{\theta}^{(l)}, & \text{otherwise.} \end{cases}$$

Mathematically, Gibbs sampling could be seen as a special type of MH sampling in that the proposal distribution q in Gibbs sampling is the full conditional distribution for each θ_m and the acceptance ratio α is always 1.

Gibbs sampling, together with many other sampling methods, are nowadays implemented in many statistical programs and packages, such as WinBUGS (Bayesian inference Using Gibbs Sampling) (Spiegelhalter et al., 2003), JAGS (Just Another Gibbs Sampler) (Plummer, 2003), Mplus (Muthén and Muthén, 2010), etc. These programs may differ in the default sampling method for a specific type of parameters, thus may behave differently.

1.5 Factor analytic models

A questionnaire is a common research instrument in a survey to collect information about subjects regarding all kinds of behavior, feelings, etc. Take the burnout measurement in the RN4CAST nurse survey as an example. Burnout is a syndrome of emotional exhaustion, depersonalization and reduced personal accomplishment that can occur among individuals who do "people work" of some kind (Maslach and Jackson, 1986). Burnout is measured indirectly via a series of questions that reflect all these aspects. In fact, the classic questionnaire used for burnout contains the 22-item Maslach Burnout Inventory (MBI, (Maslach and Jackson, 1981)) that has proved to measure the above mentioned three dimensions well. We call these three dimensions latent constructs, while the 22 items are manifest measures. The relationship between the latent constructs and the manifest measures is typically studied by a factor analytic (FA) model, which is a special type of a multivariate analysis. FA models are classified into two types, i.e. the exploratory FA (EFA) and the confirmatory FA (CFA) model.

EFA and CFA models

The EFA model is usually used to identify a number of latent constructs underlying a relatively larger set of observed variables. An EFA model is especially useful when we have no a priori hypothesis on the latent factor structure. Note the difference of an EFA model with principle component analysis (PCA) which is an exploratory variable reduction technique. Among other differences, PCA does not assume any particular statistical model for the data, while an EFA model is defined as:

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{f}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1, 2, ..., N, \\ \boldsymbol{f}_i &\sim N(\mathbf{0}, \Sigma_f), \quad \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \Sigma_\varepsilon), \quad \boldsymbol{f}_i \perp \boldsymbol{\varepsilon}_i, \end{aligned} \tag{1.10}$$

where $\boldsymbol{y}_i$ represents the p-dimensional response for individual i and $\boldsymbol{\mu}$ is the intercept vector with the same length p, $\boldsymbol{f}_i$ represents the q-dimensional common factors ($q < p$) following

a multivariate normal distribution with covariance matrix Σ_f and $\boldsymbol{L}$ is the corresponding factor loading matrix of size $p \times q$, $\boldsymbol{\varepsilon}_i$ is the residual vector for each individual, having a multivariate normal distribution with covariance matrix Σ_ε, it is assumed to be independent with the common factors and is also called the unique factors in an FA model.

The aim of a CFA model is to test the underlying factor structure that we have a priori in mind. This hypothesis may come from previous studies or is based on theory. A CFA model could also represent a further simplification of the factor structure after an EFA model has been fitted. With a reasonable model fit, the CFA model could provide evidence to confirm an assumed factor structure. A typical CFA model has the same form as an EFA model shown in model (1.10). The factor loading matrix $\boldsymbol{L}$ in a CFA model, however, is different from that of an EFA model in that some elements are fixed at constant values. That is, the cross-loadings are fixed to zero since we have in mind a priori a particular factor structure as our testing hypothesis.

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

$$\Sigma = \boldsymbol{L}\Sigma_f \boldsymbol{L}^T + \Sigma_\varepsilon. \tag{1.11}$$

Note that by implementing a factor model the covariance matrix of the observed responses could be rebuilt through the factor loadings $\boldsymbol{L}$, the covariance matrix of factors Σ_f and the covariance matrix of residuals Σ_ε.

1.5.1 Identification

The FA model (1.10) resembles a multivariate linear regression model except that $\boldsymbol{f}_i$ are not observed covariates but unknown latent factors. This causes an identification problem meaning that more than one set of parameter estimates satisfies model (1.10). Further constraints are required for the common factors $\boldsymbol{f}_j$ and/or the loading matrix $\boldsymbol{L}$, in a CFA and an EFA model.

The identifying constraints for CFA and EFA models are different as they have different model assumptions and are used for different purposes. There are plenty of ways to set these constraints and here we only display one of them. For a more detailed description of the identification issues, we refer to Thompson (2004).

In an EFA model, the following constraints are used in addition for model (1.10):

- Set the covariance matrix for the common factors to be identity: $\Sigma_f = I$.
- Estimate only the diagonal covariance matrix of the unique factors.

This is also called the orthogonal FA model because the common factors are orthogonal with each other.

For a CFA model, we use the following constraints in model (1.10):

- Fix one loading to 1 for each common factor.
- Estimate only the diagonal covariance matrix for the unique factors.

This choice of constraints implies that the general covariance matrix Σ_f of the common factors can be estimated.

Further, we would like to highlight here some of the differences between Bayesian and frequentist approaches in identifying an FA model. Firstly, for the part $\boldsymbol{Lf}$ in the factor model, the distribution of the common factors $\boldsymbol{f}$ is usually symmetric with mean zero, e.g. a multivariate normal distribution. This causes a unique identification issue in Bayesian approach called the "flipping states" issue (Maydeu-Olivares and McArdle, 2005), whereby both $\boldsymbol{L}$ and $-\boldsymbol{L}$ are the solutions for the factor loadings if no further constraints are set for $\boldsymbol{L}$. The frequentist approach finds only one of the solutions while the Bayesian approach, which is simulation-based, may move between the two solutions and may never get converged (Browne, 2012). Further constraints, therefore, are required for Bayesian factor analytic modeling. Secondly, some of the identifying constraints on parameters in the frequentist approach could be to fix these at a particular value. The same strategy could be applied in the Bayesian approach, but there is an alternative solution by introducing reasonable informative priors for the parameters needed to constraint (Muthén and Asparouhov, 2012). This approach is applied in this thesis.

1.5.2 Multilevel FA model

When the data show a multilevel structure, e.g. nurses within hospitals, the correlated nature should also be taken into account in the FA models to obtain valid estimates (Longford and Muthén, 1992). This gives rise to the multilevel FA (MFA) model. A two-level MFA model can be written as:

$$\begin{gathered}
\boldsymbol{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{L}_B \boldsymbol{f}_j + \boldsymbol{u}_j + \boldsymbol{L}_W \boldsymbol{f}_{ij} + \boldsymbol{\varepsilon}_{ij}, \\
\boldsymbol{f}_j \sim N(\boldsymbol{0}, \Sigma_{fB}), \quad \boldsymbol{u}_j \sim N(\boldsymbol{0}, \Sigma_u), \\
\boldsymbol{f}_{ij} \sim N(\boldsymbol{0}, \Sigma_{fW}), \quad \boldsymbol{\varepsilon}_{ij} \sim N(\boldsymbol{0}, \Sigma_\varepsilon), \\
i = 1, 2, ..., n_j;\ j = 1, 2, ..., k, \quad \boldsymbol{f}_j \perp \boldsymbol{u}_j \perp \boldsymbol{f}_{ij} \perp \boldsymbol{\varepsilon}_{ij},
\end{gathered} \tag{1.12}$$

where $\boldsymbol{y}_{ij}$ represents the p-dimensional response for individual i from group j, $\boldsymbol{f}_j$ is the q_B-dimensional between-level common factor vector with the factor loading matrix $\boldsymbol{L}_B$ having the dimension of $p \times q_B$, $\boldsymbol{f}_{ij}$ is the q_W-dimensional within-level common factor vector with the factor loading matrix $\boldsymbol{L}_W$ having the dimension $p \times q_W$, $\boldsymbol{u}_j$ is the p-dimensional between-level unique factor with covariance matrix Σ_u, $\boldsymbol{\varepsilon}_{ij}$ is the p-dimensional within-level unique factor with covariance matrix Σ_ε and all the common and unique factors are assumed mutually independent with each other. The implied covariance matrix for the MFA model (1.12) is then:

$$\Sigma = \boldsymbol{L}_B \Sigma_{fB} \boldsymbol{L}_B^T + \Sigma_u + \boldsymbol{L}_W \Sigma_{fW} \boldsymbol{L}_W^T + \Sigma_\varepsilon. \tag{1.13}$$

1.6 Structural equation modeling

Structural equation modeling (SEM) has a close relationship with factor analytic models. The standard SEM consists of two parts: a CFA part (called the measurement part) and a regression model among the latent common factors (called the structural part). SEM aims to 1) understand the patterns of covariances among a set of observed variables and 2) explain as much of their variance as possible with the researcher's model (Kline, 2010). It is especially useful for describing the complex causal relationships among the latent constructs. A typical model for SEM is:

$$\begin{gathered} \boldsymbol{y}_i = \boldsymbol{B}\boldsymbol{x}_i + \boldsymbol{L}\boldsymbol{f}_i + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{f}_i = (\boldsymbol{\eta}_i^T, \boldsymbol{\xi}_i^T)^T, \\ \boldsymbol{\eta}_i = \Gamma\boldsymbol{\xi}_i + \boldsymbol{\delta}_i, \qquad i = 1, 2, ..., N, \\ \boldsymbol{f}_i \sim N(\mathbf{0}, \Sigma_f), \quad \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \Sigma_\varepsilon), \quad \boldsymbol{\delta}_i \sim N(\mathbf{0}, \Sigma_\delta), \quad \boldsymbol{f}_i \perp \boldsymbol{\varepsilon}_i \perp \boldsymbol{\delta}_i, \end{gathered} \tag{1.14}$$

where the common factors $\boldsymbol{f}_i$ can be further partitioned into dependent latent factors $\boldsymbol{\eta}_i$ and independent latent factors $\boldsymbol{\xi}_i$, which are further modeled together.

The multilevel SEM extends the MFA model in that it further models the latent constructs at each level. The cross-level interactions can also be modeled properly. One specific class of (multilevel) SEM is called the (multilevel) MIMIC (multiple indicators multiple causes (Jöreskog and Goldberger, 1975)) model, which models the latent constructs in the measurement part with other fixed and/or random effects. The multilevel MIMIC model allows cross-group comparisons while assessing measurement invariance with respect to subject grouping (Muthén, 1989).

References

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics*, 33(1):38–44.

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

Browne, W. (2012). *MCMC Estimation in MLwiN, v2.25*. Centre for Multilevel Modelling, University of Bristol.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*. Wiley, 3rd edition.

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.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 1 edition.

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.

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.

Kline, R. B. (2010). *Principles and Practice of Structural Equation Modeling (Methodology in the Social Sciences)*. The Guilford Press, 3rd edition.

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

Lesaffre, E. and Lawson, A. B. (2012). *Bayesian Biostatistics (Statistics in Practice)*. Wiley, 1st edition.

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

Maslach, C. and Jackson, S. E. (1981). The measurement of experienced burnout. *Journal of Organizational Behavior*, 2(2):99–113.

Maslach, C. and Jackson, S. E. (1986). Maslach burnout inventory. University of California, Palo Alto, CA.

Maydeu-Olivares, A. and McArdle, J. J. (2005). *Contemporary Psychometrics (Multivariate Applications Series)*. Psychology Press.

Muthén, B. and Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, 17(3):313–335.

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

Muthén, L. and Muthén, B. (2010). *Mplus User's guide*. Los Angeles: Muthén & Muthén, 6th edition.

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*.

Raudenbush, S. W. and Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, 5(2):199–213.

Ripley, B. D. (1987). *Stochastic Simulation (Wiley Series in Probability and Statistics)*. Wiley, 1st edition.

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003). *WinBUGS User manual (version 1.4.3)*.

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

2

AIMS AND OUTLINE OF THE THESIS

2.1 Introduction

We describe in this chapter the motivating RN4CAST data set in more detail. This data set is used in the majority of the other chapters, the clinical and statistical aims, and the outline of this thesis.

2.2 Motivating data set

The data set used in Chapters 4 to 8 was extracted from the RN4CAST (registered nurse forecasting) project (Sermeus et al., 2011). This three-year (2009-2011) nurse workforce study was funded by the Seventh Framework Program of the European Union. For the RN4CAST project with research teams from 12 countries, a multilevel observational design was used to determine how system-level features in the organization of nursing care (work environment, education, and workload) impact individual measures of nurse wellbeing (burnout, job satisfaction, and turnover) and patient safety outcomes and care satisfaction. This resulted in a large and unique data set involving 33,731 registered nurses in 2,169 nursing units in 486 hospitals in 12 European countries. This rich data set provides ample opportunities for statistical modeling, as well as challenges. The burnout measurement, which has three dimensions, is the focus of our proposed multilevel covariance regression model.

2.3 Clinical aims

The clinical aims of our analyses in this thesis are to study the relationship between the multivariate burnout measurements and other relevant covariates, as well as the interplay of the burnout dimensions. To be specific, it is of interest to know:

- How much variability does each of the three burnout measurements show across countries, hospitals (within countries), nursing units (within countries and hospitals) and nurses (within countries, hospitals and nursing units)?
- How much of this variability can be explained with the covariates recorded at the different levels?
- Does the covariance matrix (and more precisely the correlation) between the three burnout dimensions remain the same across countries, hospitals, nursing units and even nurses after accounting for a rich set of confounders at the different levels?

2.4 Statistical aims

Inspired by these research questions, we introduce in this thesis a novel way of handling both the mean and the covariance matrix of the three-dimensional burnout response properly for the multilevel-structured RN4CAST data set. That is, we model both the multivariate mean structure and the heteroscedasticity hierarchically. The following models were developed:

- A multivariate multilevel model with covariates at each level that quantifies how much of the variation can be explained by the level-specific fixed and random effects.
- A model, whereby the covariance matrix is expressed in terms of fixed and random effects at each level.
- A model that combines a factor analytic model with the previous model.

The second development results in the multilevel covariance regression (MCR) model. The third development results in an extension of the MCR model, called the multilevel higher-order factor (MHOF) model.

2.5 Outline of the thesis

The remaining chapters of the thesis are outlined as follows.

In Chapter 3, we review the current sofware/packages that can deal with the logistic random effects regression models (with both binary and ordinal outcomes), and perform comparisons in terms of both their efficacy and efficiency. Both frequentist and Bayesian approaches are modeled.

Chapter 4 applies a two-level logistic regression model to the nursing tasks from the RN4CAST data set. We compare the differences between the domestically trained nurses and foreign trained nurses in the performance of the nursing task below their skill level.

In Chapter 5, a Bayesian two-level MIMIC model is applied to the Belgian data from the RN4CAST project. The focus is on the differences in the opinions of the nursing unit managers and staff nurses towards the nursing work environment. It is measured through an internationally validated multidimensional instrument with 32 items on the most important aspects of nurses' work environment.

Chapter 6 uses the burnout data from the RN4CAST project. The original measurement contains 22 items, and we use the sum scores for the three burnout dimensions, which were further dichotomized. We developed a three-variate four-level $probit$ model with the correlations of the three responses being random across the units at each level. We then study the relationship of the burnout and work environment at each level, as well as the correlations within burnout.

In Chapter 7, we replace the binary burnout responses from Chapter 6 with the sum sores and further model the covariance structure with both fixed and random effects. This results in the multilevel covariance regression (MCR) model. The key assumptions, interpretations, identification issues, implied marginal models, skewness and kurtosis, and the application are described in detail.

Chapter 8 further extends the MCR model by replacing the three-dimensional burnout response with factor scores directly coming from a multilevel factor analytic model applied to the original 22 burnout items. The MCR model and a multilevel factor analytic model are therefore combined and estimated simultaneously. We call this modeling approach the multilevel higher-order factor (MHOF) model.

At the end, we give concluding remarks in Chapter 9, as well as we suggest some future research topics.

References

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):6.

3

LOGISTIC RANDOM EFFECTS REGRESSION MODELS: A COMPARISON OF STATISTICAL PACKAGES FOR BINARY AND ORDINAL OUTCOMES

Chapter 3 is based on the paper:

Li, B., Lingsma, H. F., Steyerberg, E. W., and Lesaffre, E. (2011). Logistic random effects regression models: A comparison of statistical packages for binary and ordinal outcomes. BMC Medical Research Methodology, 11(1):77.

Abstract

Logistic random effects models are a popular tool to analyze multilevel also called hierarchical data with a binary or ordinal outcome. Here, we aim to compare different statistical software implementations of these models using both frequentist and Bayesian method. Frequentist approaches included R (lme4), Stata (GLLAMM), SAS (GLIMMIX and NLMIXED), MLwiN ([R]IGLS) and MIXOR; Bayesian approaches included WinBUGS, MLwiN (MCMC), R package MCMCglmm and SAS experimental procedure MCMC. As a result, the packages gave similar parameter estimates for both the fixed and random effects and for the binary (and ordinal) models when based on a relatively large data set. However, for relatively sparse data set, i.e. when the numbers of level-1 and level-2 data units were about the same, the frequentist and Bayesian approaches showed somewhat different results. The software implementations differ considerably in flexibility, computation time and usability. To conclude, for a large data set there seems to be no explicit preference for either a frequentist or Bayesian approach (if based on vague priors). The choice for a particular implementation may largely depend on the desired flexibility, and the usability of the package. For small data sets the random effects variances are difficult to estimate. In the frequentist approaches the MLE of this variance was often estimated zero with a standard error that is either zero or could not be determined, while for Bayesian methods the estimates could depend on the chosen "non-informative" prior of the variance parameter. The starting value for the variance parameter may be also critical for the convergence of the Markov chain.

3.1 Background

Hierarchical, multilevel, or clustered data structures are often seen in medical, psychological and social research. Examples are: (1) individuals in households and households nested in geographical areas, (2) surfaces on teeth, teeth within mouths, (3) children in classes, classes in schools, (4) multicenter clinical trials, in which individuals are treated in centers, (5) meta-analyses with individuals nested in studies. Multilevel data structures also arise in longitudinal studies where measurements are clustered within individuals.
The multilevel structure induces correlation among observations within a cluster, e.g. between patients from the same center. An approach to analyze clustered data is the use of a multilevel or random effects regression analysis. There are several reasons to prefer a random effects model over a traditional fixed effects regression model (Rasbash, nd). First, we may wish to estimate the effect of covariates at the group level, e.g. type of center (university versus peripheral center). With a fixed effects model it is not possible to separate out group effects from the effect of covariates at the group level. Secondly, random effects models treat the groups as a random sample from a population of groups. Using a fixed effects model, inferences cannot be made beyond the groups in the sample. Thirdly, statistical inference may be wrong. Indeed, traditional regression techniques do not recognize the multilevel structure and will cause the standard errors of regression coefficients to be wrongly estimated, leading to an overstatement or understatement of statistical significance for the coefficients of both the higher- and lower-level covariates.
All this is common knowledge in the statistical literature (Molenberghs and Verbeke, 2005), but in the medical literature still multilevel data are often analyzed using fixed effects models (Austi et al., 2003).
In this paper we use a multilevel dataset with an ordinal outcome, which we analysed as such but also in a dichotomized manner as a binary outcome. Relating patient and cluster characteristics to the outcome requires some special techniques like a logistic (or probit, cloglog, etc) random effects model. Such models are implemented in many different statistical packages, all with different features and using different computational approaches. Packages that use the same numerical techniques are expected to yield the same results, but results can differ if different numerical techniques are used. In this study we aim to compare different statistical software implementations, with regard to estimation results, their usability, flexibility and computing time. The implementations include both frequentist and Bayesian approaches. Statistical software for hierarchical models has been compared already by Zhou et al. (1999), Guo and Zhao (2000) about ten years ago, and by The center for multilevel modeling (CMM) website (nd). Our paper is different from previous reviews in that we have concentrated on partly different packages and on more commonly used numerical techniques nowadays. Moreover, we considered a binary as well as an ordinal outcome.

3.2 Methods

3.2.1 Data

The dataset we used here is the IMPACT (International Mission on Prognosis and Clinical Trial design in TBI) database. This dataset contains individual patient data from 9,205 patients with moderate and severe Traumatic Brain Injury (TBI) enrolled in eight Randomized Controlled Trials (RCTs) and three observational studies. The patients were treated in different centers, giving the data a multilevel structure. For more details on this study, we refer to Marmarou et al. (2007), and Maas et al. (2007). The permission to access the patient data used in this study was obtained from the principle investigators of the original studies.
The outcome in our analyses is the Glasgow Outcome Scale (GOS), the commonly used outcome scale in TBI studies. GOS has an ordinal five point scale, with categories respectively dead, vegetative state, severe disability, moderate disability and good recovery. We analyzed GOS on the original ordinal scale but also as a binary outcome, dichotomized into "unfavourable" (dead, vegetative and severe disability) versus "favourable" (good recovery and moderate disability).
At patient level, we included age, pupil reactivity and motor score at admission as predictors in the model, their inclusion is motivated by previous studies (Steyerberg et al., 2008). Age was treated as a continuous variable. Motor score and pupil reactivity were treated as categorical variables (motor score: 1=none or extension, 2=abnormal flexion, 3=normal flexion, 4=localises or obeys, 5=untestable, and pupil reactivity: 1=both sides positive, 2=one side positive, 3=both sides negative). Note that treatment was not included in our analysis because of absence of a treatment effect in any of the trials. For further details, see McHugh et al. (2007).
We did include the variable trial since 11 studies were involved and the overall outcome may vary across studies. The trial effect was modelled as a fixed effect in the first analyses and as a random effect in the subsequent analyses. The 231 centers were treated as a random effect (random intercept).
Two sub-datasets were generated in order to examine the performance of the software packages when dealing with logistic random effects regression models on a smaller data set. Sample 1 (cases 2 and 5) consists of a simple random sample from the full data set and contains 500 patients. Sample 2 (cases 3 and 6) was obtained from stratified random sampling the full data set with the centers as strata. It includes 262 patients, representing about 3% of the patients in each hospital.

3.2.2 Random effects models

In random effects models, the residual variance is split up into components that pertain to the different levels in the data (Goldstein, 2011). A two-level model with grouping of patients within centers would include residuals at the patient and center levels. Thus the residual variance is partitioned into a between-center component (the variance of the center-level residuals) and a within-center component (the variance of the patient-level residuals).

The center residuals, often called "center effects", represent unobserved center characteristics that affect patients' outcomes. For the cross-classified random effects model (cases 4-6, see below for a description of the model), data are cross-classified by trial and center because some trials were conducted in more than one center and some centers were involved in more than one trial. Therefore, both trial and center were taken as random effects such that the residual variance is partitioned into three parts: a between-trial component, a between-center component and the residual. Note that for the logistic random effects model the level-1 variance is not identifiable from the likelihood; the classically reported fixed variance of pertains to the latent continuous scale and is the variance of a standard logistic density, see Snijders and Bosker (2011) and Rodrıguez and Elo (2003).

Case 1: logistic random effects model on full data set

A dichotomous or binary logistic random effects model has a binary outcome (Y=0 or 1) and regresses the log odds of the outcome probability on various predictors to estimate the probability that Y=1 happens, given the random effects. The simplest dichotomous 2-level model is given by

$$\begin{aligned} \ln\left(\frac{P(Y_{ij}=1 \mid x_{ij},\mu_j)}{P(Y_{ij}=0 \mid x_{ij},\mu_j)}\right) &= \alpha_1 + \sum_{k=1}^{K}\beta_k x_{kij} + \mu_j \\ \mu_j \sim N(0,\sigma^2) \qquad j=1,2,\ldots,J \qquad & i=1,2,\ldots,n_j \end{aligned} \tag{3.1}$$

with Y_{ij} the dichotomized GOS (with $Y_{ij}=1$ if $GOS=1,2,3$ and $Y_{ij}=0$ otherwise) of the ith subject in the jth center. Further, $x_{ij}=(x_{1ij},...,x_{Kij})$ represents the (first and second level) covariates, α_1 is the intercept and β_k is the kth regression coefficient. Furthermore, u_j is the random effect representing the effect of the jth center. It is assumed that u_j follows a normal distribution with mean 0 and variance σ^2 . Here x_{kij} represents the covariates age, motor score, pupil reactivity and trial. The coefficient β_k measures the effect of increasing x_{kij} by one unit on the log odds ratio.

For an ordinal logistic multilevel model, we adopt the proportional odds assumption and hence we assume that:

$$\begin{aligned} \ln\left(\frac{P(Y_{ij}\le m \mid x_{ij},\mu_j)}{P(Y_{ij}> m \mid x_{ij},\mu_j)}\right) &= \alpha_m + \sum_{k=1}^{K}\beta_k x_{kij} + \mu_j \qquad (m=1,2,3,4) \\ \mu_j \sim N(0,\sigma^2) \qquad j=1,2,\ldots,J \qquad & i=1,2,\ldots,n_j \end{aligned} \tag{3.2}$$

In model (1.2), Y_{ij} is the GOS of the ith subject in the jth center. This equation can be seen as a combination of 4 sub-equations. The difference of the four sub-equations is only in the intercept, and the effect of the covariates is assumed to be the same for all outcome levels (proportional odds assumption). So the coefficient β_k is the log odds ratio of a higher GOS versus a lower GOS when the predictor x_{kij} increases with one unit controlling for the other predictors and the random effect in the model.

In our basic models we assumed a logit link function and a normal distribution for both the binary and the ordinal analysis, but we checked also whether different link functions and other random effect distributions are available in the packages.
Cases 2 and 3: Case 2 is based on sample 1 (500 patients), while case 3 is based on sample 2 (262 patients). For both cases only the binary logistic random effects model (1.1) was fitted to the data.
Case 4: cross-classified logistic random effects model on full data set For this case we treated trial (describing 11 studies) as a second random effect. Since trial is not nested in center, we obtained the following cross-classified random effects model:

$$\begin{gathered} \ln\left(\frac{P(Y_{ij}=1 \mid x_{ij}, u_j, v_l)}{P(Y_{ij}=0 \mid x_{ij}, u_j, v_l)}\right) = \alpha_1 + \sum_{k=1}^{K} \beta_k x_{kijl} + u_j + v_l \\ u_j \sim N(0, \sigma_u^2) \qquad v_l \sim N(0, \sigma_v^2) \\ j = 1, 2, \ldots, J \qquad i = 1, 2, \ldots, n_j \qquad l = 1, 2, \ldots, L \end{gathered} \tag{3.3}$$

with Y_{ijl} is the GOS of the ith subject in the jth center and the lth trial, and $x_{ij} = (x_{1ij1}, ..., x_{KijL})$. Note that equations (1.3) and (1.1) differ only in the additional part v_l which represents the random effect of the lth trial. We assumed that both random effects are independently normally distributed.
Cases 5 and 6:
Case 5 is based on sample 1 and case 6 on sample 2. For both cases model (1.3) was fitted to the data.
For more background on models for hierarchical (clustered) data and also for other types of models, such as marginal Generalized Estimating Equations models the reader is referred to the review of Pendergast et al. (1996).

3.2.3 Software packages

We compared ten different implementations of logistic random effects models. The software packages can be classified according to the statistical approach upon which they are based, i.e.: frequentist or Bayesian. See Additional file 1 [1] for the different philosophy upon which frequentist and Bayesian approaches are based. We first note that both approaches involve the computation of the likelihood or quasi-likelihood. In the frequentist approach parameter estimation is based on the marginal likelihood obtained from expression (1.2) and (1.3) by integrating out the random effects. In the Bayesian approach all parameters are estimated via MCMC sampling methods.
The frequentist approach is included in the R package lme4, in the GLLAMM package of Stata (Rabe-Hesketh et al., 2004), in the SAS procedures GLIMMIX (The GLIMMIX procedure, 2009) and NLMIXED (The NLMIXED procedure, 2009), in the package MLwiN

[1] All additional files in this chapter can be found in the website: http://www.biomedcentral.com/1471-2288/11/77/additional.

([R]IGLS) (Rasbash et al., 2000) and in the program MIXOR (the first program launched for the analysis of a logistic random effects model).
The frequentist approaches differ mainly in the way the integrated likelihood is computed in order to obtain the parameter estimates called maximum likelihood estimate (MLE) or restricted maximum likelihood estimate (REML) depending on the way the variances are estimated. Performing the integration is computationally demanding, especially in the presence of multivariate random effects. As a result, many approximation methods have been suggested to compute the integrated (also called marginal) likelihood. The R package lme4 is based on the Laplace technique, which is the simplest Adaptive Gaussian Quadrature (AGQ) technique based on the evaluation of the function in a well chosen quadrature point per random effect. In the general case, AGQ is a numerical approximation to the integral over the whole support of the likelihood using Q quadrature points adapted to the data (Bates et al., 2009). We used the "adapt" option in GLLAMM in Stata to specify the AGQ method (Rabe-Hesketh et al., 2004). The SAS procedure GLIMMIX allows for several integration approaches and we used AGQ if available (The GLIMMIX procedure, 2009). The same holds for the SAS procedure NLMIXED (The NLMIXED procedure, 2009). The package MLwiN ([R]IGLS) adopts Marginal Quasi-Likelihood (MQL) or Penalised quasi-Likelihood (PQL) to achieve the approximation. Both methods can be computed up to the 2nd order (Rasbash et al., 2000), here we chose the 2nd order PQL procedure. Finally, in MIXOR, only Gauss-Hermite quadrature, also known as a non-AGQ method, is available. Again the number of quadrature points Q determines the desired accuracy (Hedeker and Gibbons, 1996). However Lesaffre and Spiessens (2001) indicated that this method can give a poor approximation to the integrated likelihood when the number of quadrature points is low (say 5, which is the default in MIXOR). Therefore in our analyses we have taken 50 quadrature points but we also applied MIXOR with 5 quadrature points to indicate the sensitivity of the estimation procedure to the choice of Q.
With regard to the optimization technique to obtain the (R)MLE, a variety of techniques are available. R package lme4 uses the NLMINB method which is a local minimiser for the smooth nonlinear function subject to bound-constrained parameters.
Newton-Raphson is the only optimization technique in the GLLAMM package. SAS procedures GLIMMIX and NLMIXED have a large number of optimization techniques. We chose the default Quasi-Newton approach for GLIMMIX and the Newton-Raphson algorithm for NLMIXED. The package MLwiN ([R]IGLS) adopts iterative generalized least squares (IGLS) or restricted IGLS (RIGLS) optimization methods. We used IGLS although it has been shown that RIGLS yields less biased estimates than IGLS Goldstein (1989), we will return to this below. Finally, in MIXOR, the Fisher-scoring algorithm was used.
It has been documented that quasi-likelihood approximations such as those implemented in MLwiN ([R]IGLS) may produce estimates biased towards zero in certain circumstances. The bias could be substantial especially when data are sparse (Lin and Breslow, 1996; Rodriguez and Goldman, 1995). On the other hand, (adaptive) quadrature methods with an adequate number of quadrature points produce less biased estimates (Ng et al., 2006). Note that certain integration and optimization techniques are not available in some software for

a cross-classified logistic random effects model. This will be discussed later.
The other four programs we studied are based on a Bayesian approach. The program most often used for Bayesian analysis is WinBUGS (latest and final version is 1.4.3). WinBUGS is based on the Gibbs Sampler, which is one of the MCMC methods (The BUGS project, nd). The package MLwiN (using MCMC) allows for a multilevel Bayesian analysis, it is based on a combination of Gibbs sampling and Metropolis-Hastings sampling (Browne and Rasbash, 2009), both examples of MCMC sampling. The R package MCMCglmm is designed for fitting generalised linear mixed models and makes use of MCMC techniques that are a combination of Gibbs sampling, slice sampling and Metropolis-Hastings sampling (Hadfield, 2010). Finally, the recent experimental SAS 9.2 procedure MCMC is a general purpose Markov Chain Monte Carlo simulation procedure that is designed to fit many Bayesian models using the Metropolis-Hastings approach (The MCMC procedure, 2009).
In all Bayesian packages we used "non-informative" priors for all the regression coefficients, i.e. a normal distribution with zero mean and a large variance (104). Note that, the adjective "non-informative" prior used in this paper is the classical wording but does not necessarily mean the prior is truly non-informative, as will be seen below. The random effect is assumed to follow a normal distribution and the standard deviation of the random effects is given a uniform prior distribution between 0 and 100. MLwiN, however, uses the Inverse Gamma distribution for the variance as default. Since the choice of the non-informative prior for the standard deviation can seriously affect the estimation of all parameters, other priors for the standard deviation were also used. The total number of iterations for binary models in all cases (except for cases 3 and 6) was 10,000 with a burn-in of 3,000. More iterations (106) were used in cases 3 and 6 in order to get convergence for the small data set. For the ordinal model in case 1, the total number of iterations was 100,000 and the size of the burn-in part was 30,000. We checked convergence of the MCMC chain using the Brooks-Gelman-Rubin (BGR) method (Brooks and Gelman, 1998) in WinBUGS. This method compares within-chain and between-chain variability for multiple chains starting at over-dispersed initial values. Convergence of the chain is indicated by a ratio close to 1. In MLwiN (MCMC) the Raftery-Lewis method was used (Browne and Rasbash, 2009). For MCMCglmm, we used the BGR method by making use of the R-package CODA. The SAS procedure MCMC offers many convergence diagnostic tests, we used the Geweke diagnostic.
The specification of starting values for parameters is a bit different across packages. Among the six frequentist packages, lme4, NLMIXED and MIXOR allow manual specification of the starting values, while in the other packages default starting values are chosen automatically. NLMIXED uses 1 as starting value for all parameters for which no starting values have been specified. For lme4 and MIXOR the choice of the starting values is not clear, while GLIMMIX and GLLAMM base their default starting values on the estimates from a generalized linear model fit. In MLwiN ([R]IGLS) the 2nd order PQL method uses MQL estimates as starting values. Note that for most Bayesian implementations the starting values should be specified by the user. Often the choices of starting values, if not taken too extreme, do not play a great role in the convergence of the MCMC chain but care needs to be exercised for the variance parameters, as seen below.

3.2.4 Analysis

As outlined above, binary and ordinal logistic random effects regression models were fitted to the IMPACT data. All packages are able to deal with the binary logistic random effects model. Furthermore, the packages GLLAMM, GLIMMIX, NLMIXED, MLwiN ([R]IGLS), MIXOR, WinBUGS, MLwiN (MCMC) and SAS MCMC are able to analyze ordinal multilevel data. MCMCglmm only supports the probit model for an ordinal outcome, so that program was not used for the ordinal case. The packages R, GLIMMIX, MLwiN ([R]IGLS), WinBUGS, MLwiN (MCMC) and MCMCglmm can handle the cross-classified random effects model. Syntax codes for the analysis of the IMPACT data with the different packages are provided in Additional file 2.
We compared the packages with respect to the estimates of the parameters and the time needed to arrive at the final estimates. Further, we compared extra facilities, output and easy handling of the programs. Finally, we looked at the flexibility of the software, i.e. whether it is possible to vary the model assumptions made in (1.1) and (1.2), e.g. replacing the logit link by other link functions such as probit and log(-log) link functions or relaxing the assumption of normality for the random effects.

3.3 Results

3.3.1 Descriptive statistics

From the 9,205 patients in the original database, we excluded the patients with a missing GOS at 6 months (n=484) or when there was only partial information available on GOS (n=35), or when the age was missing (n=2) or if the patient was younger than 14 (n=175). This resulted in 8,509 patients in 231 centers in the analysis, of whom 2,396 (28%) died and 4,082 (48%) had an unfavourable outcome six months after injury (see Table 3.1). The median age was 30 (interquartile range 21-45) years, 3522 patients (41%) had a motor score of 3 or lower (none, extension or abnormal flexion), and 1,989 patients (23%) had bilateral non-reactive pupils. The median number of patients per center was 19, ranging from 1 to 425.

Table 3.1: IMPACT study: Descriptive statistics of the study population

	TINT	TIUS	SLIN	SAP	PEG	HITI	UK4	TCDB	SKB	EBIC	HITII	Total
Type	RCT	RCT	RCT	RCT	RCT	RCT	Obs.	Obs.	RCT	Obs.	RCT	
Year of study	1992-1994	1991-1994	1994-1996	1995-1997	1993-1995	1987-1989	1986-1988	1984-1987	1996	1995	1989-1991	
No. of patients	1131	1155	409	924	1574	351	988	667	139	1005	852	8509
No. of centers	50	36	50	57	29	6	4	4	31	67	21	231
Outcome(GOS)												
dead	278(25%)	225(22%)	94(23%)	212(23%)	362(24%)	99(28%)	359(45%)	264(44%)	34(27%)	281(34%)	188(23%)	2396(28%)
vegetative	44(4%)	42(4%)	14(3%)	24(3%)	114(8%)	10(3%)	13(2%)	34(6%)	6(5%)	18(2%)	32(4%)	351(4%)
severe disability	134(12%)	128(12%)	69(17%)	142(16%)	298(20%)	62(18%)	146(19%)	95(16%)	30(24%)	123(15%)	108(13%)	1335(16%)
moderate disability	171(15%)	180(17%)	84(21%)	174(19%)	374(25%)	64(18%)	130(16%)	104(17%)	27(21%)	159(19%)	199(24%)	1666(20%)
good recovery	491(44%)	466(45%)	148(36%)	367(40%)	362(24%)	115(33%)	143(18%)	107(18%)	29(23%)	241(29%)	292(36%)	2761(32%)
Predictor(age)												
Median(IQ range)	30(21-45)	30(23-41)	28(21-43)	32(23-47)	27(20-38)	34(21-47)	36(22-55)	26(21-40)	27(20-39)	37.5(24-59)	33(22-49)	30(21-45)
Predictor(motor)												
none	5(0%)	9(1%)	0(0%)	141(15%)	475(32%)	122(35%)	113(14%)	136(23%)	34(27%)	150(18%)	210(26%)	1395(16%)
extension	136(12%)	143(14%)	55(13%)	123(13%)	180(12%)	41(12%)	85(11%)	107(18%)	22(18%)	80(10%)	70(9%)	1042(12%)
abnormal flexion	237(21%)	132(13%)	91(22%)	143(16%)	165(11%)	45(13%)	37(5%)	74(12%)	14(11%)	55(7%)	92(11%)	1085(13%)
normal flexion	327(29%)	300(29%)	127(31%)	223(24%)	334(22%)	56(16%)	141(18%)	122(20%)	16(13%)	113(14%)	181(22%)	1940(23%)
localises	384(34%)	406(39%)	134(33%)	286(31%)	309(21%)	77(22%)	191(24%)	113(19%)	21(17%)	182(22%)	199(24%)	2302(27%)
obeys command	29(3%)	51(5%)	2(1%)	0(0%)	47(3%)	0(0%)	30(4%)	21(4%)	2(2%)	99(12%)	8(1%)	289(3%)
untestable & not available	0(0%)	0(0%)	0(0%)	3(0%)	0(0%)	9(3%)	194(25%)	31(6%)	17(14%)	143(18%)	59(7%)	456(5%)
Predictor(pupil)												
both side positive	806(72%)	703(68%)	315(77%)	619(67%)	784(52%)	232(66%)	427(54%)	300(50%)	70(56%)	535(65%)	585(71%)	5376(63%)
one side positive	177(16%)	118(11%)	79(19%)	178(19%)	156(10%)	53(15%)	115(15%)	55(9%)	35(28%)	79(10%)	99(12%)	1144(13%)
both side negative	135(12%)	220(21%)	15(4%)	122(13%)	570(38%)	65(19%)	249(32%)	249(41%)	21(17%)	208(25%)	135(17%)	1989(23%)

3.3.2 Case 1: binary and ordinal logistic random effects model on full data set

Binary model

Fitting the dichotomous model in the different packages gave similar results (see Table 3.2). For the frequentist approaches the R package lme4, the Stata package GLLAMM, the SAS procedures GLIMMIX and NLMIXED, and the programs MLwiN ([R]IGLS) and MIXOR provided almost the same results for the fixed effects and the variance of the random effects. One example is age, with estimated coefficients of 0.623, 0.623, 0.618, 0.623, 0.623 and 0.623, respectively for the different programs and all estimated SDs close to 0.028. Estimates for the variance of the random effects were also similar: 0.101, 0.102, 0.107, 0.102, 0.101 and 0.102, respectively. As can be noticed from Table 3.2, lme4 did not give an estimate for the SD of the variance of the random effects. The reason was provided by the developer of the package in his book (Bates D: lme4: Mixed-effects modelling with R, submitted) stating that the sampling distribution of the variance is highly skewed which makes the standard error nonsensical.
The Bayesian programs WinBUGS, MLwiN (MCMC), MCMCglmm and the SAS procedure MCMC gave similar posterior means and these were also close to the MLEs obtained from the frequentist software. For example, the posterior mean (SD) of the regression coefficient of age was 0.626 (0.028), 0.625 (0.029), 0.636 (0.028) and 0.630 (0.025) for WinBUGS, MLwiN (MCMC), MCMCglmm and SAS procedure MCMC, respectively. The posterior mean of the variance of the random effects was estimated as 0.119, 0.113, 0.110 and 0.160, respectively with SD close to 0.30.
The random effects estimates of the 231 centers could easily be derived from all packages except for MIXOR and were quite similar. For example the Pearson correlation for the estimated random effects from WinBUGS and R was 0.9999.

Table 3.2: IMPACT study: Results of the binary model in case 1 (full data set). The variance of the random effects with its standard error is given.

		R(lme4)		GLLAMM		GLIMMIX		NLMIXED		MLwiN([R]IGLS)		MIXOR		WinBUGS		MLwiN(MCMC)		MCMCglmm		MCMC	
Computing time		34s		7min		9s		15min		2s		30s		14min		4min		2min		37h	
Random Effects		Variance: 0.101		Variance: 0.102(0.027)		Variance: 0.107(0.027)		Variance: 0.102(0.027)		Variance: 0.101(0.025)		Variance: 0.102(0.032)		Variance: 0.119(0.030)		Variance: 0.113(0.030)		Variance: 0.110(0.031)		Variance: 0.160(0.034)	
	covar	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Fixed Effects	const	-0.014	0.114	-0.014	0.114	-0.014	0.114	-0.014	0.114	-0.014	0.114	-0.014	0.126	-0.026	0.115	-0.003	0.110	-0.019	0.121	-0.103	0.099
	pupil2	0.656	0.074	0.656	0.074	0.65	0.074	0.656	0.075	0.657	0.074	0.656	0.089	0.659	0.075	0.656	0.075	0.674	0.072	0.666	0.071
	pupil3	1.404	0.069	1.404	0.07	1.392	0.069	1.404	0.07	1.405	0.069	1.404	0.075	1.410	0.069	1.406	0.068	1.434	0.068	1.424	0.069
	age	0.623	0.028	0.623	0.028	0.618	0.028	0.623	0.028	0.623	0.028	0.623	0.029	0.626	0.028	0.625	0.029	0.636	0.028	0.630	0.029
	motor2	0.618	0.106	0.618	0.106	0.612	0.105	0.618	0.106	0.618	0.106	0.618	0.126	0.623	0.106	0.617	0.104	0.623	0.110	0.654	0.103
	motor3	-0.154	0.097	-0.154	0.097	-0.153	0.097	-0.154	0.097	-0.154	0.097	-0.154	0.101	-0.152	0.098	-0.158	0.096	-0.159	0.105	-0.131	0.096
	motor4	-0.782	0.086	-0.782	0.086	-0.775	0.086	-0.782	0.087	-0.782	0.086	-0.782	0.103	-0.781	0.088	-0.786	0.084	-0.811	0.089	-0.757	0.076
	motor5	-1.404	0.088	-1.404	0.089	-1.394	0.088	-1.404	0.089	-1.405	0.088	-1.404	0.108	-1.409	0.090	-1.412	0.086	-1.449	0.097	-1.394	0.070
	motor6	-1.591	0.166	-1.591	0.167	-1.577	0.166	-1.591	0.167	-1.592	0.166	-1.591	0.186	-1.598	0.168	-1.602	0.168	-1.642	0.177	-1.593	0.166
	motor9	-0.534	0.136	-0.534	0.136	-0.529	0.136	-0.534	0.136	-0.534	0.136	-0.534	0.156	-0.535	0.136	-0.536	0.136	-0.561	0.150	-0.533	0.129
	trial2	-0.073	0.125	-0.073	0.126	-0.071	0.126	-0.073	0.126	-0.073	0.125	-0.073	0.132	-0.061	0.129	-0.081	0.121	-0.058	0.131	-0.007	0.115
	trial3	0.218	0.139	0.217	0.139	0.216	0.139	0.218	0.139	0.218	0.138	0.217	0.136	0.222	0.140	0.210	0.136	0.229	0.141	0.240	0.139
	trial4	-0.192	0.116	-0.192	0.117	-0.189	0.117	-0.192	0.117	-0.192	0.116	-0.192	0.099	-0.184	0.117	-0.195	0.115	-0.174	0.122	-0.116	0.128
	trial5	0.107	0.114	0.107	0.115	0.107	0.115	0.107	0.115	0.107	0.114	0.107	0.128	0.119	0.117	0.099	0.114	0.114	0.117	0.184	0.112
	trial6	-0.039	0.173	-0.039	0.174	-0.039	0.174	-0.039	0.174	-0.039	0.173	-0.039	0.202	-0.034	0.175	-0.046	0.172	-0.048	0.187	0.049	0.188
	trial7	0.686	0.170	0.686	0.17	0.68	0.171	0.686	0.17	0.687	0.17	0.686	0.151	0.693	0.172	0.680	0.172	0.704	0.184	0.755	0.182
	trial8	0.672	0.176	0.672	0.176	0.665	0.177	0.672	0.176	0.673	0.176	0.672	0.175	0.682	0.181	0.652	0.172	0.691	0.182	0.744	0.198
	trial9	0.373	0.231	0.373	0.232	0.368	0.231	0.373	0.232	0.373	0.231	0.373	0.229	0.382	0.234	0.368	0.232	0.382	0.248	0.408	0.223
	trial10	0.090	0.123	0.09	0.123	0.09	0.123	0.09	0.123	0.09	0.123	0.090	0.112	0.099	0.124	0.083	0.118	0.097	0.127	0.149	0.125
	trial11	-0.239	0.125	-0.238	0.127	-0.233	0.126	-0.238	0.127	-0.239	0.125	-0.238	0.144	-0.225	0.127	-0.239	0.123	-0.230	0.134	-0.128	0.121

Ordinal model-proportional odds model

Fitting the proportional odds model in the different packages also gave similar results (see Table 3.3). For the frequentist approach, the Stata package GLLAMM, the two SAS procedures GLIMMIX and NLMIXED, the packages MLwiN ([R]IGLS) and MIXOR gave very similar estimates for the fixed effects parameters and the variance of the random effects. The estimate (SD) of e.g. the regression coefficient of age was 0.591 (0.023), 0.588 (0.023), 0.591(0.023), 0.592 (0.023) and 0.591 (0.027), respectively. The estimate of the variance (SD) of the random effects were 0.085 (0.020), 0.090 (0.021), 0.085 (0.020), 0.085 (0.019), and 0.085 (0.024), respectively. The MIXOR results were somewhat different from those of the other packages when based on 5 quadrature points, but this difference largely disappeared when 50 quadrature points were used, see Table 3.3. However, the SDs did not change much by increasing Q from 5 to 50 and we are not sure about the reason behind.
For the Bayesian approaches, WinBUGS and MLwiN (MCMC) produced similar results as the frequentist approaches. The posterior mean of the regression coefficient of age in WinBUGS was 0.551 and 0.592 in MLwiN (MCMC), with SD = 0.023 in both cases (same as the SAS frequentist result). The posterior mean of the variance of the random effects was 0.096 in WinBUGS and 0.093 in MLwiN (MCMC) and for both SD = 0.022, very close to the frequentist estimates. We stopped running the SAS MCMC procedure after 2,000 iterations because this already took 19 hours and the chains based on the last 1,000 iterations were far from being converged.
Finally, the estimated random effects for the 231 centers were quite the same across the different packages (except for MIXOR) with correlation again practically 1.

Table 3.3: IMPACT study: Results from the ordinal model in case 1 (full data set). The variance of the random effects with its standard error is given

		GLLAMM		GLIMMIX		NLMIXED		MLwiN([R]IGLS)		MIXOR		WinBUGS		MLwiN(MCMC)	
Computing time		11min		11s		24min		6s		3min		8h		15min	
Random Effects		Variance: 0.085(0.020)		Variance: 0.090(0.021)		Variance: 0.085(0.020)		Variance: 0.085(0.019)		Variance: 0.085(0.024)		Variance: 0.096(0.022)		Variance: 0.093(0.022)	
	covar	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Fixed Effects	pupil2	0.705	0.062	0.702	0.062	0.705	0.062	0.707	0.062	0.705	0.082	0.703	0.062	0.708	0.063
	pupil3	1.401	0.057	1.396	0.057	1.401	0.057	1.406	0.057	1.401	0.062	1.405	0.057	1.406	0.057
	age	0.591	0.023	0.588	0.023	0.591	0.023	0.592	0.023	0.591	0.027	0.551	0.023	0.592	0.023
	motor2	0.277	0.083	0.275	0.083	0.277	0.083	0.279	0.083	0.277	0.091	0.276	0.082	0.282	0.086
	motor3	-0.296	0.081	-0.295	0.081	-0.296	0.081	-0.296	0.081	-0.296	0.077	-0.305	0.080	-0.292	0.084
	motor4	-0.846	0.072	-0.843	0.072	-0.846	0.072	-0.848	0.072	-0.846	0.074	-0.847	0.072	-0.843	0.074
	motor5	-1.369	0.073	-1.365	0.073	-1.369	0.073	-1.373	0.073	-1.369	0.080	-1.368	0.073	-1.367	0.076
	motor6	-1.572	0.137	-1.565	0.133	-1.572	0.137	-1.577	0.133	-1.572	0.156	-1.567	0.137	-1.574	0.138
	motor9	-0.630	0.111	-0.628	0.112	-0.630	0.111	-0.632	0.112	-0.630	0.115	-0.640	0.112	-0.629	0.112
	trial2	-0.067	0.107	-0.066	0.106	-0.067	0.107	-0.067	0.105	-0.067	0.112	-0.075	0.109	-0.054	0.113
	trial3	0.252	0.117	0.251	0.116	0.252	0.117	0.253	0.116	0.252	0.114	0.245	0.117	0.260	0.120
	trial4	-0.122	0.099	-0.120	0.098	-0.122	0.099	-0.122	0.097	-0.122	0.083	-0.121	0.099	-0.111	0.103
	trial5	0.189	0.097	0.190	0.097	0.189	0.097	0.190	0.096	0.189	0.106	0.177	0.098	0.204	0.103
	trial6	0.051	0.146	0.051	0.147	0.051	0.146	0.051	0.146	0.051	0.138	0.083	0.147	0.062	0.149
	trial7	0.772	0.142	0.768	0.144	0.772	0.142	0.775	0.143	0.772	0.132	0.783	0.144	0.781	0.145
	trial8	0.901	0.148	0.900	0.149	0.901	0.148	0.904	0.148	0.900	0.259	0.888	0.150	0.917	0.151
	trial9	0.341	0.190	0.339	0.193	0.341	0.190	0.342	0.193	0.341	0.183	0.343	0.192	0.352	0.195
	trial10	0.265	0.102	0.264	0.102	0.265	0.102	0.266	0.101	0.265	0.088	0.302	0.102	0.275	0.105
	trial11	-0.047	0.106	-0.044	0.105	-0.047	0.106	-0.047	0.104	-0.047	0.092	-0.030	0.106	-0.033	0.111
	Inter1	-1.190	0.098	-1.188	0.098	-1.190	0.098	-1.197	0.097	-1.190	0.094	-1.186	0.098	-1.208	0.111
	Inter2	-0.931	0.098	-0.930	0.097	-0.931	0.098	-0.937	0.097	-0.931	0.106	-0.928	0.098	-0.949	0.111
	Inter3	-0.040	0.098	-0.040	0.097	-0.040	0.098	-0.041	0.096	-0.040	0.117	-0.042	0.100	-0.056	0.109
	Inter4	1.026	0.098	1.025	0.097	1.026	0.098	1.031	0.097	1.026	0.121	1.007	0.103	1.012	0.109

3.3.3 Cases 2 and 3: binary logistic random effects models on samples 1 and 2

The conclusions for case 2 are the same as for case 1 (see Table 3.4), but not for case 3 (see Table 3.5). The results for the Bayesian analyses are rather different from the results of the frequentist implementations but similar to each other, in particular with regard to the posterior standard errors. For the frequentist approaches, the variance of the random effects was estimated zero and the standard error was estimated as zero or could not be estimated. What is more important in case 3 is that the posterior means depended much on the choice of the non-informative priors for the variance component, i.e. uniform (0,1) and Inverse Gamma (0.001,0.001), but we have tried more priors and elaborated on this in the discussion section of the paper.

Table 3.4: IMPACT study: Results from the binary model in case 2 (sample 1). The variance of the random effects with its standard error is given

		R(lme4)		GLLAMM		GLIMMIX		NLMIXED		MLwiN([R]IGLS)		MIXOR		WinBUGS		MLwiN(MCMC)		MCMCglmm	
Computing time		1s		1m		1s		49s		1s		1s		3min		1min		9s	
Random Effects		Variance: 0.046		Variance: 0.051(0.150)		Variance: 0.051(0.150)		Variance: 0.051(0.150)		Variance: 0.047(0.127)		Variance: 0.051(0.237)		Variance: 0.450(0.336)		Variance: 0.348(0.289)		Variance: 0.279(0.275)	
	covar	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Fixed Effects	const	-0.157	0.455	-0.156	0.458	-0.157	0.458	-0.156	0.458	-0.157	0.455	-0.156	0.598	-0.093	0.621	-0.106	0.420	-0.175	0.525
	pupil2	0.432	0.320	0.431	0.321	0.431	0.321	0.431	0.321	0.432	0.320	0.431	0.343	0.438	0.352	0.447	0.340	0.456	0.356
	pupil3	1.357	0.284	1.359	0.289	1.359	0.289	1.359	0.289	1.358	0.284	1.359	0.354	1.512	0.314	1.504	0.314	1.553	0.303
	age	0.638	0.115	0.638	0.117	0.638	0.117	0.638	0.117	0.638	0.115	0.638	0.146	0.706	0.128	0.699	0.125	0.727	0.126
	motor2	0.040	0.435	0.039	0.437	0.039	0.437	0.039	0.437	0.040	0.435	0.039	0.591	0.032	0.496	0.054	0.449	0.112	0.484
	motor3	-0.066	0.432	-0.068	0.438	-0.068	0.438	-0.068	0.438	-0.066	0.432	-0.068	0.574	-0.169	0.505	-0.124	0.441	-0.136	0.496
	motor4	-0.772	0.369	-0.773	0.375	-0.773	0.375	-0.774	0.375	-0.772	0.369	-0.774	0.506	-0.914	0.452	-0.860	0.373	-0.863	0.417
	motor5	-0.928	0.361	-0.930	0.368	-0.930	0.368	-0.930	0.368	-0.928	0.361	-0.930	0.499	-1.076	0.445	-1.032	0.373	-1.089	0.407
	motor6	-1.600	0.733	-1.602	0.737	-1.601	0.737	-1.602	0.737	-1.600	0.733	-1.602	0.967	-1.825	0.822	-1.794	0.781	-1.868	0.809
	motor9	0.469	0.580	0.469	0.582	0.469	0.582	0.469	0.582	0.469	0.580	0.469	0.749	0.526	0.650	0.547	0.628	0.529	0.599
	trial2	-0.315	0.446	-0.315	0.448	-0.315	0.448	-0.315	0.448	-0.315	0.446	-0.315	0.457	-0.365	0.527	-0.380	0.490	-0.283	0.518
	trial3	0.634	0.510	0.635	0.513	0.636	0.513	0.635	0.513	0.634	0.510	0.635	0.561	0.712	0.588	0.667	0.543	0.729	0.577
	trial4	-0.226	0.423	-0.226	0.425	-0.226	0.425	-0.226	0.425	-0.226	0.423	-0.226	0.489	-0.258	0.511	-0.285	0.442	-0.229	0.445
	trial5	0.542	0.405	0.542	0.407	0.542	0.408	0.542	0.408	0.542	0.405	0.542	0.428	0.538	0.496	0.515	0.423	0.590	0.460
	trial6	0.077	0.648	0.078	0.651	0.078	0.651	0.078	0.651	0.078	0.648	0.077	0.776	0.079	0.769	0.043	0.691	0.177	0.739
	trial7	0.680	0.492	0.681	0.497	0.682	0.497	0.682	0.497	0.680	0.492	0.681	0.638	0.718	0.611	0.687	0.577	0.843	0.578
	trial8	0.628	0.504	0.627	0.507	0.627	0.507	0.627	0.507	0.628	0.504	0.627	0.499	0.582	0.639	0.528	0.583	0.619	0.655
	trial9	1.553	0.927	1.555	0.932	1.555	0.932	1.555	0.932	1.554	0.927	1.555	1.402	1.800	1.097	1.787	1.047	1.963	1.052
	trial10	-0.017	0.458	-0.018	0.462	-0.018	0.462	-0.018	0.462	-0.017	0.458	-0.018	0.556	-0.126	0.535	-0.133	0.468	-0.110	0.543
	trial11	0.093	0.442	0.094	0.444	0.094	0.444	0.094	0.444	0.093	0.442	0.093	0.580	0.124	0.541	0.099	0.468	0.154	0.509

Table 3.5: IMPACT study: Results from the binary model in case 3 (sample 2). The variance of the random effects with its standard error is given

		R(lme4)		GLLAMM		GLIMMIX		NLMIXED		MLwiN([R]IGLS)		MIXOR		WinBUGS		MLwiN(MCMC)	
Computing time		1s		40s		1s		14s		1s		1s		75min		30min	
Random Effects		Variance: 0		Variance: 0(0)		Variance: 0( )		Variance: 0( )		Variance: 0(0)		Variance:		Variance: 28.040(22.130)		Variance: 19.892(17.236)	
	covar	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Fixed Effects	const	-0.549	0.618	-0.549	0.618	-0.549	0.618	-0.549	0.618	-0.549	0.618	-0.549	0.674	-1.725	1.644	-1.448	1.473
	pupil2	0.322	0.451	0.322	0.451	0.322	0.451	0.322	0.451	0.322	0.451	0.322	0.472	1.946	1.184	1.647	1.109
	pupil3	1.187	0.408	1.187	0.408	1.187	0.408	1.187	0.408	1.187	0.408	1.187	0.471	3.141	1.148	2.774	1.085
	age	0.729	0.175	0.729	0.174	0.729	0.175	0.729	0.175	0.729	0.174	0.729	0.189	2.195	0.719	1.913	0.675
	motor2	1.231	0.653	1.231	0.652	1.231	0.653	1.231	0.653	1.231	0.652	1.231	0.791	3.933	1.851	3.400	1.713
	motor3	0.226	0.562	0.226	0.562	0.226	0.562	0.226	0.562	0.226	0.562	0.226	0.608	1.571	1.332	1.304	1.231
	motor4	-0.714	0.539	-0.714	0.539	-0.714	0.539	-0.714	0.539	-0.714	0.539	-0.714	0.569	-1.553	1.128	-1.434	1.062
	motor5	-1.700	0.533	-1.700	0.533	-1.700	0.533	-1.700	0.533	-1.700	0.533	-1.700	0.624	-3.742	1.312	-3.382	1.243
	motor6	-1.238	0.860	-1.238	0.859	-1.238	0.860	-1.238	0.860	-1.238	0.859	-1.238	0.898	-0.484	1.978	-0.800	1.790
	motor9	-1.423	0.775	-1.423	0.775	-1.423	0.775	-1.423	0.775	-1.423	0.775	-1.423	0.917	-0.969	1.415	-1.107	1.333
	trial2	0.433	0.579	0.433	0.579	0.433	0.579	0.433	0.579	0.433	0.579	0.433	0.606	-0.900	1.883	-0.600	1.650
	trial3	1.083	0.781	1.083	0.781	1.083	0.781	1.083	0.781	1.083	0.781	1.083	1.021	1.381	2.091	1.292	1.847
	trial4	-0.523	0.692	-0.523	0.692	-0.523	0.692	-0.523	0.692	-0.523	0.692	-0.523	0.671	-2.070	1.898	-1.813	1.715
	trial5	0.580	0.588	0.580	0.588	0.580	0.588	0.580	0.588	0.580	0.588	0.580	0.606	1.936	1.845	1.598	1.601
	trial6	1.100	0.853	1.100	0.853	1.100	0.853	1.100	0.853	1.100	0.853	1.100	1.524	2.633	2.053	2.398	1.876
	trial7	1.769	0.678	1.769	0.678	1.769	0.678	1.769	0.678	1.769	0.678	1.769	1.776	1.657	1.919	1.678	1.744
	trial8	0.714	0.671	0.714	0.671	0.714	0.671	0.714	0.671	0.714	0.671	0.714	1.356	-0.515	2.412	-0.226	2.135
	trial9	1.805	1.033	1.805	1.033	1.805	1.033	1.805	1.033	1.805	1.033	1.805	0.923	4.520	3.447	3.968	3.005
	trial10	0.689	0.655	0.689	0.655	0.689	0.655	0.689	0.655	0.689	0.655	0.689	0.745	0.727	1.579	0.748	1.425
	trial11	0.322	0.668	0.322	0.668	0.322	0.668	0.322	0.668	0.322	0.668	0.322	0.741	0.412	1.405	0.382	1.296

3.3.4 Case 4: cross-classified binary logistic random effects model based on full data set

Only lme4 in R, GLIMMIX, MLwiN ([R]IGLS), WinBUGS, MLwiN (MCMC) and MCMCglmm could handle this analysis. The results for these packages were quite similar, as shown in Table 3.6. For example for age the estimates (SD) were 0.623 (0.028), 0.617 (0.028), 0.623 (0.028), 0.624 (0.028), 0.624 (0.027) and 0.635 (0.028), for lme4, GLIMMIX, MLwiN ([R]IGLS), WinBUGS, MLwiN (MCMC) and MCMCglmm, respectively. The variances for the random effect of center were 0.116, 0.113, 0.116, 0.119, 0.120 and 0.106, respectively and for the random effect of trial they were 0.067, 0.075, 0.067, 0.114, 0.095 and 0.094, respectively.

Table 3.6: IMPACT study: Results from the cross-classified model in case 4 (full data set). The variance of the random effects with its standard error is given

			R(lme4)		GLIMMIX		MLwiN([R]IGLS)		WinBUGS		MLwiN(MCMC)		MCMCglmm	
Computing time			18s		3s		5s		17min		3min		3min	
Random Effects	Center		0.116		0.113(0.028)		0.116(0.028)		0.119(0.031)		0.120(0.032)		0.106(0.031)	
	Trial		0.067		0.075(0.042)		0.067(0.035)		0.114(0.079)		0.095(0.065)		0.094(0.113)	
		covar	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Fixed Effects		const	0.105	0.111	0.107	0.114	0.105	0.112	0.126	0.132	0.14	0.127	0.110	0.140
		pupil2	0.657	0.074	0.650	0.074	0.658	0.074	0.656	0.075	0.657	0.075	0.681	0.080
		pupil3	1.411	0.069	1.396	0.069	1.412	0.069	1.412	0.069	1.413	0.07	1.436	0.073
		age	0.623	0.028	0.617	0.028	0.623	0.028	0.624	0.028	0.624	0.027	0.635	0.028
		motor2	0.620	0.105	0.613	0.105	0.620	0.105	0.623	0.108	0.61	0.106	0.631	0.106
		motor3	-0.155	0.097	-0.154	0.097	-0.156	0.097	-0.152	0.100	-0.165	0.1	-0.157	0.096
		motor4	-0.782	0.086	-0.773	0.086	-0.782	0.086	-0.780	0.087	-0.793	0.093	-0.798	0.089
		motor5	-1.406	0.088	-1.392	0.087	-1.408	0.088	-1.406	0.091	-1.417	0.093	-1.432	0.087
		motor6	-1.579	0.166	-1.563	0.165	-1.581	0.166	-1.584	0.169	-1.602	0.167	-1.617	0.168
		motor9	-0.502	0.135	-0.498	0.135	-0.502	0.135	-0.509	0.136	-0.52	0.141	-0.524	0.132

3.3.5 Cases 5 and 6: cross-classified binary logistic random effects models on samples 1 and 2

As for case 2, we obtained in case 5 essentially the same results with all packages. For case 6, the frequentist results were similar but the Bayesian results were different and were much affected by the prior of the variance parameter as in case 3 (tables for cases 5 and 6 are not shown).

3.3.6 Usability, flexibility and speed

The packages greatly differed in their usability, by which we mean the availability of diagnostic tools/plots; ease of displaying/extracting parameter estimates and exporting results, etc. But it must be stated that all packages require a sound statistical knowledge in multilevel modelling in order to analyze such data in a reliable manner.
SAS is based on procedures for which certain options can be turned on and off. Understanding the different options in the statistical SAS procedures often requires a great deal of statistical background since the procedures are based on the most advanced and computationally powerful methods. Also SAS data management is quite powerful but is also associated with a steep learning curve. The SAS procedures NLMIXED and MCMC offer some programming facilities.
The package R has gained a lot of attention in the last decade and is becoming increasingly popular among statisticians and non-statisticians. It requires programming skills and has many basic functions. In addition, R offers great graphics to the user. For the MCMCglmm package in R, we experienced difficulties in understanding the syntax for specifying the prior of the variance parameters as explained in the manual.
Stata is very handy for analyzing simple as well as complicated problems. It has a command-line interface and also includes a graphical user interface since version 8.0. The software allows user-written packages just as in R and provides some programming facilities. The package GLLAMM is powerful in dealing with a large range of complex problems.
WinBUGS is the most popular general purpose package for Bayesian analysis with now more than 30,000 registered users. The package allows for a great variety of analyses using a programming language that resembles to some extent that of R. WinBUGS requires about the same programming skills as R.
MIXOR needs no programming but provides very limited output. Furthermore, MLwiN has a clear and intuitive interface to specify a random effects model, but lacks a simple syntax file structure.
The packages also differ in what they offer as standard output besides the parameter estimates. WinBUGS allows for the most extensive output, including diagnostic plots for model evaluation and plots of the individual random center effects. All packages except MIXOR can provide estimates of the random effects. In Figure 1 we show the box plots of the sampled random effects in WinBUGS for the first 10 centers of the binary logistic random effects model applied to the IMPACT data. Of course with packages like SAS and R the output of

the statistical procedures can be saved and then processed by some other procedure or function to deliver the required graph or additional diagnostic analysis. For example, Figure 2 is produced with R and shows the histogram of the random effects of the binary IMPACT logistic random effects model.

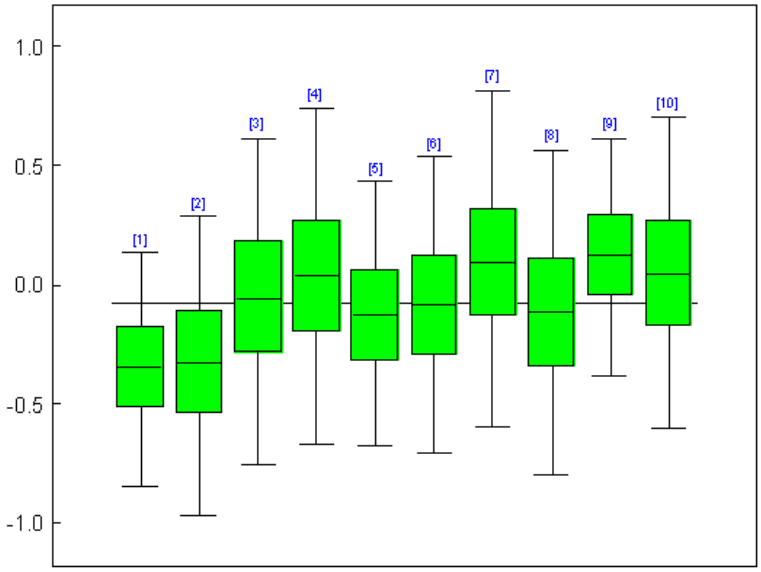

Figure 3.1: IMPACT study: Box plot of a sample of the random effects (for center 1 to 10). Each box represents a center with its random effects estimate and confidence interval.

Flexibility differs somewhat in the packages. All packages could handle a probit model and a log(-log) model except lme4 and MCMCglmm (MCMCglmm allows for logit or probit link functions for a binary model but only the probit link function for the ordinal model). But, only WinBUGS allows for changing the distribution of the random effects. Table 1 shows that WinBUGS has the greatest flexibility in adapting the model assumptions.

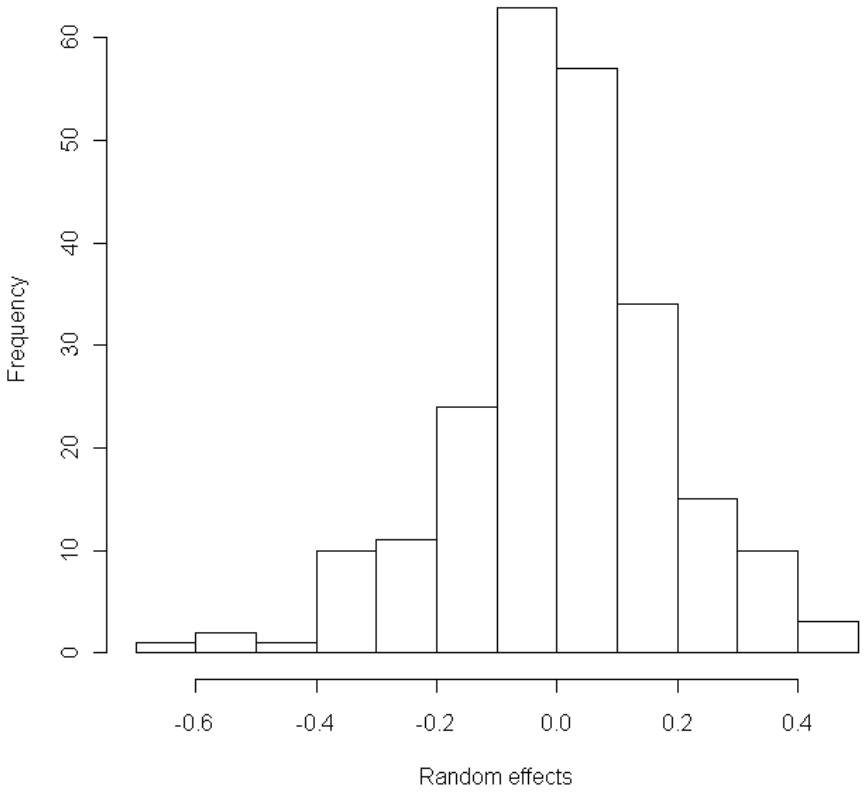

Figure 3.2: IMPACT study: Histogram of the random effects in the binary model in R

Table 3.7: Extra abilities of different packages. $\star$: Only probit model is available in MCMCglmm for ordinal model.

Package	Program/function	Link function		Obtaining the random effects	Handling ordinal proportional odds model	Modeling cross-classified model	Other than normal random effects
		Probit model	Log(-log) model				
R	LME4			X		X	
	MCMCglmm	X		X	X*	X	
MIXOR	MIXOR	X	X		X		
STATA	GLLAMM	X	X	X	X		
SAS	NLMIXED	X	X	X	X		
	GLIMMIX	X	X	X	X	X	
	MCMC	X	X	X			
MLwiN	PQL or MCMC	X	X	X	X	X	
WinBUGS	MCMC	X	X	X	X	X	X

The speed of the computations varied widely. All computations were done on an Intel Core(TM) 2 Duo E8400 processor with 3.0 GHz CPU and 3.21 GB internal memory. For case 1, only a few seconds were needed to provide the estimates with the frequentist approaches to fit the binary logistic random effects, except for SAS NLMIXED and Stata GLLAMM which needed 15 minutes and 7 minutes, respectively. The MLwiN ([R]IGLS) procedure (using 2nd order PQL) was the fastest, and GLIMMIX was almost as fast followed by lme4 and MIXOR. The Bayesian approaches were considerably slower, which is not surprising since MCMC sampling is time consuming. However, a major handicap to perform an honest comparison with regard computational speed is that the checking for convergence of MCMC methods is far more difficult than in a frequentist sense (Gelman et al., 2003) and not standardized. Nevertheless, MCMCglmm was the winner this time, but we considered all computation times as acceptable, except for the SAS MCMC procedure which took 37 hours for the binary model. Similar findings were obtained for the ordinal logistic random effects model, but compared to the binary model, the time to converge increased considerably for some software. Now the winner in the frequentist software was GLIMMIX closely followed up by MIXOR. For the Bayesian software, MLwiN (MCMC) was the winner, much faster than WinBUGS. The SAS procedure MCMC never got to convergence (we stopped it) and as mentioned above, the MCMCglmm program does not allow the ordinal logistic random effects model.

3.4 Discussion

3.4.1 Performance of each package

Although the parameter estimates were very similar between the ten software implementations, we found considerable variations in computing time, usability and flexibility.
Speed: Most of the frequentist approaches were very fast, taking only seconds, with the SAS NLMIXED procedure and the Stata package GLLAMM as exceptions. Overall, the SAS procedure GLIMMIX, the program MIXOR and the package MLwiN ([R]IGLS) were the winners. The fact that NLMIXED and GLLAMM took much longer time has much to do with that they are general purpose programs suitable for fitting a variety of complex random effects models and that they both use the AGQ method. The Bayesian approaches were invariably slower than the frequentist approaches, which is due to the computational intensive MCMC approach and that convergence is much harder to judge than in a classical frequentist sense. The speed of the Bayesian procedures appears to depend also more on the sample size than the frequentist approaches. As a result, long processing times as in WinBUGS (14 minutes for binary and 8 hours for ordinal model, respectively) may prevent the user to do much on exploratory statistical research. The R package MCMCglmm and MLwiN (MCMC) were much faster than WinBUGS, taking only a few minutes for both binary and ordinal cases. Hence, from a computational point of view, MCMCglmm and MLwiN (MCMC) are our software of choice for multilevel modeling.
In our experience, the SAS procedure MCMC was inefficient in dealing with mixed models.

It was far too time consuming (37 hours for the binary model) and it did converge neither for the regression coefficients nor for the variance of the random effects. At this moment, we cannot recommend this SAS procedure for fitting logistic random effects regression models.
Usability and flexibility: The packages differ much in nature, like working interface and data management. MLwiN and MIXOR are menu-driven although writing syntax is also allowed in both packages. SAS is supposed to work in batch mode with some procedures and macros. The others, WinBUGS, R and STATA, are embedded in a programming language. Which package to prefer from the usability viewpoint is difficult to say since it very much depends on the user but also on whether the logistic random effects model fitting is a stand-alone exercise. We know that in practice this is often not the case since we would like to process output of such an analysis to produce e.g. nice graphs. From this viewpoint MIXOR and WinBUGS score lower since they require the user to switch to other software, such as R, to produce additional output or better quality graphs. However, in recent years some packages, like R2WinBUGS in R, can combine WinBUGS and R (or other software) nicely. See the BUGS website (The BUGS project, nd) to get more information.
For the cross-classified random effects model and the sub-dataset analysis, some integration methods and optimization techniques were not available in some software. For example, in GLIMMIX, AGQ is not available for the cross-classified random effects model and we had to change to Residual Subject-specific Pseudo-Likelihood.
In the R package MCMCglmm, by default the residual variance should be explicitly specified for random effects models. But, as this variance parameter is not identifiable for the logistic random effects model, as seen above, it has to be fixed at a particular value. MCMCglmm uses arbitrary values larger than zero, while the other packages ignore the residual variance since it does not play a role in the estimation process. In order to make the results comparable, the posteriors had to be rescaled which worked most often. But one should be aware that the prior specification will be different after rescaling the posteriors, so there will be differences between MCMCglmm and other Bayesian packages if the prior considerably influences the posterior which happened here for cases 3 and 6.
RIGLS is the restricted version of IGLS in a similar way as REML is a restricted maximum likelihood procedure, with RIGLS less biased especially in linear models, as mentioned before. In logistic random effects models, IGLS was chosen for MLwiN ([R]IGLS) in our study as all other frequentist packages allow for the ML method but not all allow for REML estimation. An additional MLwiN analysis using RIGLS did show somewhat different results. For case 1, the results from RIGLS and IGLS were basically identical, only the variance estimator was 10% higher with RIGLS. For case 3, the regression estimates differed more and the RIGLS estimator of variance was not zero anymore. For more information on ML, REML, etc in different multilevel models, see Browne and Draper (2006).
WinBUGS demonstrates much flexibility. Different distributions for the random effects (e.g. gamma, uniform, t-distribution) and different link functions such as probit and log(-log) model are possible. Different link functions are also possible in the SAS procedures GLIMMIX and NLMIXED, but none of these two packages allow for other than normal distributions for the random effects. Note that in our study the binary logistic random effects model

was superior to the probit and log(-log) models according to Akaike Information Criterion (using GLIMMIX).

3.4.2 Problems with small data sets

When the data set is small and the variance of the random effects is near zero, or the ICC (intra-class correlation) is very small as in cases 3 and 6, both frequentist and Bayesian methods can give quite different estimates especially for the variance. The MLE approach might have difficulties estimating small but non-zero variance estimates. The variance was estimated zero with lme4 in R. GLLAMM also estimated the variance as well as its standard error as quite small. GLIMMIX and NLMIXED produced very small estimates for the variance but no output for the standard error. MLwiN ([R]IGLS) estimated the variance and the standard error as zero. Finally, MIXOR gave no output for either the variance or the standard error.
For the Bayesian methods, the posterior means depended much on the choice of prior for the variance component. In order to check their impact, we offered WinBUGS the following three priors for the standard deviation of the random effects: uniform (0,1), uniform (0,10), uniform (0,100), and a uniform (0, 10^6) as well as an inverse Gamma distribution (0.001,0.001) for the variance. We also offered two priors for the variance in MLwiN (MCMC): inverse Gamma distribution (0.001,0.001) and uniform (0,). This uniform distribution is actually an improper prior which might lead to an improper posterior. Further, it is not the default choice in MLwiN. However, when the default procedure was taken for the improper uniform prior, i.e. starting values are taken from an initial IGLS run, the starting value for the variance parameter was taken too small and remained so until the MCMC sampling was stopped thereby affecting severely all parameters. For this reason we restarted this MLwiN run with 1 as the starting value which solved this problem. The total number of iterations was 1,100,000 with a burn-in of 100,000 iterations with thinning applied every 10 iterations. Convergence was checked and obtained using the criteria offered in each software package. The results are shown in Table 3.8. We can see that most of the Bayesian estimates are larger than the frequentist MLE from Table 3.5, especially for the variance parameter. The reason is that the posterior distribution is highly skewed for the variance therefore the posterior mean is much larger than the posterior mode whose frequentist counterpart is the MLE. We also notice in Table 3.8 that the variance estimates in MLwiN and WinBUGS using a uniform prior on the variance are greater than the WinBUGS results with uniform priors on the standard deviation, which was mentioned by Gelman (2006). To conclude, for small data sets the choice of the prior matters for the posterior estimates of the parameters, as was also shown by e.g. Spiegelhalter et al. (2004).

Table 3.8: Impact of variance component priors on the posterior means in WinBUGS and MLwiN in case 3 (sample 2). The variance of the random effects with its standard error is given. The first three uniform distributions in WinBUGS are for the standard deviation of the random effects and the rest four prior distributions are for the variance of the random effect

		WinBUGS										MLwiN (MCMC)			
Distribution		Uniform(0,1)		Uniform(0,10)		Uniform(0,100)		Uniform(0,10^6)		IG (0.001,0.001)		IG (0.001,0.001)		Uniform(0,infinity)	
Random Effects		Variance: 0.489(0.312)		Variance: 26.450(18.250)		Variance: 28.040(22.130)		Variance: 36.950(29.120)		Variance: 20.310(17.570))		Variance: 19.892(17.236)		Variance: 36.954(28.406)	
	covar	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Fixed Effects	const	-0.599	0.695	-1.720	1.610	-1.725	1.644	-1.942	1.830	-1.463	1.482	-1.448	1.473	-1.970	1.833
	pupil2	0.435	0.510	1.927	1.161	1.946	1.184	2.177	1.267	1.662	1.121	1.647	1.109	2.198	1.270
	pupil3	1.384	0.461	3.098	1.101	3.141	1.148	3.455	1.236	2.789	1.093	2.774	1.085	3.457	1.224
	age	0.869	0.199	2.160	0.679	2.195	0.719	2.428	0.781	1.928	0.686	1.913	0.675	2.437	0.777
	motor2	1.484	0.737	3.897	1.799	3.933	1.851	4.384	2.008	3.436	1.728	3.400	1.713	4.396	2.017
	motor3	0.319	0.630	1.557	1.310	1.571	1.332	1.777	1.433	1.328	1.243	1.304	1.231	1.785	1.431
	motor4	-0.838	0.602	-1.526	1.121	-1.553	1.128	-1.627	1.205	-1.445	1.052	-1.434	1.062	-1.637	1.205
	motor5	-1.953	0.594	-3.685	1.277	-3.742	1.312	-4.026	1.407	-3.406	1.237	-3.382	1.243	-4.032	1.400
	motor6	-1.473	0.986	-0.519	1.931	-0.484	1.978	-0.158	2.143	-0.784	1.795	-0.800	1.790	-0.161	2.131
	motor9	-1.561	0.868	-0.974	1.404	-0.969	1.415	-0.852	1.495	-1.099	1.328	-1.107	1.333	-0.851	1.486
	trial2	0.391	0.659	-0.846	1.823	-0.900	1.883	-1.248	2.120	-0.612	1.672	-0.600	1.650	-1.212	2.115
	trial3	1.131	0.888	1.362	2.033	1.381	2.091	1.407	2.299	1.298	1.854	1.292	1.847	1.415	2.321
	trial4	-0.682	0.781	-2.034	1.860	-2.070	1.898	-2.312	2.080	-1.809	1.722	-1.813	1.715	-2.286	2.085
	trial5	0.639	0.667	1.913	1.792	1.936	1.845	2.219	2.100	1.614	1.636	1.598	1.601	2.242	2.078
	trial6	1.318	0.975	2.640	2.019	2.633	2.053	2.781	2.193	2.428	1.893	2.398	1.876	2.786	2.200
	trial7	1.889	0.822	1.663	1.906	1.657	1.919	1.603	2.087	1.696	1.753	1.678	1.744	1.631	2.072
	trial8	0.698	0.818	-0.469	2.338	-0.515	2.412	-0.869	2.692	-0.242	2.142	-0.226	2.135	-0.811	2.679
	trial9	2.084	1.190	4.445	3.357	4.520	3.447	4.956	3.877	3.994	3.052	3.968	3.005	4.988	3.888
	trial10	0.744	0.742	0.742	1.560	0.727	1.579	0.665	1.693	0.742	1.439	0.748	1.425	0.670	1.693
	trial11	0.311	0.747	0.422	1.394	0.412	1.405	0.417	1.496	0.389	1.298	0.382	1.296	0.421	1.496

3.4.3 Comparison with previous studies

Zhou et al (1999) compared 5 packages for generalized linear multilevel models. They compared the estimates, the computing time and the features of the packages. In our study, we compared 10 (popular) packages on similar features. Also Bayesian methods were included. Guo and Zhao (2000) compared statistical software for multilevel modelling of binary data, and they put much emphasis on PQL and MQL. Furthermore, the SAS macro GLIMMIX as well as MLn, the DOS predecessor of MLwiN, were included in their comparison. The latter packages are not in use anymore which makes this comparison now outdated.
The CMM website published an online report (multilevel modelling software reviews) which compared almost 20 packages for the normal linear model, the binary response model, the ordered category model and the cross-classified model (The center for multilevel modeling (CMM) website, nd). But the packages lme4, MCMCglmm and the SAS procedures GLIMMIX and MCMC were not considered in this review. In addition, we evaluated here also the usability and flexibility of the packages.

3.5 Conclusions

We conclude from our study that for relatively large data sets, the parameter estimates from logistic random effects regression models will probably not be much influenced by the choice of the statistical package. In that case the choice of the statistical implementation should depend on other factors, such as speed and desired flexibility. Based on our study, we conclude that if there is no prior acquaintance with a certain package and preference is given to a frequentist approach, the following packages are to be recommended: MLwiN ([R]IGLS), the R package lme4 and the SAS procedure GLIMMIX. For a Bayesian implementation, we would recommend MLwiN (MCMC) because of its efficiency. If the user is also interested in (perhaps more complicated) statistical analyses other than multilevel modelling then he/she could choose WinBUGS.
Finally, a cautionary remark is necessary, i.e. a "large data set" can still be sparse and hence "large" should be interpreted with some caution. For example, a large data set with many lowest-level units nested within nearly as many higher-level units will act as a "small" data set when a multilevel model is fit. For such data sets the result of the fitting exercise might very much depend on the chosen approach: frequentist or Bayesian. In case a Bayesian package is chosen, the parameter estimates might be much influenced by the priors for the variance of the random effects. Since some packages offer only a quite restricted set of priors (such as MLwiN) for this parameter, the choice of the Bayesian package may have a large impact on the posterior estimates of all parameters for "small" data sets. Finally, also the performance of a Bayesian analysis might very much depend on the chosen starting value for the variance parameter, e.g. when chosen (close to) zero the MCMC might be stuck around zero for a very long time (which happened with MLwiN) and thus affect severely the convergence of the Markov chain.

References

Austi, P. C., Alte, D. A., et al. (2003). Comparing hierarchical modeling with traditional logistic regression analysis among patients hospitalized with acute myocardial infarction: should we be analyzing cardiovascular outcomes data differently? *American Heart Journal*, 145(1):27–35.

Bates, D., Maechler, M., and Ben, B. (2009). *Package lme4*.

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

Browne, W. J. and Draper, D. (2006). A comparison of bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1(3):473–514.

Browne, W. J. and Rasbash, J. (2009). *MCMC estimation in MLwiN*.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Aanalysis*, 1(3):515–534.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 2 edition.

Goldstein, H. (1989). Restricted unbiased iterative generalized least-squares estimation. *Biometrika*, 76(3):622–623.

Goldstein, H. (2011). *Multilevel Statistical Models*, volume 922. Wiley.

Guo, G. and Zhao, H. (2000). Multilevel modeling for binary data. *Annual Review of Sociology*, 26:441–462.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software*, 33(2):1–22.

Hedeker, D. and Gibbons, R. D. (1996). MIXOR: A computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine*, 49(2):157–176.

Lesaffre, E. and Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random effects model: an example. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(3):325–335.

Lin, X. and Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91(435):1007–1016.

Maas, A. I., Marmarou, A., Murray, G. D., Teasdale, S. G. M., and Steyerberg, E. W. (2007). Prognosis and clinical trial design in traumatic brain injury: The impact study. *Journal of Neurotrauma*, 24(2):232–238.

Marmarou, A., Lu, J., Butcher, I., McHugh, G. S., Mushkudiani, N. A., Murray, G. D., Steyerberg, E. W., and Maas, A. I. (2007). IMPACT database of traumatic brain injury: Design and description. *Journal of Neurotrauma*, 24(2):239–250.

McHugh, G. S., Butcher, I., Steyerberg, E. W., Lu, J., Mushkudiani, N., Marmarou, A., Maas, A. I., and Murray, G. D. (2007). Statistical approaches to the univariate prognostic analysis of the IMPACT database on traumatic brain injury. *Journal of Neurotrauma*, 24(2):251–258.

Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer.

Ng, E. S., Carpenter, J. R., Goldstein, H., and Rasbash, J. (2006). Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood. *Statistical Modelling*, 6(1):23–42.

Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M., and Fisher, M. R. (1996). A survey of methods for analyzing clustered binary response data. *International Statistical Review/Revue Internationale de Statistique*, 64(1):89–118.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004). *GLLAMM manual*.

Rasbash, J. (n.d.). What are multilevel models and why should I use them? [http://www.cmm.bristol.ac.uk/learning-training/multilevel-models/what-why.shtml].

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I., and Lewis, T. (2000). *A User's Guide to MLwiN*.

Rodrıguez, G. and Elo, I. (2003). Intra-class correlation in random-effects models for binary data. *The Stata Journal*, 3(1):32–46.

Rodriguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(1):73–89.

Snijders, T. A. and Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage Publications Limited.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*. John Wiley & Sons.

Steyerberg, E. W., Mushkudiani, N., Perel, P., Butcher, I., Lu, J., McHugh, G. S., Murray, G. D., Marmarou, A., Roberts, I., Habbema, J. D. F., et al. (2008). Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLoS Medicine*, 5(8):e165.

The BUGS project (n.d.). Calling WinBUGS 1.4 from other programs. [http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/remote14.shtml].

The center for multilevel modeling (CMM) website (n.d.). Comparative timings. [http://www.cmm.bristol.ac.uk/learning-training/multilevel-m-software/tables.shtml].

The GLIMMIX procedure (2009). *SAS/STAT Users Guide. Version 9.2*.

The MCMC procedure (2009). *SAS/STAT Users Guide. Version 9.2.*

The NLMIXED procedure (2009). *SAS/STAT Users Guide. Version 9.2.*

Zhou, X.-H., Perkins, A. J., and Hui, S. L. (1999). Comparisons of software packages for generalized linear multilevel models. *The American Statistician*, 53(3):282–290.

4

A MULTI-COUNTRY PERSPECTIVE ON NURSES TASKS BELOW THEIR SKILL LEVEL: REPORTS FROM DOMESTICALLY TRAINED NURSES AND FOREIGN TRAINED NURSES FROM DEVELOPING COUNTRIES

Chapter 4 is based on the paper:

Bruyneel, L., Li, B., Squires, A., Aiken, L., Lesaffre, E., Van den Heede, K., and Sermeus, W. (2013). A multi-country perspective on nurses tasks below their skill level: Reports from domestically trained nurses and foreign trained nurses from developing countries. International Journal of Nursing Studies, 50(2):202-209.

Abstract

Several studies have concluded that the use of nurses time and energy is often not optimized. Given widespread migration of nurses from developing to developed countries, it is important for human resource planning to know whether nursing education in developing countries is associated with more exaggerated patterns of inefficiency. In this paper we aim to describe nurses reports on tasks below their skill level as well as to examine the association between nurses migratory status (domestically trained nurse or foreign trained nurse from a developing country) and reports on these tasks. The data set for the Registered Nurse Forecasting Study was used which was a cross-sectional study having 33,731 nurses in 486 hospitals in twelve European counties. Logistic random effects models were applied to estimate the effect of nurses migratory status on reports of the tasks below their skill level they performed. The findings suggest that there remains much room for improvement to optimize the use of nurses time and energy. Special attention should be given to raising the professional level of practice of foreign trained nurses from developing countries. Further research is needed to understand the influence of professional practice standards, skill levels of foreign trained nurses from developing countries and values attached to these tasks resulting from previous work experiences in their home countries. This will allow us to better understand the conditions under which foreign trained nurses from developing countries can optimally contribute to professional nursing practice in developed country contexts.

4.1 Background

System-level interventions like increasing nurse staffing and creating superior work environments have been associated with improved patient safety outcomes and a higher degree of nurse wellbeing (Aiken et al., 2012; Kelly et al., 2012). Also central to the efficient structuring of nurses work is optimizing the use of their time and effort. When asked about their last shift however, nurses across three countries (US, Canada, Germany) consistently reported high percentages of non-nursing tasks performed, including transporting of patients, delivering or retrieving of food trays, and performing of housekeeping activities. At the same time, they reported many nursing tasks that were necessary but left undone because they lacked the time to complete them (Aiken et al., 2001). Al-Kandari and Thomas (2009) used the list of non-nursing tasks from the study of Aiken and colleagues among 780 Kuwaiti nurses. Increased non-nursing task workload was positively correlated to incompletion of nursing activities. Two recent time-and-motion studies found that nurses spent considerable amounts of time in non-nursing activities. A 36-hospital time-and-motion study found that activities considered by nurses to be waste (waiting, looking, retrieving, and delivering) consumed 6.6% of reported time per 10-h shift (Hendrich et al., 2008). Another time-and-motion study showed that nurses spent 9.0% of their time during their last shift on non-nursing tasks, including replenishing charts and forms, tidying up rooms, making beds, answering phones, searching for people, gathering linen, and answering call bells (Desjardins et al., 2008).

The employment of internationally trained nurses may suggest a shortage of nurses at the institution or national level. Thus it is particularly important to optimize the full scope of professional nursing practice in institutions that employ nurses educated in other countries. Studies have shown that migrant nurses sometimes experience discrimination by means of lower wage and less upward mobility, and may be employed as nursing aids rather than as nurses, which negatively impacts their wellbeing (Kline, 2003; Center for Health Workforce Studies, 2008; Organisation for Economic Cooperation and Development, 2010). Other research suggests that nurses trained abroad aspire to the same professional nursing practice standards common to their country of current employment (Flynn and Aiken, 2002). In light of the increasing international mobility of nurses, Humphries et al. (2009) finds the evaluation of how migrant nurses skills are utilized a prerequisite to incorporating nurse migration into workforce planning.

The twelve-country Registered Nurse Forecasting (RN4CAST) study measured and linked organizational features of nurses work places to nurse wellbeing and patient outcomes to challenge assumptions underpinning previous nurse workforce planning efforts (Sermeus et al., 2011). The aim of this study is to determine whether there is a difference between domestically trained and foreign trained nurses from developing countries in nurses reports on tasks below their skill level performed during their last shift. We consider the implications of our findings for human resources management.

4.2 Methods

Study design

The RN4CAST study favoured a rigorous quantitative multi-country cross-sectional design on the basis of research methods used in a five-nation study of critical issues in nurse staffing and the impact on patient care (Aiken et al., 2001). Data were gathered via four data sources (nurse, patient and hospital profile surveys and routinely collected hospital discharge data). The design of the RN4CAST-study is described in detail by Sermeus et al. (2011). For this analysis, nurse-reported information on migratory status and tasks below skill level performed during their last shift was used.

Ethical approval

Depending on national legislation, the study protocol was approved by either central ethical committees (e.g. nation or university) or local ethical committees (e.g. hospitals). All nurses received an information letter explaining the design of the study.

Study sample

A total of 486 hospitals were sampled as primary sampling units in twelve European countries, with at least 30 hospitals per country. In two countries, the selected hospitals represent all of the relevant institutions in the country (Ireland, Norway). In Belgium, Germany, the Netherlands, Switzerland, England and Spain, a stratified random selection (geographical location within the countries, hospital size, and hospital type) was done. Additionally, the Belgian and German research teams also gave the opportunity for hospitals to participate on a voluntary basis. In Finland, Poland and Greece, hospitals were selected via purposive sampling (i.e. geographical spread, hospital size, hospital type). In Sweden, all nurses were approached via the Swedish Nursing Association, which covers about 85% of all nurses working in Sweden. Nurses were then asked to identify the hospital in which they work.

In each of the selected hospitals at least two general medical and surgical nursing units were randomly selected from a master list of nursing units. All staff nurses involved in direct patient care activities served as informants on organization of nursing care, nurse wellbeing, patient safety and quality of care. Nurses were defined in each country as those meeting the European Union definition of trained and licensed nurses according to directive 2005/35/EC. The sample consists of 33,731 nurses (62% response rate) from Belgium (n = 3186), England (n = 2990), Finland (n = 1131), Germany (n = 1508), Greece (n = 367), Ireland (n = 1406), the Netherlands (n = 2217), Norway (n = 3752), Poland (n = 2605), Spain (n = 2804), Sweden (n = 10,133), and Switzerland (n = 1632).

Study measures

A self-administered questionnaire was distributed. Nurses were asked to indicate whether they had received their training in the country they are currently working in and if not, in

which country they did receive their training. Based on the World Economic Outlook classification of countries (International Monetary Fund, 2010), nurses were categorized into domestically trained, foreign trained in a country with an emerging or developing economy (further referred to as foreign trained in a developing country), or foreign trained in a country with an advanced economy (further referred to as foreign trained in a developed country). The IMF list of emerging and developing economies (150 out of 184 countries) includes countries from all over the world. Some recent entrants to the European Union for example have remained classified as emerging economies (e.g. Latvia, Poland).

Within a series of questions about their last shift, nurses were asked to report on a list of tasks below their skill level whether they had performed these tasks never, sometimes, or often during their last shift. The following nine tasks were presented to nurses: routine phlebotomy/blood draw for tests, transporting of patients within hospitals, performing non-nursing care, performing non-nursing services not available on off-hours, delivering and retrieving food trays, answering phones/clerical duties, arranging discharge referrals and transportation, obtaining supplies or equipment, and cleaning patient rooms and equipment.

Three types of variables were used to control for confounders: nurses type of last shift worked (morning, evening, night), number of years worked as a nurse and level of education (bachelor degree or not).

Statistical analysis

We first described for each country the share of foreign trained nurses and the share of nurses from developing and developed countries. We provide detailed data on the country of origin. We also assess whether there are statistically significant differences between domestically trained nurses and foreign trained nurses from developing countries in reporting type of last shift worked, number of years worked, and level of education. Second, we described nurses reports on the list of nine tasks performed during their last shift. Third, we compared reports on tasks performed by domestically trained nurses and foreign trained nurses from developing countries. For analytic purposes, we dichotomized nurses responses as never performed and sometimes/often performed. A heat map (Sneath, 1957) was used to graphically compare these reports, with a system of colour-coding where a dark grey square indicates that a higher proportion of foreign trained nurses from developing countries reported this task compared to domestically trained nurses (and light grey square vice versa). A composite measure of tasks performed during nurses last shift (min = 0, max = 9) was calculated for each individual nurse by taking the sum of the nine dichotomized nursing tasks. This composite measure had a binomial distribution. The overall effect (i.e. over all countries) of nurses migratory status on this composite measure was estimated using a two-level logistic random effects regression. The country effect was modelled as a fixed effect. The hospitals were treated as a random effect. We calculated the intraclass correlation coefficient at the hospital level as an indication of the degree of homogeneity. The analysis was adjusted for nurses type of last shift worked, number of years worked as a nurse and level of education. We analyzed the consistency of the overall effect by speci-

fying interaction effects between the countries under study and migratory status. We also constructed a series of similar two-level random effects regression models to analyze the overall effect of migratory status on each task separately. Despite all efforts to get random effects models with interaction effects to converge, this proved to be hard for four out of nine tasks because of computational problems. Descriptive findings for these tasks showed repetitive high proportions of both domestically trained nurses and foreign trained nurses from developing countries indicating they had performed these tasks during their last shift. We repeated our analysis comparing nurses reports on tasks never/sometimes performed and often performed, and found similar findings. We also compared the difference in tasks reported between domestically trained nurses and foreign trained nurses from a developed country and found no statistically significant differences. The data analysis for this paper was generated using SAS/STAT software, Version 9.3 of the SAS System for Windows. Copyright 2011 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

4.3 Findings

Foreign trained nurses

2107 nurses (6.2% of total sample) indicated they were trained in another country than where they were currently employed, of which 832 were trained in a developing country (2.5% of total sample). There was large variation in the share of foreign trained nurses between countries: Ireland (38.6%), Switzerland (22.1%) and England (16.7%), Norway (5.5%), Germany (5.1%), Greece (5.1%), Belgium (3.1%), Netherlands (2.4%), Sweden (2.3%), Spain (1.3%), Finland (.9%). In Poland, all nurses that participated in the study were domestically trained nurses and in Greece there were no foreign trained nurses from developing countries. The share of foreign trained nurses varied considerably between hospitals in the top three countries with foreign trained nurses, ranging from 16% to 56% (Ireland), 4% to 50% (Switzerland) and 1% to 52% (England). Countries with low numbers of foreign trained nurses from developing countries (Finland, Greece, Poland) or high missing values on country of training (Belgium) were dropped for further analysis, which resulted in a total of 813 foreign trained nurses from developing countries remaining for further analysis. Figure 4.1 presents the large variation in the share of nurses from developing countries employed in the sample of eight remaining European countries. While in Switzerland only 11% of the share of foreign trained nurses were trained in developing countries, this ran up as high as 80% in England. In many countries, a large part of the share of foreign trained nurses could be explained by mobility between neighboring countries or countries in the region. 31.6% (n = 112) of the foreign-trained nurses in Switzerland were trained in Germany, 30% (n = 107) were trained in France, and 12% (n = 41) were trained in Italy. Nurses trained in developing countries now working in Switzerland included nurses from India (2.0%, n = 7), Bahrain (1.4%, n = 4), the Philippines (1.1%, n = 3) and Bosnia and Herzegovina (.85%, n = 3), among others. In Sweden 26.8% (n = 62) of the foreign trained nurses had obtained their training in Finland, and 11.7% (n = 27) in Germany. The share of foreign trained nurses from

developing countries was ethnically very diverse, with most nurses trained in Bosnia and Herzegovina (6.5%, n = 15). In Spain a different image emerged, with a large share of nurses trained in South- American countries, mainly in Peru (21.6%, n = 8). Norways largest share of the foreign trained nurse workforce was composed of 26.1% (n = 53) nurses trained in Sweden, 15.3% (n = 31) in Australia, and 15.3% (n = 31) in Denmark. Like in Switzerland, nurses from developing countries came from the Philippines (2.5%, n = 5), Bahrain (1.5%, n = 3) and Bosnia and Herzegovina (1.5%, n = 3). In the Netherlands, next to Belgian (12.9%, n = 7) and German (9.3%, n = 5) nurses, there was a substantial percentage of nurses from the former Dutch colonies of Suriname (22.2%, n = 12) and Indonesia (14.1%, n = 8). Polish-trained nurses accounted for 17.1% (n = 13) of the German foreign trained nurse workforce, and Kazakh nurses accounted for 5.3% (n = 4). In England, the main source countries were the Philippines (31.0%, n = 153) and India (22.7%, n = 117) but also nurses from sub-Saharan Africa (Ghana, Kenia, Nigeria, South Africa, Uganda, Zambia and Zimbabwe) accounted for a large proportion (15.9%, n = 78). Like in England, the use of overseas recruiters is widespread in Ireland. Contrary to England however, Irelands share of nurses from developing countries was almost completely accounted for by nurses from India (20.9%, n = 111) and the Philippines (17.3%, n = 92) only. The share of European foreign trained nurses in Ireland (54.5% of total) was almost exclusively made up of nurses having received their training in the UK (51.5% of total).

In all eight countries, foreign trained nurses from developing countries had more years of experience in working as a nurse. These differences were statistically significant across all countries. Statistically significant differences were found for the level of education in England and Ireland, where the share of foreign trained nurses from developing countries reporting they had obtained a bachelor level degree in their home country was higher than the share of domestically trained nurses.

Nurses reports on tasks performed during their last shift

Across countries, a high proportion of nurses reported having sometimes or often performed tasks below their skill level during their last shift. Most reported tasks (country-weighted average) were answering phones/clerical duties (97.4%), performing non-nursing care (90.1%), and obtaining supplies or equipment (71.2%). There was large variability between countries in nurses reports. For example in Spain, only 16.8% reported having cleaned patient rooms and equipment, while in England this was 90% (see Table 4.1).

Comparison of reports from domestically trained nurses and nurses trained in developing countries

The heat map shows that in 62 out of 72 cases, higher percentages of nurses from developing countries reported they performed these nine tasks, compared to domestically trained nurses (Table 4.1 for detailed findings). Findings were consistent between hospitals and in the case of nurses from the same developing country working in different countries under study here. For example, 25 English trusts had a total of 153 Philippines employed. In 24

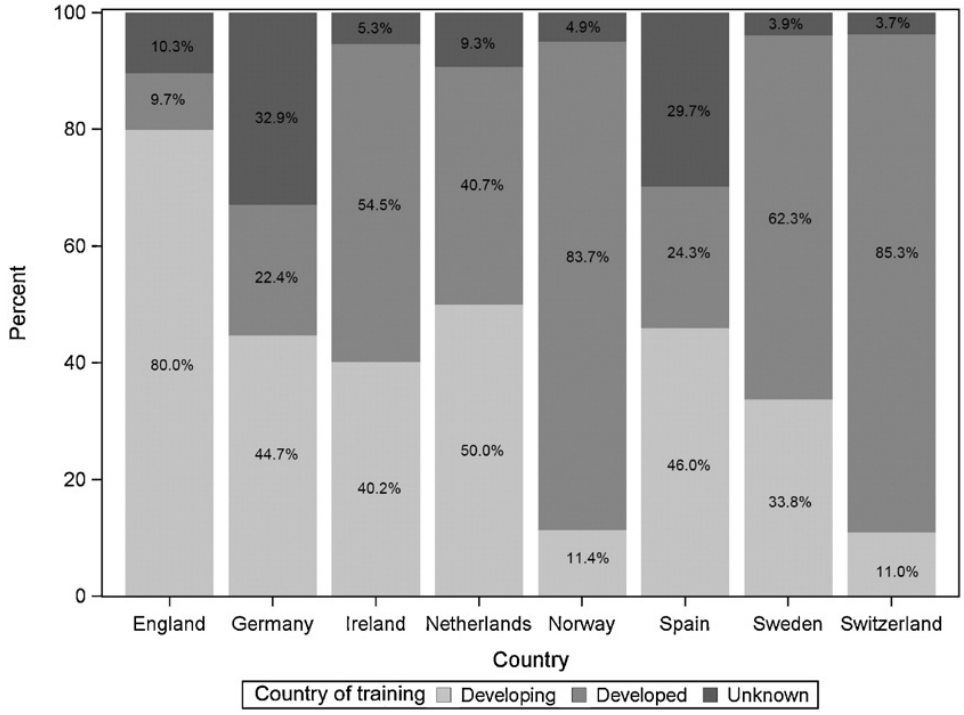

Figure 4.1: Foreign trained nurses: share of foreign trained nurses from developing and developed countries

Table 4.1: Nurses reports of tasks below their skill level performed during their last shift[a]: overall percentages and percentages by migratory status (trained in a developing country (De) versus domestically trained (Do)[b]

		Delivering and retrieving food trays	Performing non-nursing care	Arranging discharge referrals	Routine phlebotomy/blood draw for tests	Transport of patients within the hospital	Cleaning patient rooms and equipment	Filling in for non-nursing services not available on off-hours	Obtaining supplies or equipment	Answering phones, clerical duties
	All (3038)	83.8	96.8	76.9	85.8	69.9	82.6	47.6	71.6	97.9
BE	Do (3021)	83.7	96.9	76.9	85.8	69.9	82.7	47.5	71.5	97.9
	De (17)	100	94.1	76.5	88.2	64.7	70.6	58.8	88.2	100
	All (1274)	76.7	97.2	59.8	74.3	59.3	57	58.6	65.8	96.9
CH	Do (1246)	76.6	97.3	59.7	74.1	58.9	57.2	58.6	65.5	96.9
	De (28)	78.6	92.9	64.3	85.2	77.8	46.4	59.3	80	96.3
	All (1448)	82.4	98	74.1	41.7	70.9	63.9	65.8	85.4	98.7
DE	Do (1414)	82.3	97.9	74	41.5	71.2	63.8	65.7	85.1	98.6
	De (34)	85.3	100	76.5	50	55.9	67.6	70.6	97.1	100
	All (2746)	44.1	91	57.7	86.2	45.2	16.8	22.5	72.8	98.5
ES	Do (2729)	44	91	57.5	86.1	45.2	16.7	22.5	72.7	98.4
	De (17)	68.8	88.2	76.5	94.1	52.9	37.5	23.5	87.5	100
	All (1070)	63.3	87.2	41.6	12.8	31.7	56.4	72.1	38.6	97.8
FI	Do (1068)	63.3	87.2	41.6	12.8	31.5	56.5	72.1	38.6	97.8
	De (2)	50	100	0	0	100	0	50	50	100
GR	All (335*)	37.7	77.2	79.4	93.8	63.6	64.9	65.5	86.2	94.8
	All (1061)	64.2	95.2	80.7	28.5	67.5	81.6	69.2	84.3	99.1
IE	Do (847)	58.7	94.9	79	26.3	64.1	78.9	70	85.1	99.1
	De (214)	86.3	96.7	87.5	37.1	81	92.8	65.9	81.1	99.5
	All (2180)	57.4	93.2	76.2	29.8	68.5	63.2	40.7	49.8	98.7
NL	Do (2153)	57.1	93.1	76	29.7	68.3	63	40.5	49.3	98.7
	De (27)	85.2	96.3	85.2	42.3	81.5	77.8	59.3	84.6	96.3
	All (3516)	78.8	71.5	64.3	39.9	42.6	68.1	59.1	63.9	97.4
NO	Do (3493)	78.8	71.5	64.2	39.7	42.6	68.1	59	63.7	97.4
	De (23)	82.6	78.3	69.6	69.6	34.8	73.9	65.2	87	100
PL	All (2593*)	75	94.1	59	97.7	90.4	85.7	62.6	70	97.9
	All (9913)	64.5	84.2	58.1	79.5	53.4	69.5	37.7	81.3	94.6
SE	Do (9837)	64.4	84.1	58	79.4	53.3	69.4	37.6	81.2	94.6
	De (76)	75.7	98.6	63	91.9	63	82.4	49.3	89.2	94.7
	All (2866)	66.7	96	83.1	54.4	60.7	90	63.2	85.5	99.7
UK	Do (2472)	63.6	96	82.1	51.8	57.9	89	62.1	85.2	99.7
	De (394)	86.5	95.6	89.1	71	78.3	96.4	70.3	87.8	99.7

[a]: Nurses responses were dichotomized as never performed and sometimes/often performed

[b]: Based on the World Economic Outlook classification of countries (International Monetary Fund).

out of 25 trusts, Philippine-trained nurses compared to domestically trained nurses more often reported they had delivered and retrieved food trays during their last shift. This was also the case in 19 out of 20 Irish hospitals where Philippine-trained nurses were working.

The intraclass correlation coefficient for the nine items varied from .08 to .35, and was .21 for the composite measure, justifying the need for specifying a multilevel model. Table 4.2 displays that after adjusting for last shift worked, years of experience, and level of education, there remained a pronounced overall effect of being a foreign trained nurse from a developing country and an increase in reports of tasks performed during the last shift. This overall effect was found for the model testing the association between nurses migratory status and the composite measure of tasks performed during the last shift. The interaction effect for this analysis was non-significant. The series of models to analyze the overall effect of migratory status on each task separately showed that for eight out of nine tasks there was an overall effect of being a foreign trained nurse from a developing country and an increase in reporting those tasks. Being a foreign trained nurse from a developing country was a significant predictor of all five tasks for which an interaction effect was specified. The interaction effect was non-significant for three tasks (arranging discharge referrals, routine phlebotomy/blood draw for tests, cleaning patient rooms and equipment). For delivering and retrieving food trays and obtaining supplies or equipment, the interaction effect was significant. For three out of four tasks for which no interaction effect could be specified, being a foreign trained nurse from a developing country was a significant predictor (performing non-nursing care, transport of patients within the hospital, filling in for non-nursing services not available on off-hours). Migratory status failed to predict the task of answering phones, clerical duties, for which in each country at least 90% of both domestically trained nurses and foreign trained nurses reported they had performed this task during their last shift (Figure 4.2).

4.4 Discussion

This study documented high proportions of nurses across twelve countries indicating they had performed tasks below their skill level during their last shift. These findings support the previous studies of Aiken et al. (2001), Desjardins et al. (2008); Hendrich et al. (2008) in which nurses reported much time spend on non-nursing tasks or much time wasted during their last shift.

Findings also revealed that, while a high share of all nurses reported having performed tasks below their skill level during their last shift, being a foreign trained nurse from a developing country was a significant predictor of performing tasks below skill level. These findings resulted from a two-level logistic random effects regression model testing the overall effect of migratory status on a composite measure of tasks performed, and persisted for a series of two-level random effects regression models to analyze the overall effect of migratory status on each task separately. The consistency in results across countries and hospitals makes these findings compelling.

In 2010, the World Health Assembly adopted the WHO Global Code of Practice on the

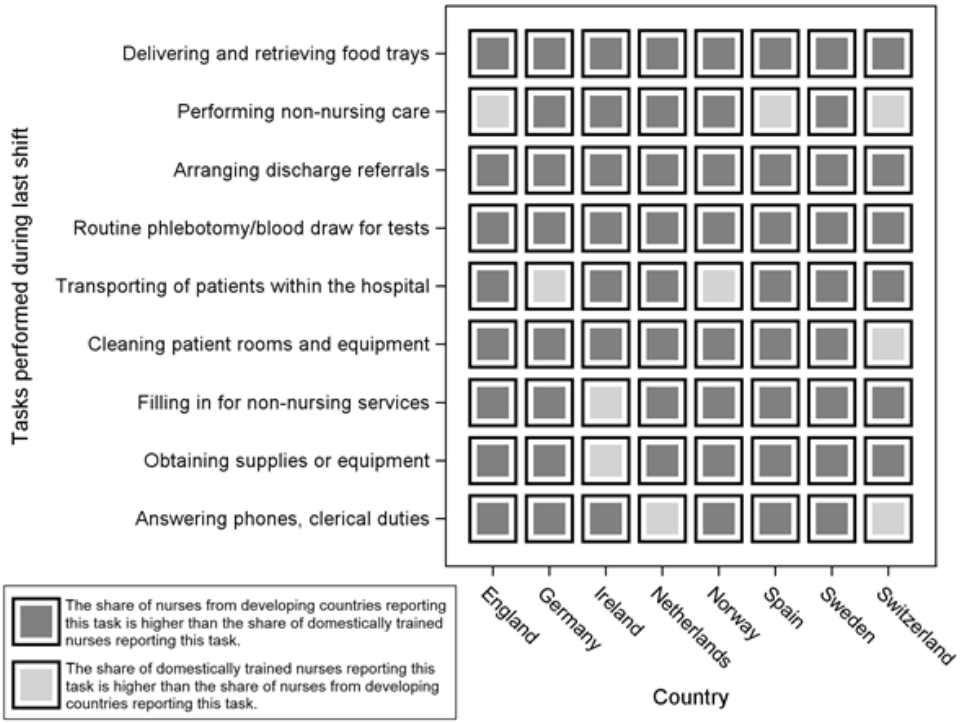

Figure 4.2: Nurses reports of tasks below their skill level performed during their last shift. Heat map comparing reports from domestically trained nurses and nurses trained in a developing country.

Table 4.2: Logistic random effects models[a] estimating the overall effect of nurses migratory status (trained in a developing country versus domestically trained[b] across eight countries[c] on task below skill level performed during nurses last shift

Tasks performed during nurses last shift	Estimate	Odds ratio (95% CI)	p-Value
Composite measure of nine nursing tasks	0.74	2.10 (1.68–2.61)	<0.0001
Delivering and retrieving food trays[d]	1.65	5.21 (4.04–6.72)	<0.0001
Performing non-nursing care[e]	0.53	1.70 (1.13–2.56)	0.014
Arranging discharge referrals[d]	0.89	2.44 (1.92–3.08)	<0.0001
Routine phlebotomy/blood draw for tests[d]	0.90	2.46 (1.91–3.17)	<0.0001
Transport of patients within the hospital[e]	0.73	2.08 (1.71–2.52)	<0.0001
Cleaning patient rooms and equipment[d]	0.64	1.90 (1.44–2.50)	<0.0001
Filling in for non-nursing services not available on off-hours[e]	0.19	1.21 (0.99–1.47)	0.048
Obtaining supplies or equipment[d]	0.30	1.35 (1.03–1.78)	0.033
Answering phones, clerical duties[e]	0.53	1.70 (0.70–4.10)	0.235

[a]: Adjusted for last shift worked (morning, evening, night as reported by the nurses), number of years worked as a nurse and degree obtained (bachelor degree or not)
[b]: Based on the World Economic Outlook classification of countries (International Monetary Fund)
[c]: England, Germany, Ireland, Netherlands, Norway, Spain, Sweden, Switzerland
[d]: Interaction effect specified
[e]: No interaction effect specified due to computational problems

International Recruitment of Health Personnel (World Health Organization, 2010). The ambition of this first code global in scope is for WHO Member States to refrain from the active recruitment of health personnel from developing countries facing critical shortages of health workers. The code also emphasizes the importance of equal treatment for migrant health workers and the domestically trained health workforce.

The RN4CAST data provided an opportunity to contribute to our understanding of this limited topic of research. The mix of countries participating in this study reflects the diversity of health systems in Europe, ensuring a rich perspective of nursing workforce issues from all angles. We used robust statistical techniques to analyze the differences among domestically trained nurses and foreign trained nurses from developing countries. The proportion of foreign trained nurses from developing or from developed countries corresponded closely to that observed by the OECD (Organisation for Economic Cooperation and Development, 2010).

Several limitations warrant consideration. First, our measure of migratory status may not have captured adequately the nationality of the nurse since only the country of training was known. Second, the proportion of foreign trained nurses to the total sample was rather small. From the twelve countries under study, we had to drop from our analysis three countries with low numbers of foreign trained nurses from developing countries (Finland, Greece, Poland) and one country with high missing values on the variable of country of training. Third, the list of items on tasks below nurses skill level was investigator-developed (Aiken et al., 2001). However, it has shown to have predictive validity to the incompletion

of nursing activities (Al-Kandari and Thomas, 2009). Also, other authors used similar items to describe waste (Hendrich et al., 2008) or non-nursing tasks (Desjardins et al., 2008). An early work measurement study from Connor (1961) already identified activities such as housekeeping and dietary tray delivery as non-nursing activities. It are exactly these tasks that could be delegated to non-nursing personnel (O'Brien-Pallas et al., 2004). It is however conceivable that some tasks surveyed here in certain situations of care were indeed nursing tasks. It is also possible that we have not captured all tasks below nurses skill level. Last, in this multi-country European context, consideration is warranted since the context in which nurses performed these tasks can be very diverse. The influence of professional practice standards, skill levels of foreign trained nurses from developing countries and values attached to these tasks resulting from previous work experiences in their home countries was unknown. We did not know for example whether foreign trained nurses from developing countries were more likely than domestic trained nurses to be assigned to perform tasks below their skill level or whether foreign trained nurses were more task oriented and brought the customs and roles of nursing from their developing country backgrounds into developed countries and are thus more prone to voluntarily take on tasks below their skill level. The differences we found between reports from domestically trained nurses and foreign trained nurses were, however, not attributable to lower level of education or less years of experience. To the contrary, in each country the foreign trained nurses from developing countries had significantly more experience in working as a nurse compared to domestically trained nurses. We did not know however how long they had been working as a nurse in their destination country. Lastly, future research should assess whether performing more tasks below the skill level is a barrier to providing good patient care or results in lower nurse wellbeing.

4.5 Conclusion

The findings suggest that there remains much room for improvement to optimize the use of nurses time and energy. Human resources management should give more attention to professional socialization and life-long learning for nurses to improve their priority setting and time management as well as ensuing that non-nursing resources are designated to carry out tasks that do not require the unique training of professional nurses. Nurses from developing countries may be particularly in need of continuing education on professional nurse roles and responsibilities in complex healthcare settings. Further research is needed to understand the influence of professional practice standards, skill levels of foreign trained nurses from developing countries and values attached to these tasks resulting from previous work experiences in their home countries. This will allow us to better understand the conditions under which foreign trained nurses from developing countries performed these tasks and to support improved structuring their work.

References

Aiken, L. H., Clarke, S. P., Sloane, D. M., Sochalski, J. A., Busse, R., Clarke, H., Giovannetti, P., Hunt, J., Rafferty, A. M., and Shamian, J. (2001). Nurses' reports on hospital care in five countries. *Health Affairs*, 20(3):43–53.

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: British Medical Journal*, 344:e1717.

Al-Kandari, F. and Thomas, D. (2009). Factors contributing to nursing task incompletion as perceived by nurses working in kuwait general hospitals. *J Clin Nurs*, 18(24):3430–3440.

Center for Health Workforce Studies (2008). The hospital nursing workforce in New York: Findings from a survey of hospital registered nurses.

Connor, R. J. (1961). A work sampling study of variations in nursing work load. *Hospitals*, 35:40–41.

Desjardins, F., Cardinal, L., Belzile, E., and McCusker, J. (2008). Reorganizing nursing work on surgical units: a time-and-motion study. *Nurs Leadersh (Tor Ont)*, 21(3):26–38.

Flynn, L. and Aiken, L. H. (2002). Does international nurse recruitment influence practice values in us hospitals? *Journal of Nursing Scholarship*, 34(1):67–73.

Hendrich, A., Chow, M. P., Skierczynski, B. A., and Lu, Z. (2008). A 36-hospital time and motion study: how do medical-surgical nurses spend their time? *The Permanente Journal*, 12(3):25–34.

Humphries, N., Brugha, R., and McGee, H. (2009). Career progression of migrant nurses in ireland: Nurse migration project policy brief 5. Technical report, Royal College of Surgeons in Ireland, Dublin.

Kelly, L. A., McHugh, M. D., and Aiken, L. H. (2012). Nurse outcomes in magnet and non-magnet hospitals. *J Nurs Adm*, 42(10 Suppl):S44–S49.

Kline, D. S. (2003). Push and pull factors in international nurse migration. *J Nurs Scholarsh*, 35(2):107–111.

O'Brien-Pallas, L., Thomson, D., Hall, M., Pink, G., Kerr, M., Wang, S., Li, X., and Meyer, R. (2004). *Evidence-based standards for measuring nurse staffing and performance*. Canadian Health Services Research Foundation (Fondation canadienne de la recherche sur les services de santé).

Organisation for Economic Cooperation and Development (2010). Policy Brief February 2010. International Migration of Health Workers. Improving International Co-operation to Address the Global Health Workforce Crisis.

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):6.

Sneath, P. H. (1957). The application of computers to taxonomy. *Journal of general microbiology*, 17(1):201–226.

World Health Organization (2010). The WHO Global Code of Practice on the International Recruitment of Health Personnel.

5

NURSING UNIT MANAGERS AND STAFF NURSES OPINIONS OF THE NURSING WORK ENVIRONMENT: A BAYESIAN MULTILEVEL MIMIC MODEL FOR CROSS-GROUP COMPARISONS

Chapter 5 is based on the paper:

Bruyneel, L., Li, B., Squires, A., Gilmartin, M., Spotbeen, S., Meuleman, B., Lesaffre, E., and Sermeus, W. (2014). Nursing unit managers' and staff nurses' opinions of the nursing work environment: a Bayesian multilevel mimic model for cross-group comparisons. Research in Nursing & Health (submitted)

Abstract

The objective of this study is to effectively compare nursing unit managers and staff nurses opinions of the nursing work environment (measured by the PES-NWI) by means of a state-of-the- art statistical approach. A Bayesian two-level MIMIC model is performed to evaluate measurement invariance as a logical prerequisite to addressing cross-group comparisons. Our findings provide evidence that the PES-NWI is to a great degree invariant in evaluating nursing work environment perceptions across nurse managers and staff nurses, and across language groups. Further, nursing unit managers evaluated certain important nurse work environment dimensions more positively when compared to their staff nurses. This might influence organizational change implementation. Implications for both analytic methods in cross-sectional organizational analyses and cross-cultural research are discussed.

5.1 Introduction

Health services researchers frequently use comparative surveys to detect and explain differences between (sub)groups of respondents. An important research domain where group comparisons are a core component is nursing outcomes research. This research studies how the organization of nursing care (workload, education, work environment) is associated with patient safety and nurse wellbeing (Lake, 2007). Studies on the quality of nurses work environment have recently received significant attention in the leading medical and health policy journals. The field defines the nursing work environment as not only the resources and physical environment where a nurse works, but also includes nurses in governance activities, quality of workplace relationships, and quality of nurse management. Many valuable insights about the organizational dynamics behind nursing role implementation in the acute care setting have resulted from research on nurses work environment. Good hospital work environments are associated with higher rates of nurse wellbeing (Aiken et al., 2012; Kelly et al., 2011), superior patient experiences with hospital care (Aiken et al., 2012; Kutney-Lee et al., 2009) and lower patient in-hospital mortality (Aiken et al., 2008; Friese et al., 2008). Moreover, a positive work environment is a prerequisite for benefiting from better nurse staffing (Aiken et al., 2011).

One of the critical factors in the nursing work environment that has been consistently identified in explaining improved nurse wellbeing, is that of quality of nurse management (Gunnarsdttir et al., 2009; Duffield et al., 2011; Kleinman, 2004). This evidence results from staff nurses perceptions of their managers. Other hospital organizational research studies used a different and perhaps more interesting study design by bringing together both frontline workers and managers views on the topic of interest. These studies suggest divergent views between hospital frontline workers and management on several features in the organization of hospital care. Price et al. (2007) showed that clinical nurses and their managers offered divergent views of deficiencies in quality improvement implementation. Kalisch and Lee (2012) found that nurse leaders reported more missed nursing care and higher teamwork, but less problems with having adequate material and labor resources than did nursing staff members. Studies including several types of hospital staff found that managers have a more positive perception of the patient safety climate (Singer et al., 2008) and perceive greater improvements in patient safety culture, but lower improvements on the timeliness of care delivery (Parand et al., 2011). Similar chasms in staff nurses and nursing unit managers perceptions of the nursing work environment have not been studied previously.

A main issue in comparative research is that the observed differences may be biased as a result of the measurement instrument not functioning invariantly across groups being compared (Vandenberg and Lance, 2000). Measurement invariance pertains to the extent to which respondents across groups perceive and interpret the content of items in the same way (Byrne and Watkins, 2003). Lack of measurement invariance is also referred to as differential item functioning (DIF). This could lead to serious erroneous inference and thus incorrect research findings and implications. Health services researchers increasingly advo-

cate for the incorporation of measurement invariance detection techniques in the analytic strategy (Cherepanov et al., 2011; Borsboom, 2006), yet such evaluations have been notably absent from the literature. Recent advancements in computational statistics for measurement invariance evaluation have greatly enhanced the feasibility of studies focusing on this important part of psychometric evaluation of measurement tools.

The aim of this study is to effectively evaluate cross-group differences in nursing unit managers and staff nurses opinions of the nursing work environment by means of a state-of-the-art statistical approach. Nurse managers have an immensely important role in taking immediate action to remedy work environment issues reported by staff nurses. For optimal patient outcomes, nurse managers and their staff nurses must, therefore, have concordant views about the quality of the nursing work environment. In line with the aforementioned research findings, in which managers generally tend to have more positive views on features of hospital care, we posit that nurse managers compared to staff nurses have more positive views of nurses work environment. When comparing groups, the assumption of measurement invariance must be tested in order to adequately capture the scope of these differences. No prior empirical work has tested the measurement invariance assumption among nurse managers and staff nurses ratings of the nursing work environment. We therefore had no a priori hypotheses regarding work environment items exhibiting DIF.

5.2 Method

Study setting

This study analyzes a convenience sample of one large Belgian hospital entity. Hospital management identified an opportunity for a hospital-wide assessment of nurses practice environment, staffing, productivity, and their wellbeing, using previously extensively validated instruments from the Registered Nurse Forecasting (RN4CAST) study, a large European nurse workforce study funded by the Seventh Framework Programme of the European Union (Sermeus et al., 2011). Dutch and French versions of the questionnaire, with translations previously validated (Squires et al., 2013), were distributed to nursing unit managers and staff nurses on 118 patient care nursing units, with exactly one nurse manager per unit. Eighty-seven out of 118 nurse managers completed the questionnaire (75.0% response rate). On these 87 nursing units, 821 out of 1159 staff nurses completed the questionnaire (70.8% response rate). The final data set thus consists of 908 observations. For the purpose of this study, a nursing unit manager is defined as a registered nurse who is largely responsible for administratively and clinically managing the nursing staff on a given nursing unit.

Study design

Nursing unit managers and staff nurses perceptions of the nursing work environment were measured using the practice environment scale of the nursing work index (PES-NWI) (Lake, 2002). This instrument was derived from a long line of successful research on nurses working conditions that began with research by Kramer and Hafner (1989) on magnet hospitals. Their initial instrument measured elements of nurse job satisfaction and nurse perceptions

of the quality of care. Aiken and Patrician (2000) further refined the instrument into the Revised Nursing Work Index (NWI-R), to determine the aspects of the professional work environment. The PES-NWI consists of 32 statements measured on a 4-point Likert scale anchored between 1=completely disagree to 4=completely agree. In the factor analytic stage of the development of the PES-NWI, Lake (2002) proposed five dimensions of the nursing work environment. These include collegial nurse-physician relations; nurse managerial abilities, leadership, and support of nurses; nurse participation in hospital affairs; nursing foundations for quality of care; and staffing and resource adequacy. The United States National Quality Forum endorses this instrument for use in nursing-sensitive care performance measurement and globally, it has shown good internal consistency and validity (Warshawsky and Havens, 2011).

Statistical methods

A two-level structural equation model is applied as the analytic strategy. The methods are explained within a methodological framework that is largely based on recent innovative statistical advancements by Muthén and Asparouhov (2012). These techniques may apply equally well to many other studies in the field of health services research and beyond.

We propose the following arguments for supporting our analytic approach. First, our instrument that measures the work environment, the PES-NWI, consists of various latent variables (nursing work environment dimensions) which should be accounted for in the statistical analysis. Although exploratory factor analytic techniques are often used to assess the psychometric properties of the PES-NWI (Warshawsky and Havens, 2011), the common practice is to calculate subscale scores by taking the average of the item scores. These newly created variables are subsequently used in regression analysis. Structural equation modeling provides an alternative approach. This model is able to test complete and simultaneous relationships between complex and multidimensional phenomena (Tabachnick and Fidell, 2012).

Second, the research design of most comparative, cross-sectional studies involves clustered data. Nurse workforce studies, for example, involve nurses clustered in various organizational levels, such as nursing units within hospitals. Ignoring a multilevel structure in the analysis can be problematic since the fundamental independence assumption underlying many commonly used statistical techniques is violated (Muthén, 1991). Gabriel et al. (2013); Li et al. (2013) recently conducted multilevel PES-NWI analyses, with two and four levels respectively, showing how different PES-NWI dimensions predict nurse wellbeing at different levels. The PES-NWI has also previously been analyzed in a two-level confirmatory factor analytic framework by Gajewski et al. (2010) and Gabriel et al. (2013).

To compensate for the aforementioned analytic issues, we handled the clustered and latent nature of our measures by estimating two-level factor models with nurses clustered in nursing units. In these models we accounted for the categorical nature of the data. First, to assess the multidimensionality of the PES-NWI, we adopted a two-level exploratory factor analytic model (EFA) with Geomin rotation. EFA is extremely useful to understand the factor structure and evaluate whether items are misbehaving. An exploratory analysis was

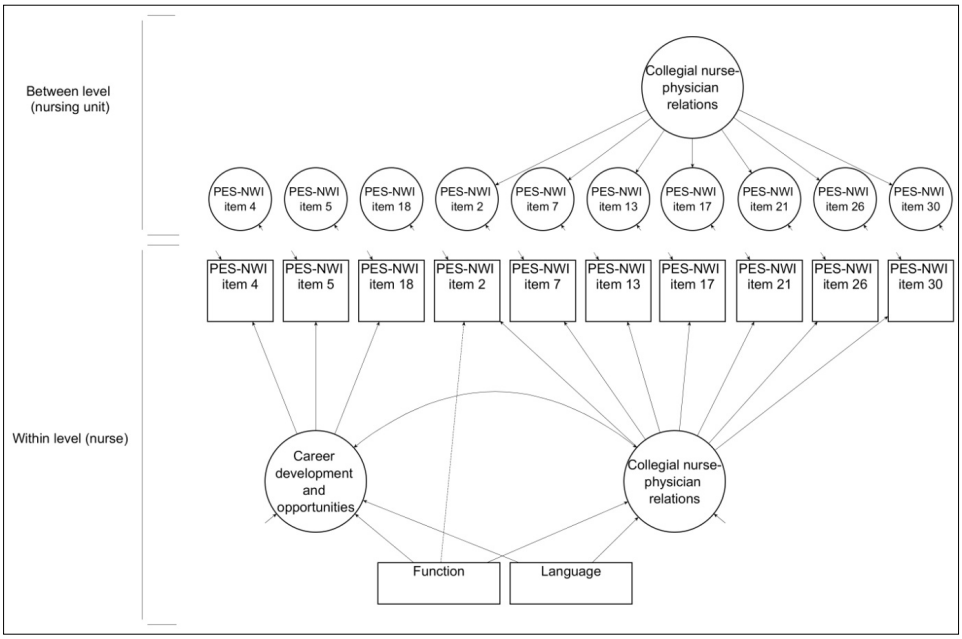

Figure 5.1: Two-level MIMIC model

first used since a previous Belgian study showed that the PES-NWI latent variables differed from Lakes solution (Van Bogaert et al., 2009). In the second step, we used the EFA findings to specify a two-level confirmatory factor analytic (CFA) model. By regressing the latent and observed variables on covariates, a multiple indicators multiple causes (MIMIC) model is established. This model allows cross-group comparisons while assessing measurement invariance with respect to subject grouping (Muthén, 1989). A hypothetical two-level MIMIC model with two latent variables and two covariates at the lowest level, one latent variable at the highest level, and one potential DIF effect, is presented in Figure 1. The considered MIMIC model is given by:

$$\begin{aligned} \boldsymbol{y}_{ij} &= \boldsymbol{B}\boldsymbol{x}_{ij} + \boldsymbol{L}_B\boldsymbol{\eta}_j + \boldsymbol{u}_j + \boldsymbol{L}_W\boldsymbol{\eta}_{ij} + \boldsymbol{\varepsilon}_{ij}, \\ \boldsymbol{\eta}_{ij} &= \boldsymbol{B}^*\boldsymbol{x}_{ij} + \boldsymbol{\delta}_{ij}, \\ bfeta_j \sim N(\mathbf{0}, \Sigma_B), \quad \boldsymbol{u}_j &\sim N(\mathbf{0}, diag(\sigma^2_{u1}, \sigma^2_{u2}, ..., \sigma^2_{uP},)), \\ \boldsymbol{\delta}_{ij} \sim N(\mathbf{0}, \Sigma_W), \quad \boldsymbol{\varepsilon}_{ij} &\sim N(\mathbf{0}, diag(\sigma^2_{\varepsilon 1}, \sigma^2_{\varepsilon 2}, ..., \sigma^2_{\varepsilon P},)) \end{aligned} \tag{5.1}$$

where $\boldsymbol{y}_{ij}$ represents a vector of P items coming from the ith nurse in the jth nursing unit. B is a $P \times Q$ matrix of the direct effects associated with the Q-dimensional vector of individual-level covariates $\boldsymbol{x}_{ij}$. $\boldsymbol{B}^*$ is a $m_W \times Q$ matrix of the indirect effects associated with the same Q-dimensional vector $\boldsymbol{x}_{ij}$ where m_W is the number of common factors at the nurse level. $\boldsymbol{\eta}_j$ is the m_B-dimensional nursing unit level common factor following a multivariate normal

distribution with mean zero and a covariate matrix Σ_B, and $\boldsymbol{L}_B$ is its $P \times m_B$ loading matrix with m_B the number of common factors at the nursing unit level. $\boldsymbol{\eta}_{ij}$ is the m_W-dimensional nurse level common factors with a multivariate normal distribution with mean zero and a covariance matrix Σ_W, and $\boldsymbol{L}_W$ is its loading matrix with a $P \times m_W$ dimension. $\boldsymbol{u}_j$ is the P-dimensional random intercept with independent normal distribution for each element with mean zero and a variance σ^2_{up}, while $\boldsymbol{\varepsilon}_{ij}$ is the P-dimensional residual, each element of which follows an independent normal distribution with mean zero and a variance $\sigma^2_{\varepsilon p}$.

From the MIMIC model, one can capture cross-group differences if there is a significant effect of a covariate on a latent variable, indicating that the factor means are different for different levels of the covariates. For this study, for example, an effect of the main covariate of interest, nurses function (nursing unit manager versus staff nurse) on any of the PES-NWI latent variables would indicate that the latent variable mean is different for nursing unit managers and staff nurses. DIF can be examined by estimating direct effects between a covariate and the observed variables. A covariate having a direct effect on an observed variable (over and above the indirect effect via the factors), indicates that the average response differs across respondents with different values for the covariate but the same score on the latent factor. In other words, direct effects represent violations of the assumption of measurement invariance [1] (Muthén, 1989). For this study, for example, a direct effect between the covariate on nurses function and one of the 32 PES-NWI items, over and above the indirect effect via the factors, would indicate that this particular item does not behave the same for nursing unit managers and staff nurses.

Previous research using the PES-NWI has not always adequately accounted for confounding variables. Jak et al. (2010) advise researchers investigating measurement bias to include as many possible violator variables as available. Possible confounders and respective coding included in the analysis are nurses language (0 =French, 1 =Dutch), degree (0 =diploma degree, 1 =bachelor degree), type of employment (0 =part-time, 1 =full-time), migratory status (0 =domestically trained, 1 =foreign trained) and gender (0 =female, 1 =male).

We adopted a Bayesian structural equation modeling (BSEM) approach to identify the model Muthén and Asparouhov (2012). This Bayesian approach was recently applied by Fong and Ho (2013) to evaluate the latent structure of a measurement instrument to assess and screen symptoms of anxiety and depression. In contrast to a frequentist approach, substantive information can be included in the model. Here, it is useful to include such information in the form of small-variance priors. Informative priors for the cross-loadings can restrict them stochastically, thereby overcoming possible identification problems. The final model is a Bayesian two-level MIMIC model with cross-loadings and direct effects from the covariates to the items with zero mean and small-variance priors, 0.05 and 0.1 respectively, based on prior factor analytic findings and drawing on Muthén and Asparouhov (2012)

[1] Such differences in conditional item means i.e. scalar non-invariance (Steenkamp and Baumgartner, 1998) represent only one possible form of measurement non-invariance, besides for example metric non-invariance (inequality of the strength of the relationships between indicators and latent factors Steenkamp and Baumgartner (1998)). Testing metric invariance within our MIMIC approach is less straightforward, and would imply including an impractically large number of interaction effects. Yet, scalar invariance implies stronger assumptions and thus a more stringent test than metric invariance, which justifies our approach (see also Muthén (1989))

simulation modeling approach. Muthén and Asparouhov (2013) described the importance of prior variance choosing in detecting non-invariance. A larger value for a prior variance may result in an increase in both the standard error of the parameter and the chances that the estimate can escape from the invariance value. Models with different values for the prior variance were therefore performed. Findings across models turned out to be very stable.

For both EFA and CFA entries the standardized model results are presented. The bolded entries presented in EFA and CFA findings are loadings that are largest for the item. An additional requirement for the entry to be bolded is that the loading must exceed a value of .3 and no severe cross-loadings can be present. Significant effects, meaning that the Bayesian credibility interval did not cover zero, are marked with an asterisk. Intra class correlations (ICC) are presented, which reflect the proportion of a single variables variance that can be accounted for by the "between" level. In addition we looked at the design effect, which is a function of the average cluster size and ICC, and is calculated as $\varphi = 1 + (\bar{n} - 1) * ICC$, where $\bar{n}$ equals the average group size. A design effect larger than two indicates that the clustering in the data needs to be taken into account during estimation Satorra and Muthen (1995). Evaluation of good EFA model fit between the hypothesized model and the observed data is based on Hu and Bentler (1999) suggestions to have cutoff values close to or above 0.95 for Tucker Lewis Index (TLI) and Comparative Fit Index (CFI), below or close to 0.08 for the standardized root mean square residual (SRMR), and below or close to 0.06 for the root mean square error of approximation (RMSEA). SAS Version 9.3 of the SAS System for Windows was used for descriptive analysis and to prepare data for further analyses. Mplus Version 7 was used for structural equation modeling (Muthén and Muthén, 2010).

5.3 Results

We begin this section with a description of the final sample size and follow with results specific to each type of analysis we conducted. The results begin with findings from the two-level exploratory factor analysis. Then, we describe the results from the MIMIC model.

Sample characteristics

The response rate at the unit level varied from 23.1% to 100%. Seventy-eight out of 87 nursing units had a response rate above 50%. The number of nurses per nursing unit varied from 3 to 27 (mean= 9.4, median= 9).

Table 5.1 displays the personal characteristics for nursing unit managers and staff nurses separately. The majority of nurses were female, spoke French, had obtained a bachelor degree, were domestically educated and worked full-time. A considerably higher proportion of nursing unit managers compared to staff nurses worked full-time and were domestically educated.

Two-level exploratory factor analysis

Table 5.2 displays the findings of the two-level exploratory factor analysis. Also included in this table are the factor structure of Lake (2002) as well as the ICC values. Twenty-one

Table 5.1: Sample characteristics

	Parameter	Nurse managers (n=87)	Staff nurses (n=901)
Language	French	75.86% (n=66)	86.36% (n=709)
	Dutch	24.14% (n=21)	13.64% (n=112)
	Missing	-	-
Degree	Diploma degree	5.75% (n=5)	18.76% (n=154)
	Bachelor degree	90.80% (n=79)	74.06% (n=608)
	Missing	3.45% (n=3)	7.19% (n=59)
Type of employment	Part-time	8.05% (n=7)	38.37% (n=315)
	Full-time	91.95% (n=80)	58.95% (n=484)
	Missing	-	2.68% (n=22)
Migratory status	Domestically trained	98.85% (=86)	85.38% (n=701)
	Foreign trained	1.15% (n=1)	10.23% (n=84)
	Missing	-	4.38% (n=36)
Gender	Female	89.66% (n=78)	83.43% (n=685)
	Male	9.20% (n=8)	13.89% (n=114)
	Missing	1.15% (n=1)	2.68% (n=22)

out of 32 items have a design effect larger than two, indicating the need for applying a two-level model. CFI (0.96) and TLI (0.95) indicated good model fit, as did RMSEA (0.022, 90% CI: 0.020-0.026) and SRMR (0.036), for a solution with six within-level factors and three between-level factors. This solution seemed logical with respect to its clear interpretability and parsimony. The within-level part of the model describes the factor structure for how the nurses PES-NWI item perceptions covary within nursing units. The six within-level factors are: career development and opportunities (3 items), collegial nurse-physician relations (7 items), nurse staffing (2 items), frontline nurse management (2 items), support for nurses (4 items), and nursing foundations for quality of care (4 items). Ten out of 32 items are not included in any within-level dimension. The between-level part of the model describes the factor structure for how the nursing unit PES-NWI item means covary. The between-level factors are: career development, opportunities, and support for nurses (11 items), collegial nurse-physician relations (7 items), and frontline nurse management, nurse quality foundations and staffing and resources adequacy (11 items). Thus, at the nurse level, several underlying dimensions are much more clearly manifested than at the nursing unit level, where more general factors appear, except for the factor of collegial nurse-physician relations which at both levels is identical.

5.3.1 MIMIC model

Results for the confirmatory factor analytic part of the MIMIC model (not shown) were close to the EFA findings. Inclusion of the covariates did not cause distortion in the solution,

Table 5.2: Parameter estimates for the two-level exploratory factor analysis

Item	Lake (2002)*	ICC	loadings (within)						loadings (between)		
			Fw1	Fw2	Fw3	Fw4	Fw5	Fw6	Fb1	Fb2	Fb3
1	STAF	0.045	0.027	0.058	0.191*	0.066	0.105*	0.013	0.013	-0.375	**0.673***
2	RNMD	0.194	0.016	**0.687***	-0.089	0.066	0.027	-0.029	0.027	**1.039***	-0.088
3	MANA	0.183	0.282*	0.012	-0.034	0.152*	**0.542***	-0.130	**0.847***	0.241	0.024
4	QUAL	0.168	0.394*	-0.052	-0.079*	0.169*	0.095	0.190*	**0.813***	0.150	0.057
5	PART	0.104	0.953*	-0.016	-0.009	-0.135*	0.000	0.011	**0.779***	0.105	0.225
6	PART	0.068	0.259*	0.109*	0.148*	0.038	0.335*	-0.027	**1.206***	0.371	-0.121
7	RNMD	0.147	0.086*	**0.695***	0.080*	0.001	0.014	-0.086*	0.013	**0.926***	0.145
8	STAF	0.124	0.144*	0.295*	0.323*	0.039	0.020	-0.006	-0.154	0.352	**0.741***
9	STAF	0.230	0.012	-0.062*	**0.903***	0.015	-0.025	0.037	0.006	0.198	**0.586***
10	MANA	0.262	0.014	-0.025	0.047*	**0.934***	-0.035	-0.009	-0.038	0.139	**0.745***
11	PART	0.180	0.112*	0.059	0.020	0.204*	**0.505***	-0.088	**0.932***	-0.047	-0.271
12	STAF	0.325	-0.161*	0.019	**0.709***	-0.024	0.277*	-0.014	0.102	0.278	**0.511***
13	RNMD	0.167	0.031	**0.845***	0.070*	0.034	0.015	-0.180*	-0.056	0.974*	0.041
14	MANA	0.241	0.194*	0.273*	0.183*	0.046	0.233*	0.035	0.343*	0.541*	0.397*
15	QUAL	0.080	0.171*	0.034	0.022	0.164*	0.050	0.194*	**0.680***	-0.205	0.048
16	PART	0.106	0.092	0.025	0.134*	0.081	0.203*	0.051	**0.807***	-0.068	-0.006
17	RNMD	0.214	0.064	**0.692***	0.032	-0.026	-0.042	0.037	-0.056	**0.979***	-0.031
18	PART	0.146	**0.690***	0.039	0.042	-0.005	0.032	0.103*	**0.874***	0.010	0.198
19	QUAL	0.108	0.205*	0.069*	0.118*	0.140*	0.114*	0.322*	0.313*	-0.014	**0.527***
20	QUAL	0.109	0.052	0.202*	0.177*	0.057	-0.029	0.277*	0.033	0.326	**0.579***
21	RNMD	0.223	-0.108*	**0.762***	-0.035	0.026	-0.005	0.125*	-0.028	**0.947***	0.182
22	MANA	0.260	-0.030	0.063	0.004	**0.659***	0.022	0.110*	-0.088	0.065	**0.775***
23	PART	0.119	0.016	-0.035	-0.020	0.162*	**0.740***	0.068	**0.996***	0.152	-0.109
24	QUAL	0.061	0.127*	0.038	0.092*	-0.033	0.363*	0.354*	**0.777***	-0.168	0.336
25	PART	0.034	-0.022	-0.007	0.042	-.113*	**0.647***	0.317*	0.540*	-0.230	0.545
26	RNMD	0.277	-0.036	**0.821***	-0.043	-0.045	-0.021	0.196*	-0.012	**0.948***	0.016
27	QUAL	0.216	0.027	0.018	0.087*	0.134*	0.182*	**0.410***	0.405*	-0.047	0.429*
28	QUAL	0.058	-0.041	0.096	-0.087	0.206*	0.081	0.379*	0.016	-0.002	**0.795***
29	PART	0.100	0.000	-0.044	-0.026	-0.087	0.511*	0.446*	**0.699***	0.168	0.044
30	RNMD	0.178	-0.059	**0.562***	0.-017	-0.065	0.248*	0.080*	0.007	**0.939***	0.210
31	QUAL	0.161	0.083	-0.005	0.027	0.107	-0.091	**0.509***	0.038	-0.030	**0.409***
32	QUAL	0.182	0.205*	0.100*	-0.021	0.023	-0.067	**0.455***	0.243	0.067	**0.372***

Correlations (within)							Correlations (between)			
	Fw1	Fw2	Fw3	Fw4	Fw5	Fw6		Fb1	Fb2	Fb3
Fw1	1						Fb1	1		
Fw2	0.348*	1					Fb2	-0.111	1	
Fw3	0.327*	0.309*	1				Fb3	0.388*	0.229*	1
Fw4	0.402*	0.363*	0.281*	1						
Fw5	0.493*	0.338*	0.346*	0.259*	1					
Fw6	0.327*	0.377*	0.190*	0.203*	0.266*	1				

RNMD: Collegial nurse-physician relations; **PART**: Nurse participation in hospital affairs; **MANA**: Nurse manager ability, leadership and support of nurses; **QUAL**: Nursing foundations for quality of care; **STAF**: Staffing and resource adequacy

Fw1: Career development and opportunities; **Fw2**: Collegial nurse-physician relations; **Fw3**: Nurse staffing; **Fw4**: Frontline nurse management; **Fw5**: Support for nurses; **Fw6**: Nursing foundations for quality of care; **Fb1**: Career development, opportunities, and support for nurses; **Fb2**: Collegial nurse-physician relations; **Fb3**: Frontline nurse management, nurse quality foundations, and staffing and resource adequacy

except for the third between-level factor. Here, items relating to quality of care, that were included from the EFA model, are no longer significant.

Table 3 displays the changes in the intercepts of the factors for different levels of the covariates (cross-group comparison). Nursing unit managers perceived three out of six latent nurse work environment variables significantly more positive than staff nurses. Also, compared to domestically educated nurses, foreign educated nurses held more positive views of career development opportunities.

Two items exhibited DIF across nurses and nursing unit managers. Five items exhibited DIF across Dutch and French speaking nurses, of which one item also exhibited DIF across domestically trained and foreign trained nurses. Finally, one item exhibited DIF across male and female respondents. These findings are more thoroughly discussed in the next section.

5.4 Discussion

Many substantive research questions imply a study of the extent to which measurement items exhibit DIF with respect to subject grouping. The large majority of organizational researchers in various research domains have so far largely undervalued this type of psychometric evaluation. An important intent of this study is a call for health services researchers to examine DIF in their evaluation of cross-group comparisons, rather than assuming measurement invariance. This study extends PES-NWI research by using the instrument among nurse managers for the first time and simultaneously attempting to ascertain measurement invariance of the instrument across nurse managers and staff nurses. Our analysis indicated that the PES-NWI is to a great degree an invariant measure for evaluating perceptions of the nursing work environment across chief and staff nurses, although some evidence of DIF was found. As derived by Sass (2011) from the work of Cheung and Rensvold (1999) and Millsap and Kwok (2004), there are three appropriate options when items exhibit DIF: (a) only use the items for which no evidence of DIF was found; (b) apply a partial measurement invariance model; (c) assume that for the items exhibiting DIF, the differences are too small to influence the results and proceed using all the items. Sass (2011) states that option c is feasible when the degree of DIF is minimal and the majority of items are invariant. As 30 out of 32 item for our main covariate of interest (nurses function) are invariant, we feel confident that our study indeed demonstrated that nursing unit managers at this facility, when compared to staff nurses, hold more positive views of career development and opportunities, collegial nurse-physician relations, and support for nurses. As suggested by Parand et al. (2011), such diverging views indicate the importance of consulting both frontline staff as well as managers in organizational decision-making.

The presence of DIF across language groups is of particular interest, given the extensive cross-cultural content validation our instruments underwent in the RN4CAST study, from which the questionnaires used in this study originated. In that study, the translation process standardized the interpretation of the items across all twelve participating countries (Squires et al., 2013). A translation manager was appointed to ascertain high standards of instrument translation that reduce item bias. Construct bias was further reduced by as-

Table 5.3: Bayesian two-level MIMIC: effects of covariates

	Function	Language	Degree	Working percentage	Migratory status	Gender
Factors	**Estimate**	**Estimate**	**Estimate**	**Estimate**	**Estimate**	**Estimate**
Fw1	0.136*	0.034	-0.058	0.023	0.102*	-0.061
Fw2	0.135*	-0.013	0.014	0.038	-0.012	-0.034
Fw3	0.082	0.032	-0.078	0.033	0.032	0.037
Fw4	0.088	0.045	-0.041	-0.037	-0.019	0.106
Fw5	0.216*	0.009	-0.024	0.015	0.098	-0.027
Fw6	0.106	-0.095	-0.052	0.078	0.042	-0.025
Item						
1	-0.015	-0.084*	0.022	0.028	-0.046	0.037
2	-0.039	0.002	0.000	-0.038	-0.064	-0.028
3	0.063	0.191*	0.007	0.000	-0.071*	0.018
4	0.021	-0.092*	0.050	0.005	0.001	-0.047
5	-0.002	-0.010	0.001	-0.020	0.007	0.002
6	0.019	-0.079*	-0.046	0.017	0.006	0.001
7	-0.011	0.028	0.034	-0.007	-0.043	0.046
8	-0.050	0.061	0.000	-0.035	-0.037	0.013
9	-0.022	-0.034	0.005	0.006	0.008	-0.024
10	-0.032	0.005	-0.015	-0.017	-0.008	0.001
11	0.009	-0.084	0.007	0.024	0.006	0.000
12	0.026	0.007	0.011	-0.008	0.003	0.006
13	0.025	0.043	-0.018	-0.002	-0.007	0.018
14	-0.006	0.017	-0.025	-0.009	0.031	0.015
15	0.084*	0.105*	0.026	-0.066	-0.035	0.030
16	0.041	0.046	0.008	-0.044	0.017	-0.094*
17	-0.035	-0.010	0.016	0.036	-0.022	0.017
18	0.002	0.033	-0.023	0.053	-0.007	-0.006
19	-0.003	-0.011	0.020	0.018	0.059	0.009
20	-0.006	-0.070	0.012	0.023	0.015	-0.014
21	0.005	-0.074	-0.014	0.001	0.011	-0.032
22	-0.003	-0.011	0.020	0.018	0.059	0.009
23	-0.006	-0.070	0.012	0.023	0.015	-0.014
24	-0.030	0.037	0.014	-0.036	0.070	0.022
25	-0.005	-0.023	-0.022	-0.040	0.013	0.032
26	0.014	0.004	0.022	0.017	0.041	-0.040
27	0.071	0.018	-0.052	0.031	0.015	0.056
28	0.053	0.023	-0.003	0.029	-0.070	-0.063
29	-0.093*	0.012	0.005	0.021	0.016	-0.058
30	0.009	-0.012	-0.037	-0.032	0.033	0.043
31	-0.012	-0.045	0.003	-0.060	-0.014	0.052
32	-0.041	0.001	0.049	0.066	-0.039	0.010

sessing dissimilarity of constructs in the investigated countries through the application of content validity indexing procedures. Method bias was decreased by a common sampling and survey administration protocol. It is, therefore, important to further examine which factors underlie these findings of DIF.

Limitations

There are a number of methodological challenges that could affect the interpretation of our findings. First, with regards to the interpretation of the exploratory factor analytic model, we used Hu and Bentler (1999) recommendations for cut-off values, however, there is still disagreement in setting acceptable levels of fit in goodness-of-fit indexes (Marsh et al., 2004). Another drawback related to model fit evaluation in this study, is that the posterior predictive p-value that is used in Bayesian analysis, is not yet implemented in Mplus for multilevel analyses.

Second, although our factor analytic models provided clearly interpretable factors, two of our factors consisted of two indicators only. To encompass the scope of the construct, Tabachnick and Fidell (2012) recommend at least three indicators per factor since the interpretation of the factor defined by less than three variables might be hazardous. It has however also been stated that two items per factor can be sufficient, and that this is widely applied (Kenny and McCoach, 2003). Important is that the number of included items encompass the scope of the construct (Bandalos and Finney, 2010). Although specifying only two variables per factor does not necessarily affect model identification, it might be preferable to have a larger number of variables per factor to better encompass the scope of the work environment construct. For the factor nurse staffing, only the two items that clearly referred to having enough nurses had high factor loadings. In Lakes development of the PES-NWI factor loadings for these two items were also much higher than factor loadings for the other two items. The same is true for the two items with highest salience on the factor of frontline nurse management. Although the conceptual interpretation for these factors can be viewed as perfect subfactors of Lakes original factors, both might benefit from adding additional items.

Third, we studied measurement invariance in the form of a MIMIC model. An alternative approach to study measurement invariance is a multiple group confirmatory factor analytic (MGCFA) model. MGCFA entails the simultaneous analysis of two or more measurement models while a MIMIC model involves a single measurement model and input matrix . Advantages of the MGCFA model over the MIMIC model are therefore that it is more flexible in that it allows more parameters to represent non-invariance. However, the MIMIC model is more parsimonious and is better suited for analyses where the researcher considers many covariates jointly, which was the case in this study. Another advantage of MIMIC models is that it usually has smaller sample size requirements than MGCFA.

Recommendations for research practice

Two important recommendations emerge from this study for both analytic methods in cross-sectional organizational analyses and cross-cultural research.

First, our findings showed interesting instances of between language group variations in the measurement properties of the PES-NWI. This evidence of measurement non-invariance, even after the many precautions taken in the early stages of the RN4CAST study, suggests that researchers need to take precautions for the potential threat of DIF in all stages of their study. An additional tool for examining measurement invariance worth considering in studies that could be affected by cultural dynamics is the technique of anchoring vignettes (King et al., 2004). These are descriptions of hypothetical people or situations, included in the data collection. Anchoring vignettes provide a common scale of measurement, of which the information is used to account for response category differences. As such, they improve the problems of interpersonal and cross-cultural incomparability in survey research. Anchoring vignettes are highly recommended for research and positive experiences with this technique have been reported by various authors (Johnson, 2006; Salomon et al., 2004; Van Soest et al., 2011).

Second, in this study we aligned a multilevel analysis strategy to the level of theory (Klein et al., 1994). Although we took into account the effects of clustered observations, this can be further modeled in innovative ways related to the concept of measurement invariance. Jak et al. (2013) recently proposed a method for investigating measurement invariance across clusters, illustrating how cluster bias is caused by between-level variables. Also, consider the recent article of Davidov et al. (2012). Using data from the second round of the European Social Survey, the authors demonstrated how measurement non-invariance evidenced by MGCFA can be explained by using multilevel SEM. More specifically, country-level covariates allowed the authors to explain why one of their indicators was non-invariant across countries. Including between-level variables should be of particular interest to PES-NWI research, since nurse work environment policy decisions are often informed by PES-NWI score comparisons between hospitals (McHugh et al., 2013) or countries (Aiken et al., 2012). There have been, and will be, numerous studies where these important concepts should be central to the examination of measurement invariance.

5.5 Conclusion

This study showed that nursing unit managers evaluate certain important nurse work environment dimensions more positively when compared to their staff nurses. Such chasms between frontline workers and management could destabilize work places, contribute to negative work environments, and could ultimately hinder any possible solutions and strategic direction to issues raised by frontline nurses to management. Our findings therefore support endeavors to better understand the connection between nursing unit managers and staff nurses perceptions of their work environment and other organizational features of hospital care. Methodologically, evidence of differential item functioning was found through several items that behaved differently for different language groups included in this study.

This underscores the importance of a rigorous translation procedure and cultural adaptation of items for research studies using instruments developed for measurement in a different healthcare system. We hope our application of a method to assess measurement invariance to health outcomes research linked to healthcare workers, will encourage organizational researchers to take on the challenge of applying advanced statistical techniques to verify the suitability of their analysis to the studys design and implementation challenges.

References

Aiken, L. H., Cimiotti, J. P., Sloane, D. M., Smith, H. L., Flynn, L., and Neff, D. F. (2011). The effects of nurse staffing and nurse education on patient deaths in hospitals with different nurse work environments. *Medical Care*, 49(12):1047–1053.

Aiken, L. H., Clarke, S. P., Sloane, D. M., Lake, E. T., and Cheney, T. (2008). Effects of hospital care environment on patient mortality and nurse outcomes. *The Journal of Nursing Administration*, 38(5):223–229.

Aiken, L. H. and Patrician, P. A. (2000). Measuring organizational traits of hospitals: The revised nursing work index. *Nursing Research*, 49(3):146–153.

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: British Medical Journal*, 344:e1717.

Bandalos, D. L. and Finney, S. J. (2010). *The reviewers guide to quantitative methods in the social sciences*, chapter Factor analysis: Exploratory and confirmatory, pages 93–114. Routledge.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, 44(11 Suppl 3):S176–181.

Byrne, B. M. and Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology*, 34(2):155–175.

Cherepanov, D., Palta, M., Fryback, D. G., Robert, S. A., Hays, R. D., and Kaplan, R. M. (2011). Gender differences in multiple underlying dimensions of health-related quality of life are associated with sociodemographic and socioeconomic status. *Medical Care*, 49(11):1021–1030.

Cheung, G. W. and Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, 25(1):1–27.

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

Duffield, C. M., Roche, M. A., Blay, N., and Stasa, H. (2011). Nursing unit managers, staff retention and the work environment. *J Clin Nurs*, 20(1-2):23–33.

Fong, T. C. T. and Ho, R. T. H. (2013). Factor analyses of the hospital anxiety and depression scale: a bayesian structural equation modeling approach. *Quality of Life Research*, 22(10):2857–2863.

Friese, C. R., Lake, E. T., Aiken, L. H., Silber, J. H., and Sochalski, J. (2008). Hospital nurse practice environments and outcomes for surgical oncology patients. *Health Services Research*, 43(4):1145–1163.

Gabriel, A. S., Erickson, R. J., Moran, C. M., Diefendorff, J. M., and Bromley, G. E. (2013). A multilevel analysis of the effects of the practice environment scale of the nursing work index on nurse outcomes. *Research in Nursing & Health*, 36(6):567–581.

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. *Nurs Res*, 59(2):147–153.

Gunnarsdttir, S., Clarke, S. P., Rafferty, A. M., and Nutbeam, D. (2009). Front-line management, staffing and nurse-doctor relationships as predictors of nurse and patient outcomes. a survey of icelandic hospital nurses. *Int J Nurs Stud*, 46(7):920–927.

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.

Jak, S., Oort, F. J., and Dolan, C. V. (2010). Measurement bias and multidimensionality; an illustration of bias detection in multidimensional measurement models. *Advances in Statistical Analysis*, 94(2):129–137.

Jak, S., Oort, F. J., and Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, 20(2):265–282.

Johnson, T. P. (2006). Methods and frameworks for crosscultural measurement. *Medical Care*, 44(11 Suppl 3):S17–S20.

Kalisch, B. J. and Lee, K. H. (2012). Congruence of perceptions among nursing leaders and staff regarding missed nursing care and teamwork. *J Nurs Adm*, 42(10):473–477.

Kelly, L. A., McHugh, M. D., and Aiken, L. H. (2011). Nurse outcomes in magnet and non-magnet hospitals. *J Nurs Adm*, 41(10):428–433.

Kenny, D. A. and McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 10(3):333–351.

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.

Klein, K. J., Dansereau, F., and Hall, R. J. (1994). Levels issues in theory development, data collection, and analysis. *Academy of Management Review*, 19(2):195–229.

Kleinman, C. S. (2004). Leadership: a key strategy in staff nurse retention. *The Journal of Continuing Education in Nursing*, 35(5):128.

Kramer, M. and Hafner, L. P. (1989). Shared values: Impact on staff nurse job satisfaction and perceived productivity. *Nurs Res*, 38(3):172–177.

Kutney-Lee, A., McHugh, M., Sloane, D., Cimiotti, J., Flynn, L., Neff, D., and Aiken, L. (2009). Nursing: A key to patient satisfaction. *Health Affairs*, 28(4):w669–w677.

Lake, E. T. (2002). Development of the practice environment scale of the nursing work index. *Res Nurs Health*, 25(3):176–188.

Lake, E. T. (2007). The nursing practice environment: Measurement and evidence. *Med Care Res Rev*, 64(2 Suppl):104S–122S.

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.

Marsh, H. W., Hau, K.-T., and Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing hu and bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, 11(3):320–341.

McHugh, M., Kelly, L. A., Smith, H. L., Wu, E. S., Vanak, J. M., and Aiken, L. H. (2013). Lower mortality in magnet hospitals. *J Nurs Adm*, 43(10 Suppl):S4–10.

Millsap, R. E. and Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, 9(1):93.

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

Muthén, B. and Asparouhov, T. (2013). BSEM measurement invariance analysis. Mplus Web Notes: No. 17.

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

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, 28(4):338–354.

Muthén, L. and Muthén, B. (2010). *Mplus User's guide*. Los Angeles: Muthén & Muthén, 6th edition.

Parand, A., Burnett, S., Benn, J., Pinto, A., Iskander, S., and Vincent, C. (2011). The disparity of frontline clinical staff and managers' perceptions of a quality and patient safety initiative. *J Eval Clin Pract*, 17(6):1184–1190.

Price, M., Fitzgerald, L., and Kinsman, L. (2007). Quality improvement: The divergent views of managers and clinicians. *J Nurs Manag*, 15(1):43–50.

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

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, 29(4):347–363.

Satorra, A. and Muthen, B. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, 25:267–316.

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):6.

Singer, S. J., Falwell, A., Gaba, D. M., and Baker, L. C. (2008). Patient safety climate in us hospitals: Variation by management level. *Med Care*, 46(11):1149–1156.

Squires, A., Aiken, L. H., van den Heede, K., Sermeus, W., Bruyneel, L., Lindqvist, R., Schoonhoven, L., Stromseng, I., Busse, R., Brzostek, T., Ensio, A., Moreno-Casbas, M., Rafferty, A. M., Schubert, M., Zikos, D., and Matthews, A. (2013). A systematic survey instrument translation process for multi-country, comparative health workforce studies. *Int J Nurs Stud*, 50(2):264–273.

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

Tabachnick, B. G. and Fidell, L. S. (2012). *Using Multivariate Statistics (6th Edition)*. Pearson, 6 edition.

Van Bogaert, P., Clarke, S., Vermeyen, K., Meulemans, H., and Van de Heyning, P. (2009). Practice environments and their associations with nurse-reported outcomes in belgian hospitals: Development and preliminary validation of a dutch adaptation of the revised nursing work index. *Int J Nurs Stud*, 46(1):54–64.

Van Soest, A., Delaney, L., Harmon, C., Kapteyn, A., and Smith, J. P. (2011). Validating the use of anchoring vignettes for the correction of response scale differences in subjective

questions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(3):575–595.

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.

Warshawsky, N. E. and Havens, D. S. (2011). Global use of the practice environment scale of the nursing work index. *Nurs Res*, 60(1):17–31.

6

GROUP-LEVEL IMPACT OF WORK ENVIRONMENT DIMENSIONS ON BURNOUT EXPERIENCES AMONG NURSES: A MULTIVARIATE MULTILEVEL PROBIT MODEL

Chapter 6 is based on the paper:

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):281-91.

Abstract

High grades of burnout among nurses are a global problem. In the multi-country RN4CAST study, burnout was questioned to more than 30,000 nurses, from about 2,000 nursing units in more than 400 hospitals from 12 countries. Three binarized measurements of burnout are captured, as well as three nurse working environmental measurements. Previous works showed significant association between the two kinds of measurements based on simple regressions. Then whether the association remained the same across all levels became the next question. There was also interest in verifying whether the relationship between the three burnout outcomes remained the same over countries and hospitals. Therefore, on top of the mixed effects mean structure, we added a mixed effects structure in the correlation matrix. In the current paper, we propose a Bayesian tri-variate four-level probit factor model to estimate the relationship between the three burnout outcomes and the working environmental variable in each level, as well as a flexible correlation structure via a common latent factor with structured loadings. Despite the complex structure of the data, the model converged well in WinBUGS. We obtained significant negative relationships between the working environment and the burnout variables in each level, with different magnitude. Further, we found a positive correlation structure varying across countries but staying quite stable across hospitals and nursing units within a country. We conclude that the multivariate multilevel probit factor model provides an elegant manner to flexibly model the multivariate binary data in a multi-level context. The implementation in WinBUGS was successful and the extension to categorical, ordinal and mixture outcomes presents no difficulties.

6.1 Background

The context in which nursing outcomes research is undertaken is often multilevel in nature. Multilevel complexity can be caused by naturally occurring dependencies (e.g. nurses in hospitals) or imposed-by-design dependencies (multistage sampling). These complex structures imply an explicit multilevel analysis to take into account the correlated nature of the data. Several studies introduced the conceptual and statistical background in multilevel analysis for nursing research and portrayed examples of the application of two-level techniques for meta-analysis (Wu, 1997), confirmatory factor analysis (Gajewski et al., 2010) and regression analysis (Adewale et al., 2007; Cho, 2003; Park and Lake, 2005). They detailed how features of these two-level techniques overcome the fallacies of conventional single level models in the analysis of clustered data. These studies vastly contributed to illustrating the basics in multilevel modelling for patient and organizational outcomes research.

The current article takes the application of multilevel regression techniques in nursing research a step further by analyzing the association between nurses' work environment and burnout in a four-level data set (country, hospital, nursing unit, nurse) resulting from the Registered Nurse Forecasting (RN4CAST) project. This multi-country nurse workforce study has provided a unique data set on organizational features of nursing care and measures of nurse wellbeing, patient satisfaction with care, and quality and safety of patient care (Sermeus et al., 2011). Understanding research problems in this data structure dictates more complicated multilevel analysis strategies than have been used in previous efforts.

Nurse work environment and burnout, the constructs of interest studied in the current article, have been well researched previously. Large-scale studies have shown that nurses working in both post-industrial (Aiken et al., 2001; Hasselhorn, 2003) and developing countries (Poghosyan et al., 2009) are susceptible to burnout. Burnout in turn impacts patient satisfaction with nursing care (Vahey et al., 2004) and plays a mediational role in nurses' reports on quality of care and adverse events, job dissatisfaction and turnover intentions (Laschinger and Leiter, 2006; Leiter and Maslach, 2009; Van Bogaert et al., 2009). The consequences of burnout thus potentially negatively affect nurses, patients, organizations and health systems in general. Of interest is that the large majority of nurse researchers studying burnout have mainly focused on the emotional exhaustion dimension of the syndrome. This dimension refers to feelings of being overextended and depleted of one's emotional and physical resources, and has indeed been described by the world leading researchers in the field of burnout as the key aspect of burnout (Maslach et al., 1996). However, they have also repeatedly emphasized the significance of the three-dimensional burnout model in that it 'clearly places the individual strain experience within the social context of the workplace and involves the person's conception of both self and others' (Maslach, 1993). Measuring emotional exhaustion only 'fails to capture the critical aspects of the relationships that people have with their work' (Maslach and Leiter, 2008).

We therefore study in this article all three burnout dimensions: emotional exhaustion, depersonalisation and personal accomplishment. Depersonalisation refers to negative, callous, or excessively detached responses to various aspects of the job. Feelings of incom-

petence and a lack of achievement and productivity in work are captured by the personal accomplishment dimension (Maslach and Leiter, 2008). A point of departure for our line of research presented here is the well-documented evidence on the causes of burnout. Studies across countries worldwide found that modifiable dimensions of nurses' work environment and workload predict burnout rates among nurses (Bruyneel et al., 2009; Kelly et al., 2011; Nantsupawat et al., 2011). Such dimensions include staffing and resource adequacy, managerial support for nursing, nurse participation in hospital affairs, doctor-nurse collegial relations, and promotion of care quality. Nurses' reports on their work environment and burnout experience, that are both multidimensional constructs, provide an excellent opportunity to introduce nurse researchers to advanced multilevel regression analyses.

The aim of this study is two-fold. First, to explore and investigate the effect of the nursing unit, hospital, and country level variability on the relationship between dimensions of nurses' work environment and dimensions of burnout. Second, to explore the significance of the nursing unit, hospital, and country level variability among the burnout dimensions.

6.2 Methods

6.2.1 Data Sources

The data used in this study come from the RN4CAST project, a three year (2009-2011) nurse workforce study funded by the Seventh Framework Programme of the European Union. For the RN4CAST project, research teams from across twelve countries used a multilevel observational design to determine how system-level features in the organization of nursing care (work environment, education, workload) impact individual measures of nurse well-being (burnout, job satisfaction, turnover) and patient safety outcomes and care satisfaction. The design of the RN4CAST project is described in detail by Sermeus et al. (2011). The relevant data for the current analysis include nurses' ratings of their work environment and reports on burnout experiences.

6.2.2 Ethical considerations

In all but one country, depending on national legislation, the study protocol was approved by either central ethical committees (e.g. nation or university) or local ethical committees (e.g. hospitals). In the Netherlands no ethical approval was required.

6.2.3 Study sample

A four-level hierarchical structure is the form of the sampling strategy used in the RN4CAST project (Figure 1). The study encompasses data from 33,731 nurses (level 1) in 2089 nursing units (level 2) in 486 hospitals (level 3) in 12 countries (level 4). The participating countries are Belgium, England, Finland, Germany, Greece, Ireland, the Netherlands, Norway, Poland, Spain, Sweden, and Switzerland. A minimum of 30 hospitals participated in each country. In most of the countries, the selected hospitals either represented all hospitals in the

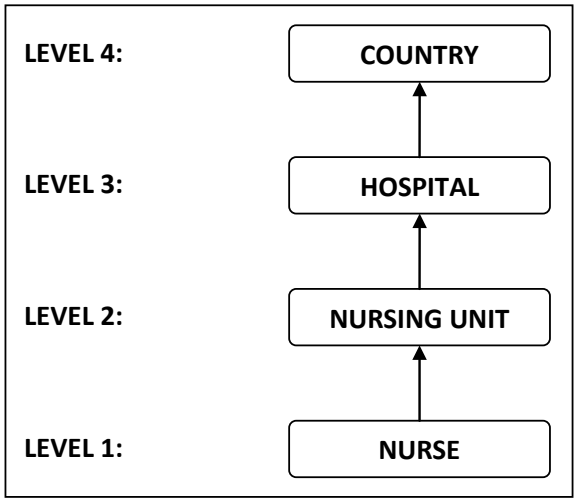

Figure 6.1: Classification diagram of the four-level RN4CAST data structure

country (Ireland and Norway) or were random samples of all general (non-specialized) hospitals. In Finland, Poland and Greece, the purposive sampling was used which was based on the geographical spread, hospital size and hospital type. At least two general medical and surgical nursing units for each hospital were randomly selected, of which all nurses involved in direct patient care activities were invited to participate in the study. A different sampling strategy was used in Sweden that nurses were selected via the Swedish Nursing Association, and the hospitals in which they work were then identified. The overall response rate of 62% compares favourably with rates seen in other nursing outcomes research studies and was for most countries consistently high across nursing units and hospitals. England (38.6%), Finland (46.2%) and Germany (41.6%) had lower response rates. Swedish data were excluded as no unit identifiers were available from the Swedish sampling design. The collected data have the characteristics of a strict hierarchical structure. First, lower level units are nested within one and only one unit at the next higher level. Second, lower level units present repeated samples of higher level units. Third, there was successive sampling from each level of the hierarchical population. Fourth, as can be expected, the sample size within higher level units was imbalanced, albeit there were a sufficient number of respondents for analysis in the sampled units.

6.2.4 Study measures

The nurse work environment was measured using the Practice Environment Scale of the Nursing Work Index (PES-NWI), an internationally validated organizational measure (Warshawsky and Havens, 2011) that reflects the multidimensionality of nurses' work environment. The PES-NWI operationalizes five dimensions that facilitate or constrain nursing practice. Nurses therefore score statements about the work environment on a four point

Likert scale ('Totally agree'=4, 'Agree'=3, 'Not agree'=2, 'Totally not agree'=1). The five dimensions, example item and number of items of each dimension are: managerial support for nursing ('A nurse manager who is a good manager and leader'; 4 items), nurse participation in hospital affairs ('Career development/clinical ladder opportunity'; 8 items), doctor-nurse collegial relations ('Physicians respect nurses as professionals'; 7 items), staffing and resource adequacy ('Enough registered nurses on staff to provide quality patient care'; 4 items), and promotion of care quality ('Working with nurses who are clinically competent'; 9 items). The Pearson coefficient correlation matrix showed relatively higher correlations between the dimensions of staffing and resource adequacy and nurse participation in hospital affairs and other dimensions. This has the potential to adversely affect regression estimates. The multicollinearity checking showed a potential problem for the dimension nurse participation in hospital affairs. In addition, another paper by Kutney-Lee et al. (2009) used the same three environment dimensions. We therefore did not include these two dimensions for further analyses.

The response variable for this analysis is the multidimensional burnout phenomenon. Burnout is a syndrome of emotional exhaustion, depersonalisation and reduced personal accomplishment that can occur among individuals who do "people work" of some kind (Maslach and Jackson, 1986). We evaluated burnout using the 22-item Maslach Burnout Inventory (MBI) that has been extensively used to capture the three dimensions of burnout. On a seven point Likert scale, nurses assessed the frequency (ranging from never to every day) of burnout experiences. Degrees of burnout are calculated separately for the dimensions of emotional exhaustion (9 items), depersonalisation (5 items) and reduced personal accomplishment (8 items) by using the numerical cut-off points listed on Maslach and Jackson (1986) scoring key. This key categorizes respondents into low, average and high ranges of experienced burnout for each dimension. We dichotomize respondents as experiencing high burnout or not, thus requiring a binary modelling. A probit regression was chosen in this study.

About 10% of the nurses, at least one data value was missing on either the work environment or the burnout items. For the work environment dimensions, any missing data values were completed with the mean of the non-missing data values. For the dimensions of burnout, missing data values were imputed using the multinomial distribution of frequencies per hospital. That is, each missing was replaced by a random value from the observed multinomial distribution in each hospital. After applying the missing data imputation strategies in R (version 2.13.0), the final data set contained 23446 nurses coming from 2087 nursing units, within 352 hospitals and 11 countries.

6.2.5 Statistical analysis

6.2.5.1 Intraclass Correlation Coefficient

The basic statistical prerequisite for the appropriate application of multilevel analyses includes clustered data with a positive intraclass correlation coefficient (ICC). A (true) positive ICC violates the independent observations assumption of ordinary least squares esti-

mation, resulting in downwardly biased standard error estimates, overly large test statistics, and inflated type I error rates (Krull and MacKinnon, 2001). We estimated the ICC to get an idea of the degree of variation in the burnout dimensions that were explained at each level. For the four-level model, the ICC_s are defined as follows:

$$ICC_c = \frac{\sigma_c^2}{\sigma_{all}^2},\ ICC_h = \frac{\sigma_c^2 + \sigma_h^2}{\sigma_{all}^2},\ ICC_h = \frac{\sigma_c^2 + \sigma_h^2 + \sigma_u^2}{\sigma_{all}^2}, \tag{6.1}$$

with σ_{all}^2 the sum of all variances, i.e. $\sigma_{all}^2 = \sigma_c^2 + \sigma_h^2 + \sigma_u^2 + \sigma_n^2$. The subscripts c, h, u, n represent country, hospital, nursing unit and nurse levels, respectively. For the probit model, the lowest level variance, i.e. σ_n^2 , is set to be one using the latent variable strategy. Regarding the interpretation of the intra-class correlation, we note that ICC_c is the correlation of two nurses' responses within the same country (different hospitals and nursing units), and ICC_h is the correlation of two nurses' responses within the same hospital (same country but different nursing units), while ICC_u is the correlation of two nurses' responses within the same nursing unit (same country and hospital). The higher the ICC scores, the higher the degree of homogeneity among nurses clusters. In order to get the partitioned proportion of the total variation into each level, we subtracted the higher level ICC from the lower level ICC, that is:

$$P_c = \frac{\sigma_c^2}{\sigma_{all}^2},\ P_h = \frac{\sigma_h^2}{\sigma_{all}^2},\ P_u = \frac{\sigma_u^2}{\sigma_{all}^2},\ P_n = \frac{\sigma_n^2}{\sigma_{all}^2}. \tag{6.2}$$

6.2.5.2 Model specification

The outcomes of interest, i.e. the three burnout subscales, take place at the individual level. These are called level-one outcomes. The work environment dimensions were conceptualized to influence these level-one outcomes at higher organizational levels. In multilevel jargon, such variables are called ecological variables. We will continue to use this term and the term work environment dimension throughout this paper. Contrary to previous nurse workforce studies, we will avoid to name these variables environmental variables. Ecological variables were calculated as the average of the item responses of nurses within nursing units, hospitals and countries. In a multilevel context, the effect of a covariate can be decomposed into effects on different levels, which is recommended by Neuhaus and Kalbfleisch (1998). This decomposition allows us to learn the difference of the relationship at each level, which renders the modelling more flexible. We now rewrite each of the ecological variables as follows:

$$\bar{X}_u = (\bar{X}_u - \bar{X}_h) + (\bar{X}_h - \bar{X}_c) + \bar{X}_c, \tag{6.3}$$

where $\bar{X}$ is the aggregated average value of one of the work environment dimensions and the subscripts u, h, c represent the nursing unit, hospital and country levels respectively. This representation partitions the nursing unit level covariate into a sum of three parts: the nursing unit level deviation from the hospital level mean, the hospital level deviation from the country level mean and the country level mean. The rationale is that, by partitioning

the covariates we can estimate the pure effect of the covariate at each level. For example, by subtracting the hospital level mean from the nursing unit level variable, we keep only the unit level effect thereby removing the higher level effects.

6.2.5.3 Step-by-step multilevel modelling approach

We propose a step-by-step approach towards building a model for multilevel regression analysis of the relationship between a multidimensional covariate and a multidimensional dichotomous outcome. To relate the work environment dimensions to burnout dimensions, we first build a series of nine univariate simple multilevel probit models. Here, we consider the impact of each work environment dimension on each burnout dimension separately. Second, we build a series of three univariate multiple multilevel probit models. Here, we consider the joint impact of the work environment dimensions on each of the burnout dimensions separately. The univariate simple multilevel probit model is described in appendix A. The extension to the multiple model only involves adding more covariates at each level.

The probit model assumes that there is an underlying latent variable Z that follows a normal distribution with standard deviation one, conditional on all the fixed effects. This latent variable expresses the true feeling of the nurse and is assumed that for $Z > 0$, burnout is expressed on a manifest scale indicated by $Y = 1$, otherwise zero. In the univariate simple multilevel probit models, we have one ecological covariate decomposed as in equation (3), augmented with a random intercept for each higher level. The random effects follow a normal distribution with mean zero and a specific variance. No random slopes were included into the model following exploratory analyses (using Akaike's information criterion (Akaike, 1974)), indicating that the relationship between the work environment and burnout differs in different levels, but stays constant within each level. We used the same settings for the random effects in the univariate multiple multilevel probit models.

The final outcome of our step-by-step approach is a multivariate multilevel probit model. The combination of the three univariate multiple multilevel probit models results in a three-variate four-level probit model, which could also be called, in general, the multilevel multivariate probit model (MVP) . In this modelling, a common factor was introduced to construct the correlations among the three burnout dimensions (cf. three-variate). Similar to the univariate models, all three covariates are partitioned into three parts (unit, hospital, and country level pure effects). There are three random intercepts vectors corresponding to the three higher levels for each outcome, as well as the three random factor loadings, which imply a varying correlation structure. See appendix B for more details on the modelling of multilevel MVP.

6.2.5.4 Computational aspects

We used the R lme4 package (Bates et al., 2009) to fit the univariate simple and multiple multilevel probit models. However, the multivariate multilevel probit model is beyond the scope of this package and we are not aware of any frequentist software that can handle this

model. For this reason, we used the popular WinBUGS package. This software is based on the Bayesian paradigm and uses Markov Chain Monte Carlo sampling techniques to arrive at the parameter estimates. A Bayesian analysis needs prior distributions for all its parameters. We have used here the following priors. For the regression coefficients we have taken a normal distribution with mean zero and a large variance. The factor loadings (λs) were given a (multivariate) normal distribution with hyper-parameters, i.e. the variance (matrix), which has a vague conjugate inverse Wishart distribution. The random intercepts at each level followed the same priors as the factor loadings. For the posterior statistics, we calculated the posterior mean, median, standard error, and the 95% equal tail credibility interval. This credibility interval is the Bayesian equivalent of the classical 95% confidence interval, which indicates a significant non-zero estimate if the interval does not include zero, and a non-significant estimate if the interval includes zero. The hierarchical centering strategy of (Gelfand et al., 1995) was applied to improve the convergence of the MCMC iterations. Three chains were initialized with different starting values. We obtained posterior means and 95% credible intervals based on 10,000 iterations after having removed a burn-in part of 20,000 iterations. The Brooks-Gelman-Rubin diagnostic plot (Brooks and Gelman, 1998), which tests the within- and between-chains variation, was used to check the convergence of all parameters. The WinBUGS program is available from the first author.

6.3 Results

6.3.1 General description

The mean estimates of emotional exhaustion and depersonalisation burnout rates in Greece are higher than for the other countries, while Poland has the highest rate of reduced personal accomplishment. Greece shows the widest interquartile ranges for all three dimensions of burnout, while the Netherlands shows the narrowest. The burnout rates thus vary greatly across Greece hospitals but are stable across Dutch hospitals. Swiss nurses' ratings of their work environment are the highest for all three dimensions. Greek and Polish nurses' ratings of their work environment are lowest. The full descriptive findings were published previously by Aiken et al. (2012).

6.3.2 Intra-class Correlation Coefficients

Table 1 shows the proportion of total variance that could be explained at each level for the three burnout dimensions and environment dimensions. The country level explained about 22% of the variation in emotional exhaustion, 13% in depersonalisation and 6% in personal accomplishment. The hospital level explains the least variation. Less than 5% in the variation of all three outcomes can be explained at the hospital level. The nursing unit level contributes about 10%, 6% and 2% for the three outcomes respectively. These multilevel variances decomposition indicate the modelling for multilevel analyses. For the variances decomposition of the three environment dimensions, the different proportions at each level suggest different ranges of the environment variations. The hospital level variation for each

environment dimension is the smallest among the three levels respectively. This will be further discussed in the discussion part.

Table 6.1: Proportion of total variance explained of the three burnout dimensions and environment dimensions at the country, hospital, nursing unit and individual level

Outcomes/Covariates	Country	Hospital	Nursing unit	Nurse
Emotional exhaustion	22.4	3.8	9.5	64.4
Depersonalization	13.0	3.2	6.3	77.5
Personal accomplishment	6.0	2.3	2.3	89.4
Managerial support for nursing	7.3	3.1	89.6	-
Doctor-nurse collegial relations	14.8	2.0	83.2	-
Promotion of care quality	15.0	4.6	80.4	-

6.3.3 Relationship between the work environment and burnout

All nine univariate simple multilevel models gave significant negative effect estimates for the ecological variables at almost all levels (Table 2). The effect is most pronounced for emotional exhaustion, while personal accomplishment shows the weakest effect. An exemplary graph of a univariate simple multilevel model is given in Figure 2. This figure displays a clear negative trend between the country level ecological variable of managerial support of nurses and emotional exhaustion. The negative regression line is the adjusted line that takes into account the number of nurses in each country, which is represented by the area of the circle. Greece has the smallest sample size and appears to be an outlier. Table 2 displays the results for the three univariate multiple multilevel probit models. There is a pronounced ecological effect of the nursing unit level variability in the relationship between the work environment dimensions of doctor-nurse collegial relations and promotion of care quality and all three burnout dimensions. The effect of the nursing unit level variability for the dimension of managerial support of nursing is only present for emotional exhaustion. At the hospital level, the latter effect is present for both emotional exhaustion and depersonalisation. Doctor-nurse collegial relations have no effect on either burnout dimension at the hospital level. Promotion of care quality is significantly related to all three burnout dimensions at the hospital level. At the country level, we found only an effect for doctor-nurse collegial relations on personal accomplishment. This effect was absent in the three-variate four-level probit model (Table 3). The other fixed effects and the standard deviations in the final model are similar to those of univariate multiple multilevel models. The 95% equal tail credibility interval is the Bayesian equivalent of the classical 95% confidence interval. That is, the estimate is significantly larger/smaller than zero if the interval does not include zero, and not significant if the interval includes zero.

Table 6.2: Univariate simple and multiple probit model estimates

Outcomes	Covariates	Levels	Univariate simple models			Univariate multiple models		
			EST	SE	P-value	EST	SE	P-value
Emotional exhaustion	Managerial support for nursing	Nursing Unit	-0.623	0.039	<0.001	-0.27	0.05	<0.001
		Hospital	-0.672	0.079	<0.001	-0.54	0.107	<0.001
		Country	-2.33	0.708	0.001	-2.604	1.341	0.052
	Doctor-nurse collegial relations	Nursing Unit	-0.483	0.05	<0.001	-0.108	0.052	0.037
		Hospital	-0.496	0.119	<0.001	0.073	0.133	0.582
		Country	-1.114	0.729	0.126	0.634	0.877	0.469
	Promotion of care quality	Nursing Unit	-1.109	0.061	<0.001	-0.789	0.078	<0.001
		Hospital	-0.684	0.097	<0.001	-0.298	0.125	0.017
		Country	-1.559	0.661	0.018	-0.421	0.743	0.571
Depersonalization	Managerial support for nursing	Nursing Unit	-0.406	0.037	<0.001	-0.083	0.048	0.082
		Hospital	-0.489	0.071	<0.001	-0.27	0.097	0.005
		Country	-1.634	0.494	0.001	-1.839	0.94	0.05
	Doctor-nurse collegial relations	Nursing Unit	-0.419	0.047	<0.001	-0.172	0.05	0.001
		Hospital	-0.493	0.104	<0.001	-0.085	0.12	0.48
		Country	-0.793	0.504	0.116	0.429	0.615	0.485
	Promotion of care quality	Nursing Unit	-0.838	0.058	<0.001	-0.671	0.075	<0.001
		Hospital	-0.597	0.084	<0.001	-0.354	0.113	0.002
		Country	-1.081	0.467	0.021	-0.273	0.521	0.6
Personal accomplishment	Managerial support for nursing	Nursing Unit	-0.257	0.032	<0.001	-0.031	0.041	0.453
		Hospital	-0.302	0.059	<0.001	-0.087	0.08	0.277
		Country	-1.008	0.331	0.002	-0.279	0.557	0.616
	Doctor-nurse collegial relations	Nursing Unit	-0.343	0.039	<0.001	-0.2	0.043	<0.001
		Hospital	-0.35	0.085	<0.001	-0.087	0.1	0.387
		Country	-0.9	0.228	<0.001	-0.727	0.364	0.046
	Promotion of care quality	Nursing Unit	-0.537	0.049	<0.001	-0.407	0.065	<0.001
		Hospital	-0.437	0.069	<0.001	-0.336	0.093	<0.001
		Country	-0.448	0.341	0.189	-0.025	0.31	0.936

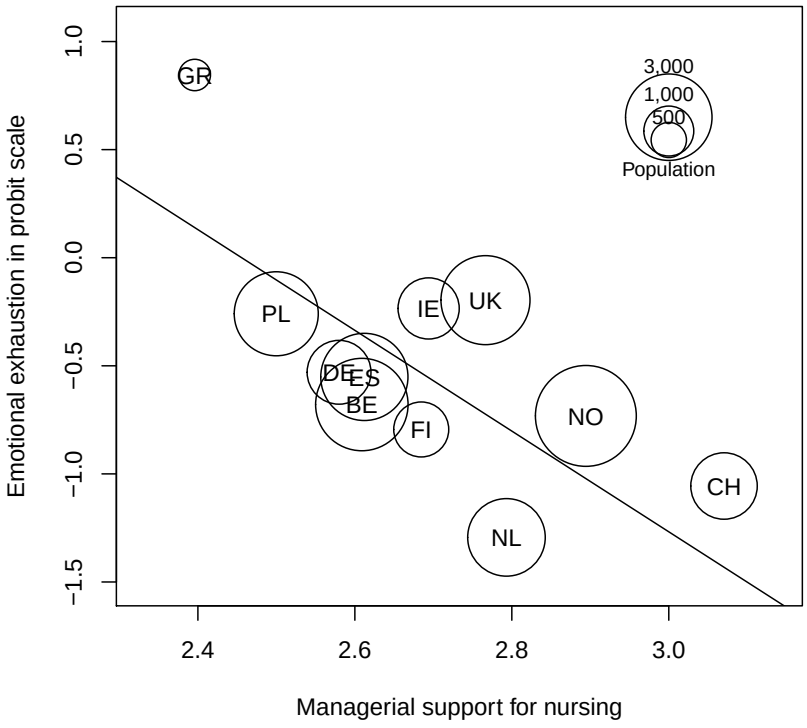

Figure 6.2: Relationship between emotional exhaustion and managerial support for nursing at the country level

Table 6.3: Bayesian multivariate multilevel probit model estimates

Outcomes	Covariates	Levels	Mean	SE	2.50%	Median	97.50%
Emotional exhaustion	Managerial support for nursing	Nursing unit	-0.277	0.05	-0.373	-0.278	-0.178
		Hospital	-0.532	0.115	-0.754	-0.531	-0.31
		Country	-2.572	2.283	-7.138	-2.571	1.935
	Doctor-nurse collegial relations	Nursing unit	-0.108	0.054	-0.214	-0.108	-0.002
		Hospital	0.056	0.142	-0.219	0.057	0.335
		Country	0.628	1.477	-2.295	0.619	3.587
	Promotion of care quality	Nursing unit	-0.783	0.079	-0.935	-0.783	-0.628
		Hospital	-0.284	0.135	-0.545	-0.286	-0.019
		Country	-0.451	1.293	-3.043	-0.44	2.074
Depersonalization	Managerial support for nursing	Nursing unit	-0.094	0.049	-0.188	-0.094	0.002
		Hospital	-0.265	0.107	-0.478	-0.264	-0.059
		Country	-1.856	1.892	-5.73	-1.822	1.769
	Doctor-nurse collegial relations	Nursing unit	-0.168	0.05	-0.265	-0.167	-0.071
		Hospital	-0.105	0.133	-0.362	-0.105	0.154
		Country	0.442	1.237	-1.949	0.425	2.93
	Promotion of care quality	Nursing unit	-0.667	0.077	-0.818	-0.665	-0.516
		Hospital	-0.344	0.126	-0.588	-0.346	-0.093
		Country	-0.297	1.066	-2.356	-0.304	1.832
Personal accomplishment	Managerial support for nursing	Nursing unit	-0.031	0.043	-0.114	-0.031	0.053
		Hospital	-0.101	0.093	-0.281	-0.101	0.081
		Country	-0.236	1.581	-3.351	-0.244	2.953
	Doctor-nurse collegial relations	Nursing unit	-0.204	0.045	-0.293	-0.204	-0.116
		Hospital	-0.084	0.114	-0.31	-0.083	0.137
		Country	-0.755	1.034	-2.799	-0.76	1.291
	Promotion of care quality	Nursing unit	-0.41	0.068	-0.545	-0.41	-0.277
		Hospital	-0.319	0.107	-0.531	-0.319	-0.113
		Country	-0.042	0.883	-1.841	-0.036	1.737

6.3.4 Relationship among the burnout responses

Figure 3 displays the partitioned level-specific correlation structures among the three burnout dimensions. The dots are the posterior means and the lines are the 95% confidence interval in Bayesian way (for the country level or hospital level, this confidence interval is actually the median interval within the country or hospital). A sample of 20 is randomly selected at hospital and nursing unit levels respectively to make the figure readable. At the country level (first column in Figure 3), the correlations varied vastly, with some significant differences between countries. At the hospital level (second column in Figure 3), all remaining correlation structures stayed close to zero after removing the country level correlations. Similar findings are seen for the nursing unit level after removing the country and hospital level correlations (third column in Figure 3). This indicates that the correlation structure among the three outcomes was quite different between countries, but stayed stable between hospitals within countries and between nursing units within hospitals. Greece again performed much different from the other countries.

6.4 Discussion

In this paper, we investigated the relationship between nurse work environment dimensions (managerial support for nursing, doctor-nurse collegial relations, and promotion of care quality) and burnout dimensions (emotional exhaustion, depersonalisation, personal accomplishment) using an advanced multilevel approach. We aimed to explore and investigate the effect of the nursing unit, hospital, and country level variability on the relationship between dimensions of nurses' work environment and dimensions of burnout. We also explored the significance of the nursing unit, hospital, and country level variability among the burnout dimensions. We first specified ecological measures of the nurse work environment dimensions at the three organizational levels (nursing unit, hospital, country). The effect of the covariate was decomposed into effects on different levels. This so called partitioning strategy allowed us to specify the pure effect of the covariate at each level. We then combined these ecological measures with individual-level burnout experiences within a series of multilevel statistical models that would allow us to model the complex contextuality and heterogeneity.

Our approach towards building a model for multilevel regression analyses of the relationship between such multidimensional covariate and multidimensional dichotomous outcome took three steps. We first fitted univariate simple multilevel probit models where we considered the impact of each work environment dimension on each burnout dimension separately. Second, we fitted univariate multiple multilevel probit models where we considered the joint impact of the work environment dimensions on each of the burnout dimensions separately. Last, we fitted a multivariate multilevel probit model where we considered the joint impact of the work environment dimensions on the three burnout dimensions. Not surprisingly, our results showed a negative relationship between work environment dimensions and burnout experiences among nurses. However, by maintaining in our advanced analyses the social context in which the data were collected, we added some interesting

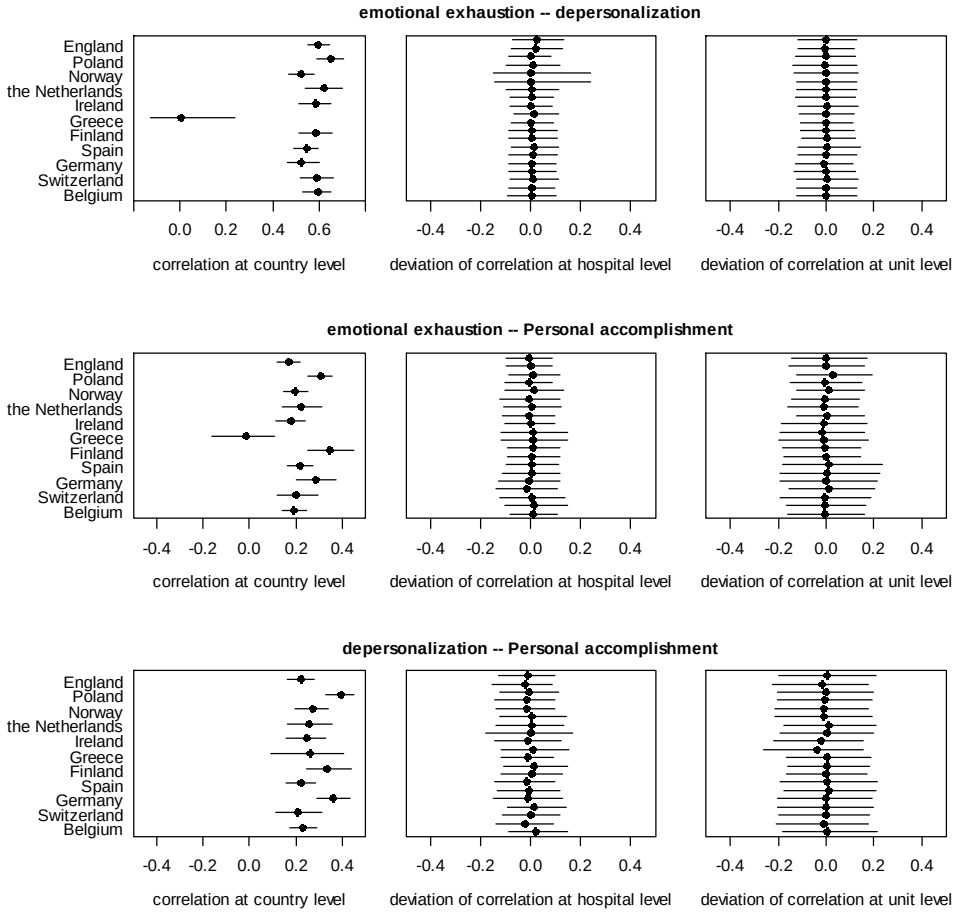

Figure 6.3: Partitioned level-specific correlation structures among the three burnout outcomes. For the hospital and nursing unit level deviation of correlations (the 2nd and 3rd columns), a sample of 20 is randomly selected, respectively.

findings to what was already known about this relationship. By using the partition strategy and modelling the burnout and work environment dimensions jointly, we now have a more detailed view of their relationship. The final model showed no country-level effect for either work environment dimension on any of the three burnout dimensions. Doctor-nurse collegial relations affected all burnout dimensions at the nursing unit level only. For the dimension of promotion of care quality, the effect of the ecological exposure on burnout was pronounced at both the nursing unit and the hospital level for all three burnout dimensions. The magnitude of this effect was consistently stronger at the nursing unit. Findings for the dimension of managerial support for nursing were ambiguous. The effect on emotional exhaustion was less pronounced at the nursing unit level than at the hospital level. An effect on depersonalization was only present at the hospital level. There was no effect on personal accomplishment at either level. In developing the PES-NWI, Lake (2002) had already identified that substantive domains of the subscales ranged from the broad hospital context to the immediate nursing unit context, leading her to conclude that the subscales exhibit multilevel range in hospital contexts. The varying magnitude in effects found at different levels in this study pleas for the use of a multilevel analysis in future studies.

The results should however be interpreted with caution. Previous efforts from our research group have shown that aggregating features of nursing care of all nursing units at the hospital level might obscure the hospital level effects on outcome measures (Van den Heede et al., 2009). For this study, that means that even though there is no hospital effect for some work environment dimensions on certain burnout dimensions, human resources management should not conclude that interventions at the hospital level are by definition not effective. Statistical support for this is given by the larger variance of work environment and burnout dimensions at the unit than at the hospital level, as seen from the intraclass correlations (Table 1). The results indicate that all three work environment dimensions deserve the attention of human resources management to secure better outcomes. The pronounced effects of the dimensions of promotion of care quality and managerial support for nurses at both the nursing unit and the hospital level point to a need for leaders from bedside to boardroom to further develop managerial skills and share goals for achieving positive health care environments. Front-line, middle and upper managers need to move towards an integrated vision on promotion of care quality in tune with the workforce. As shown by our empirical findings, at the unit level, nurses should partner up with physicians. The multivariate multilevel probit model allowed for a flexible hierarchical correlation structure. We found a positive correlation matrix among the three burnout variables. This varied across countries, but was stable across hospitals and nursing units within a specific country.

There is a large body of literature that has described the relationship between burnout and work environment. Although some of them used different measure instruments, they came up with similar findings. Melchior et al. (1997) analyzed the relationship between burnout and work environment at nurse level and nursing unit level separately. They found significant relationships at each level. However, their modeling is not very efficient (separate analysis for each level) and suffers from a small sample size at the nursing unit level. Van Bogaert et al. (2009) explored the nursing unit level relationship between work envi-

ronment dimensions and burnout using a 2-level linear mixed model for each of the three burnout dimensions separately. They found significant relationships for all environment coefficients. O'Mahony (2011) studied the relationship between work environment and two burnout dimensions (emotional exhaustion and depersonalization), and found a significant correlation through simple linear regressions. Liu et al. (2012) analyzed data from South China using a logistic regression model and concluded that improving the unit level work environment from poor to better leads to a moderate (33%) decrease in job-related emotional exhaustion. As can be inferred in the results part, our 3-variate 4-level probit model detected similar findings as previous works at the nursing unit level. However, we provide more detailed information, also at higher levels, i.e. hospital and country levels.

Stepping back from what this article adds, it is not free from statistical, practical and conceptual limitations. First, we encountered multicollinearity between dimensions of the work environment. This multicollinearity could be due to specific items of different dimensions correlating highly, rather than the whole dimensions. A confirmatory factor analysis (CFA) is needed to study the factor structure underlying these items. The four-level RN4CAST data structure would require a more complicated multilevel CFA to detect the potentially different factor structure in each level. Such approach requires the application of new goodness-of-fit tests for verifying the statistical assumptions made at the different levels of the hierarchy. This analysis was beyond the scope of this study.

Second, the Bayesian multivariate multilevel probit model included no fixed effects in the correlation structure among the three burnout dimensions, although adding covariates is theoretically possible. In practice, such models with both fixed and random effects in correlation structure causes rather slow convergence and needs to be improved further.

A third possible limitation is that country was treated as a random effect throughout this paper. However, since country is not chosen at random (in contrast to hospital and nursing unit), we could have assumed that it has a fixed effect, involving an index variable for each country.

In studies that involve multiple levels, researchers should be cautious of four types of fallacies that potentially arise when the methods fail to fit the conceptual model (Diez-Roux, 1998). Ecological fallacies arise when drawing inferences at the individual level based on group-level data. Atomistic fallacies occur when drawing inferences at the group level based on individual level data. Ecological and atomistic fallacies are both types of inferential fallacies that can be overcome by ensuring that the data collected match the level at which inferences are to be made, as was accomplished by the design of the RN4CAST project. The psychological fallacy would arise when ignoring the relevant group-level covariates in a study of individual level outcomes. In this article we have considered the nursing unit, hospital and country level variability in the relationship under study. Fourth, the sociological fallacy would arise when ignoring the role of individual level factors in a study of groups. This brings us to a fourth potential limitation of the study. We have shown that social contexts shape burnout experiences among nurses by including group-level variables. By not including possible confounding individual level variables like nurses' age and gender, it might appear that we have perpetuated the idea that burnout experiences

are absolutely socially determined rather than leaving room for individual determinants. Combining group-level and individual-level covariates in the proposed models is methodologically challenging. It would be meaningful in future papers to analyze the joint impact of social context and individual characteristics on burnout experiences.

As described in the study sample section, the RN4CAST project accommodated within the framework of a strict hierarchy. A fifth potential limitation is that it is plausible that, although participating nurses strictly worked in the sampled nursing unit and hospital, both covariates and outcomes may be conditioned through social processes operating between nursing units in hospitals.

Last, excluding Sweden from the final analysis may be considered misleading and inefficient. We therefore did a sensitivity analysis to see the influence of the Swedish data on the model estimates. These tests consist of three models which are: the hospital level univariate random effects model with country as the random effects, the country level univariate regression and the nurse level univariate random effects model with country and hospital as the random effects. All these models were ran using both datasets with and without Swedish data to detect differing estimates. The difference was minimal, and the estimates from the two data sets for all the three models were close, both for fixed and random effects.

6.5 Conclusions

Nurse work environment dynamics are related to nurses' burnout experiences at both the nursing unit and the hospital level. The correlation structure among the three burnout outcomes varies across countries, but is stable between hospitals within countries and between nursing units within hospitals. The findings provide a motivation for nurses and physicians within nursing units to partner up and for nurse leaders from bedside to boardroom to further develop their managerial skills. There is a clear need towards an integrated vision on promotion of care quality in tune with the workforce. The results also imply that, in evaluating health care organizations, researchers should sample the different levels of the organization under study and maintain this structure in analyzing the data.

References

Adewale, A. J., Hayduk, L., Estabrooks, C. A., Cummings, G. G., Midodzi, W. K., and Derksen, L. (2007). Understanding hierarchical linear models: Applications in nursing research. *Nursing Research*, 56(4):S40–S46.

Aiken, L. H., Clarke, S. P., Sloane, D. M., Sochalski, J. A., Busse, R., Clarke, H., Giovannetti, P., Hunt, J., Rafferty, A. M., and Shamian, J. (2001). Nurses reports on hospital care in five countries. *Health Affairs*, 20(3):43–53.

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: British Medical Journal*, 344:e1717.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723.

Bates, D., Maechler, M., and Ben, B. (2009). *Package lme4*.

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

Bruyneel, L., den Heede, K. V., Diya, L., Aiken, L., and Sermeus, W. (2009). Predictive validity of the international hospital outcomes study questionnaire: An rn4cast pilot study. *J Nurs Scholarsh*, 41(2):202–210.

Cho, S.-H. (2003). Using multilevel analysis in patient and organizational outcomes research. *Nursing Research*, 52(1):61–65.

Diez-Roux, A. V. (1998). Bringing context back into epidemiology: Variables and fallacies in multilevel analysis. *American Journal of Public Health*, 88(2):216–222.

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.

Gelfand, A., Sahu, S., and Carlin, B. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika*, 82(3):479–488.

Hasselhorn, H., T. P. M. B. (2003). Working conditions and intent to leave the profession among nursing staff in europe. Stockholm, National Institute for Working Life.

Kelly, L. A., McHugh, M. D., Aiken, L. H., et al. (2011). Nurse outcomes in magnet® and non-magnet hospitals. *The Journal of Nursing Administration*, 41(10):428.

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

Kutney-Lee, A., McHugh, M., Sloane, D., Cimiotti, J., Flynn, L., Neff, D., and Aiken, L. (2009). Nursing: A key to patient satisfaction. *Health Affairs*, 28(4):w669–w677.

Lake, E. T. (2002). Development of the practice environment scale of the nursing work index. *Res Nurs Health*, 25(3):176–188.

Laschinger, H. K. S. and Leiter, M. P. (2006). The impact of nursing work environments on patient safety outcomes: The mediating role of burnout engagement. *Journal of Nursing Administration*, 36(5):259–267.

Leiter, M. P. and Maslach, C. (2009). Nurse turnover: the mediating role of burnout. *Journal of Nursing Management*, 17(3):331–339.

Liu, K., You, L.-M., Chen, S.-X., Hao, Y.-T., Zhu, X.-W., Zhang, L.-F., and Aiken, L. H. (2012). The relationship between hospital work environment and nurse outcomes in guangdong, china: A nurse questionnaire survey. *Journal of Clinical Nursing*, 21(9-10):1476–1485.

Maslach, C. (1993). Burnout: A multidimensional perspective.

Maslach, C. and Jackson, S. E. (1986). Maslach burnout inventory. University of California, Palo Alto, CA.

Maslach, C., Jackson, S. E., and Leiter, M. P. (1996). *Maslach Burnout Inventory Manual*. Consulting Psychologists Pr, 3rd edition.

Maslach, C. and Leiter, M. P. (2008). Early predictors of job burnout and engagement. *Journal of Applied Psychology*, 93(3):498.

Melchior, M. E., van den Berg, A. A., Halfens, R., Abu-Saad, H. H., Philipsen, H., and Gassman, P. (1997). Burnout and the work environment of nurses in psychiatric long-stay care settings. *Soc Psychiatry Psychiatr Epidemiol*, 32(3):158–164.

Nantsupawat, A., Srisuphan, W., Kunaviktikul, W., Wichaikhum, O.-A., Aungsuroch, Y., and Aiken, L. H. (2011). Impact of nurse work environment and staffing on hospital nurse and quality of care in thailand. *Journal of Nursing Scholarship*, 43(4):426–432.

Neuhaus, J. M. and Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 54(2):638–645.

O'Mahony, N. (2011). Nurse burnout and the working environment. *Emerg Nurse*, 19(5):30–37.

Park, S. and Lake, E. T. (2005). Multilevel modeling of a clustered continuous outcome: Nurses' work hours and burnout. *Nursing research*, 54(6):406.

Poghosyan, L., Aiken, L. H., and Sloane, D. M. (2009). Factor structure of the maslach burnout inventory: An analysis of data from large scale cross-sectional surveys of nurses from eight countries. *Int J Nurs Stud*, 46(7):894–902.

Sermeus, W., Aiken, L. H., den Heede, K. V., Rafferty, A. M., Griffiths, P., Moreno-Casbas, M. T., Busse, R., Lindqvist, R., Scott, A. P., Bruyneel, L., Brzostek, T., Kinnunen, J., Schubert, M., Schoonhoven, L., Zikos, D., and consortium, R. N. A. S. T. (2011). Nurse forecasting in europe (rn4cast): Rationale, design and methodology. *BMC Nurs*, 10:6.

Vahey, D., Aiken, L., Sloane, D., Clarke, S., and Vargas, D. (2004). Nurse burnout and patient satisfaction. *Medical Care*, 42(2 Suppl):II57.

Van Bogaert, P., Meulemans, H., Clarke, S., Vermeyen, K., and Van de Heyning, P. (2009). Hospital nurse practice environment, burnout, job outcomes and quality of care: Test of a structural equation model. *Journal of Advanced Nursing*, 65(10):2175–2185.

Van den Heede, K., Lesaffre, E., Diya, L., Vleugels, A., Clarke, S. P., Aiken, L. H., and Sermeus, W. (2009). The relationship between inpatient cardiac surgery mortality and nurse numbers and educational level: Analysis of administrative data. *International Journal of Nursing Studies*, 46(6):796–803.

Warshawsky, N. E. and Havens, D. S. (2011). Global use of the practice environment scale of the nursing work index. *Nursing Research*, 60(1):17.

Wu, Y.-W. B. (1997). An application of hierarchical linear models to meta-analysis in nursing research. *Nursing Research*, 46(5):295–298.

Appendix

A. Univariate simple multilevel probit model

Let Z_{nuhc} be one of the three latent normally distributed responses, representing the nth nurse, within the uth nursing unit, the hth hospital and the cth country. This variable expresses the true burnout feeling of the nurse. It is assumed that for $Z_{nuhc} > 0$, burnout is expressed on a manifest scale indicated by $Y_{nuhc} = 1$, otherwise 0.

$$\begin{gathered} Z_{nuhc} = \beta_0 + \beta_1 X_{uhc} + \beta_2 X_{hc} + \beta_3 X_c + b_{0c} + b_{0hc} + b_{0uhc} + \varepsilon_{nuhc} \\ b_{0c} \sim N(0, \sigma_1^2),\ b_{0hc} \sim N(0, \sigma_2^2),\ b_{0uhc} \sim N(0, \sigma_3^2),\ \varepsilon_{nuhc} \sim N(0,1), \\ n = 1, 2, ..., N_u,\ u = 1, 2, ..., N_h,\ h = 1, 2, ..., N_c,\ c = 1, 2, ..., 11 \end{gathered} \tag{6.4}$$

B. Multivariate multilevel probit model

For a better understanding, we first introduce the classical multivariate probit model (MVP) (single-level structure). The classical MVP has been widely studied by many researchers with different theories and solutions, see e.g. Bock and Gibbons (1996), Chib and Greenberg (1998), Lawrence et al. (2008). Here we adopt Bock and Gibbons' solution of factor modlling which is formally defined as follows:

$$\begin{gathered} Z_i = \beta_0 + \beta_1 X_i + \lambda_0 F_i + \varepsilon_i, \\ F_i \sim N(0,1),\ \varepsilon_i \sim N(0, \Sigma_\varepsilon),\ i = 1, 2, ..., N, \end{gathered} \tag{6.5}$$

where in our three-variate case, Z_i is a 3×1 vector of latent continuous responses at the ith observation, with the same definition of latent variable as in the univariate case. The vectors of regression coefficients β_0 and β_1 , and the vector of factor loadings λ_0 are of length 3, corresponding to the three outcomes. The common factor Fi serves to model the correlations of the three outcomes in combination with the covariance structure of the residuals ε_i , Σ_ε . We assumed that this covariance matrix is diagonal (errors independent). The covariance matrix of the random part, i.e. $cov(\lambda_0 F_i + \varepsilon_i) = \lambda_0 \lambda_0^T + \Sigma_\varepsilon$, is called the factor analytic representation of the covariance which can reproduce any 3×3 covariance matrix by an appropriate choice of λ_0 and Σ_ε . Note that here the covariance matrix equals the correlation matrix because the variances are assumed to be one. This model assumes that the correlation matrix is the same across all countries, hospitals and nursing units, which might be not a realistic assumption. In order to vary the correlation across countries one could replace $\lambda_0 F_i$ by $(\lambda_0 + \lambda_c) F_i$, whereby λ_c changes with country. In this way there are as many correlation matrices as countries. We can further extend this expression to let the correlation vary also with hospital and nursing unit resulting in a term $(\lambda_0 + \lambda_c + \lambda_{hc} + \lambda_{uhc}) F_{nuhc}$. In this way we have a different correlation matrix for each nursing unit. Because of the large number

of nursing units (2087) we have assumed that for λ_c , λ_{hc} and λ_{uhc} , each has a normal distribution with mean zero and a variance to be estimated, which reduces drastically the number of parameters to estimate but also expresses that we do expect that the correlations across nursing units, hospitals and countries do not vary wildly. We then implemented the four-level structure into the MVP model. This is defined as follows (similar notation as in the univariate multilevel probit model in appendix A):

$$\begin{gathered} Z_{nuhc} = \beta_0 + \beta_1 X_{uhc} + \beta_2 X_{hc} + \beta_3 X_c + b_{0c} + b_{0hc} + b_{0uhc} + \\ (\lambda_0 + \lambda_c + \lambda_{hc} + \lambda_{uhc}) F_{nuhc} + \varepsilon_{nuhc}, \\ b_{0c} \sim N_3(0, \Sigma_1),\ b_{0hc} \sim N_3(0, \Sigma_2),\ b_{0uhc} \sim N_3(0, \Sigma_3), \\ \lambda_c \sim N(0, \Sigma_4),\ \lambda_{hc} \sim N(0, \Sigma_5),\ \lambda_{uhc} \sim N(0, \Sigma_6) \\ F_{nuhc} \sim N(0,1),\ \varepsilon_{nuhc} \sim N(0, \Sigma_6), \\ n = 1, 2, ..., N_u,\ u = 1, 2, ..., N_h,\ h = 1, 2, ..., N_c,\ c = 1, 2, ..., 11 \end{gathered} \tag{6.6}$$

In this model, all observed and latent variables have a multilevel structure with multiple subscripts defined in the same way as before. The covariates are partitioned into three parts, as was done in the univariate model. There are three random intercepts vectors corresponding to the three higher levels for each outcome, as well as the three random factor loadings. As the factor is introduced to model the correlations among the three outcomes, the varying factor loadings imply a varying correlation structure. Restrictions are needed to render the model identifiable, which are of the same type as above.

7

A multivariate multilevel Gaussian model with a mixed effects structure in the mean and covariance part

Chapter 7 is based on the paper:

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.

Abstract

A traditional Gaussian hierarchical model assumes a nested multilevel structure for the mean and a constant variance at each level. We propose a Bayesian multivariate multilevel factor model that assumes a multilevel structure for both the mean and the covariance matrix. That is, in addition to a multilevel structure for the mean we also assume that the covariance matrix depends on covariates and random effects. This allows to explore whether the covariance structure depends on the values of the higher levels and as such models heterogeneity in the variances and correlation structure of the multivariate outcome across the higher level values. The approach is applied to the three-dimensional vector of burnout measurements collected on nurses in a large European study to answer the research question whether the covariance matrix of the outcomes 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. Simulations illustrate the performance of our modeling approach.

7.1 Introduction

In this paper we are interested in modeling a multivariate multilevel Gaussian data structure. Our modeling approach is inspired by research questions that were formulated in the context of the Registered Nurse Forecasting (RN4CAST) project (Sermeus et al., 2011). This European FP7-funded nurse workforce study was conducted from 2009 to 2011 in twelve countries in Europe and involved a large number of hospitals, nursing units and nurses. The aim of the project was to study the impact of system-level features in the organization of nursing care (work environment, educational level, workload) on individual measures of nurse wellbeing (burnout, job satisfaction, turnover) and patient safety outcomes and satisfaction with care. As outcome we have chosen three classically used burnout scores developed about twenty years ago in the US from a 22-item questionnaire (Maslach et al., 1996). These scores have been used intensively in the nursing research literature and are perceived to represent well the burnout process.

It was of interest to know how each of the burnout outcomes depend on country, hospital, nursing unit and nurse characteristics. Such dependencies can be explored by fitting a multilevel model to each of the three outcomes. However such a model only explores the dependence of the mean on the covariates. It is equally important to see whether these factors also impact the variability of burnout. Indeed a high variability of burnout within, say a hospital, may also affect the quality of care in that hospital. Furthermore, we were curious whether the relationship of the three burnout outcomes remained constant across the values of the different levels of the multilevel structure. In particular we were interested in looking for factors that alter the correlation structure of the burnout measurements. Such determinants may shed light on whether the dimensions of burnout vary with covariates. Finally, it is important to realize that in the RN4CAST study, the 22-item questionnaire was translated into eleven languages. Although translation was done by experts, it may still happen that the interpretation of the questions depends on local, say ethnic or cultural, differences. Since no information on such factors was recorded in the RN4CAST study, we wished to explore the variability of the correlation structure of the burnout outcomes across the different levels of the data. To explore the dependence of the covariance matrix on recorded and not-recorded covariates we propose here to extend the classical multivariate multilevel model with a covariance matrix that may depend on covariates and random effects.

For the 2-level model, various approaches were suggested in the literature to allow the variance function of the random effect and/or of measurement error to depend on covariates, see e.g. Ibáñez et al. (1999); Foulley et al. (1990); Lin et al. (1997). Approaches were also suggested whereby the variance structure contains random effects. Foulley et al. (1992) proposed a Bayesian linear mixed model whereby the measurement error variance also has a mixed model structure. Foulley and Gianola (1996) then further extended this model to generalized linear mixed models. Kizilkaya and Tempelman (2005) applied a full Bayesian structural mixed effects multiplicative model for residual variances in a generalized linear mixed model. Lee and Nelder (2006) proposed a DHGLM (Double Hierarchical Generalized Linear Model), which models both mean and the residual variance (overdispersion)

with random effects. Further, Lee and Noh (2012) extended the DHGLM to model mean, residual variance, and also variance of random effects, with random effects.

For a multivariate Gaussian response, extensions of the classical regression model have been suggested whereby the covariance matrix is allowed to depend on covariates. In this respect, one may model the covariance matrix directly as a function of covariates or first split up the covariance matrix into the correlation and variance part and model both parts separately. The major challenge is to ensure that the covariance matrix remains positive definite for all covariate values. Chiu et al. (1996) proposed a way of directly modeling the logarithmic transformation of the covariance matrix. Although it is a flexible approach without causing positive definite problems, the interpretation of the model parameters is often quite difficult. A popular approach was suggested by Pourahmadi (1999) who used the separation strategy and suggested a modified Cholesky decomposition of the covariance matrix, i.e. $T\Sigma T^T = D$, and then modeled the elements of T and D as a function of covariates. The interpretation of the parameters is meaningful, however, only when there exists a natural ranking of the responses as in longitudinal studies or time series studies. For applications of this approach, see e.g. Daniels and Pourahmadi (2002) and Cecere et al. (2006). Barnard et al. (2000) suggested another decomposition which separates the covariance matrix into the (classical) variance and correlation part, i.e. $\Sigma = \text{diag}(S)\, R\, \text{diag}(S)$. It also allows for heterogeneous covariance matrices across groups, e.g. gender or age groups. The computation, however, is intensive for moderate and large sample size. More recently, Hoff and Niu (2012) proposed a covariance regression model to allow for heterogeneity in the variance part of the classic multivariate regression model. They suggested the model $\Sigma_{\boldsymbol{x}} = A + B\boldsymbol{x}\boldsymbol{x}^T B^T$, whereby A is a "baseline" positive definite matrix and B a matrix of regression coefficients. Fox and Dunson (2011) suggested a Bayesian non-parametric covariance regression that could efficiently reduce the high parameter dimensionality, which is especially useful when the dimensionality of response is high.

For multivariate multilevel models, the literature lacks modeling approaches that allow the covariance matrix to depend on covariates and/or random effects. In this paper we generalized the approach of Hoff and Niu by specifying a factor analytic model to the three-dimensional response with the factor loadings depending on covariates and random effects. The covariance matrix could then be built up through the factor loadings to have a complex structure. This somewhat resembles structural equation modeling (SEM) which aims at 1) understand the patterns of covariances among a set of observed variables and 2) explain as much of their variance as possible with the researcher's model (Kline, 2010). The difference of our approach with SEM will be discussed at the end.

In Section 7.2 we provide further details on the motivating data set and introduce the research questions that triggered our modeling approach(es). In Section 7.3 we elaborate on a factor-analytic approach to model the covariance matrix in a hierarchical way. Both covariates as well as random effects are incorporated into the covariance structure. In Section 7.4 we indicate how our approach can be generalized to more than three responses. The MCMC procedure to estimate the model parameters is discussed in Section 7.5. Section 7.6 focuses on the impact of the non-response issue and the handling of missingness in the

response and the covariates. In Section 7.7 the multivariate multilevel model is applied to the motivating RN4CAST data set and the research questions posed in Section 7.2 are addressed. Simulation studies to illustrate the performance of our modeling approach(es) are described in Section 7.8. We give concluding remarks in Section 7.9.

7.2 Motivating data set: the RN4CAST project

7.2.1 Description of the project

The RN4CAST project is a three year (2009-2011) nurse workforce study involving 33731 registered nurses in 2169 nursing units in 486 hospitals in 12 European countries. Multilevel sampling was conducted such that within each of the 12 countries, a minimum of 30 general (non-specialized) hospitals were randomly selected, except for Ireland and Norway where the selected hospitals represented all of the relevant institutions. At least two adult general medical and surgical nursing units for each hospital were randomly selected, since the link of nurses' workload and work environment to patient safety and clinical outcomes is best documented in these types of nursing units. All nurses involved in direct patient care activities were then invited to the study. The overall response rate was 62%. While there is a considerable non-response rate, it compares favorably with rates seen in other nursing outcomes research studies of this scale (Aiken et al., 2002). Hospital level response rates exceeded 50% for all countries except for Greece (42%) and the Netherlands (37%). All nursing units that were randomly selected within the participating hospitals agreed to participate to the study. Individual nurse response rates across countries were consistently high, except for England (38.6%), Finland (46.2%) and Germany (41.6%). Swedish data were excluded as no unit identifiers were available from the Swedish sampling design. For more details on the sampling strategy, see Sermeus et al. (2011). In Section 7.6 we return to the possible impact of the non-response on scientific conclusions from the RN4CAST study.

Burnout was measured using the 22-item Maslach Burnout Inventory (MBI) with each item having a seven-point Likert scale (from never to every day coded from 0 to 6) on the frequency of burnout experiences, e.g. "I feel emotionally drained from my work". Maslach et al. (1996) extracted three main dimensions of burnout: *emotional exhaustion (EE)*, *depersonalization (DP)* and *reduced personal accomplishment (PA)*. These three dimensions are sum scores obtained from the original MBI scale. There are about 10% of the nurses having at least one of the 22 items missing. In Section 7.6 we elaborate on how we dealt with the missing part in the response. Burnout is indicated by higher scores on *EE* and *DP*, and lower scores on *PA*, but we reversed the code for *PA* for interpretational purposes (for the three burnout measurements a large value means then more burnout). The crude correlations among these dimensions are 0.56 for *EE* and *DP*, 0.28 for *EE* and *PA*, and 0.32 for *DP* and *PA*, somewhat higher than 0.52, 0.22 (reversed) and 0.26 (reversed), as reported by Maslach et al. (1996). It is assumed in the literature that the three dimensions describe relatively well burnout.

The survey battery including the MBI was translated into eleven languages from its original American English version while ensuring its relevance to the nursing practice and health

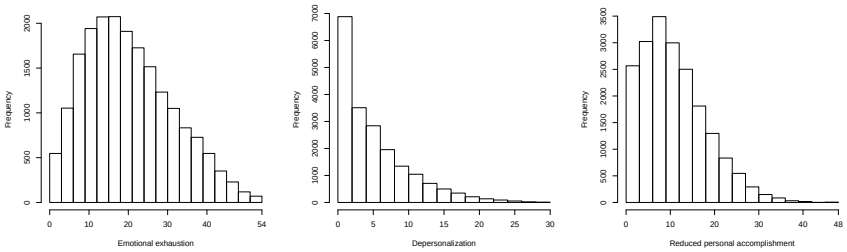

Figure 7.1: Histograms of the three burnout dimensions

care contexts of twelve countries (Squires et al., 2013). Reasonable methodological attempts were therefore taken to reduce bias to allow for comparability of concepts across countries. A translation manager was appointed to ascertain high standards of instrument translation that reduce item bias. Questions were worded in a similar manner and direction across all countries. Construct bias was reduced by assessing dissimilarity of constructs in the investigated countries through the application of content validity indexing procedures including bilingual nurse workforce experts. It was argued that these procedures allowed for a sound and rigorous qualitative examination of the meaning of items.

7.2.2 Descriptive statistics

The three burnout measurements are sums of scores on items recorded on a Likert scale and have therefore a discrete nature. The histograms of the three burnout measurements are shown in Figure 7.1 and indicate quite skewed distributions.

We are interested in establishing the relationship between the burnout measurements and nurse, nursing unit, hospital and country characteristics. Some descriptive statistics of these covariates are shown in Table 8.1. For the variables *working experience* and *work environment*, we report in that table for each level the mean of each covariate (taken as mean of the means at the lower level, i.e. are aggregated at each of the higher levels) and the range of its (mean) values. The covariate *working experience* is a nurse level variable, expressing the working years of being a registered nurse. From Table 8.1 we can conclude that at nurse level, working experience ranges from less than one year to about 50 years. The variation does not narrow much at nursing unit level; while at hospital level, the minimum average working experience for nurses is around 5 years, and increases to 9 years at country level. The overall mean working experience is around 14 years. The *work environment* covariate is an overall average summary of work environment based on the Practice Environment Scale of the Nursing Work Index (PES-NWI) (Lake, 2002). For each item of this covariate (e.g. "Praise and recognition for a job well done") a score on a four-point Likert scale must be supplied, i.e. "Totally agree"=4, "Agree"=3, "Not agree"=2, "Totally not agree"=1 such that high values reflect a positive environment. The mean work environment is around 2.5, which represents an overall neutral feeling about work environment. This covariate is

quite stable across countries but varies more across nurses and nursing units. The covariate *size* represents the total number of beds at hospital level (varying from 30 to 3,000) and the number of participating nurses at the nursing unit level (varying between 1 and 80). The type of hospital and type of nursing unit appear to play a role for burnout (Servellen and Leake, 1993) and are therefore considered here as possible factors that influence burnout. We considered two hospital characteristics: *teaching status* (a university hospital = 1, or not = 0) and *technology level* (heart surgery and/or transplant surgery available = 1, or not = 0). These two types of hospitals constitute around 24% and 29% of all hospitals, respectively. The nursing units were classified as either surgical (about 50%) or medical, whereby around 4.7% of the nursing units recorded as both surgical and medical were classified as medical.

Working experience was not reported for about 6.3% of the nurses. For the other covariates there was less than 2% missing. In Section 7.6 we detail on how we treated the missing covariate values. Finally, there were only 7% male nurses and burnout might be strongly related to gender, we preferred to consider a more homogeneous group and restricted our analysis to the female nurses. As a result, Table 8.1 is based on 21016 nurses, coming from 2023 nursing units within 345 hospitals in 11 countries.

Table 7.1: Descriptive statistics of the considered covariates in the statistical models.

	Working experience(yrs)*	Work environment*	Size*†	Teaching hospital‡	Technical hospital‡	Surgery nursing unit‡
Country	13.9 (9.1,18.8)	2.5 (2.3,2.9)	–	–	–	–
Hospital	14.3 (5.1,27.8)	2.5 (1.7,3.3)	483.6 (30,3213)	23.8%	29.0%	–
Nursing unit	13.9 (0.3,41.0)	2.5 (1.4,3.6)	11.4 (1,71)	–	–	49.9%
Nurse	13.9 (0.1,50.0)	–	–	–	–	–

∗: Mean (and range)
†: No. of beds at hospital level and No. of available nurses at nursing unit level
‡: Percentage

7.2.3 Research questions

The following research questions regarding burnout among nurses emerged:

- Question 1: How much variability does each of the three burnout measurements show across countries, hospitals (within countries), nursing units (within countries and hospitals) and nurses (within countries, hospitals and nursing units)?

- Question 2: How much of the variability can be explained with the covariates recorded at the different levels?

- Question 3: Does the covariance matrix (and more precisely the correlation) between the three burnout dimensions remain the same across countries, hospitals, nursing

units and even nurses after accounting for a rich set of confounders at the different levels?

To answer these research questions, we gradually developed altogether five models thereafter denoted as *Models 1* to *5*. The relatively simpler models, i.e. *Models 1* and *2* will be described in the rest of this section, while the more complex models, i.e. *Models 3* to *5* constituting the innovated part of our analysis, will be elaborated in the next section.

The first question involves a classical Gaussian multilevel analysis for each burnout measurement separately. Note that the burnout measurements have, in principle, a discrete nature but with many possible values. In Section 7.7 we use a latent continuous scale to analyze the data. To simplify matters, we assume for now that the response is continuous and has a Gaussian distribution. A classical four-level hierarchical structure may be considered here in which nurses were selected from nursing units within hospitals within countries. When the three burnout measurements are analyzed jointly, the model turns into

$$\textit{Model 1:} \qquad \boldsymbol{y}_{ijkl} = \boldsymbol{\mu}_0 + \boldsymbol{u}_{jkl} + \boldsymbol{u}_{kl} + \boldsymbol{u}_l + \boldsymbol{\delta}_{ijkl}, \tag{7.1}$$

$$\textit{with} \qquad \boldsymbol{u}_{jkl} \sim N(\boldsymbol{0}, \Sigma_u), \boldsymbol{u}_{kl} \sim N(\boldsymbol{0}, \Sigma_h), \boldsymbol{u}_l \sim N(\boldsymbol{0}, \Sigma_c), \boldsymbol{\delta}_{ijkl} \sim N(\boldsymbol{0}, \Sigma_\delta),$$

where $\boldsymbol{y}_{ijkl}$ represents the vector of three burnout measurements, taken on the ith nurse coming from the jth nursing unit in the kth hospital in the lth country. The subscripts u, h and c represent the nursing unit, hospital and country level, respectively. In addition to the normality assumption in model (7.1) the random components $(\delta_{ijkl}, u_{jkl}, u_{kl}, u_l)$ are assumed to be statistically independent. Σ_u is the 3×3 covariance matrix of the random vector $\boldsymbol{u}_{jkl}$ (and similarly for Σ_h, Σ_c and Σ_δ). The classical multivariate multilevel random effects model is well introduced by Goldstein (2010), among others. In the remainder of the paper we refer to the above multivariate model as *Model 1*.

In Question 2 we are interested to see whether the variability in the means of burnout measurements across countries, hospitals and nursing units can be explained by demographic variables or organizational features of nursing care, such as those listed in Table 8.1. To account for these covariates, *Model 1* is extended as follows:

$$\textit{Model 2:} \qquad \boldsymbol{y}_{ijkl} = \boldsymbol{B}\boldsymbol{x}_{ijkl} + \boldsymbol{u}_{jkl} + \boldsymbol{u}_{kl} + \boldsymbol{u}_l + \boldsymbol{\delta}_{ijkl}, \tag{7.2}$$

with $\boldsymbol{x}_{ijkl}$ representing a vector of nurse-specific, nursing unit-specific, hospital-specific and country-specific covariates with regression matrix $\boldsymbol{B}$. We will refer to this model as *Model 2*.

Question 3 refers to the covariance matrix of $\boldsymbol{\delta}_{ijkl}$, expressing a so-called intrinsic variability and relationship of the three burnout measurements not explained by the covariates and random effects in Equation (7.2). In *Models 1* and *2*, it is assumed that this variability can be represented by Σ_δ. Thus the associated correlation matrix is assumed constant across countries, hospitals and nursing units. However, it is of interest to know whether the correlation matrix varies with countries and hospitals. To illustrate this, suppose that the three

burnout measurements are basically uncorrelated in one country while they are highly correlated in another country. Then this could imply that, due to cultural or other differences, the 22-item MBI is interpreted differently in these countries. Variability in Σ_δ (and its correlation matrix) can be addressed by allowing it to depend on covariates, which leads to *Model 3* (defined in Section 7.3.1). But, since typically such covariates are measured with error, and/or we may not know all relevant covariates, additional random effects might be needed to explain this variability. Therefore we introduce *Model 4* in Section 7.3.1. To the best of our knowledge these extensions have not been suggested in the literature. The focus in this paper is therefore on exploring the behavior of *Model 3* and even more of *Model 4*, in general and in particular on the RN4CAST data. More specifically, we are interested to see how the correlation matrix of the three burnout dimensions depends on covariates and random effects representing the unexplained variability at higher levels. This enables us to judge the stability of the correlations across the levels and to evaluate the so-called intrinsic correlations (in the absence of the previously mentioned fixed and random confounders) of the three burnout dimensions.

An alternative approach to a 4-level multilevel model could be a multiple group 3-level multilevel model, whereby country is treated as a fixed effect rather than a random effect. Treating country as a random effect may be problematic when there are a few number of countries, as in our case (Maas and Hox, 2005; Meuleman and Billiet, 2009). Therefore, we also applied the multiple group model, referred to as *Model 5* (defined in Section 7.3.1), to the motivating data set.

Note that we could have assessed also the heterogeneity of Σ_v (with $v = u, h, c$), i.e. $\Sigma_v = \Sigma_v(\boldsymbol{x}^*)$. However, in this paper, homogeneity of Σ_v will be assumed. We return to this possible extension in the discussion section of the paper.

7.3 A single factor Model

To explore the covariance structure of the burnout dimensions, as required to address Question 3 of the previous section, the classical multilevel model needs to be extended. For multivariate multilevel models, the literature lacks modeling approaches that allow the covariance matrix to depend on covariates and/or random effects. In this section, we suggest a possible way to incorporate structure in the covariance matrix of the three-variate multilevel model of burnout variables of the motivating data set. Our model is a combination of a Gaussian hierarchical model and a factor model to allow for multilevel structures in both the means and the covariance matrix. While a likelihood method can be invoked to estimate the model parameters, we aim in this paper for a Bayesian approach using a Markov chain Monte Carlo (MCMC) technique.

7.3.1 Definition of models

A classical multilevel model assumes a mixed effects structure only in the mean part of the model. However, it makes also sense to allow for a multilevel structure in the variance part of the model. This was done by Foulley et al. (1992) for a univariate Gaussian hierarchical

structure. Their model assumes:

$$\begin{aligned} y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{u}_j + \delta_{ij}, \quad \boldsymbol{u}_j \sim N(\mathbf{0}, \Sigma_u), \quad \delta_{ij} \sim N(0, \sigma_{ij}^2), \\ \sigma_{ij}^2 = \exp(\boldsymbol{x}_{ij}^{*T}\boldsymbol{\beta}^* + \boldsymbol{z}_{ij}^{*T}\boldsymbol{u}_j^*), \quad \boldsymbol{u}_j^* \sim N(\mathbf{0}, \Sigma_u^*), \end{aligned} \tag{7.3}$$

where y_{ij} represents the response of the ith subject in the jth group. Both the mean as well as the variance part of the model are expressed as a mixed model. In the mean part, $\boldsymbol{x}_{ij}$ is a vector containing covariates from each level with $\boldsymbol{\beta}$ its associated vector of fixed effects and $\boldsymbol{z}_{ij}$, $\boldsymbol{u}_j$ represent the covariates and random effects respectively in the random part. To allow for heterogeneity in the measurement error that is partly explained by covariates $\boldsymbol{x}_{ij}^*$ and partly unexplained, the logarithm of the residual variance is regressed on fixed and random effects. Note that the covariates $\boldsymbol{x}_{ij}^*$ and $\boldsymbol{z}_{ij}^*$ in the variance part may differ from the corresponding covariates in the mean part of the model.

In a multivariate but single-level Gaussian regression model, Hoff and Niu (2012) introduced heterogeneity in the covariance part by allowing the covariance matrix of the response to depend on covariates. Their proposed $rank - 1$ covariance regression model is given by:

$$\begin{aligned} \boldsymbol{y}_i = \boldsymbol{B}\boldsymbol{x}_i + F_i \times \boldsymbol{B}^*\boldsymbol{x}_i^* + \boldsymbol{\varepsilon}_i, \\ \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \Sigma_\varepsilon), \quad F_i \sim N(0, 1), \end{aligned} \tag{7.4}$$

where $\boldsymbol{y}_i$ represents the multivariate response for the ith subject, $\boldsymbol{B}$ is the regression matrix associated with the covariates and $\boldsymbol{x}_i$ is the covariate vector of the ith subject. F_i is assumed to be common for each dimension of $\boldsymbol{y}_i$, and follows a standard normal distribution. $\boldsymbol{B}^*\boldsymbol{x}_i^*$ could be seen as the coefficient of F_i, which is referred to as the factor loading. The covariance matrix for the response $\boldsymbol{y}_i$ is then:

$$\Sigma_i = \boldsymbol{B}^*\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}\boldsymbol{B}^{*T} + \Sigma_\varepsilon. \tag{7.5}$$

The covariance matrix Σ_i could be interpreted as the sum of a baseline covariance matrix Σ_ε and a part that depends on covariate $\boldsymbol{x}_i^*$.

Model 3 extends model (7.4) to the multilevel context. Heterogeneity in the covariance part is then expressed by incorporating covariates into the covariance matrix. *Model 3* is given by:

$$\begin{aligned} \textit{Model 3:} \qquad \boldsymbol{y}_{ijkl} = \boldsymbol{B}\boldsymbol{x}_{ijkl} + \boldsymbol{u}_{jkl} + \boldsymbol{u}_{kl} + \mathbf{u}_l + \boldsymbol{\delta}_{ijkl}, \\ \boldsymbol{\delta}_{ijkl} = \boldsymbol{\Lambda}_{ijkl}F_{ijkl} + \boldsymbol{\varepsilon}_{ijkl}, \quad \boldsymbol{\Lambda}_{ijkl} = \rho(\boldsymbol{B}^*\boldsymbol{x}_{ijkl}^*), \\ \textit{with} \qquad \boldsymbol{u}_{jkl} \sim N(\mathbf{0}, \Sigma_u), \quad \boldsymbol{u}_{kl} \sim N(\mathbf{0}, \Sigma_h), \quad \boldsymbol{u}_l \sim N(\mathbf{0}, \Sigma_c), \\ F_{ijkl} \sim N(0, 1), \quad \boldsymbol{\varepsilon}_{ijkl} \sim N(\mathbf{0}, \Sigma_\varepsilon), \end{aligned} \tag{7.6}$$

where $\boldsymbol{y}_{ijkl}$ represents a vector of p responses coming from the ith nurse within the jth nursing unit from the kth hospital in the lth country, $\boldsymbol{B}$ is a $p \times q$ matrix of fixed effects associated with the q-dimensional vector $\boldsymbol{x}_{ijkl}$ gathering information from all levels, while $\boldsymbol{u}_{jkl}$, $\boldsymbol{u}_{kl}$ and $\boldsymbol{u}_l$ represent the p-dimensional random intercepts at each higher level with general covariance matrices $\Sigma_u, \Sigma_h, \Sigma_c$, respectively. The within-nursing unit residuals $\boldsymbol{\delta}_{ijkl}$ are decomposed into a fixed part assumed constant across nurses with general covariance matrix Σ_ε and a part that varies with q^* characteristics $\boldsymbol{x}^*_{ijkl}$ possibly different from $\boldsymbol{x}_{ijkl}$. $\boldsymbol{B}^*$ is a $p \times q^*$ matrix of fixed effects associated with $\boldsymbol{x}^*_{ijkl}$. The link function $\rho(\cdot)$ applies elementwise on the $q^* \times 1$ vector. When taken the identity function, *Model 3* is a generalization of Hoff and Niu's $rank - 1$ covariance regression model (Hoff and Niu, 2012) to a multilevel setting. Other functions are possible such as the (elementwise) exponential function, see also Sections 7.3.3 and 7.3.4. We assume in addition that all random components in above model are mutually independent. The factor analytic representation of the covariance matrix has the advantage that the impact of covariates is easily included in the covariance matrix via the factor loadings while retaining the positive definiteness property. In addition, the interpretation of the impact of covariates on the covariance matrix is relatively easy as seen in Section 7.3.2.

The covariance matrix for the residual part of *Model 3*, i.e. of $\boldsymbol{\delta}_{ijkl}$, (conditional on the random effects) is given by:

$$\Sigma_{ijkl} = \rho(\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl}) \rho(\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl})^T + \Sigma_\varepsilon. \tag{7.7}$$

It is readily seen that Σ_{ijkl} is positive definite when Σ_ε satisfies this property.

A next extension consists of including random effects into the covariance structure of the residual part for reasons stated in Section 7.2.3. Therefore we extend *Model 3* to *Model 4* which involves adapting the factor loadings matrix as follows:

$$\begin{aligned} \textit{Model 4:} \qquad & \boldsymbol{y}_{ijkl} = \boldsymbol{B}\boldsymbol{x}_{ijkl} + \boldsymbol{u}_{jkl} + \boldsymbol{u}_{kl} + \mathbf{u}_l + \boldsymbol{\delta}_{ijkl}, \\ & \boldsymbol{\delta}_{ijkl} = \boldsymbol{\Lambda}_{ijkl} F_{ijkl} + \boldsymbol{\varepsilon}_{ijkl}, \quad \boldsymbol{\Lambda}_{ijkl} = \rho(\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl} + \boldsymbol{u}^*_{jkl} + \boldsymbol{u}^*_{kl} + \boldsymbol{u}^*_l), \end{aligned} \tag{7.8}$$

with random effects $\boldsymbol{u}^*_{jkl} \sim N(\mathbf{0}, \Sigma^*_u), \boldsymbol{u}^*_{kl} \sim N(\mathbf{0}, \Sigma^*_h), \boldsymbol{u}^*_l \sim N(\mathbf{0}, \Sigma^*_c)$. Again we assume mutual independence of all random components of the model. This model allows the covariance matrix of $\boldsymbol{\delta}_{ijkl}$ to depend on the different levels beyond what is explained by the covariates $\boldsymbol{x}^*_{ijkl}$. Now the covariance matrix for the residual component of the model, $\boldsymbol{\delta}_{ijkl}$, given the fixed and random effects is given by:

$$\Sigma_{ijkl} = \rho(\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl} + \boldsymbol{u}^*_{jkl} + \boldsymbol{u}^*_{kl} + \boldsymbol{u}^*_l) \rho(\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl} + \boldsymbol{u}^*_{jkl} + \boldsymbol{u}^*_{kl} + \boldsymbol{u}^*_l)^T + \Sigma_\varepsilon. \tag{7.9}$$

Note that including random effects into the covariance structure may allow to model abrupt changes in the variability and/or correlation of the burnout measurements across the units at the different levels.

In expressions (7.7) and (7.9), the covariance matrix at nurse level is split up into a part that is "explained" by covariates or random effects at the higher level. It is of main interest

to see how variability is explained by these factors, but also how stable the residual covariance matrix Σ_ε remains. The latter covariance matrix could be considered representing the intrinsic variances and correlations of the three burnout dimensions in the absence of the fixed and random confounders.

In the RN4CAST study, there are only 12 countries (and 11 involved in the study). An alternative approach to a four-level model is a three-level multiple group model. This is our *Model 5* which differs from *Model 4* in that the *country* variable is treated as a categorical covariate and belongs to the fixed effects part in both the mean and the covariance part. Hence *Model 5* is of the form of *Model 4* but with one level less and one categorical covariate (implying 10 binary covariates) extra:

$$\begin{aligned} \textit{Model 5:} \qquad & \boldsymbol{y}_{ijk} = \boldsymbol{B}\boldsymbol{x}_{ijk} + \boldsymbol{B}_c I_k + \boldsymbol{u}_{jk} + \boldsymbol{u}_k + \boldsymbol{\delta}_{ijk}, \\ & \boldsymbol{\delta}_{ijk} = \boldsymbol{\Lambda}_{ijk} F_{ijk} + \boldsymbol{\varepsilon}_{ijk}, \quad \boldsymbol{\Lambda}_{ijk} = \rho(\boldsymbol{B}^* \boldsymbol{x}^*_{ijk} + \boldsymbol{B}^*_c I_k + \boldsymbol{u}^*_{jk} + \boldsymbol{u}^*_k), \end{aligned} \tag{7.10}$$

where I_k represents the vector of the 10 binary covariates indicating the country hospital k belongs to, with $\boldsymbol{B}_c$ and $\boldsymbol{B}^*_c$ its coefficient matrices in the mean part and the loadings, respectively. See Table 7.2 for an overview of the five considered models.

Table 7.2: Expressions of *Models 1* to *5*

Models	Mean part	Covariance part
Model 1	$\boldsymbol{y}_{ijkl} = \boldsymbol{\mu}_0 + \boldsymbol{u}_{jkl} + \boldsymbol{u}_{kl} + \boldsymbol{u}_l + \boldsymbol{\delta}_{ijkl}$	–
Model 2	$\boldsymbol{y}_{ijkl} = \boldsymbol{B}\boldsymbol{x}_{ijkl} + \boldsymbol{u}_{jkl} + \boldsymbol{u}_{kl} + \boldsymbol{u}_l + \boldsymbol{\delta}_{ijkl}$	–
Model 3	$\boldsymbol{y}_{ijkl} = \boldsymbol{B}\boldsymbol{x}_{ijkl} + \boldsymbol{u}_{jkl} + \boldsymbol{u}_{kl} + \boldsymbol{u}_l + \boldsymbol{\delta}_{ijkl}$	$\boldsymbol{\delta}_{ijkl} = \rho(\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl}) F_{ijkl} + \boldsymbol{\varepsilon}_{ijkl}$
Model 4	$\boldsymbol{y}_{ijkl} = \boldsymbol{B}\boldsymbol{x}_{ijkl} + \boldsymbol{u}_{jkl} + \boldsymbol{u}_{kl} + \boldsymbol{u}_l + \boldsymbol{\delta}_{ijkl}$	$\boldsymbol{\delta}_{ijkl} = \rho(\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl} + \boldsymbol{u}^*_{jkl} + \boldsymbol{u}^*_{kl} + \boldsymbol{u}^*_l) F_{ijkl} + \boldsymbol{\varepsilon}_{ijkl}$
Model 5	$\boldsymbol{y}_{ijk} = \boldsymbol{B}\boldsymbol{x}_{ijk} + \boldsymbol{B}_c I_k + \boldsymbol{u}_{jk} + \boldsymbol{u}_k + \boldsymbol{\delta}_{ijk}$	$\boldsymbol{\delta}_{ijk} = \rho(\boldsymbol{B}^* \boldsymbol{x}^*_{ijk} + \boldsymbol{B}^*_c I_k + \boldsymbol{u}^*_{jk} + \boldsymbol{u}^*_k) F_{ijk} + \boldsymbol{\varepsilon}_{ijk}$

We note that Li et al. (2013) used the same data set but their 3-variate 4-level probit model model was based on binarized burnout measurements and only allowed for heterogeneity of the covariance part via random effects. In addition, in that paper the properties of that model were not explored. We also note that the idea of implementing a factor model into a multivariate regression modeling is not new. Bock and Gibbons (1996), Gibbons and Lavigne (1998) and Gibbons and Wilcox-Gök (1998) analyzed the multivariate probit model via a factor analytic model but only for a single level model. Muthén (1994) proposed a maximum likelihood analysis of the covariance structure via a two-level factor model. The novelty of our approach is that the factor loadings are now allowed to depend on covariates and random effects and in a multilevel context.

7.3.2 Interpretation of model parameters

The effect of covariates and random effects is assumed to be linear (up to the link function ρ) on the factor loadings. Hoff and Niu (2012) provided a geometrical interpretation of the co-

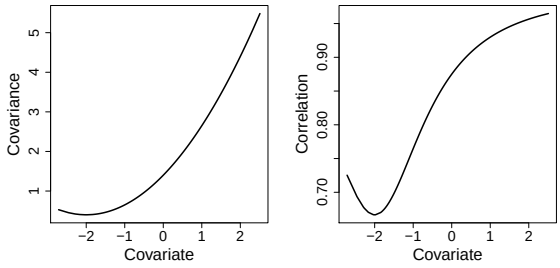

Figure 7.2: Relationship between covariance/correlation and covariate based on expression (7.7)

variates in their multivariate regression model. In addition they showed that, by adding an intercept term to the covariate structure $\boldsymbol{x}^*_{ijkl}$, the variance of $\boldsymbol{\delta}_{ijkl}$ can be either increasing or decreasing with covariates. The geometrical interpretation applies also to our extensions, at least for the identity link. It is straightforward to see from expressions (7.7) and (7.9) that the relationship between the covariance entries and the covariates and/or random effects is quadratic. This is illustrated in the LHS of Figure 7.2, obtained from a simulated data set based on *Model 3* with a linear input of the covariate. Changing the scale of the covariate (say by taking the square root) allows for other (non-linear) relationships. The relationship between the correlation and the covariates is somewhat more complicated (Figure 7.2), but one can state that roughly the same behavior is seen, i.e., when the covariance increases or decreases with a covariate so does the correlation.

7.3.3 Identifiability of the model

It is well-known that there are identifiability issues in the above factor model. The reason is that in $\boldsymbol{\Lambda}_{ijkl}F_{ijkl}$, $\boldsymbol{\Lambda}_{ijkl}$ is known up to the sign since the common factor F is assumed to have a standard normal distribution and hence can be positive or negative. This implies for the identity link, e.g., that if $\boldsymbol{B}^*$ is a solution for *Model 3*, then also $-\boldsymbol{B}^*$. This is called the "flipping states" issue, see Maydeu-Olivares and McArdle (2005), which means that the factor loadings could be either positive or negative but with the same absolute value. A maximizing algorithm finds only one of the two possible values of $\boldsymbol{B}^*$, and as such does not pose numerical complications. However, there are more problems with a simulation-based method, like for the MCMC sampling approach we used here. Indeed, the sampling algorithm may move between the two solutions and never converge (Browne, 2012). This is an issue with *Models 3* and *4*, when ρ is the identity link. In Section 7.3.4 we suggest to use a mixture prior for this choice of link function. For an exponential link function, the factor loadings are always positive and there is no "flipping states" problem. In that case, it is assumed that the covariates and random effects have a multiplicative effect on the factor loadings.

When there is enough variability in the covariates, Hoff and Niu proved that in their

model all parameters are identifiable up to a sign. For *Model 3*, we sketch in Appendix 7.9 that this result must also hold. For *Model 4*, things are more complex. To prove likelihood identifiability in mixed models is generally complex, as indicated in Wu (2010). In addition, the random effects for *Model 4* are only determined up to the sign, which can be seen from the trace plots. Post-processing of the Markov chain of each random effect in the covariance part was therefore necessary to obtain its posterior mean (up to the sign), for this we used the R function *normalmixEM* from the *mixtools* package. Then a QQ-plot based on a folded normal distribution was used to assess the normality assumption (up to the sign) of these random effects. More details can be found in the Supplementary Material [1].

7.3.4 Priors

We have opted for a Bayesian approach to estimate the model parameters, which requires a prior distribution for all its parameters. One of the benefits of applying a Bayesian approach is that prior information can be incorporated into the analysis, if available. Two problems were encountered to include prior information into the model. The first problem relates to the way we deal with the bounded outcome score response (more details in Section 7.7). The parameter estimates from this approach are difficult to compare with those obtained from a logistic model (using binary or categorical responses) or from a linear model using the original burnout scores. Secondly, it is hard to imagine what should constitute reasonable values of the parameters in the factor analytic model of the covariance structure. However, because of the huge sample size of our study, practically each prior information on the parameters has a negligible effect on the posterior estimates. That is why we finally chose a vague normal prior (with mean zero and variance 10^6) for regression parameters ($\boldsymbol{B}$-parameters) and a vague inverse Wishart prior (with small diagonal values, say 0.01, for the scale matrix and degrees of freedom equal to the dimension of the matrix (Lesaffre and Lawson, 2012)) for the covariance matrices of the random effects in the mean and the loadings, i.e. Σ_v and Σ_v^* (with $v = u, h, c$). Note that constraining the off-diagonal elements to zero may speed up the convergence of the chain considerably.

It is, however, more difficult to specify the prior distribution for the $\boldsymbol{B}^*$-parameters in the factor loadings because of the "flipping states" identifiability problem mentioned in Section 7.3.3. This identification problem needs to be taken into account when sampling the random effects in the factor loadings. In Appendix 7.9 we show the JAGS (Just Another Gibbs Sampler) program (based on the R package rjags (Plummer, 2013)) for a simple version of *Model 4*, i.e. $\boldsymbol{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{u}_j + (\boldsymbol{\beta}_0^* + \boldsymbol{\beta}_1^* \boldsymbol{x}_{ij}^* + \boldsymbol{u}_j^*) F_{ij} + \boldsymbol{\varepsilon}_{ij}$. For $\mathbf{L}_j = \boldsymbol{\beta}_0^* + \boldsymbol{u}_j^*$, we have taken the mixture prior $0.5\, N(-\boldsymbol{\mu}_0^*, \Sigma_u^*) + 0.5\, N(\boldsymbol{\mu}_0^*, \Sigma_u^*)$ with $\boldsymbol{\mu}_0^*$ given a classical (vague) independent normal prior. Half of the $\mathbf{L}_j$ will fluctuate around $\boldsymbol{\mu}_0^*$ while the other half around $-\boldsymbol{\mu}_0^*$. The slopes $\boldsymbol{\beta}_1^*$ and $-\boldsymbol{\beta}_1^*$ will also be sampled from a mixture, but the sign will be determined from the sign of $\boldsymbol{\beta}_0^*$ when the covariates show enough variation as can be deduced from Section 7.3.3 and Appendix 7.9. Note that the MCMC procedure with the mixture prior for the factor loading parameters is needed when the identity link is used, but not for the exponential

[1] All Supplementary materials in this chapter can be found in the website: http://onlinelibrary.wiley.com/doi/10.1002/sim.6062/suppinfo.

link function where all factor loadings are positive. In that case both the fixed and random effects in the factor loadings are assumed to have a multiplicative effect. However, experience showed that convergence of the MCMC algorithm is much more difficult to achieve with the exponential link function.

7.3.5 Implied marginal models, skewness and kurtosis

Marginalized over the random effects, *Models 1* to *3* correspond to a multivariate normal model possibly with some heteroscedasticity described by covariates (*Model 3*) and with marginal covariance matrix given by expression (7.7). The skewness and kurtosis of the marginal normal densities are both zero. The marginalized *Model 4* has covariance matrix (see Appendix 7.9):

$$\Psi_{ijkl} = (\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl})(\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl})^T + \Sigma_u + \Sigma_h + \Sigma_c + \Sigma^*_u + \Sigma^*_h + \Sigma^*_c + \Sigma_\varepsilon, \tag{7.11}$$

but does not represent a normal model anymore. For each marginal density skewness is again zero since all random effects are mutually independent, but there is an excess of the kurtosis for the qth marginal density equal to (see Appendix 7.9):

$$\text{kurtosis}_q = \frac{6a_q^{*2} + 12a_q^* b_q}{(a_q + a_q^* + b_q + c_q)^2}. \tag{7.12}$$

In expression (8.6), a_q, a_q^*, b_q, c_q are the qth diagonal elements of $(\Sigma_u + \Sigma_h + \Sigma_c)$, $(\Sigma^*_u + \Sigma^*_h + \Sigma^*_c)$, $(\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl})(\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl})^T$ and Σ_ε, respectively. From this expression we can conclude that the marginal densities are leptokurtic unless the variance of the random effects in the factor loadings is zero ($a_q^* = 0$), and that the kurtosis also depends on the covariates in the factor loadings.

In Figure 7.3, we show the 2-dimensional joint distribution and the 1-dimensional marginal distributions for three scenarios with different kurtosis (0.24,1.50 and 3.60) by varying the values of a_q, a_q^*, b_q and c_q. To this end, we have simulated a 2-dimensional *Model 4* without covariates in the factor loadings. For each panel in Figure 7.3, both the fitted curve from 500,000 observations from *Model 4* and the best fitting normal curve are plotted. We notice the ability of *Model 4* to fit heavier-tailed distributions.

7.3.6 Model selection

Classical Bayesian selection criteria such as the deviance information criterion (DIC) (Spiegelhalter et al., 2002) and the pseudo-Bayes factor (PSBF) (Geisser and Eddy, 1979) can be applied to choose between models. JAGS cannot provide DIC because of the BOS strategy we applied, such that we needed to calculate DIC outside JAGS. To compute the PSBF we used the approach given in e.g. Lesaffre and Lawson (2012). The logarithm of the PSBF to compare *Model 1* with *Model 2* is denoted as $\ell PSBF_{1,2}$ whereby positive values indicate preference for the second model.

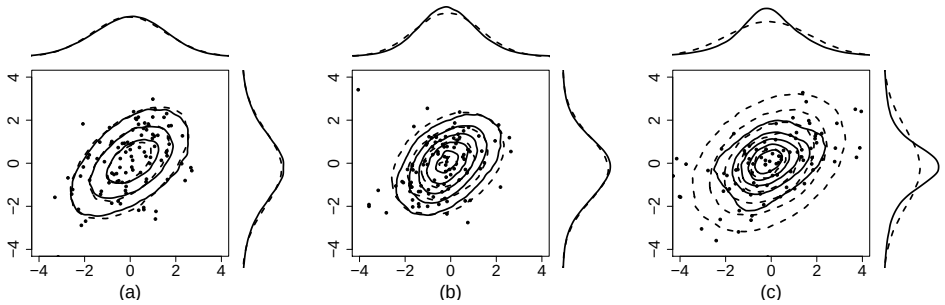

Figure 7.3: Joint and marginal distributions of *Model 4* (solid line) and best fitting normal curve (dashed line) when kurtosis in both marginals is 0.24 (a), 1.50 (b), and 3.60 (c)

The computation of PSBF and DIC is more complicated in the presence of missing data (response and/or covariates). In Celeux et al. (2006) several versions of DIC have been proposed in the presence of missing data. In this case, the authors suggested to use the DIC as reported by default by WinBUGS and JAGS, called the "complete DIC". While more research needs to be done to make a justified choice of the DIC version when the data are plagued by missing data, we decided to adhere to this suggestion for all models (*Models 1 to 5*). The calculation of PSBF is based on the same likelihood that is used for the complete DIC. Note that these models have the same response variables and covariates with the same missing data except for *Model 1*. For *Model 1* we used a trick to compare its DIC and PSBF to the models with covariates. That is, we included the imputation model for the covariates (used in the other models) in the JAGS program, such that it makes sense to compare the complete DIC and PSBF for the five models.

7.4 Multiple factors model

The single factor model works well when there are only a few responses. For three responses, one can easily show that each 3×3 covariance matrix can be represented by the single factor model. When $p > 3$, the single factor model is not sufficient to represent all $p \times p$ covariance matrices and an extension to the multiple factors model may be needed. Extending *Model 4* to $p > 3$ dimensions involves m factors $F_{ijkl.f}$ such that

$$\boldsymbol{\delta}_{ijkl} = \sum_{f=1}^{m} \boldsymbol{\Lambda}_{ijkl.f} F_{ijkl.f} + \boldsymbol{\varepsilon}_{ijkl}, \quad with \quad \boldsymbol{\Lambda}_{ijkl.f} = \rho(\boldsymbol{B}_f^* \boldsymbol{x}_{ijkl}^* + \boldsymbol{u}_{jkl.f}^* + \boldsymbol{u}_{kl.f}^* + \boldsymbol{u}_{l.f}^*),$$

and

$$\boldsymbol{u}_{jkl.f}^* \sim N(\mathbf{0}, \Sigma_{u.f}^*), \quad \boldsymbol{u}_{kl.f}^* \sim N(\mathbf{0}, \Sigma_{h.f}^*), \quad \boldsymbol{u}_{l.f}^* \sim N(\mathbf{0}, \Sigma_{c.f}^*), \quad F_{ijkl.f} \sim N(0,1),$$

with mutual independence of the random components as before. This model is similar to

Hoff and Niu's $rank - m$ model but now in a multilevel context.

As for Hoff and Niu's $rank - m$ model, the intercept matrix $\boldsymbol{B}_0^*$ in the factor loadings, which has $p \times m$ elements with p the dimension of the outcome, should have orthogonal columns. Besides, same as in *Model 4*, the mixture prior for the factor loadings should be applied for each factor. In addition, the identifiability issues are similar to those of *Model 4* but more involved. This is a topic of future research.

Fox and Dunson (2011) pointed out that with Hoff and Niu's $rank - m$ model the total number of parameters dramatically increase with increasing m when the dimensionality of the outcome is high. They suggested an alternative approach, which could be adopted also here. Reducing the high dimensionality of the outcome is, however, not the focus in this paper as this model is not needed for our motivating data set.

7.5 Computational procedure

A JAGS program of each model was written for the analysis of the motivating data set. *Models 1* and *2* were estimated with the R package *rjags*, while the *dclone* package was used for the other models, which is based on JAGS with multiple cores. The DIC calculation was based on the `dic.samples()` function, which corrects for overoptimism in computing the classical DIC (Plummer, 2002). PSBF was calculated following the way described in Lesaffre and Lawson (2012). A 2-level version of the program for *Model 4* can be found in Appendix 7.9. The 4-level program can be derived from this program but is available from the first author upon request.

Three chains were initialized with different starting values for all models. For *Models 1* and *2*, 10,000 iterations were conducted with the first half as the burn-in part, while for *Models 3* to *5* 100,000 iterations were set with the first 70,000 as burn-in part. Convergence was decided when the Brooks-Gelman-Rubin plots showed good behavior indicated by the estimate Rhat < 1.1 (Brooks and Gelman, 1998). In addition, the Monte Carlo error of the posterior mean should be around or less than 5% of the posterior SD. Upon convergence, we computed the posterior median for the variance components and the posterior mean for the remaining parameters together with the equal tail 95% credible interval (CI). Both DIC and $\ell PSBF$ were used to select the most appropriate model.

7.6 Missing data

The RN4CAST project is plagued with a relatively large non-response rate, but also with missingness in the response and the covariates. Therefore, consideration is needed on the effect of missing information on the substantive conclusions of the analysis, but also on what can be done to reduce possible bias due to lacking data. Below we consider the three problems separately, suggest solutions if possible and/or reflect on what might the possible impact of the non-response/missingness. Although trivial, we note that our suggested multivariate multilevel model remains valid whether or not the study is plagued with missing data.

7.6.1 Non-response

In Section 7.2.1 we described that there is considerable non-response in the RN4CAST study especially at hospital and nurse level. Around six out of ten (63.55%) hospitals that were invited to this study agreed to participate. At the planning stage of the study, representative checks (hospital type, size) were carried out in each country to assure that the sample represents the population appropriately. When necessary, corrective actions (such as extra motivating hospitals to participate) were taken to improve representativeness of the participating hospitals. The response rates of the nurses were around 40% in the different countries. Thus, at best we can claim that all reasonable efforts (this was a huge study) were undertaken to improve the representativeness of the study, but bias due to non-response cannot be excluded. One referee suggests to implement a corrective action in estimating the model parameters making use of the approach described in e.g. Kott (1994). However, this kind of correction needs covariate information of all subjects, which was not available in the RN4CAST study for the non-responders. We note that in a similar survey among US nurses (Smith, 2008) the author assessed the non-response bias. More particularly, the author randomly sampled non-respondents and was able to motivate them to fill in (at the very end) the questionnaires. He found indeed some differences in the non-responders' demographic characteristics compared to the initial responders. However, no differences were found with regard to nurses' assessments of their work environment and burnout. This supported our hope and belief that chances are low that our findings are dramatically affected by systematic tendencies of certain respondents to have opted into our survey or to have opted out.

7.6.2 Missingness in the burnout measurements

The three burnout outcomes are sum scores of the items within each dimension respectively. Around 10% of the nurses have at least one of the 22 items missing. This implies that for these subjects some or all burnout scores were too low. Imputation of the missing item was done by treating the scores with missing items as interval censored data with the current value as the lower limit of the interval (equivalent to assigning zeros to the missing items) and the upper limit of the interval obtained by taking the largest value for each missing item, which is 6 in this case. It can easily be seen that the interval censored trick combines well with the BOS approach we utilized across the whole analysis.

7.6.3 Missing covariates

Missingness also plagued the covariates, but primarily *working experience* with about 6.3% of the subjects not filled in this item. A classical approach to deal with missing covariates is to make use of multiple imputation. This is a fairly straightforward approach with MCMC software. Indeed, it only involves to sample at each iteration, in parallel to the main estimation program, the missing covariate values using an appropriate imputation model and plug-in the sampled value into the main model. The chosen model for imputation is a Gaussian linear regression model including all the other covariates in Table 8.1 as predictors. See

the Supplementary Material for the actual models chosen to estimate the missing covariate values.

7.7 Analysis of the RN4CAST burnout data

7.7.1 Choice of response and covariates

Figure 7.1 shows a skewed distribution for each of the three burnout measurements. Because the measurements are in fact discrete with a non-trivial portion of zeros they cannot be transformed to normality with common transformations. However, since the burnout measurements are examples of a bounded outcome score (BOS), one may apply the technique suggested by Lesaffre et al. (2007). This technique assumes that a standardized version of the burnout measurement is a coarsened latent continuous variable which has a Gaussian distribution after a logit transformation. More 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 $\log[z/(1-z)] \sim N(\mu, \sigma^2)$ and such that y is obtained by coarsening z. In our analysis, we applied this technique on each of the three burnout measurements. *Models 1* to *5* were then defined on the latent continuous outcomes.

The candidate covariates involved in the mean and/or variance structures in later analyses, are listed in Table 8.1. In order to make the regression coefficients of these variables comparable and to improve computational properties, standardized covariates (mean=0, SD=1) were used in later analyses. In addition, in order to investigate the level-specific effects of the covariates the following decomposition was made, as suggested by Neuhaus and Kalbfleisch (1998):

$$\begin{aligned} x_{ijkl} &= (x_{ijkl} - \bar{x}_{jkl}) + (\bar{x}_{jkl} - \bar{x}_{kl}) + (\bar{x}_{kl} - \bar{x}_{l}) + \bar{x}_{l} \\ &= x_n + x_u + x_h + x_c, \end{aligned} \tag{7.13}$$

$\bar{x}_{jkl} = \frac{1}{n_{jkl}} \sum_{i=1}^{n_{jkl}} x_{ijkl}$ and similarly for the other means. In expression (8.7), the lowest level variable is partitioned into four parts corresponding to the four levels, i.e. n for nurse, u for nursing unit, h for hospital and c for country. By doing so, we study the "pure" effect of the covariates at each level (for *work environment*, there is no nurse-level partition because the lowest level measurement is nursing unit).

We now address the three research questions mentioned in Section 7.2. For each model fitted to the data, we describe the variability of the means of the burnout measurements at the different levels separately from the variability (and correlation) of these measurements at the nurse level. The parameter estimates for all five models can be found in the Supplementary Material of the paper.

7.7.2 Question 1: 4-level model without covariates

The first analysis is based on *Model 1* and aims to measure the variability of each of the three burnout measurements according to countries, hospitals (within countries), nursing units (within countries and hospitals) and nurses (within countries, hospitals and nursing units). For this model we obtained a DIC = 222,447.

Variability of means: Table 7.3 shows the variance components for each burnout dimension, which are the diagonal elements of the covariance matrix of the random effects at each level, representing the variation at each level. The burnout measurements show little variability at the hospital and nursing unit level. Diya et al. (2013) also concluded little hospital variability in their analysis of the Belgian data of the RN4CAST project. Figure 7.4 shows the mean of the random effects at country level for the three burnout dimensions for *Model 1*. The variation of *PA* across countries is much smaller than for *EE* and *DP*, which could also be seen from Table 7.3. This holds for the hospital and nursing unit levels as well, indicating a limited variation of *PA* at these three levels.

(Co)Variance at nurse level: Table 7.3 shows that the burnout measurements show most variability at the nurse level. There is particular interest here in Σ_{ijkl} and its associated correlation matrix. In *Model 1*, homogeneity of the residual covariance structure is assumed, i.e. $\Sigma_{ijkl} \equiv \Sigma_\delta$. In Table 7.4 it is seen that the three burnout dimensions are moderately correlated, especially *EE* and *DP*, but all correlations are somewhat less than the data-based correlations obtained by ignoring the multi-level structure of the data.

Table 7.3: Variance (and percentage) of the burnout measurements at each level for *Models 1* to *5*. For *Models 1* and *2*, the diagonal elements of Σ_c (country), Σ_h (hospital), Σ_u (nursing unit) and Σ_δ (nurse) are reported. For *Models 3* to *5*, Σ_δ is replaced by its median value (defined in the text)

Burnout	Model	Country	Hospital	Nursing unit	Nurse
EE	*1*	0.359 (24.3%)	0.060 (4.1%)	0.122 (8.3%)	0.935 (63.3%)
	2	0.265 (20.3%)	0.050 (3.8%)	0.059 (4.5%)	0.932 (71.4%)
	3	0.263 (21.9%)	0.049 (4.1%)	0.058 (4.8%)	0.832 (69.2%)
	4	0.256 (24.8%)	0.047 (4.5%)	0.049 (4.7%)	0.682 (66.0%)
	5	–	0.047 (6.0%)	0.049 (6.3%)	0.682 (87.7%)
DP	*1*	0.305 (12.1%)	0.070 (2.8%)	0.139 (5.5%)	2.003 (79.6%)
	2	0.257 (10.9%)	0.057 (2.4%)	0.073 (3.1%)	1.980 (83.7%)
	3	0.255 (11.2%)	0.057 (2.5%)	0.073 (3.2%)	1.896 (83.1%)
	4	0.254 (12.2%)	0.057 (2.7%)	0.067 (3.2%)	1.702 (81.8%)
	5	–	0.057 (3.1%)	0.068 (3.7%)	1.700 (93.2%)
PA	*1*	0.150 (11.1%)	0.034 (2.5%)	0.044 (3.3%)	1.118 (83.1%)
	2	0.141 (10.7%)	0.028 (2.1%)	0.030 (2.3%)	1.114 (84.8%)
	3	0.140 (11.9%)	0.028 (2.4%)	0.030 (2.5%)	0.979 (83.2%)
	4	0.143 (16.7%)	0.027 (3.1%)	0.024 (2.8%)	0.664 (77.4%)
	5	–	0.027 (3.8%)	0.024 (3.4%)	0.664 (92.9%)

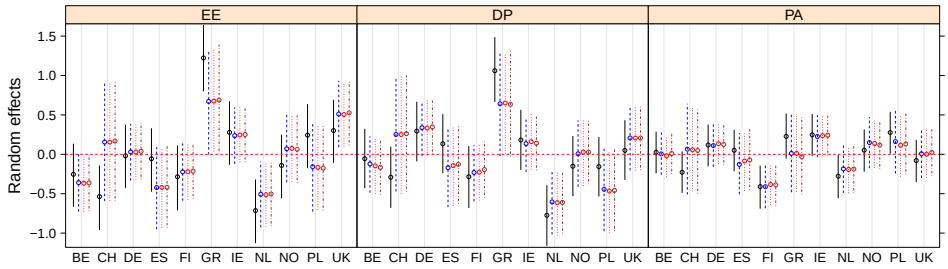

Figure 7.4: Mean (represented by a circle) and equal tail 95% CI of the random effects at country level for *EE*, *DP* and *PA* for *Model 1* (black solid line), *Model 2* (blue dashed line), *Model 3* (red dotted line) and *Model 4* (brown dashed-dotted line)

7.7.3 Question 2: 4-level model with covariates in the mean structure

In Question 2 we look for the covariates that explain best the means of each burnout dimension in each nursing unit, hospital and country and hence also explain best their variability. This analysis involves *Model 2*. The large sample size of the study in combination with a computationally intensive MCMC procedure forced us to use a relatively simple variable selection strategy. At first we included all covariates mentioned in Section 7.2.2 into the model. We then kept those for which the 95% CI did not include 0. *Model 2* appears to give a better fit to the data since $\ell PSBF_{1,2} = 426.3$, confirmed by a lower DIC = 221,782.

Variability of means: Figure 7.5 shows the impact of the selected covariates: *working experience* and *work environment* at the different levels and *type of the nursing unit*. Burnout appears to be less in a positive environment, and in nursing units and hospitals with more experienced nurses. There is also more burnout in surgical units. When these three covariates were included in the model, the unexplained residual variance at each level dropped for each of the three burnout dimensions (see Table 7.3) with the greatest reduction for the nursing unit level (variances dropped to about half). In Figure 7.4 it is seen that, compared to *Model 1*, the mean country effects have shrunken towards zero.

(Co)Variance at nurse level: From Table 7.3 we notice that in absolute terms the variability at nurse level is about the same as for *Model 1*, but since part of the variability of the means is explained by the included covariates, the relative contribution to the variability at nurse level increased considerably. Again homogeneity of the residual covariance structure is assumed, i.e. $\Sigma_{ijkl} \equiv \Sigma_\delta$. In Table 7.4 we notice that the estimated correlations of *Models 1* and *2* are basically the same.

7.7.4 Question 3: 4-level model with residual covariance matrix depending on covariates

In *Model 3* covariates are included in the factor loadings to check whether the residual covariance matrix Σ_δ depends on (some of) the recorded covariates at the different levels. The aim is then also to check whether the intrinsic correlations obtained from Σ_ε are fairly con-

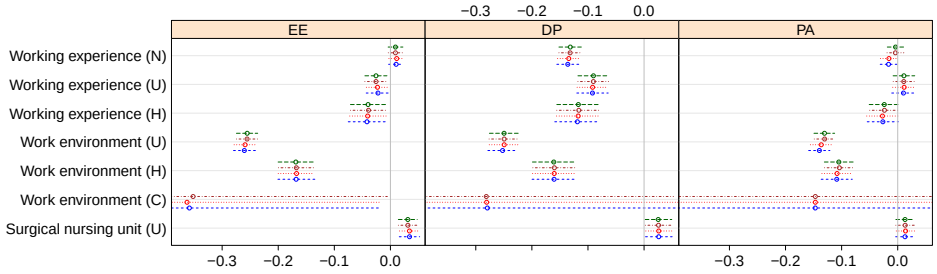

Figure 7.5: Mean (represented by a circle) and equal tail 95% CI for fixed effects estimates (based on standardized covariates) in the mean structure for *Model 2* (blue dashed line), *Model 3* (red dotted line), *Model 4* (brown dashed-dotted line) and *Model 5* (darkgreen long-dashed line). The symbols N, U, H and C refer to the split up in expression (8.7)

Table 7.4: Correlations in burnout measurements ignoring the multilevel structure (Data-based), and based on residual correlation matrix for *Models 1* to *5*, which correspond to Σ_δ for *Models 1* and *2*, and Σ_ε for the other models

Model	Correlations		
	(*EE*, *DP*)	(*EE*, *PA*)	(*DP*, *PA*)
Data-based	0.565	0.286	0.324
1	0.485	0.268	0.329
2	0.488	0.266	0.328
3	0.460	0.224	0.301
4	0.464	0.222	0.314
5	0.463	0.221	0.315

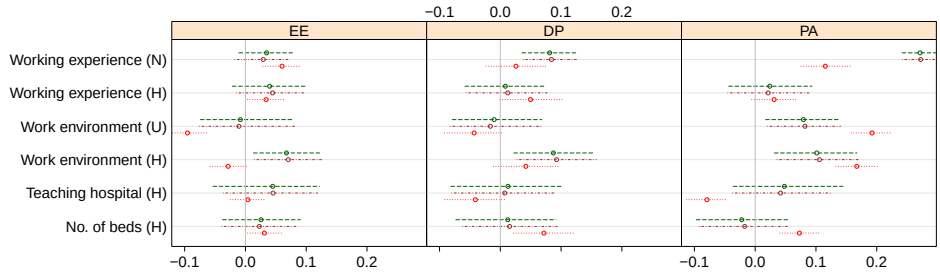

Figure 7.6: Mean (represented by a circle) and equal tail 95% CI for fixed effects estimates (based on standardized covariates) in the factor loadings for *Model 3* (red dotted line), *Model 4* (brown dashed-dotted line), and *Model 5* (darkgreen long-dashed line)

stant across countries, hospitals, etc. From this exploration we might better understand the factors that determine the burnout process.

We used the same variable selection method as for the mean structure. No country level covariates were now included in the factor loadings because of the small sample size at country level and the convergence problem caused by this. In the end, four variables were significantly impacting the covariance matrix, i.e. *work environment* (both at hospital and nursing unit level), *working experience* (both at hospital and nurse level), *No. of beds* at hospital level and *teaching hospital*. Both PSBF and DIC showed a preference of *Model 3* over *Model 2* with $\ell PSBF_{2,3} = 167.8$ and DIC = 220,810, respectively.

Variability of means: All estimated regression coefficients in the mean structure are quite close to those of *Model 2*, as can be seen in Figure 7.5. Random effects estimates at country level are shown in Figure 7.4 which are also close to those of *Model 2*. This indicates that modeling the covariance structure does not affect much the mean part, which is known in the Gaussian case.

In *Model 3*, the elements δ_{ijkl} are regressed on covariates and therefore Σ_{ijkl} is no longer constant across the levels. It is not straightforward to compare the variability of the means for this model. We have chosen to report (the proportion of) variation at each of the different levels in Table 7.3 when taking the median covariance matrix for δ_{ijkl} defined by the covariance matrix Σ_δ for which all covariates are given their mean values.

(Co)Variance at nurse level: The correlations obtained from the matrix Σ_ε are shown in Table 7.4. They represent the intrinsic correlations when confounders on the different levels are accounted for. We note that all of the three correlations dropped slightly from those of *Model 2*. The posterior mean and the 95% CI of the coefficients of covariates in the factor loadings are shown in Figure 7.6. From Figure 7.2 we know that the relationship between each covariate separately and the (co)variance is quadratic (irrespective of the sign of the regression coefficient). The impact of the covariate *working experience at nurse level* on the marginal covariance matrix and correlation matrix, when varying from its minimal to max-

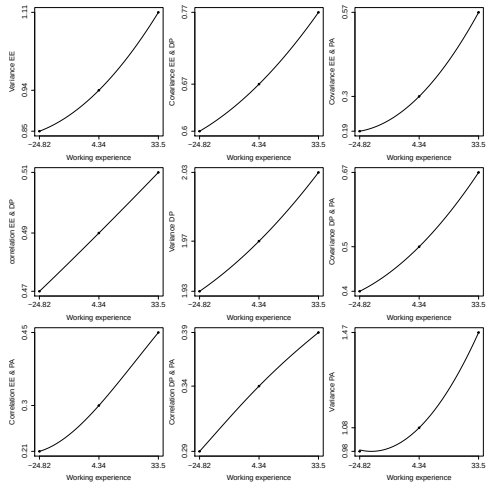

Figure 7.7: Dependence of covariance matrix Σ_{ijkl} in equation (7.7) (upper triangle)/associated correlation matrix and *working experience at nurse level*. The maximum, minimum and the median values of the covariate are marked

imal value, is shown in Figure 7.7. Especially the correlations with *PA* seem to depend highly on *working experience at nurse level*, implying that the longer the professional career of the nurse the more emotional exhaustion and depersonalization the less self-esteemed the nurse is. When varying $\widehat{\boldsymbol{B}}_f^* \boldsymbol{x}_{ijkl}^*$ from its minimal to its maximal value, e.g. the correlation between *EE* and *PA* now ranges from -0.25 to 0.55 and the variance of *PA* ranges from 0.97 to 2.52.

7.7.5 Question 3: 4-level model with residual covariance matrix depending on covariates and random effects

Finally, we included random effects into the covariance structure to accommodate for the measurement error in the included covariates in the factor loadings, for not-included (and possibly not measured) important covariates and to protect against potential outlying entries. This analysis involves *Model 4*. With $\ell PSBF_{3,4} = 1255.7$ there is a strong indication that *Model 4* fits the data better than *Model 3*. This is confirmed by a much lower DIC = 213,908.

Variability of means: From Figure 7.5 we see that the fixed effects estimates are basically the same as those of *Models 2* and *3*, which is also the case for the random effects at country level (Figure 7.4). The entries in Table 7.3 for *Model 4* again correspond to the median value determined in Σ_δ, i.e. by taking the mean value for each covariate, as well as zero for the random effects in the factor loadings.

(Co)Variance at nurse level: We see from Table 7.4 that there is remarkable stability of the correlations between *Model 4* and the previous model. From Figure 7.6 we note that, adding random effects in the factor loadings, renders three hospital-level covariates non-significant,

i.e. *working experience*, *teaching hospital* and *No. of beds*. In general, we observe that both the estimates as well as the 95% CI are reasonably affected by including the random effects. The interpretation of this result is not straightforward, but definitely the included random effects appear to be associated with these three covariates. Judging from the DIC and PSBF *Model 4* seems to be preferable, which could imply that ignoring random effects in the factor loadings results in biasedly estimated regression coefficients (and SEs) in the factor loadings. We also plotted the mean random effects at the country level in the factor loadings (not shown), again Greece showed some outlying behavior but to a lesser extent.

7.7.6 A multiple group 3-level model with residual covariance matrix depending on covariates and random effects

As there are only 11 countries, an alternative is to run a 3-level multiple group model (*Model 5*), treating the 11 countries as fixed effects both in the mean structure and the factor loading structure. Except for this change all other settings are exactly the same as for *Model 4*. Belgium was taken as reference country. With $\ell PSBF_{4,5} = 82.1$ and DIC = 213,885, *Model 5* is preferred to *Model 4*. However, the changes in PSBF and DIC are relatively small, especially for DIC that dropped only 23. Considering the large sample size, the difference in performance between the two models is not great.

Variability of means: The variances partitioned at each level are listed in Table 7.3, again the median estimates are shown. Note that as country was treated as a fixed effect, we actually took the median values for the reference country, i.e. Belgium. Figure 7.5 shows that the fixed effects (except for the country effects) estimates in the mean structure are quite close to those from other models. For the country fixed effects (not shown), we noticed that *Switzerland*, *the Netherlands* and *Finland* score better on burnout than the reference country *Belgium*, while *Spain, Poland, Ireland, Greece* and *Germany* score worse.

(Co)Variance at nurse level: The parameter estimates in the factor loadings are quite close to those in *Model 4*, see Figure 7.6. The intrinsic correlations between the burnout dimensions are close to those of *Model 3*.

7.7.7 Model assessment

The assessment of *Model 4* was performed in this section. More specifically, we did three kinds of assessments. Firstly, a posterior predictive check (PPC) based on Gelman's chi-square statistic (Gelman et al., 2013) was performed to find evidence of the goodness-of-fit of the models for each of the burnout responses separately. Secondly, we conducted a normality check for the random effects to justify the model assumption and further ran the model with a multivariate t(3)- distribution for the random effects. Thirdly, the current strategy of handling missing data was compared with an alternative way, i.e. imputing the missingness for the response variables prior to the analysis and removing the data with missing covariate. Not only the parameter estimates of *Model 4*, but also the model selection based on DIC and PSBF were compared using the two strategies of handling missing data. More details for the second and third model assessments are described in the Supplemen-

tary Material. As a result, we found that the current model showed an acceptable fit to the data, with PPC values for the three burnout dimensions being 0.36, 0.28 and 0.29, respectively. The normality assumption basically holds for all random effects in both the mean and the loadings as evidenced from the QQ-plots described in Section (7.3.3). Further, no qualitative changes were found in the model with multivariate t distribution. Finally, the two strategies of dealing with missing data for both the response and the covariate provided quite close estimates. Both DICs and $\ell PSBFs$ are highly consistent with each other for the two strategies except for $\ell PSBF_{2,3}$. Nevertheless, we believe that the current approach of dealing with the missing information is preferable.

7.7.8 Some clinical conclusions

The primary clinical finding of this study is that all proposed models posit the importance of nurses' work environment, both at the hospital and nursing unit level, in explaining all three dimensions of burnout. This pinpoints the key role that hospital health resources management and front-line nurse leaders play in developing positive work environments to prevent burnout. Maslach and Leiter (2008) in their (2008) longitudinal study on burnout changes over a 1-year interval highlighted the importance of customized preventive interventions. As proposed by these authors, such interventions should take place at an organizational level, since burnout tends to cluster within particular groups. In dealing with burnout, hospitals and health care organizations in general should therefore continuously evaluate burnout and predicting features within the organizational climate.

The second finding is that the intrinsic correlations, as reported in Table 7.4 are relatively stable across the five models. But we could also notice that several factors determine the covariance structure and hence will have an impact on the association structure of the three burnout dimensions. We have illustrated this by showing that, as working experience increases, the correlation between *EE* and *PA* increases. This might corroborate the finding that *EE* and *PA* develop in parallel within a problematic organizational environment without any major causal links between the two (Schaufeli et al., 1996).

7.8 Simulation study

To evaluate the performance of the proposed multivariate multilevel factor model, we performed a limited simulation study. We investigated:

1. the robustness of the parameter estimates of *Models 3* and *4* in the presence of outliers;
2. a comparison of *Models 4* and *5* when the number of entries at each level are varied.

Further details of this simulation study can be found in the Supplementary Material of this paper. The performance of the models is evaluated by computing the standardized bias and the frequentist coverage of the equal tail 95% CI. The former is calculated, according to Collins et al. (2001), as $100(\bar{\hat{\beta}}-\beta)/SE(\hat{\beta})$, where β is the true value for each parameter, and $\bar{\hat{\beta}}$ and $SE(\hat{\beta})$ are the mean and the standard deviation of the estimates across all simulations, respectively.

In the first simulation study, we created various types of outliers and compared the performance of *Models 3* and *4*. As a result, we obtained that *Model 4* is generally more robust against outliers at each level than *Model 3* for the fixed effects estimates in the factor loadings. The standardized bias was relatively low for *Model 4* and the coverage of the equal tail 95% CI was relatively close to 95%. This means that the relation between the covariance matrix and covariates is rather robustly against outliers at each level. The same is true for the estimates of Σ_ε, when the outliers are not taken at the lowest level.

The choice between *Models 4* and *5* has been intensively discussed (Maas and Hox, 2005; Kreft and de Leeuw, 1998; Browne and Draper, 2000). Maas and Hox (2005) suggest a minimum size of 50 for the highest level to obtain unbiased standard error estimates. In our case, however, there are only 11 countries at the highest level. Therefore, the performance of the model estimates under different sample size was of great interest here. Our simulations indicated that the multivariate multilevel model performs relatively well for a reasonable sample size at each level. Unbiased point estimates could be obtained from a relatively small sample size (only 10) at the highest level, but the coverage could be distorted somewhat, even for a moderately sized data set (around 30 highest level observations).

7.9 Discussion

We proposed in this paper a multivariate multilevel model with a structured covariance matrix that depends on both fixed and random effects. Our approach is an extension of Hoff an Niu's covariance regression model to a multilevel setting including random effects in the covariance matrix analyzed in a Bayesian way. Our modeling approach makes use of the mixture normal prior for the factor loadings and leads to identified model estimates. While the model was developed for three responses, it is easily extended to p responses. Better results might be obtained by using the multi-factor extension of Hoff and Niu or the suggestion of Fox and Dunson (2011). Being an extension of Hoff an Niu's approach, the impact of the covariates and the random effects on the covariance matrix is also intuitive. However, the identifiability issues associated with our factor analytic approach to model heterogeneity in the covariance structure complicates the evaluation of the model assumptions.

Our modeling approach was applied to the multilevel data of the RN4CAST study, which is a large European study conducted in twelve European countries. While the study was carefully planned, the non-response rate turned out to be considerable. Nevertheless, there is reason to believe that the substantive conclusions from this study are relatively robust, but of course we cannot rule out bias due to lacking data.

An alternative approach to tackle the research questions is to involve the SEM strategy, of which the interest lies also in modeling the covariance matrix. Recently multi-level SEM approaches have been suggested (Kline, 2010). In this respect, the statistical package Mplus (Muthén and Muthén, 2012), is particularly convenient and powerful for latent variable analysis, e.g. SEM, path analysis, survival analysis, etc. Mplus provides also Bayesian software, which is increasingly used for handling complex SEMs. The latest version (version

7.0) has the option for a Bayesian analysis of a model somewhat similar to *Model 4*. However, the package can only handle a two-level model for random loading models and no lowest level covariates in the factor loadings are allowed. Besides these shortcomings it is not possible in Mplus to utilize the BOS theory as we used here. In SEM, usually the focus is on the structure equations, i.e. the relationship among the latent variables and with covariates. This, however, is not the case of our modeling approach where we fix the latent variables (common factors) as standard normally distributed and model the relationship between the responses and the common factors, i.e. the factor loadings. In this way, we can reconstruct the covariance matrix among the responses without any structural equation among latent variables. The difference between our modeling and SEM could also be explained this way: with $f = \Lambda F$, SEM models the mean of f, up to the scale Λ which is assumed to be constant, while our modeling focuses on the standard deviation of f, i.e. Λ, leaving F standard normally distributed.

As mentioned at the end of Section 7.2, our model can be extended with the covariance matrix of the random effects depending on covariates and random effects. This could be considered as a multivariate version of the DHGLM by Lee and Nelder (2006) in the sense that not only the residual covariance matrix, but also the covariance matrix of the random effects could be given a mixed effects structure. This extension is theoretically straightforward. However, we did not achieve convergence in our MCMC analysis of the RN4CAST data when we let both the residual covariance matrix as the covariance matrix depend on covariates (and random effects).

We conclude that the 3-variate 4-level factor model proposed in this paper performed well when applying it to the RN4CAST study and we believe it is a useful addition to the current models for analyzing multivariate multilevel models. With regard to the application to the RN4CAST data, additional to the current findings, we have dug deeper to the intrinsic correlation among the three dimensions and provide detailed interplay of the correlation with covariates at different levels and the multilevel structure itself via random effects.

References

Aiken, L. H., Clarke, S. P., Sloane, D. M., Sochalski, J., and Silber, J. H. (2002). Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *JAMA: the Journal of the American Medical Association*, 288(16):1987–1993.

Barnard, J., McCulloch, R., and Meng, X. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4):1281–1312.

Bock, R. and Gibbons, R. (1996). High-dimensional multivariate probit analysis. *Biometrics*, 52(4):1183–1194.

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

Browne, W. (2012). *MCMC Estimation in MLwiN, v2.25*. Centre for Multilevel Modelling, University of Bristol.

Browne, W. and Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, 15(3):391–420.

Cecere, S., Jara, A., and Lesaffre, E. (2006). Analyzing the emergence times of permanent teeth: an example of modeling the covariance matrix with interval-censored data. *Statistical Modelling*, 6(4):337–351.

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.

Chiu, T., Leonard, T., and Tsui, K. (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.

Daniels, M. and Pourahmadi, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, 89(3):553–566.

Diya, L., Li, B., Van Den Heede, K., Sermeus, W., and Lesaffre, E. (2013). Multilevel factor analytic models for assessing the relationship between nurse reported adverse events and patient safety. *Journal of the Royal Statistical Society: Series A*. (*in press*).

Foulley, J. and Gianola, D. (1996). Statistical analysis of ordered categorical data via a structural heteroskedastic threshold model. *Genetics Selection Evolution*, 28(3):249–273.

Foulley, J., Gianola, D., San Cristobal, M., and Im, S. (1990). A method for assessing extent and sources of heterogeneity of residual variances in mixed linear models. *Journal of Dairy Science*, 73(6):1612–1624.

Foulley, J., San Cristobal, M., Gianola, D., and Im, S. (1992). Marginal likelihood and Bayesian approaches to the analysis of heterogeneous residual variances in mixed linear Gaussian models. *Computational Statistics & Data Analysis*, 13(3):291–305.

Fox, E. and Dunson, D. (2011). Bayesian nonparametric covariance regression. *arXiv:1101.2017v2 [stat.ME]*.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition.

Gibbons, R. and Lavigne, J. (1998). Emergence of childhood psychiatric disorders: A multivariate probit analysis. *Statistics in Medicine*, 17(21):2487–2499.

Gibbons, R. and Wilcox-Gök, V. (1998). Health service utilization and insurance coverage: a multivariate probit analysis. *Journal of the American Statistical Association*, 93(441):63–72.

Goldstein, H. (2010). *Multilevel Statistical Models (Wiley Series in Probability and Statistics)*. Wiley, 4th edition.

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

Ibáñez, M., Carabaño, M., and Alenda, R. (1999). Identification of sources of heterogeneous residual and genetic variances in milk yield data from the Spanish Holstein-Friesian population and impact on genetic evaluation. *Livestock Production Science*, 59(1):33–49.

Kizilkaya, K. and Tempelman, R. (2005). A general approach to mixed effects modeling of residual variances in generalized linear mixed models. *Genetics Selection Evolution*, 37(1):31–56.

Kline, R. B. (2010). *Principles and Practice of Structural Equation Modeling, (Methodology in the Social Sciences)*. The Guilford Press, 3rd edition.

Kott, P. S. (1994). A note on handling nonresponse in sample surveys. *Journal of the American Statistical Association*, 89(426):693–696.

Kreft, I. G. G. and de Leeuw, J. (1998). *Introducing Multilevel Modeling (Introducing Statistical Methods Series)*. SAGE Publications Ltd.

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

Lee, Y. and Nelder, J. A. (2006). Double hierarchical generalized linear models (with discussion). *Applied Statistics*, 55(2):139–185.

Lee, Y. and Noh, M. (2012). Modelling random effect variance with double hierarchical generalized linear models. *Statistical Modelling*, 12(6):487–502.

Lesaffre, E. and Lawson, A. B. (2012). *Bayesian Biostatistics (Statistics in Practice)*. Wiley, 1st edition.

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

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.

Lin, X., Raz, J., and Harlow, S. (1997). Linear mixed models with heterogeneous within-cluster variances. *Biometrics*, 53(3):910–923.

Maas, C. and Hox, J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 1(3):86–92.

Maslach, C., Jackson, S. E., and Leiter, M. P. (1996). *Maslach Burnout Inventory Manual*, 3rd edition.

Maslach, C. and Leiter, M. P. (2008). Early predictors of job burnout and engagement. *Journal of Applied Psychology*, 93(3):498.

Maydeu-Olivares, A. and McArdle, J. J. (2005). *Contemporary Psychometrics (Multivariate Applications Series)*. Psychology Press.

Meuleman, B. and Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methods*, 3(1):45–58.

Muthén, B. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, 22(3):376–398.

Muthén, L. and Muthén, B. (2012). *Mplus User's Guide*, 7th edition.

Neuhaus, J. M. and Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 54(2):638–645.

Plummer, M. (2002). Discussion of the paper by spiegelhalter et al. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

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

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

Schaufeli, W. B., Maslach, C., and Marek, T. (1996). *Professional Burnout: Recent Developments in Theory and Research (Series in Applied Psychology: Social Issues and Questions)*. CRC Press, 1st edition.

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):6.

Servellen, G. and Leake, B. (1993). Burn-out in hospital nurses: a comparison of acquired immunodeficiency syndrome, oncology, general medical, and intensive care unit nurse samples. *Journal of Professional Nursing*, 9(3):169–177.

Smith, H. L. (2008). A double sample to minimize skew due to non-response in a mail survey. In Ruiz-Gazen, A., Guilbert, P., Haziza, D., and Tille, Y., editors, *Survey methods: Applications to longitudinal investigations, health, electoral investigations, and investigations in the developing countries*, pages 334–339. Paris: Dunod.

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.

Squires, A., Aiken, L. H., van den Heede, K., Sermeus, W., Bruyneel, L., Lindqvist, R., Schoonoven, L., Stromseng, I., Busse, R., Brozstek, 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.

Wu, L. (2010). *Mixed Effects Models for Complex Data (Monographs on Statistics and Applied Probability 113)*. CRC Press, 1st edition.

Appendix

Identification of Model 3 *and* Model 4

Hoff and Niu's showed that, when there is enough variation in the covariate values, their model is identifiable. Although a formal proof of identifiability of *Models 3* and *4* is beyond the scope of this paper, we present below some arguments that all parameters are estimable for *Model 3*. In addition, we never experienced identifiability problems (non-convergence if the Markov chain) when analyzing the RN4CAST data with the included covariates for both *Models 3* and *4*. Let us assume a simplified 2-level *Model 4*, given by:

$$\begin{aligned} \boldsymbol{y}_{ij} &= \boldsymbol{B}\boldsymbol{x}_{ij} + \boldsymbol{u}_j + \boldsymbol{\delta}_{ij}, \\ \boldsymbol{\delta}_{ij} = \boldsymbol{\Lambda}_{ij} F_{ij} + \boldsymbol{\varepsilon}_{ij}, &\quad \boldsymbol{\Lambda}_{ij} = \boldsymbol{B}^* \boldsymbol{x}^*_{ij} + \boldsymbol{u}^*_j, \\ \boldsymbol{u}_j \sim N(\boldsymbol{0}, \Sigma_u), &\quad \boldsymbol{u}^*_j \sim N(\boldsymbol{0}, \Sigma^*_u), \\ F_{ij} \sim N(0,1), &\quad \boldsymbol{\varepsilon}_{ij} \sim N(\boldsymbol{0}, \Sigma_\varepsilon) \end{aligned} \tag{7.14}$$

Model 3 is then obtained by dropping the $\boldsymbol{u}^*_j$ terms. For this model, we argue that identifiability of all model parameters follows from the following reasoning:

- The regression coefficients $\boldsymbol{B}$ in the mean part are estimable if there are no linear dependencies among the covariates;
- The matrix $\Sigma_u + \boldsymbol{B}^* \boldsymbol{x}^*_{ij} (\boldsymbol{B}^* \boldsymbol{x}^*_{ij})^T + \Sigma_\varepsilon$ can be estimated from the marginal covariance matrix of $\boldsymbol{y}_{ij}, \boldsymbol{y}_{i^*j^*}$, with $i = i^*, j = j^*$;
- The matrix Σ_u can be estimated from the marginal covariance matrix of $\boldsymbol{y}_{ij}, \boldsymbol{y}_{i^*j^*}$, with $i \neq i^*, j = j^*$;
- The matrix $\boldsymbol{B}^* \boldsymbol{x}^*_{ij} (\boldsymbol{B}^* \boldsymbol{x}^*_{ij})^T$ can be estimated from the marginal covariance matrix of $\boldsymbol{y}_{ij}, \boldsymbol{y}_{i^*j^*}$, with $i \neq i^*, j \neq j^*$. Then the elements $\boldsymbol{B}^*$ can be estimated up to a sign, if there is enough variation in the covariate values.

A more formal proof of the identifiability of *Models 3* and *4* is a topic of future research.

JAGS model code

The JAGS code is for a 3-variate 2-level model, whereby nurses (i) are nested in nursing units (j) is given by:

$$
\begin{aligned}
\boldsymbol{y}_{ij} &= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_{1ij} + \boldsymbol{\beta}_2 x_{2j} + \boldsymbol{u}_j + \boldsymbol{\delta}_{ij}, \\
\boldsymbol{\delta}_{ij} &= \boldsymbol{\Lambda}_{ij} F_{ij} + \boldsymbol{\varepsilon}_{ij}, \quad \boldsymbol{\Lambda}_{ij} = \boldsymbol{\beta}_0^* + \boldsymbol{\beta}_1^* x_{1ij} + \boldsymbol{\beta}_2^* x_{2j} + \boldsymbol{u}_j^*, \\
\boldsymbol{u}_j &\sim N(\mathbf{0}, \Sigma_u), \quad \boldsymbol{u}_j^* \sim N(\mathbf{0}, \Sigma_u^*), \\
F_{ij} &\sim N(0,1), \quad \boldsymbol{\varepsilon}_{ij} \sim N(\mathbf{0}, \Sigma_\varepsilon),
\end{aligned}
\tag{7.15}
$$

```
model
{
  for (i in 1:N)
  # N is the total number of observations
  {
    Y[i,1:3]~dmnorm(mu[i,],tau[,])
    # Y has three dimensions
    mu[i,1:3]<-beta1[]*x1[i]+u[unit[i],]+
               (lamx1[B[unit[i]],]*x1[i]+u.l[unit[i],])*F[i]
    # u is the random intercept, including beta0
    # x1 is within-group covariate
    # B is the indicator for mixture prior of u.l
    # u.l is the random intercept in the factor loadings
    F[i]~dnorm(0,1)
    # F is the common factor with standard normal distribution
  }

  for (i in 1:nu)
  # nu is the total number of nursing unit
  {
    u[i,1:3]~dmnorm(mu.u[i,],tau.u[,])
    B0[i]~dbern(0.5)
    B[i]<-B0[i]+1
    # B is the indicator for mixture prior of u.l
    u.l[i,1:3]~dmnorm(mu.l[B[i],i,],tau.ul[,])
    mu.u[i,1:3]<-beta0[]+beta2[]*x2[i]
    for (k in 1:2)
    {
      mu.l[k,i,1:3]<-lamx0[k,]+lamx2[k,]*x2[i]
      # x2 is between-group covariate
    }
  }

  for (i in 1:3)
```

```
  {
    beta0[i]~dnorm(0,0.000001)
    beta1[i]~dnorm(0,0.000001)
    beta2[i]~dnorm(0,0.000001)

    lamx0[1,i]<-lam0[i]
    lam0[i]~dnorm(0,0.000001)
    lamx0[2,i]<-lam0[i]*(-1)

    lamx1[1,i]<-lam1[i]
    lam1[i]~dnorm(0,0.000001)
    lamx1[2,i]<-lam1[i]*(-1)

    lamx2[1,i]<-lam2[i]
    lam2[i]~dnorm(0,0.000001)
    lamx2[2,i]<-lam2[i]*(-1)
  }

  tau[1:3,1:3]~dwish(T[,],3)
  sigma2[1:3,1:3]<-inverse(tau[,])

  tau.u[1:3,1:3]~dwish(T[,],3)
  sigma2.u[1:3,1:3]<-inverse(tau.u[,])

  tau.ul[1:3,1:3]~dwish(T[,],3)
  sigma2.ul[1:3,1:3]<-inverse(tau.ul[,])

  # T[,] is a 3x3 identity matrix defined in the data part
}
```

***Implied marginal covariance matrix and kurtosis of* Model 4**

Calculating the second central moment of $\boldsymbol{y}_{ijkl}$ in *Model 4* results in the marginal covariance matrix Ψ_{ijkl}:

$$\Psi_{ijkl} = E[(\boldsymbol{y}_{ijkl} - E(\boldsymbol{y}_{ijkl}))(\boldsymbol{y}_{ijkl} - E(\boldsymbol{y}_{ijkl}))^T], \tag{7.16}$$

where

$$\boldsymbol{y}_{ijkl} - E(\boldsymbol{y}_{ijkl}) = \boldsymbol{u}_l + \boldsymbol{u}_{kl} + \boldsymbol{u}_{jkl} + (\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl} + \boldsymbol{u}^*_l + \boldsymbol{u}^*_{kl} + \boldsymbol{u}^*_{jkl}) F_{ijkl} + \boldsymbol{\varepsilon}_{ijkl}. \tag{7.17}$$

Because of the mutual independence of all random effects, factor (standard normal distributed) and residuals, equation (7.16) could be further written as:

$$
\begin{aligned}
\Psi_{ijkl} &= E[(\boldsymbol{u}_l + \boldsymbol{u}_{kl} + \boldsymbol{u}_{jkl} + (\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl} + \boldsymbol{u}^*_l + \boldsymbol{u}^*_{kl} + \boldsymbol{u}^*_{jkl}) F_{ijkl} + \boldsymbol{\varepsilon}_{ijkl}) \\
&\quad (\boldsymbol{u}_l + \boldsymbol{u}_{kl} + \boldsymbol{u}_{jkl} + (\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl} + \boldsymbol{u}^*_l + \boldsymbol{u}^*_{kl} + \boldsymbol{u}^*_{jkl}) F_{ijkl} + \boldsymbol{\varepsilon}_{ijkl})^T] \\
&= E(\boldsymbol{u}_l \boldsymbol{u}_l^T + \boldsymbol{u}_{kl} \boldsymbol{u}_{kl}^T + \boldsymbol{u}_{jkl} \boldsymbol{u}_{jkl}^T + (\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl} + \boldsymbol{u}^*_l + \boldsymbol{u}^*_{kl} + \boldsymbol{u}^*_{jkl}) \\
&\quad (\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl} + \boldsymbol{u}^*_l + \boldsymbol{u}^*_{kl} + \boldsymbol{u}^*_{jkl})^T + \boldsymbol{\varepsilon}_{ijkl} \boldsymbol{\varepsilon}^2_{ijkl}) \\
&= E(\boldsymbol{u}_l \boldsymbol{u}_l^T + \boldsymbol{u}_{kl} \boldsymbol{u}_{kl}^T + \boldsymbol{u}_{jkl} \boldsymbol{u}_{jkl}^T + (\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl})(\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl})^T + \boldsymbol{u}^*_l \boldsymbol{u}^{*T}_l \\
&\quad + \boldsymbol{u}^*_{kl} \boldsymbol{u}^{*T}_{kl} + \boldsymbol{u}^*_{jkl} \boldsymbol{u}^{*T}_{jkl}) + \boldsymbol{\varepsilon}^T_{ijkl}) \\
&= (\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl})(\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl})^T + \Sigma_u + \Sigma_h + \Sigma_c + \Sigma^*_u + \Sigma^*_h + \Sigma^*_c + \Sigma_\varepsilon
\end{aligned}
\tag{7.18}
$$

The kurtosis for each marginal distribution of $\boldsymbol{y}_{ijkl}$ is calculated from the second and fourth standard moments, as follows:

$$
kurtosis = E(y_{ijkl} - E(y_{ijkl}))^4 / (E(y_{ijkl} - E(y_{ijkl})^2))^2 - 3, \tag{7.19}
$$

where $y_{ijkl} - E(y_{ijkl})$ is the univariate version of equation (7.17). It is readily seen that:

$$
kurtosis = \frac{E(u_l + u_{kl} + u_{jkl} + (\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl} + u^*_l + u^*_{kl} + u^*_{jkl}) F_{ijkl} + \varepsilon_{ijkl})^4}{(E(u_l + u_{kl} + u_{jkl} + (\boldsymbol{B}^* \boldsymbol{x}^*_{ijkl} + u^*_l + u^*_{kl} + u^*_{jkl}) F_{ijkl} + \varepsilon_{ijkl})^2)^2} - 3. \tag{7.20}
$$

Together with some basic moment calculations for a normal random variables x with mean zero and standard deviation σ, e.g.: the first until fourth moments for x are $0, \sigma^2, 0, 3\sigma^4$ respectively, we could then work out the kurtosis that follows the expression (8.6).

8

MULTILEVEL HIGHER ORDER FACTOR MODEL: JOINT MODELING OF A MULTILEVEL FACTOR ANALYTIC MODEL AND A MULTILEVEL COVARIANCE REGRESSION MODEL

Chapter 8 is based on the paper:

Li, B., Bruyneel, L., and Lesaffre, E. (2014). Multilevel higher-order factor model: Joint modeling of a multilevel factor analytic model and a multilevel covariance regression model. Structural Equation Modeling: A Multidisciplinary Journal. (submitted)

Abstract

We propose a multilevel higher-order factor (MHOF) model that combines a multilevel factor analytic (MFA) model and a multilevel covariance regression (MCR) model, in a Bayesian context. The latter model was proposed by Li et al. (2013a) to express the covariance matrix of the responses with a mixed effects structure via a factor analytic model with structured factor loadings. The MHOF model replaces the responses in the MCR part with the factor scores coming from an MFA model, while preserving the features of the MCR model. This is quite efficient when the responses of the MCR model are not measured directly but are latent variables such as the burnout measurements in our example data set.

8.1 Introduction

Factor analytic methods are commonly used to obtain insight into high-dimensional data structures. When in addition the data have a multilevel structure, this hierarchical structure should be incorporated into the factor analysis (Longford and Muthén, 1992). Modeling high-dimensional multivariate data can therefore become quite challenging, and becomes even more demanding when some of the classical assumptions such as normality and homoscedasticity are not met. In a previous paper (Li et al., 2013a), we proposed a Bayesian multivariate multilevel model with a built-in factor analytic model for the covariance matrix (also called the multilevel covariance regression, denoted as MCR) to handle a large four-level structured data set with a complex heteroscedastic structure with a three-dimensional burnout response that has a non-normal but bounded distribution. The MCR approach proved to be a useful technique to get insight into this complex data structure.

Notwithstanding the complex nature of the data tackled in Li et al. (2013a), it was realized that the problem is even more complex. Indeed, the three-dimensional burnout vector was obtained in the early 1980s through single-level factor models based on a relatively small sample of US health and service occupation workers. From a preliminary list of 47 items measuring burnout, a 3-dimensional FA solution measured by 22 items emerged (Maslach and Jackson, 1981). To simplify the practical application of their 3-dimensional FA solution, the authors suggested to use three sum scores of the original items representing the three dimensions of the burnout measurements (Maslach et al., 1996). Although the dimensions were labeled after the factor analysis rather than deduced theoretically (Schaufeli et al., 1993), this factor solution is generally supported (Worley et al., 2008). However, FA approaches accounting for the multilevel research design that many researchers deploy, are notably absent. The data set used in Li et al. (2013a) has a four-level structure: burnout was measured from nurses within nursing units in hospitals across 12 European countries. After taking into account this four-level structure the burnout dimensions may be different from the solution suggested by Maslach and Jackson (1981). We therefore argue that the original factor solution (and the resulting practical approach based on sum scores) obtained by Maslach and Jackson (1981) may not apply to the RN4CAST population. It thus seems preferable to evaluate the configural invariance of the three-factor burnout solution to the RN4CAST data, while taking into account the multilevel structure of these data. In addition, since the MCR model applied in Li et al. (2013a) revealed that the multilevel model based on the sum scores showed heteroscedasticity in its covariance structure, it also seemed necessary to combine the multilevel FA (MFA) model with the MCR approach here. The aim of this paper is to jointly model an MCR and an MFA model with a Bayesian approach by using the factor scores from the MFA model as the responses in the MCR model. We argue that this approach overcomes several drawbacks associated with the use of sum scores.

There are typically two ways to combine the two models. The first is a two-stage approach whereby the factor solutions are determined and then used as input for the MCR model. The second is a joint approach whereby the two estimation processes are performed simultaneously. A comparison of these two approaches will be done here using a simulation

study. Joint modeling of the MFA model and the MCR model results in a multilevel higher-order factor (MHOF) model. Basic ideas of a higher-order factor model can be found in e.g. Maruyama (1997). Here, we basically extend a complex higher-order factor structure to the multilevel setting. The proposed MHOF model can be seen as a particular but complex example of multilevel structural equation modeling (MSEM) (Muthén, 1994; Hox, 2010), but also as a multiple indicators multiple causes (MIMIC) model (Jöreskog and Goldberger, 1975). However, the structural part of our modeling (i.e. MCR) has, to our knowledge, not received attention in any MSEM or MIMIC model. Our modeling approach can handle the following aspects simultaneously:

- multiple indicators with a multilevel structure;
- higher-order common factors;
- complex structured heteroscedasticity of the lower-order common factors.

We illustrate the MHOF model with the same unique RN4CAST data set that was used in Li et al. (2013a). Applying the joint model to the full multi-country set of RN4CAST data is however computationally demanding. But more importantly, the coefficients estimates and the covariance pattern of burnout may vary across countries likely necessitating an even more complex statistical approach. We therefore applied the MHOF model to the Belgian data only. The extension of our approach to the twelve countries will be the topic of interest in a subsequent paper with a more clinical focus.

The remainder of this paper is structured as follows. In Section 8.2, more details on the motivating data set are provided and the treatment of the non-responses and missing data is also elaborated. We specify the research questions that triggered the development of the MHOF model at the end of this section. The proposed MHOF model and its two components (MFA and MCR models) are introduced in Section 8.3. Computational details on fitting the models to data are discussed in Section 8.4. A limited simulation study to compare the parameter estimates between the two-stage approach and the joint approach (MHOF model) is described in Section 8.5. Section 8.6 applies the proposed MHOF model to the motivating data set thereby addressing the research questions. The two-stage model and the MCR model with sum scores as responses (proposed in Li et al. (2013a)) are also applied here to make comparisons with the MHOF model. Several model assessments are performed afterwards and a clinical interpretation of the parameter estimates is provided at the end of the section. We conclude this paper with a summary of the significant advancements in statistical analyses resulting from the MHOF model.

8.2 Motivating Data set

8.2.1 Data description

We use the Belgian sample of the three-year RN4CAST nurse workforce study. The aim of this European FP7-funded project was to study the impact of system-level features in the organization of nursing care on individual measures of nurse wellbeing and patient

safety outcomes and satisfaction with care. For the majority of the countries involved in the RN4CAST study, including Belgium, a three-stage sampling design was implemented. First, a minimum of 30 general (non-specialized) hospitals were randomly selected. Second, at least two adult general medical and surgical nursing units were randomly selected in each of the participating hospitals. Third, all nurses involved in direct patient care activities in the participating nursing units were invited to participate in the study. The rationale and design of the RN4CAST study are described in detail by Sermeus et al. (2011). The current paper focuses on the multidimensional construct of job burnout in relation to several important covariates.

The most commonly used instrument for measuring burnout is the 22-item Maslach Burnout Inventory (MBI) (Maslach et al., 1996). The MBI items are each rated on a 7-point Likert scale ranging from "never" to "every day" coded from 0 to 6, corresponding to the frequency of burnout experiences, e.g. "I feel emotionally drained from my work". The three burnout dimensions extracted by Maslach and Jackson (1981) are emotional exhaustion (*EE*), depersonalization (*DP*) and personal accomplishment (*PA*). According to the definitions proposed by Maslach et al. (1996), EE assesses feelings of being emotionally overextended and exhausted by one's work, DP measures an unfeeling and impersonal response toward recipients of one's service, care, treatment, or instruction, and PA assesses feelings of competence and successful achievement in one's work with people. *PA* thus has a reverse wording from *EE* and *DP*. That is, a higher value indicates a higher degree of burnout for *EE* and *DP*, while the reverse is true for *PA*. Although sum scores of these three dimensions are commonly used in the literature to describe burnout, here we use the full set of 22 items as responses in our analyses for reasons described above. Figure 8.1 shows the distributions of each of the 22 items. Note that many items have "L" shaped (*EE*, *DP*) or "J" shaped (*PA*) distributions. The statistical approach to handle this type of distributed data is described in Section 8.6.

The descriptive statistics of the covariates involved in the later statistical analyses are shown in Table 8.1. For the covariates *work environment*, *working experience* and *work load*, we report for each level the mean of each covariate (taken as mean of the means at the lower level, i.e. are aggregated at each of the higher levels) and the range of their (mean) values. For the other covariates, i.e. *fulltime* and *surgical nursing unit*, both of which are binary, only the percentage is reported. The covariate *work environment* reflects the overall organizational-level work environment rated by each nurse and was measured by the Practice Environment Scale of the Nursing Work Index (PES-NWI) with 32 items. For each item e.g. "Praise and recognition for a job well done", nurses scored on a four-point Likert scale from 1 to 4 indicating "totally agree", "agree", "not agree" and "totally not agree", respectively. Five dimensions were summarized: *managerial support for nursing*, *doctor-nurse collegial relations*, *promotion of care quality*, *staffing and resource adequacy*, and *nurse participation in hospital affairs*, respectively. Only the first three were used here because of a multicollinearity problem (Kutney-Lee et al., 2009; Li et al., 2013b). Each of these three sub-scales have a mean between 2.5 and 3 at each level, indicating a slightly more positive than neutral feeling about the work environment. Note that the three work environment covariates are

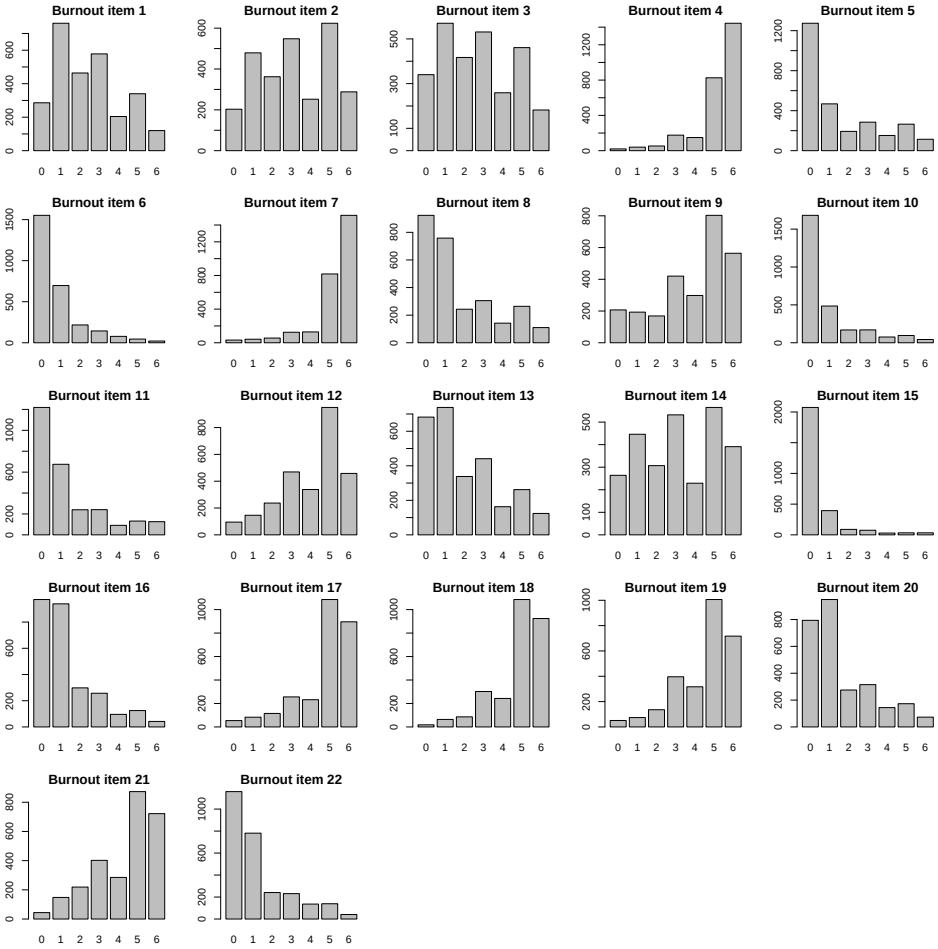

Figure 8.1: Bar plot for each of the 22 burnout items.

calculated as the mean of their non-missing components, see Section 8.2.2 on how we dealt with the missing components. A more sophisticated approach could have been used here. Namely, we could have performed again a factor analysis to determine the (five) most important dimensions of the covariate structure of 32 PES-NWI items. However, for reasons of simplicity we preferred to postpone this exercise to a next paper.

The covariate *working experience* expresses the number of years having worked as a registered nurse. The mean working experience is around 15 years, ranging from around 40 years at the nurse level to 23 years and 10 years at the nursing unit and hospital levels, respectively. The variable *fulltime* is an indicator of a full-time working nurse with more than half of the nurses working fulltime. The *work load* variable is an overall measure of the organizational-level work load in terms of the average number of patients cared for by each nurse. The work load ranges from around 4 to 22 patients at the nursing unit level, while at the hospital level the mean work load of the nursing units ranges from around 7 to 15 patients. Lastly, the variable *surgical nursing unit* is a binary covariate indicating whether a surgical nursing unit or a medical unit was examined. Almost half of the nursing units are surgical. Similar to our previous paper, we excluded the male nurses (10%) to have a more homogeneous group. Therefore Table 8.1 is based on a sample size of 2809 female nurses in 268 nursing units in 55 hospitals in Belgium.

Table 8.1: Descriptive statistics of the considered covariates in the statistical models

Covariates		Hospital	Nursing unit	Nurse
Work environment†	*ms*	2.60 (2.19,3.08)	2.59 (1.71,3.57)	–
	dn	2.58 (2.21,2.88)	2.57 (1.77,3.23)	–
	pc	2.75 (2.41,3.05)	2.75 (2.19,3.16)	–
Working experience†		15.21 (9.27,20.00)	15.28 (4.33,27.88)	15.53 (0.08,41.00)
Fulltime∗		–	–	53.14
Work load†		10.97 (7.26,15.45)	10.89 (3.62,21.33)	–
Surgical nursing unit∗		–	46.64	–

ms: managerial support for nursing; *dn*: doctor-nurse collegial relations;
pc: promotion of care quality
∗: Percentage at the corresponding level is calculated
†: Mean and range are calculated at each level, if applicable

8.2.2 Treatment of non-response and missing data

Non-response

The RN4CAST study suffered from a relatively large non-response rate, especially at the hospital and nurse levels. Namely, 56 out of the 104 invited hospitals in Belgium participated to the study. This ranks middle in terms of non-response rate across all RN4CAST countries. A check for representativeness (hospital type, size) was carried out at the planning stage. When necessary, some corrective actions (such as extra motivating hospitals to participate) were taken. This was done e.g. in the Brussels-Capital region. The response

rate at the individual nurse level was 72.07%, which is higher than the average response rate (60.2%) in the RN4CAST study. Since no information is available on the non-respondents, no corrective actions for non-response could be undertaken at the analysis stage. However, in a similar nurse survey conducted in the US (Smith, 2008), the author assessed the non-response bias by randomly selecting a number of non-respondents and motivating them to fill in the questionnaires. No differences were found with regard to nurses' assessments of their work environment and burnout. Therefore we argue our findings are likely not dramatically affected by the non-response data.

Missingness in the burnout measurements

There are about 10% nurses having at least one of the 22 burnout items missing. Technically, missing responses are easily dealt with using the Markov Chain Monte Carlo (MCMC) technique in Bayesian modeling. For instance, with WinBUGS the missing response needs to be indicated by 'NA'. In the MCMC iterative procedure, the missing response is then automatically imputed at each iteration, thereby actually performing a multiple imputation approach at convergence. Here it implies that for each iteration, the missing values of the 22 items are imputed via the likelihood based on the current parameter estimates and will be involved in the estimation of parameters for the next iteration. To this end, the parameter estimates are actually based on the observed and the multiple imputed data.

Missingness in covariates

We first note that the three work environment covariates are calculated as the mean of their non-missing components. Hence, when some of the components are missing they are not taken into account into the mean score. This assumes a too simple missing-data pattern, actually missing completely at random. For all covariates in Table 8.1, only *working experience* and *fulltime* have missing values with the missing rates about 11% and 1%, respectively. To maximally exploit the data set, we applied a Bayesian imputation scheme for each covariate. But this now also requires a statistical model for the joint distribution of the two covariates. Ibrahim et al. (2002) pointed out that the joint model specification is problematic when both the categorical and continuous covariates have missing values, which is the case here for some of the observations. This is one of the the motivations for them to come up with the solution for the joint distribution a sequence of one-dimensional conditional distributions. Applying this to our situation, which is much simpler with only two covariates that have occasionally missing data, implies that $p(x_1, x_2 \mid \boldsymbol{\psi}, \ldots) = p(x_1 \mid x_2, \psi_1, \ldots)p(x_2 \mid \psi_2, \ldots)$. That is, the joint distribution of the two covariates is written as the product of two conditional distributions with x_1 and x_2 the *working experience* and *fulltime*, respectively and $\ldots$ signifies that other covariates are included in the model specification. $\boldsymbol{\psi}$ represents all parameters involved in the joint distribution which could be divided into ψ_1 and ψ_2 belonging to the two conditional models, respectively. For the conditional distribution $p(x_1 \mid x_2, \psi_1, \ldots)$ we assumed a Gaussian linear model for x_1 with regressors all covariates in Table 8.1 (hence including *fulltime*). The conditional distribution $p(x_2 \mid \psi_2, \ldots)$ has

a logistic regression model with again all complete covariates as predictors. Obviously, the parameters of the two conditional distributions must be estimated using only the subjects with non-missing regressor values in their respective model. A more detailed specification of the models can be found in the Supplementary Materials [1]. Finally, the parameters of the two conditional imputation models are estimated simultaneously with the MHOF model for the burnout responses.

8.2.3 Research questions

In this paper we wish to examine three research questions related to the burnout measurements. We believe that the MHOF model developed in the next section in combination with the large RN4CAST data set will allow us to address these questions appropriately. Our research questions are:

- Is there any evidence of configural invariance/ non-invariance (Steenkamp and Baumgartner, 1998) comparing the factor structure of the current data set with the one proposed by Maslach and Jackson (1981)? That is, we wish to evaluate whether the two factor structures are the same in terms of the number of common factors and the items within each common factor dimension.

- Are the means of the latent burnout dimensions (factor solutions) correlated with the organizational-level and individual-level characteristics?

- Are the correlations among the latent burnout dimensions stable across hospitals, nursing units and nurses, after taking into account a rich set of confounders at different levels?

The first research question involves a multilevel FA model based on the original 22 burnout items. We will however show that for all three research questions the MHOF model is needed.

8.3 Proposed model

The Multilevel Higher-Order Factor (MHOF) model contains two parts as illustrated in Figure 8.2: an MFA model and an MCR model. In this section, we first describe these two components and then combine them to establish the MHOF model. To simplify matters, we assume in this section that in all models the responses are normally distributed. In Section 8.6 we explain how we dealt with non-normality of the 22 items in the RN4CAST study.

8.3.1 Multilevel factor analytic (MFA) model

The factor model aims to summarize a (large) number of measurements with a limited set of (latent) variables (called common factors) thereby preserving most of the original information. The MFA model is a multilevel version of the classical confirmatory factor model, see

[1] All Supplementary materials in this chapter can be found in the website where it is published.

Hox (2010) and Goldstein (2010) for a general and detailed description of the MFA model. Our motivating data set has a three-level structure, i.e. nurses within nursing units within hospitals, implying a three-level factor model. Since our interest lies in further modeling of the nurse-level common factors, we estimate an unstructured covariance matrix at the nursing unit level and hospital level respectively and keep only the nurse-level factor structure. The considered three-level factor model is therefore:

$$\begin{gathered} \boldsymbol{y}_{ijk} = \boldsymbol{\mu} + \boldsymbol{b}_{jk} + \boldsymbol{b}_k + \boldsymbol{L}\boldsymbol{z}_{ijk} + \boldsymbol{\varepsilon}_{ijk}, \\ \boldsymbol{b}_{jk} \sim N(\mathbf{0}, \Sigma_{bu}), \quad \boldsymbol{b}_k \sim N(\mathbf{0}, \Sigma_{bh}), \quad \boldsymbol{z}_{ijk} \sim N(\mathbf{0}, \Sigma_z), \\ \boldsymbol{\varepsilon}_{ijk} \sim N(\mathbf{0}, diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_P^2)), \end{gathered} \tag{8.1}$$

where $\boldsymbol{y}_{ijk}$ represents the $P = 22$-dimensional response for nurse i in nursing unit j in hospital k; $\boldsymbol{z}_{ijk}$ is the Q-dimensional nurse-level common factor following a multivariate normal distribution with an unstructured covariance matrix Σ_z, and $\boldsymbol{L}$ is its $P \times Q$-dimensional loading matrix; $\boldsymbol{b}_{jk}$ and $\boldsymbol{b}_k$ represent the higher-level random effects with an unstructured covariance matrix Σ_{bu} and Σ_{bh} respectively. $\boldsymbol{\varepsilon}_{ijk}$ represents the nurse-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 in Section 8.3.4. The upper part of Figure 8.2 displays graphically model (8.1).

8.3.2 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 analytic model with factor loadings having a linear model. The covariance matrix of the multivariate responses, determined by the factor part, depends quadratically on covariates. The MCR model extends Hoff and Niu's covariance regression model into the multilevel situation. In an MCR model, both the mean and the covariance parts are allowed to have a mixed effects model and can be modeled simultaneously. In Li et al. (2013a), a Bayesian approach was opted, as here, for estimating the parameters. The authors showed that this approach performed well on the RN4CAST data.

The MCR model for the Belgian data of the RN4CAST project is given by:

$$\begin{gathered} \boldsymbol{z}_{ijk} = \boldsymbol{B}\boldsymbol{x}_{ijk} + \boldsymbol{u}_{jk} + \boldsymbol{u}_k + \boldsymbol{\delta}_{ijk}, \\ \boldsymbol{\delta}_{ijk} = \boldsymbol{\Lambda}_{ijk} F_{ijk} + \boldsymbol{\varepsilon}_{ijk}, \quad \boldsymbol{\Lambda}_{ijk} = \boldsymbol{B}^* \boldsymbol{x}^*_{ijk} + \boldsymbol{u}^*_{jk} + \boldsymbol{u}^*_k, \\ \boldsymbol{u}_{jk} \sim N(\mathbf{0}, \Sigma_u), \quad \boldsymbol{u}_k \sim N(\mathbf{0}, \Sigma_h), \\ \boldsymbol{u}^*_{jk} \sim N(\mathbf{0}, \Sigma^*_u), \quad \boldsymbol{u}^*_k \sim N(\mathbf{0}, \Sigma^*_h), \\ F_{ijk} \sim N(0, 1), \quad \boldsymbol{\varepsilon}_{ijk} \sim N(\mathbf{0}, \Sigma_\varepsilon), \end{gathered} \tag{8.2}$$

where $\boldsymbol{z}_{ijk}$ is the Q-dimensional latent burnout variables for nurse i in nursing unit j in hospital k; $\boldsymbol{B}$ is a $Q \times m$ matrix of fixed effects associated with the m-dimensional vector $\boldsymbol{x}_{ijk}$ gathering information from all levels. $\boldsymbol{u}_{jk}$ and $\boldsymbol{u}_k$ represent the Q-dimensional random

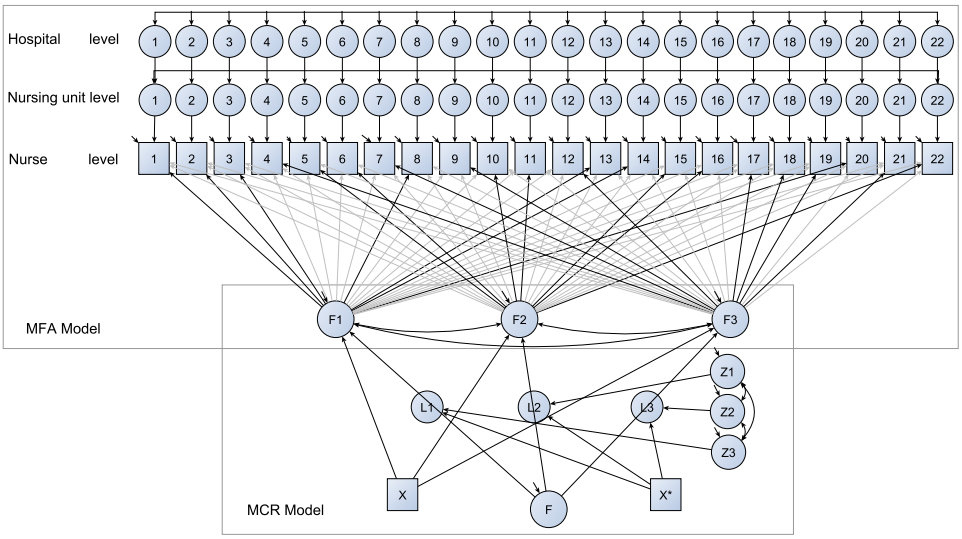

Figure 8.2: The MHOF model: The upper part describes the MFA model, which is a three-level factor analytic model and the lower part describes the MCR part, which is a three-level covariance regression model with the factor loadings having a mixed effects structure. A square represents an observed response or covariate. A circle represents a latent variable. L1-L3 represent the loadings of F1-F3 on F respectively. A unidirectional arrow connecting two objects represents a regression relationship, and a bidirectional arrow connecting two objects represents the correlation of the two objects. A small arrow pointed to a single object represents the estimation of the error term and the multiple arrows at the nursing unit and hospital levels represent the estimation of the general covariance matrix at each level respectively. The arrows with gray color represent the cross loadings.

intercepts in the mean part at each higher level with general covariance matrices Σ_u and Σ_h, respectively. The residuals $\boldsymbol{\delta}_{ijk}$ are decomposed into a fixed part assumed constant across nurses with general covariance matrix Σ_ε and a part that varies with m^* characteristics $\boldsymbol{x}^*_{ijk}$ possibly different from $\boldsymbol{x}_{ijk}$ and allowed to depend on all levels through a factor part. $\boldsymbol{B}^*$ is a $Q \times m^*$ matrix of fixed effects associated with $\boldsymbol{x}^*_{ijk}$, while $\boldsymbol{u}^*_{jk}$ and $\boldsymbol{u}^*_k$ represent the Q-dimensional random intercepts in the loadings at each higher level with general covariance matrices Σ^*_u and Σ^*_h, respectively. The lower part of Figure 8.2 displays graphically model (8.2).

8.3.3 Multilevel higher-order factor (MHOF) model

The MHOF model is a combination of an MFA and an MCR model. The combination is 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. Figure 8.2 shows graphically how the MFA and MCR models are combined into the MHOF model. The MHOF model is formally

defined as:

$$
\begin{aligned}
&\boldsymbol{The\ MFA\ part:}\\
&\qquad \boldsymbol{y}_{ijk} = \boldsymbol{\mu} + \boldsymbol{b}_{jk} + \boldsymbol{b}_{k} + \boldsymbol{L}\boldsymbol{z}_{ijk} + \boldsymbol{\varepsilon}^{FA}_{ijk},\\
&\qquad \boldsymbol{b}_{jk} \sim N(\boldsymbol{0}, \Sigma_{bu}), \quad \boldsymbol{b}_{k} \sim N(\boldsymbol{0}, \Sigma_{bh}),\\
&\qquad \boldsymbol{\varepsilon}^{FA}_{ijk} \sim N(\boldsymbol{0}, diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_P^2)).\\
&\boldsymbol{The\ MCR\ part:}\\
&\qquad \boldsymbol{z}_{ijk} = \boldsymbol{B}\boldsymbol{x}_{ijk} + \boldsymbol{\delta}_{ijk},\\
&\boldsymbol{\delta}_{ijk} = \boldsymbol{\Lambda}_{ijk}F_{ijk} + \boldsymbol{\varepsilon}^{CR}_{ijk}, \quad \boldsymbol{\Lambda}_{ijk} = \boldsymbol{B}^{*}\boldsymbol{x}^{*}_{ijk} + \boldsymbol{u}^{*}_{jk} + \boldsymbol{u}^{*}_{k},\\
&\qquad \boldsymbol{u}^{*}_{jk} \sim N(\boldsymbol{0}, \Sigma^{*}_{u}), \quad \boldsymbol{u}^{*}_{k} \sim N(\boldsymbol{0}, \Sigma^{*}_{h}),\\
&\qquad F_{ijk} \sim N(0, 1), \quad \boldsymbol{\varepsilon}^{CR}_{ijk} \sim N(\boldsymbol{0}, \Sigma_{\varepsilon}).
\end{aligned}
\tag{8.3}
$$

The MFA part is the same as in model (8.1) except for the distribution of $\boldsymbol{z}_{ijk}$, which is incorporated in the MCR model later on. We refer to it as the lower-order factor in the MHOF model. The $\boldsymbol{z}_{ijk}$ in the MCR part are output from the MFA model. All parameters have the same meaning as in model (8.2) except that now the random effects at the nursing unit and hospital levels in the mean, i.e. $\boldsymbol{u}_{jk}$ and $\boldsymbol{u}_{k}$, have been removed for identification purposes discussed in Section 8.3.4. The common factors F_{ijk} are also called the higher-order factors in the MHOF model.

Model (8.3) could be also written in a compact form as:

$$
\begin{aligned}
&\boldsymbol{y}_{ijk} = \boldsymbol{\mu} + \boldsymbol{b}_{jk} + \boldsymbol{b}_{k} + \boldsymbol{L}\boldsymbol{B}\boldsymbol{x}_{ijk} + \boldsymbol{L}(\boldsymbol{B}^{*}\boldsymbol{x}^{*}_{ijk} + \boldsymbol{u}^{*}_{jk} + \boldsymbol{u}^{*}_{k})F_{ijk} + \boldsymbol{L}\boldsymbol{\varepsilon}^{CR}_{ijk} + \boldsymbol{\varepsilon}^{FA}_{ijk},\\
&\boldsymbol{b}_{jk} \sim N(\boldsymbol{0}, \Sigma_{bu}), \quad \boldsymbol{b}_{k} \sim N(\boldsymbol{0}, \Sigma_{bh}), \quad \boldsymbol{\varepsilon}^{FA}_{ijk} \sim N(\boldsymbol{0}, diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_P^2)).\\
&\boldsymbol{u}^{*}_{jk} \sim N(\boldsymbol{0}, \Sigma^{*}_{u}), \quad \boldsymbol{u}^{*}_{k} \sim N(\boldsymbol{0}, \Sigma^{*}_{h}), \quad F_{ijk} \sim N(0, 1), \quad \boldsymbol{\varepsilon}^{CR}_{ijk} \sim N(\boldsymbol{0}, \Sigma_{\varepsilon}).
\end{aligned}
\tag{8.4}
$$

From this expression we can see that the nursing unit and hospital level-specific random intercepts, i.e. $\boldsymbol{b}_{jk}$ and $\boldsymbol{b}_{k}$, 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 $\boldsymbol{z}_{ijk}$. We will come back to this in the next section on identification issues.

The MCR model expresses the covariances through a single factor structure whereby the factor loadings are assumed to be random and modeled with a mixed effects structure. The covariance matrix built by the factor loadings then has a mixed effects structure accordingly, and the implied marginal covariance matrix for the latent common factors is given by:

$$
\Psi_{ijk} = (\boldsymbol{B}^{*}\boldsymbol{x}^{*}_{ijk})(\boldsymbol{B}^{*}\boldsymbol{x}^{*}_{ijk})^{T} + \Sigma^{*}_{u} + \Sigma^{*}_{h} + \Sigma_{\varepsilon}. \tag{8.5}
$$

Note that in our example the dimension of the response in the MCR model is $Q = 3$ and that the marginal covariance matrices Ψ_{ijk} span the whole of the 3×3 matrices at each value of $\boldsymbol{x}^{*}_{ijk}$.

Similar as in Li et al. (2013a), we can prove that the marginal distribution for each of the three common factors is not normal. While for each marginal density, the skewness is still zero when all random effects are mutually independent(assumed here), the kurtosis for the qth marginal density of the lower-order factor becomes:

$$\text{kurtosis}_q = \frac{6a_q^{*2} + 12a_q^* b_q}{(a_q^* + b_q + c_q)^2}. \tag{8.6}$$

In expression (8.6), a_q^*, b_q, c_q are the qth diagonal elements of $(\Sigma_u^* + \Sigma_h^*)$, $(\boldsymbol{B}^* \boldsymbol{x}_{ijk}^*)(\boldsymbol{B}^* \boldsymbol{x}_{ijk}^*)^T$ and Σ_ε, respectively. From this expression we can conclude that the marginal densities are leptokurtic unless the variance of the random effects in the factor loadings is zero ($a_q^* = 0$), and that the kurtosis also depends on the covariates in the factor loadings.

8.3.4 Identification issues

There are three main identification concerns for the MHOF model, coming from the MFA part, the MCR part, and the joint modeling of these two models. We describe these one by one.

Firstly, 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 lower-order factors $\boldsymbol{z}_{ijk}$. A common way is to set the first loading for each common factor to a constant value, usually 1, see, e.g. Kline (2010). Cross-loadings are in practice most often small but non-zero, that is why we estimated them as well. This, however, violates the rule of having at least Q^2 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 and thereby overcoming the identification problem. This was called Bayesian structural equation modeling (BSEM) by Muthén and Asparouhov (2012). The largest cross-loading estimates obtained by Maslach and Jackson (1981) were close to 0.4. Therefore we have chosen an informative normal prior with mean zero and standard deviation around 0.22 for all cross-loadings. This choice corresponds to a 95% prior range of the cross loadings from -0.45 to 0.45 and allowed the factor loadings to be identified. In addition we assumed that the P unique factors (residuals) are mutually independent.

Secondly, for the MCR part, the identification issue is the same as described in the paper by Li et al. (2013a). That is, we scaled the higher-order factors with a standard normal distribution, and assigned a mixture prior for the random effects $\boldsymbol{u}_{jk}^*$ and $\boldsymbol{u}_k^*$ and the coefficient $\boldsymbol{B}^*$ respectively to overcome the "flipping states" issue. Given a reasonable variation of $\boldsymbol{x}_{ijk}^*$, all parameters in the MCR part, including the whole covariance matrix of the residual $\boldsymbol{\varepsilon}_{ijk}^{CR}$, could be identified.

The final 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 lower-order factors that connect the MFA part and the MCR part. From the MHOF model expressed in both models (8.3) and (8.4), the nursing unit and hospital level-specific covariance matrices formulated

by $\boldsymbol{b}_{jk}$ and $\boldsymbol{b}_k$, respectively, are saturated, therefore adding random effects to the lower-order factor scores $\boldsymbol{z}_{ijk}$, 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 ($\boldsymbol{u}_{jk}$ and $\boldsymbol{u}_k$) when modeling the mean of burnout (i.e. the lower-order factor scores $\boldsymbol{z}_{ijk}$). There is no such constraint for the modeling the variance/covariance of the factor scores, i.e. modeling the heteroscedasticity. It is valid to assume that the variance of burnout 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 the lower-level factor(s) could be multilevel structured, even when the modeling of the mean structure does not have a hierarchical structure. We refer to the paper of Li et al. (2013a) for more details on the MCR model.

8.4 Computational procedure

We have opted for the Bayesian approach, for a variety of reasons. One is that the considered model has a relatively large number of random effects, which is known to cause problems in maximum likelihood algorithms. For the three-level MHOF model, there are $2P$ random intercepts for the P indicators, Q lower-order common factors, one higher-order common factor, and two extra random intercepts for each of the Q factor loadings for the higher-order common factor. For our analyses, the Jags (Just another Gibbs sampler) MCMC program (Plummer, 2003) was used through the R package *rjags* (Plummer, 2013). We ran 50,000 iterations for the burn-in part and another 50,000 iterations for model estimation. The Brooks-Gelman-Rubin plots with the potential scale reduction factor (Brooks and Gelman, 1998) (PSRF, which should be smaller than 1.1) was used for convergence checking. In addition, we ran the Markov chain until the Monte Carlo error was around or smaller than 5% of the posterior standard deviation.

In a Bayesian analysis, all parameters need a prior distribution to express what was already known beforehand. 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 lower-order factors to overcome the identification issue described in Section 8.3.4. 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 a 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, except for the mixture priors of some of the loading parameters (see above).

Models were compared using a pseudo Bayes factor (PSBF), see e.g. Lesaffre and Lawson (2012) for details on the computation. The classical criteria such as DIC (Spiegelhalter et al., 2002), though theoretically feasible for the BOS approach (see Section 8.6), is practically cumbersome as the BOS approach requires an integration for each observation. In Li et al.

(2013a) the DIC was determined approximately by replacing the bounded burnout response with a scaled logit transformation of the response. They demonstrated that the approximation was acceptable when the number of the categories for the sum scores of burnout is large (around 40). Here, however, there are only seven categories for each burnout item and hence the approximate solution will be too crude. Therefore, we limited ourselves to the PSBF criterion to compare models. The logarithm of the PSBF to compare model A with model B is denoted as $\ell PSBF_{A,B}$ whereby positive values indicate preference for the second model.

8.5 Comparison of the two-stage approach and the MHOF model: a limited simulation study

For the two-stage model, we first run the MFA model, extract the nurse-level factor scores, and then model these factor scores with an MCR model. For the MHOF model these two stages are done in one step. The two-stage model is clearly less computationally intensive, and therefore a possibly attractive alternative to the MHOF model. However, it may lead to biased parameter estimates if there is heteroscedasticity in the covariance part of the model and it may result in less efficient estimation because of the separation of the correlated information in the MFA model and the MCR model. To assess these statements we conducted a limited simulation study to compare the performance of the two-stage model and our proposed MHOF model. The performance was assessed by the standardized bias and the 95% coverage for each parameter, which is suggested by Burton et al. (2006). The standardized bias is calculated as $100|\bar{\hat{\beta}} - \beta|/SE(\hat{\beta})$, with β the true value for each parameter, and $\bar{\hat{\beta}}$ and $SE(\hat{\beta})$ the mean and the standard deviation of the estimates of all simulations, respectively. When the standardized bias is less than 40%, the parameter is considered to be reasonably well estimated (Collins et al., 2001). Further, we require that the coverage of the 95% credible interval should be close to 95%.

The settings and the main results can be found in the Supplementary Materials. To summarize the 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 highly overestimated the covariance matrix of the higher-order common factors.

8.6 Application to the RN4CAST data set

We now apply the MHOF model to the Belgian data from the RN4CAST project. The research questions connected to the hierarchical (nurses within nursing units within hospitals in Belgium) and multidimensional (three-dimensional burnout outcome) structure of the data requires the integration of a 3-level MFA model and a 3-level MCR model.

8.6.1 Choice of response and covariates

The "L" and "J" shaped distributions in Figure 8.1 require some advanced modeling approach. Well ordinal logistic or probit modeling may be considered, we opted here for another technique to handle these bounded outcome scores (BOS). We follow here the approach suggested by Lesaffre et al. (2007). This assumes that a standardized version of the burnout measurement is a coarsened latent continuous variable which has a Gaussian distribution after a logit transformation. More 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 $\log[z/(1-z)] \sim N(\mu, \sigma^2)$ and such that y is obtained by coarsening z. Lesaffre et al. (2007) showed that this approach can handle well a large variety of distributions defined on a finite interval. In our analysis, we applied this technique to all 22 burnout items.

The candidate covariates involved in the mean and/or variance structures in later analyses, are listed in Table 8.1. In order to make the regression coefficients of these variables comparable and to improve computational properties, standardized covariates (mean=0, SD=1) were used in later analyses. In addition, in order to investigate the level-specific effects of the covariates, the following decomposition was made, as suggested by Neuhaus and Kalbfleisch (1998):

$$\begin{aligned} x_{ijk} &= (x_{ijk} - \bar{x}_{jk}) + (\bar{x}_{jk} - \bar{x}_k) + \bar{x}_k \\ &= x_n + x_u + x_h, \end{aligned} \tag{8.7}$$

$\bar{x}_{jk} = \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} x_{ijk}$ and $\bar{x}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{jk}$. In expression (8.7), the lowest level variable is partitioned into three parts corresponding to the three levels, i.e. n for nurse, u for nursing unit and h for hospital. By doing so, we study the "pure" effect of the covariates at each level (for *work environment* and *work load*, there is no nurse-level partition because the lowest level measurement is nursing unit).

8.6.2 An exploratory MFA model

Before we fitted the MHOF model, an exploratory MFA was conducted to explore the factor structure of the 22 Maslach items, thereby addressing the first research question in Section 8.2. This analysis will also hint towards an appropriate MFA part of the MHOF model. *Mplus* v7 (Muthén and Muthén, 2010) was used for this task, making use of the robust maximum likelihood estimator (Asparouhov and Muthen, 2005) and an oblique rotation method. As a result, the three-factor solution based on the eigenvalues (larger than 1) summarizes the 22 items relatively well with reasonable RMSEA (0.027), and SRMR (0.027) values. RMSEA represents the root mean square error of approximation, with a value of smaller than 0.06 being considered good fit (Hu and Bentler, 1999). SRMR is the standardized root mean square residual with a value smaller than 0.08 representing a good fit (Hu and Bentler, 1999). We found that the configuration of the nurse-level factor structure is similar as in Maslach and Jackson (1981), i.e. with similar three factors *EE*, *DP* and *PA* having similar theoretical meanings as before, but with two major discrepancies besides some minor differences in the

cross-loadings. Namely, items 6 and 16 shifted from *EE* in the original structure to *DP* in the current analysis. Table 8.2 shows the differences for the loadings estimate of these two items in each study. That may indicate that the Belgian population of nurses interpreted the nursing work with patients as a category of depersonalization instead of emotional exhaustion.

Table 8.2: Comparison of factor loadings for two items from the Belgian RN4CAST study and the original Maslach et al. study

Item	Description	Belgian RN4CAST data			Maslach et al. study		
		EE	***DP***	*PA*	***EE***	*DP*	*PA*
6	Working with people all day is really a strain for me	0.14	**0.44**	-0.05	**0.61**	0.22	-0.10
16	Working directly with people puts too much stress on me	0.18	**0.40**	0.01	**0.54**	0.31	-0.06

8.6.3 MHOF model

Next, we applied the proposed MHOF model to the RN4CAST data set. The factor structure in the MFA part was obtained from the factor solution obtained in the exploratory step. For the MCR part, first all covariates in Table 8.1 were included as x_{ijk} and x^*_{ijk}. Next, only those covariates for which the 95% credible interval (CI) did not contain zero, were retained in the model. The final model contains several covariates at each level in x_{ijk} and only the nurse-level covariate *working experience* in x^*_{ijk}. Whether to include random effects in the factor loadings part of the covariance matrix was checked by comparing the PSBF of the two models without (model A) and with (model B) random effects. We obtained $\ell PSBF_{A,B} = 81.5$, justifying the inclusion of the random effects.

Table 8.3 shows the posterior mean estimates and the 95% CIs for the coefficients in the mean part and covariance part of modeling burnout. The upper part of Table 8.3 displays the mean part of modeling burnout. We can see that at the nurse level, fulltime nurses suffer more from emotional exhaustion, but have higher self-rating of personal accomplishment, than part-time working nurses. Less experienced nurses suffer more from emotional exhaustion and depersonalization than experienced nurses. At the nursing unit level, two subscales of the work environment, i.e. *promotion of care quality* and *doctor-nurse collegial relations*, were found significant on all burnout dimensions except for *doctor-nurse collegial relations* on *EE*. That is, nurses feel less emotional exhaustion, less depersonalization and more personal accomplishment when working in a nursing unit with a better environment. *Work load* of a nursing unit was found positively related to two out of three burnout subscales: the higher the average number of patients cared for by each nurse, the higher the feelings of emotional exhaustion and depersonalization. At the hospital level, we found that nurses from a hospital with more managerial support for nursing feel less emotionally exhausted than those from other hospitals. Also, nurses from a hospital with closer doctor-nurse collegial relations suffer more from emotional exhaustion than those from other hospitals, which

seems somewhat illogical. We come back to this in the clinical interpretation part. *Work load* of a hospital was found positively related to emotional exhaustion.

The lower part of Table 8.3 displays the covariance part of modeling burnout. We can see that experienced nurses have a larger variation of self-rating of personal accomplishment than less experienced nurses. A clinical interpretation of these results is given in Section 8.6.7. Figure 8.3 shows the relationships between the (co)variances/correlations among the three burnout dimensions with the variable *working experience* at the nurse level. All (co)variances/correlations remained almost constant with *working experience* except for the variance of personal accomplishment, which changed significantly with *working experience*, from around 0.7 for the unexperienced nurses to around 1.5 for the most experienced nurses. We did not find any significant covariates that could explain the variation of the variances of burnout at the hospital and nursing unit levels. Therefore the variation of the variances can only be 'explained' by the random effects. The variance estimates of the random effects are similar at these two levels for each of the three burnout dimensions respectively, which are around 0.15, 0.21 and 0.13 for emotional exhaustion, depersonalization and personal accomplishment, respectively.

8.6.4 Comparison with the two-stage model

We compared the final MHOF model with the parameter estimates obtained from a two-stage analysis. The MFA model is based on the factor solution from Section 8.6.2. For the MCR model, the variables selection strategy was identical as before.

We found that, for the two-stage model and the MHOF model, the same covariates showed significant effects on the mean modeling of burnout at the nurse and nursing unit levels. At the hospital level, however, no covariates were found to be significantly associated to any of the burnout dimensions for the two-stage model, while three covariates were found to be significant in the MHOF model. In the covariance part, both models found only the covariate *working experience* to be significantly associated to *PA*. With regard to the magnitude of the parameter estimates, we found that the two-stage model provided smaller estimates than the MHOF model, which was in line with what we found through the simulation study. A more detailed list of the parameter estimates of the two models was given in the Supplementary Materials. In conclusion, for our motivating data set, the MHOF model seemed to show more power and efficiency for the parameter estimates.

8.6.5 Comparison with the MCR model using sum scores

In this section we compare the parameter estimates obtained from the MHOF model with those obtained from the MCR model with the sum scores (defined in Maslach et al. (1996)) as the responses. The use of sum scores can be criticized because it does not adjust for different weighting of the individual items, but rather assume equal contributions of the items. The sum scores are also assumed to be mutually exclusive while in reality, the cross-loadings (often relatively small) are quite common. That is, one item could contribute to more than one sum scores (DiStefano et al., 2009). We now wished to see whether the results obtained

Table 8.3: Posterior median and 95% CI for the parameters in the mean and loadings for the MHOF model

Parameters	Burnout	Median	95% CI
In the mean part of modeling burnout			
At the nurse level			
Fulltime	*EE*	0.120	(0.015, 0.225)*
	DP	0.115	(-0.034, 0.266)
	PA	0.112	(0.031, 0.200)*
Working experience	*EE*	-0.065	(-0.118, -0.013)*
	DP	-0.123	(-0.201, -0.047)*
	PA	0.036	(-0.010, 0.081)
At the nursing unit level			
Promotion of care quality	*EE*	-0.176	(-0.256, -0.101)*
	DP	-0.225	(-0.325, -0.132)*
	PA	0.085	(0.033, 0.141)*
Doctor-nurse collegial relations	*EE*	-0.051	(-0.117, 0.013)
	DP	-0.137	(-0.219, -0.061)*
	PA	0.048	(0.005, 0.093)*
Work load	*EE*	0.092	(0.032, 0.151)*
	DP	0.099	(0.027, 0.175)*
	PA	-0.022	(-0.063, 0.016)
At the hospital level			
Managerial support for nursing	*EE*	-0.156	(-0.274, -0.029)*
	DP	-0.064	(-0.277, 0.137)
	PA	0.091	(-0.014, 0.193)
Doctor-nurse collegial relations	*EE*	0.124	(0.017, 0.229)*
	DP	0.038	(-0.152, 0.214)
	PA	-0.049	(-0.139, 0.049)
Work load	*EE*	0.188	(0.084, 0.291)*
	DP	0.073	(-0.117, 0.250)
	PA	-0.043	(-0.135, 0.056)
In covariance part of modeling burnout			
Working experience	*EE*	-0.007	(-0.161, 0.133)
	DP	-0.034	(-0.198, 0.137)
	PA	0.241	(0.158, 0.325)*

$*$: the 95% CI dose not include zero

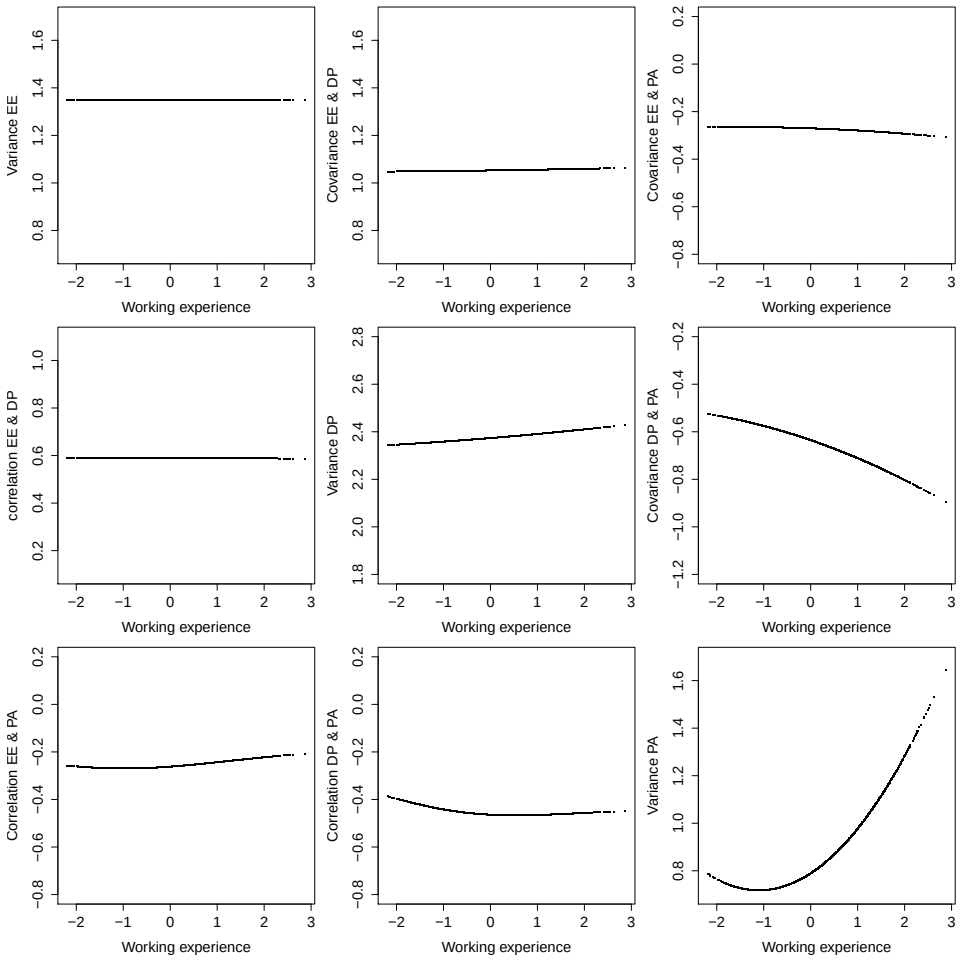

Figure 8.3: (co)Variances (upper triangle) and correlations (lower triangle) with *working experience* at the nurse level

from sum scores analysis are different on the Belgian data to those obtained from the MHOF model.

It is, however, not straightforward to compare the parameter estimates of the two models because the scales of the burnout used in the two models, namely the sum scores and the factor scores, are not exactly the same. However, we can still check whether the same covariates were detected significantly and then have an idea of the consistency of the two models. In the mean part, the MHOF model and the MCR model were highly consistent in terms of the significant covariates detected at each level, i.e. both models found significant associations between burnout dimensions and the variables *fulltime* and *working experience* at the nurse level, *promotion of care quality*, *doctor-nurse collegial relations* and *work load* at the nursing unit level, and *managerial support for nursing*, *doctor-nurse collegial relations* and *work load* at the hospital level. While in terms of the significance of the covariates on each of the burnout dimensions, there were some differences between the two models, e.g. *fulltime* was found significant on *EE* and *PA* with the MHOF model while significant on *DP* and *PA* with the MCR model. With regard to the covariance part, both models identified a strong effect of *working experience* on *PA*. A more detailed list of the parameter estimates for the MCR model with sum scores was given in the Supplementary Materials.

8.6.6 Model assessment

The following two checks were performed for model assessment. First a sensitivity analysis checked whether the multivariate normal prior for the random effects in the loadings should be replaced with a multivariate t(3) distribution, which has fatter tails. The two models were compared with PSBF, whereby $\ell PSBF_{norm,t(3)} = 10.9$ indicated that t(3)-model is preferable. However, the parameter estimates for the two models were quite close to each other: the difference of the posterior means of the two models varies around 8% of the standard deviation of the parameters (SD_p) with standard deviation of the difference 36% of SD_p.

A popular goodness of fit test in the Bayesian approach is nowadays the posterior predictive check (PPC). For the MHOF model, we used the Gelman χ^2 statistic (Gelman et al., 2013) as the discrepancy function to calculate the PPC for each of the 22 items as well as the overall PPC, 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 22 latent continuous responses instead of the observed burnout items. $\boldsymbol{\theta}$ represents all parameters including the random effects. The Bayesian P-values of PPC for the 22 burnout items range from 0.20 to 0.48, with a mean value 0.41, while the P-value for the overall PPC was 0.13. All of these P-values indicate a reasonable model fit of our proposed MHOF model.

8.6.7 Clinical interpretation

The current study extends the theoretical and empirical knowledge of burnout measured by the Maslach Burnout Inventory by examining both the mean and the covariance structure of the burnout dimensions in relation to work setting and individual factors, taking into account the latent nature of the burnout dimensions and the multilevel research design. First, with regards to the latent nature of the burnout dimensions, our findings showed that the Maslach Burnout Inventory exhibited evidence of configural invariance. The problems in factor loadings that we found with items 6 and 16 have previously been reported in a meta-analytic study of MBI factor analyses (Worley et al., 2008) and in a large multicountry nurse workforce study with a similar research design (Poghosyan et al., 2009). Second, our findings in both the mean and the variance parts of the three latent burnout dimensions have interesting clinical meanings. In the mean part, we found that both work setting and individual factors affected the burnout dimensions. *Doctor-nurse collegial relations* had a complex relationship to the burnout dimensions. At the nursing unit level, better relations between doctors and nurses are related to less feelings of depersonalization and higher rates of personal accomplishment. At the hospital level however, better relations were related to higher degrees of emotional exhaustion. Such association has to our knowledge not been reported previously and needs close attention in further research. Findings for the other two work setting factors demonstrated that these operate differently at different levels, with *promotion of care quality* and *managerial support for nursing* having effects only at the nursing unit level and hospital level respectively. Studies in which the relationship between work setting factors and outcomes are examined at multiple levels, have only emerged recently in the nursing literature (Li et al., 2013a,b; Gabriel et al., 2013). Previous research studies mostly showed findings that were likely to be very similar in terms of replicating the same relationships between aggregated work setting factors and outcomes, even in different study settings. Studies such as this allow for more refined and more context-specific insights. Although in general the implications are similar, i.e. better work settings result in better outcomes, there is a larger degree of variation between the study findings. That is, in different settings, different work setting factors at different levels have a different impact on outcomes. As such, these studies demonstrate that policy makers and human resources managers in different settings need to implement different management strategies to achieve better outcomes. In line with previously reported research findings, our findings also show that such management strategies should be very specific for nurses with different work experience and working status. Indeed, more experienced nurses tend to suffer less from emotional exhaustion and depersonalization than less experienced nurses. Higher degrees of job burnout among younger, less experienced workers have been reported before and can be the consequence of unsuccessful occupational socialization (Bakker et al., 2002). It could also be hypothesized that experienced nurses may have developed better coping mechanisms for burnout, or that workload and scope of practice are more fit to experienced nurses. Further, our results indicate that fulltime nurses suffer more from emotional exhaustion but have a higher sense of personal accomplishment, compared to part-time working nurses. In line with our findings, Burke and Greenglass (2000), who measured the effects of

hospitals restructuring on 1362 nurses, found that full-time staff were more emotionally exhausted and depersonalized, yet reported greater professional efficacy. This evidence supports the multidimensional measurement of burnout and opposes analytic strategies that solely focus on emotional exhaustion as the one and only hallmark of burnout (Schaufeli et al., 2009). It might also imply that caution is warranted in using SEM to model a causal pattern of relationships among the burnout dimensions (Bakker et al., 2002; Van Bogaert et al., 2013), in which emotional exhaustion has a direct impact on depersonalization and an indirect impact on reduced personal accomplishment. As shown here, personal accomplishment can develop independently of emotional exhaustion, which was previously theorized by Leiter (1993). Already in the development of the MBI, Maslach and Jackson (1981) stated that the personal accomplishment subscale is independent of the emotional exhaustion and depersonalization subscales and cannot be assumed to be the opposite of these subscales.

The covariance part provides additional insights into the burnout phenomenon. Findings suggest a significantly larger variation in personal accomplishment for experienced nurses. A reason might be that, over the years, differences in nurses' feelings related to intrinsic rewards (e.g. high collaboration with colleagues, positive patient experiences) and extrinsic rewards (e.g. salary, opportunities for advancement) become more pronounced. The fixed effects in the variance part reflect the temporal or demographical measurement of the variation of burnout. The justification of including random effects implies that the covariances and the correlations among burnout are different across hospitals and nursing units. A larger value indicates more heterogeneity of burnout feelings among nurses within a unit (a hospital or a nursing unit). We call this the measurement of "harmonic burden", i.e. the units with smaller variances reflect a better harmonic burden within it. Different from the fixed effects in the variance part, the random effects reflect the spatial or geographical measurement of the variation of burnout.

8.7 Conclusions

Our proposed multilevel higher-order factor model provides a way to directly assess the heteroscedasticity of the multi-dimensional lower-order factor scores in a complex situation. In a multi-center (-district, -country, etc.) study where the interested measurement might not be observed directly but is a latent construct obtained from a factor analytic model, a traditional structural equation modeling approach could be applied. The latent construct might however show heteroscedasticity because of e.g. the "cultural difference" among centers, districts, countries, etc., or because of different characteristics of the individuals. We then strongly recommend researchers to apply the proposed MHOF model that could efficiently combine the factor analytic model and a regression model, while taking into account the multilevel structure as well as the heteroscedasticity. This modeling can reveal some "hidden" information that may have never been obtained through modeling only the mean of the measurements. For example, the larger variance/covariance matrix in some nursing unit in our motivating data set may imply the unbalanced burden of nursing care within that nursing unit. Therefore this could provide valuable information for the ad-

ministration of the hospitals and nursing units. The contribution of this paper is however not confined to nursing research. Our methods apply equally well to numerous research topics in psychology, sociology and political science, to name a few, which often deal with multilevel research designs, latent constructs, and an interest in covariance regression.

References

Asparouhov, T. and Muthen, B. (2005). Multivariate statistical modeling with survey data. In *Paper Presented at the Federal Committee on Statistical Methodology (FCSM) Research Conference*.

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

Bakker, A. B., Demerouti, E., and Schaufeli, W. B. (2002). Validation of the maslach burnout inventory-general survey: An internet study. *Anxiety, Stress & Coping*, 15(3):245–260.

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

Burke, R. J. and Greenglass, E. R. (2000). Effects of hospital restructuring on full time and part time nursing staff in Ontario. *International Journal of Nursing Studies*, 37(2):163–171.

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.

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.

DiStefano, C., Zhu, M., and Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, 14(20):1–11.

Gabriel, A. S., Erickson, R. J., Moran, C. M., Diefendorff, J. M., and Bromley, G. E. (2013). A multilevel analysis of the effects of the practice environment scale of the nursing work index on nurse outcomes. *Research in Nursing & Health*, 36(6):567–581.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition.

Goldstein, H. (2010). *Multilevel Statistical Models (Wiley Series in Probability and Statistics)*. Wiley, 4th edition.

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

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

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.

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (2002). Bayesian methods for generalized linear models with covariates missing at random. *Canadian Journal of Statistics*, 30(1):55–78.

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.

Kline, R. B. (2010). *Principles and Practice of Structural Equation Modeling (Methodology in the Social Sciences)*. The Guilford Press, 3rd edition.

Kutney-Lee, A., McHugh, M., Sloane, D., Cimiotti, J., Flynn, L., Neff, D., and Aiken, L. (2009). Nursing: a key to patient satisfaction. *Health Affairs*, 28(4):w669–w677.

Leiter, M. P. (1993). Burnout as a developmental process: Consideration of models. In Schaufeli, W. B., Maslach, C., and Marek, T., editors, *Professional burnout: Recent developments in theory and research. Series in applied psychology: Social issues and questions*, pages 237–250. Philadelphia, PA, US: Taylor & Francis.

Lesaffre, E. and Lawson, A. B. (2012). *Bayesian Biostatistics (Statistics in Practice)*. Wiley, 1st edition.

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

Li, B., Bruyneel, L., and Lesaffre, E. (2013a). A multivariate multilevel Gaussian model with a mixed effeccts structure in the mean and covariance part. *Statistics in Medicine*. (accepted).

Li, B., Bruyneel, L., Sermeus, W., Van den Heede, K., Matawie, K., Aiken, L., and Lesaffre, E. (2013b). 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.

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

Maruyama, G. M. (1997). *Basics of Structural Equation Modeling*. SAGE Publications, Inc, 1st edition.

Maslach, C. and Jackson, S. E. (1981). The measurement of experienced burnout. *Journal of Organizational Behavior*, 2(2):99–113.

Maslach, C., Jackson, S. E., and Leiter, M. P. (1996). *Maslach Burnout Inventory Manual*, 3rd edition.

Muthén, B. and Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, 17(3):313–335.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, 22(3):376–398.

Muthén, L. and Muthén, B. (2010). *Mplus User's guide*. Los Angeles: Muthén & Muthén, 6th edition.

Neuhaus, J. M. and Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 54(2):638–645.

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*.

Poghosyan, L., Aiken, L. H., and Sloane, D. M. (2009). Factor structure of the Maslach burnout inventory: an analysis of data from large scale cross-sectional surveys of nurses from eight countries. *International Journal of Nursing Studies*, 46(7):894–902.

Schaufeli, W. B., Enzmann, D., and Girault, N. (1993). Measurement of burnout: A review. In Schaufeli, W. B., Maslach, C., and Marek, T., editors, *Professional burnout: Recent developments in theory and research*, pages 199–215. Washington, DC: Taylor & Francis.

Schaufeli, W. B., Leiter, M. P., and Maslach, C. (2009). Burnout: 35 years of research and practice. *Career Development International*, 14(3):204–220.

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):6.

Smith, H. L. (2008). A double sample to minimize skew due to non-response in a mail survey. In Ruiz-Gazen, A., Guilbert, P., Haziza, D., and Tille, Y., editors, *Survey methods: Applications to longitudinal investigations, health, electoral investigations, and investigations in the developing countries*, pages 334–339. Paris: Dunod.

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.

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

Van Bogaert, P., Kowalski, C., Weeks, S. M., Clarke, S. P., et al. (2013). The relationship between nurse practice environment, nurse work characteristics, burnout and job outcome and quality of nursing care: A cross-sectional survey. *International Journal of Nursing Studies*, 50(12):1667–1677.

Worley, J. A., Vassar, M., Wheeler, D. L., and Barnes, L. L. (2008). Factor structure of scores from the Maslach burnout inventory: A review and meta-analysis of 45 exploratory and confirmatory factor-analytic studies. *Educational and Psychological Measurement*, 68(5):797–823.

9

CONCLUSIONS

9.1 General conclusions

A traditional multilevel regression model assumes a constant residual variance after adjusting for the fixed and random effects. The focus of this modeling approach is on the expected value (the mean) of the response that can vary with covariates (e.g. age) and across the level of the units (e.g. medical centers in a multi-center study). When heteroscedasticity exists in the multilevel context also, the variance may depend on covariates and on random effects, resulting in a mixed effects regression model for the variance with an expression similar as for a mixed effects regression model of the mean. In the multivariate case, not only the variances, but also the covariance(s) or the correlation(s), can be regressed on fixed and random effects. The proposed models in this thesis handle this complex situation well. We called this modeling approach multilevel covariance regression (MCR). Modeling of the covariance matrix in our MCR model is done by implementing a factor analytic (FA) model to the residual part of a multivariate random effects model. The Bayesian approach is chosen as the preferred estimating method because of its flexibility in handling complex situations.

The MCR model aims at modeling heteroscedasticity with both fixed and random effects in a multivariate and multilevel situation. This provides a way of obtaining insight in the factors that determine the level of the response and their variability. The multidimensional construct of work-related nurse burnout, measured in the multi-country Registered Nurse Forecasting (RN4CAST) study, provides an excellent example of how our innovative statistical model provides new substantive insights. This is illustrated in Chapters 6, 7 and 8. With the classical multilevel regression model we demonstrated that nurses' *working experience* has a negative effect on the mean burnout measures (i.e. more experienced nurses tend to be less burnout), while with the multilevel covariance regression model we showed in addition that the variation of burnout depends on nurses' *working experience* and random effects at each level. The variance of the random effects in the variance part may indicate the extend to which there is inequality of nursing care activity within a given unit (which we refer to as "harmonic burden" in Chapter 8). Therefore it can give valuable suggestions to the management of nursing care within a country, hospital or nursing unit.

The type of the response of the multilevel covariance regression model could be binary (Chapter 6), continuous but possibly with a non-standard distribution (e.g. bounded outcome scores in Chapter 7), ordinal, etc., or a combination of these types. Further, if the response comes from another model, it is also possible to jointly estimate the two models. In Chapter 8, the burnout response of the multilevel covariance regression model is a result from the factor scores in a multilevel factor analytic (MFA) model, and estimation from simultaneously fitting two models (MFA and MCR models) yields a multilevel higher order factor (MHOF) model. An advantage of this model is that the original items from the questionnaire are used instead of predefined dimensions. This, together with modeling the heteroscedasticity, provides some clues on the differences between the factor structure of the current data set and that of the predefined structure.

To conclude, the multilevel covariance regression model with different types of response performed well when applied to the RN4CAST data set using the Bayesian approach. With

regard to nursing research, our model provides more refined analyses of the relationships between nurses' burnout and their characteristics as well as the organizational-level covariates. Findings allow insights not only for the mean measures of burnout, but also for the correlations among the three burnout dimensions. The proposed model thus uses the data more efficiently. Not confined to nursing research only, our proposed modeling approaches apply equally well to many other research fields with multilevel research designs, latent constructs, and where there is an interest in covariance regression.

9.2 Future research

In this section we suggest topics for future research on the multilevel covariance regression model. Specifically, we will discuss further extensions of the MCR and MHOF models, as well as Bayesian model assessment.

9.2.1 Model extension

The MHOF model combines the MCR model with a multilevel factor analytic (MFA) model, whereby the nurse-level factor scores from the MFA model are used as the burnout responses in the MCR model. Similarly, we could also use the factor scores for the covariate *work environment* which is measured through a 32-item questionnaire. Namely, we could fit an MFA model for *work environment*, and extract the factor scores at each level to replace the used *work environment* variables in the MHOF model. These *work environment* variables were generated from a predefined factor structure based on a different population. The combination implies a simultaneous estimation of the MHOF and the MFA models. The computation, however, might be quite demanding as more random effects and latent variables are modeled.

So far, only the nurse-level covariance matrix of the three burnout dimensions is considered depending on fixed and random effects. It is also possible to model the higher-level covariance matrices of the MCR and MHOF models with both fixed and random effects. This extension can address the following question, e.g. "do the nursing unit-level correlations of burnout remain constant across hospitals and across the characteristics of the nursing unit?".

9.2.2 Bayesian model assessment

We first discuss the selection of the number of latent factors in a FA model. As the first exploratory step of an MHOF model, a multilevel exploratory factor analytic model was conducted in Chapter 8 to find the factor structure of burnout. For this, the frequentist approach using the program *Mplus* produces many approximate indices to evaluate the fit of a FA model, such as RMSEA, TLI, CLI, SRMR, etc. These can help us determine the number of latent factors for a parsimonious FA model with a reasonable model fit. In a Bayesian context, the number of factors can be determined by comparing the goodness-of-fit indices, e.g. the posterior predictive checks with an appropriate discrepancy measure. A more formal method was proposed by Lopes and West (2004) who utilized the reversible

jump MCMC method for this purpose. Either way needs closer investigation to evaluate its performance on our proposed models.

The Bayesian model comparison for our proposed MCR and MHOF models can be quite challenging especially when the response is not of standard form, such as the BOS type of response (bounded outcome scores) used in Chapters 7 and 8. Both for DIC and PSBF (pseudo Bayes factor) we encountered difficulties with the multivariate BOS response. For PSBF, it would be of interest in general, how to incorporate model complexity and to apply and test this on the MCR and MHOF models.

References

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, 14(1):41–68.

Summary

The multilevel model is the focus for all chapters in this thesis. We have compared software for handling one type of this model, namely the logistic random effects regression model. The multilevel factor analytic (MFA) model and multilevel structural equation modeling (multilevel SEM) are also applied. Inspired by a number of clinical research questions from the Registered Nurse Forecasting (RN4CAST) study, we have further developed two novel modeling approaches based on the multilevel model to handle more complex situations. These are the multilevel covariance regression (MCR) modeling and the multilevel higher-order factor (MHOF) modeling.

Chapter 3 compares the commonly used packages/programmes for handling several logistic random effects regression models. Both frequentist and Bayesian approaches were reviewed. Frequentist approaches included R (lme4), Stata (GLLAMM), SAS (GLIMMIX and NLMIXED), MLwiN ([R]IGLS) and MIXOR; Bayesian approaches included WinBUGS, MLwiN (MCMC), R package MCMCglmm and SAS experimental procedure MCMC. We saw that most often the packages gave similar parameter estimates, but they differ considerably in flexibility, computation time and usability. The frequentist and Bayesian approaches also performed differently for the small sample problem. The Bayesian approach, though sometimes time-consuming, showed the greatest flexibility in modeling.

In **Chapters 4 to 8**, the RN4CAST data set, collected in the context of a large European nurse survey project, was the main inspiration for new statistical developments. In **Chapter 4**, a logistic random effects model was applied to a list of nursing tasks below their skill level they performed during the daily nursing work. It allowed for a comparison between tasks performed by domestically trained nurses and foreign trained nurses from developing countries.

In **Chapter 5**, the focus is on the nursing work environment rated by each nurse. A two-level multiple indicators multiple causes (MIMIC) model was applied to detect differences between nursing unit managers' and staff nurses' opinions on the work environment. We found out that for certain work environment dimensions, nursing unit managers had more positive opinions compared to nurses.

In **Chapters 6 to 8** the burnout measurement of the nurses was of interest. As the original measure of burnout was not normally distributed, a dichotomized burnout variable was used in **Chapter 6**. The relationship between burnout and the work environment was examined in a four-level context while taking into account the intra-class correlation and heteroscedasticity of the binary burnout measurement. Modeling of heteroscedasticity was achieved through a built-in factor analytic model wherein the factor loadings changed across units (countries, hospitals and nursing units). This is actually a first step and a simple version of the multilevel covariance regression model proposed in **Chapter 7**.

As an extension of the model in **Chapter 6**, we proposed the multilevel covariance regression (MCR) model in **Chapter 7**. This model can handle the original non-normally distributed burnout measures as well as modeling the heteroscedasticity with both fixed and random effects. The properties of the MCR model, the identification issues and the

interpretations were carefully described in detail. The simulation study showed good performance of the model in handling outliers at each level. The application to the data from the RN4CAST study revealed new insights that could not be found through classic mixed models.

A further extension of the MCR model was described in **Chapter 8**, which is called the multilevel higher order factor (MHOF) model. It replaces the response in the MCR model with the factor scores coming from a multilevel factor analytic model, which was estimated simultaneously with the MCR model. This brought in more challenges on e.g. computation, identification, etc. Our proposed Bayesian method could handle this complex situation well.

Samenvatting

In deze thesis hebben we het hierarchisch model toegepast op verpleegkundige gegevens en uitbreidingen voorgesteld voor het multivariate respons geval. In de eerste twee hoofdstukken hebben we de basisconcepten ingeleid nodig voor deze thesis.

Hoofdstuk 3 vergelijkt de populaire software voor de verschillende logistisch random effects regressie modellen; zowel frequentistische als Bayesiaanse benaderingen werden behandeld. De frequentistische software was: R (lme4), Stata (GLLAMM), SAS (GLIMMIX en NLMIXED), MLwiN ([R]IGLS) en MIXOR en de Bayesiaanse software was: WinBUGS, MLwiN (MCMC), R pakket MCMCglmm en de toen experimentele SAS (versie 9.2) procedure MCMC. Onze vergelijking liet zien dat vaak de meeste pakketten similaire parameter schattingen gaven, maar ook dat ze aardig konden verschillen in flexibiliteit, berekeningstijd en gebruiksvriendelijkheid. De performantie van de frequentistische en de Bayesiaanse methodes verschilden ook voor kleine steekproeven. De Bayesiaanse aanpak, weliswaar vaak meer computerintensief, vertoonde de grootste flexibiliteit in het modelleren van de gegevens.

In **hoofdstukken 4 tot 8** hebben we de RN4CAST data set intensief gebruikt. Deze data set was het resultaat van een grootschalig Europees project dat werd opgezet om meer inzicht te bekomen in de relatie werkomgeving en werkervaring bij verpleegkundigen. In **hoofdstuk 4** hebben we een logistisch random effect model gebruikt om de factoren te onderzoeken die bepalen waarom verpleegkundigen taken uitvoerden die niet tot hun oorspronkelijk takenpakket hoorden en beneden hun competentie lagen. Het model vergelijkt hierin ook de verpleegkundigen opgeleid in de respectievelijke RN4CAST landen met de verpleegkundigen opgeleid in hun ontwikkelingsland van herkomst.

In **hoofdstuk 5**, hebben we de verpleegkundige werkomgeving, zoals gescoord door de verpleegkundigen in de RN4CAST studie, onderzocht. We hebben hiervoor een hierarchisch model met twee niveaus en meerdere indicatoren (MIMIC model) toegepast om de verschillende beleving van de werkomgeving te bestuderen tussen de hoofden van de verpleegkundige afdelingen enerzijds en de verpleegkundigen in die respectievelijke afdelingen anderzijds. Onze analyse wees uit dat voor bepaalde werkomgevingsvariabelen de hoofden een duidelijk meer rooskleurig beeld hadden dan de verpleegkundigen zelf.

In **hoofstukken 6 tot 8** hebben we burnout bij verpleegkundigen onderzocht. Omdat de originele burnout maat niet normal verdeeld is, hebben we deze gedichomotomiseerd in **hoofdstuk 6**. We hebben dan de relatie tussen deze binaire burnoutmaat en de verpleegkundige werkomgeving onderzocht in een hierarchisch model met 4 niveau's rekening houdende met de heteroscedasticiteit van de binaire burnout metingen. We hebben de heteroscedasticiteit gemodelleerd met behulp van een factor analytische aanpak waarbij de factorladingen mogen verschillen tussen de verpleegeenheden (landen, hospitalen en verpleegeenheden). Dit was de eerste stap en een eenvoudige versie van hierarchisch covariantie regressie model voorgesteld het volgende hoofdstuk.

In **hoofdstuk 7** hebben we het voorgaande model uitgebreid tot het hierarchisch covariantie regressie (MCR) model. Dit model werkt op de originele niet-normaalverdeelde

burnoutmaten en modelleert de heteroscedasticiteit als een mixed effects model. In dit hoofdstuk hebben we ook de eigenschappen van het MCR model onderzocht, alsook de interpretatie van de modelparameters en de mogelijke identificatieproblemen voor het bepalen van parameterschattingen. Een simulatiestudie toonde aan dat het model geschikt is om de variabiliteit van de variantie-covariantie matrix aan het licht te brengen. Toegepast op de RN4CAST studie, bleek uit het MCR model ook de invloed van werkomgeving op de variabiliteit van de burnoutmaten.

In **hoofdstuk 8** stellen we een uitbreiding van het MCR model voor, genaamd het hierarchisch hogere orde factor (MHOF) model. Dit vervangt de hoog-dimensionele respons in het MCR model door de factor scores uit een hierarchisch factor analytisch model. De parameters van de twee deelmodellen (MFA en MCR) worden gezamenlijk geschat. Dit bracht extra uitdagingen met zich mee, zoals bijvoorbeeld bijkomende computationele en identificatie problemen. Echter, de Bayesiaanse aanpak leverde geen problemen op voor deze verdere uitbreiding.

Acknowledgements

This journey has finally come to an end for the moment. It is a pleasant journey, although it is never easy. I am quite amazed by what I've done so far when looking at the thesis in front of me, indeed. It would not have been possible to finish this without all kinds of help around me. I am very grateful to all my colleagues, my friends and my family.

First and foremost I would like to thank my supervisor Emmanuel Lesaffre who has been supporting me enormously with great patient and with brilliant ideas. And also thank you for guiding me into the beautiful new world of Bayesian. I feel so proud of being in this field.

I would also like to give my sincere thanks to all my colleagues from the Biostatistics department in EMC. Being a member of this international family is the fortune of my whole life. I would like to thank Eline, Siti, Karolina, Johan, Paul, Sten, Magdalena, Dimitris, Vironika, Susan, Elrozy, Kazem, Joost, Nicole, Betina, Maria, Lidia, Wim, Gianluca, Sara, Els, Dymph, Cibele, Anne, Kees, Ehsan, Benedict and also some new members of the family. Thank you all for creating such great atmosphere. I would like to give my special thanks to a former colleague Marek who helped me kindly whenever I had problems for my project. May you rest in peace in heaven.

I also owe my thanks to a lot of friends outside my department. I would like to thank Prof. Steyerberg, Hester, Rachel, Nano, Astrid and Adi from this university, and I would like to thank Luk, Prof. Sermeus, and Luwis from the university of Leuven for the long term pleasant and successful collaboration. I would like to thank Bonnie and Chang. Chang, you are one of the friends I knew for the longest time here and you are always so kind and willing to help.

时间过得很快，4 年多的博士生涯马上结束了。一个人漂泊在海外的生活并不容易， 所幸结识了一帮志同道合的朋友及损友，让生活变得不乏味。在这里大家一同留下了太多值得纪念的回忆。大怪路子，杀人游戏，K 歌，大大小小的聚餐，BBQ，乒乓，羽毛球，踢毽子，桌球，滑雪，钓鱼，挖生蚝，等等。在此，我要感谢我的朋友海波，稚超，徐洪，路狄非，康宁，刘凡，温蓓，孙伟，海燕，程姐，吕鹏，甜娜，瞟瞟，天石，蔡蕊，明慧，小娜夫妇，吴庆夫妇，艳楠夫妇，舟桥夫妇，商鹏夫妇，胡晗，凯音，展民，莹颖，田田，玺峰，黄龄，刘哲，朱玉，唐晖，亚迪，克若，栾莹，延伟， 亚楠，等等，感谢你们让我在这里度过了难忘的时光。

At last, I would like to thank my parents and my two elder sisters, for their selfless support all along the way. I owe my sincere thanks to my wife for being so persistent and apprehensive and for supporting me in all aspects.

Curriculum Vitae

Personal information	
Name	Baoyue Li
Date of birth	5 AUG 1982
Place of birth	LiaoNing, China
Nationality	Chinese
Education	
2001-2006	B.Sc in Public Health, Fudan University, Shanghai, China
2006-2010	M.Sc in Medicine Statistics, Fudan University, Shanghai, China
2008-2010	M.Sc in Clinic Epidemiology, Erasmus MC, Rotterdam, the Netherlands
2010-2014	PHD student of Biostatistics, Erasmus MC, Rotterdam, the Netherlands

Presentations and conferences

- Logistic random effects regression models: A comparison of statistical packages. 31st International Society for Clinical Biostatistics, Montpelier, France, 2010.
- A Bayesian multivariate multilevel probit model applied to nursing burnout data. 33rd International Society for Clinical Biostatistics, Bergen, Norway, 2012.
- A multivariate multilevel Gaussian model with a mixed effects structure in the mean and covariance part, 34th International Society for Clinical Biostatistics, Munich, Germany, 2013.
- A multivariate multilevel Gaussian model with a mixed effects structure in the mean and covariance part, 4th Bayes conference, Rotterdam, the Netherlands, 2013
- Joint modeling of a multilevel factor analytic model and a multilevel covariance regression model, 5th Bayes conference, London, UK, 2014

PHD training

- Analysis of growth data, 2010.
- The craft of smoothing, 2010.
- Bayesian methods in medical research, 2010.
- Bayesian methods and bias analysis, 2010.
- Mixture models, cluster and discriminant analysis with R package mixAK, 2010.
- Bayesian variable selection and model choice for structure additive regression, 2011.

- Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling, Sensitivity Analysis, and Causal Inference, 2011.
- New Features Mplus version 7, 2012.
- Bayesian variable selection, 2012.
- Bayesian adaptive methods for clinical trials, 2012.
- An introduction to the joint modeling of longitudinal and survival data, 2013.

Assistant teaching

- Bayesian methods, Nihes, 2012-2014.
- Modern methods, Nihes, 2011-2013.
- Classical methods, Nihes, 2010-2013.
- Repeated measurement, Nihes, 2012-2014.
- Multiple regression methods using SPSS, 2011-2014.
- Introduction to epidemiology, Nihes summer programme, 2011-2012.

Publications

Li, B., Bruyneel, L., and Lesaffre, E. (2014). Multilevel higher-order factor model: Joint modeling of a multilevel factor analytic model and a multilevel covariance regression model. *Structural Equation Modeling: A Multidisciplinary Journal*. (under review)

Bruyneel, L., Li, B., Squires, A., Gilmartin, M., Spotbeen, S., Meuleman, B., Lesaffre, E., and Sermeus, W. (2014). Nursing unit managers' and staff nurses' opinions of the nursing work environment: a Bayesian multilevel mimic model for cross-group comparisons. *Research in Nursing & Health* (under review)

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.

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.

Bruyneel, L., Li, B., Squires, A., Aiken, L., Lesaffre, E., Van den Heede, K., and Sermeus, W. (2013). A multi-country perspective on nurses tasks below their skill level: Reports from domestically trained nurses and foreign trained nurses from developing countries. *International Journal of Nursing Studies*, 50(2):202-209.

Croughs, P. D., Li, B., Hoogkamp-Korstanje, J. A. A., and Stobberingh, E. (2013). Thirteen years of antibiotic susceptibility surveillance of Pseudomonas aeruginosa from intensive care units and urology services in the Netherlands. *European Journal of Clinical Microbiology & Infectious Diseases*, 32(2):283-288.

Schubert, M., Ausserhofer, D., Desmedt, M., Schwendimann, R., Lesaffre, E., Li, B., and De Geest, S. (2013). Levels and correlates of implicit rationing of nursing care in Swiss acute care hospitals: A cross sectional study. *International Journal of Nursing Studies*, 50(2):230-239.

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):281-91.

Li, B., Lingsma, H. F., Steyerberg, E. W., and Lesaffre, E. (2011). Logistic random effects regression models: A comparison of statistical packages for binary and ordinal outcomes. *BMC Medical Research Methodology*, 11(1):77.

Lingsma, H. F., Roozenbeek, B., Li, B., Lu, J., Weir, J., Butcher, I., Marmarou, A., Murray, G. D., Maas, A. I., Steyerberg, E. W. (2011). Large between-center differences in outcome after moderate and severe traumatic brain injury in the international mission on prognosis and clinical trial design in traumatic brain injury (IMPACT) study. *Neurosurgery*, 68(3):601-608.