Balancing simple models and complex reality

Contributions to item response theory in educational measurement

Het balanceren van simpele modellen en de complexe werkelijkheid Bijdragen aan de item respons theorie voor het onderwijskundig meten

(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op vrijdag 13 mei 2016 des middags te 12.45 uur

door

Maria Bolsinova

geboren op 15 oktober 1988
te Moskou, Rusland

Promotoren: Prof. dr. H. Hoijtink
Prof. dr. G.K.J. Maris

This thesis was (partly) accomplished with financial support from CITO Dutch National Institute for Educational Measurement

Contents

Chapter 1

Introduction

> Essentially, all models are wrong,
> but some are useful.
>
> George E.P. Box

In educational measurement, data obtained using educational tests are gathered both for practical and scientific purposes, for example for individual assessment or to study the effects of educational policies. While these data often have a very complex structure, we try to capture their most important aspects with relatively simple models. The reason is that statistical models are needed to make inferences about the unobservable constructs of interest (e.g., reading ability, foreign language proficiency, and arithmetic ability) on the basis of observed test data. A general and commonly used framework for modeling data from tests is item response theory [IRT].

IRT focuses on observed item responses and using relatively simple models explains (predicts) item responses by item and person characteristics, and their interactions. Various parametric IRT models specify the so called item response functions for the relationship between the observed response and the latent variable (usually called ability or proficiency in the context of educational measurement). This allows one to estimate the level of ability given the response data. In this dissertation IRT models for dichotomous data (i.e., each response is either correct or incorrect) are considered. For a comprehensive overview of IRT, the reader is referred to Lord and Novick (1968), Lord (1980), Hambleton and Swaminathan (1985), and van der Linden and Hambleton (1997).

This dissertation presents various contributions to item response theory in educational measurement which in one way or another search for an optimal balance between simple models and complex reality. The dissertation consists of two parts: Part I presents contributions to modeling response time and accuracy and Part II presents Bayesian contributions to IRT. The final chapter concludes with an epi-

logue reflecting upon the main results of the dissertation and discussing some suggestions for further research.

1.1 Outline of Part I

Educational testing is increasingly carried out in a computerised form instead of a traditional paper-and-pencil form. This allows one to record not only the responses given by the students but also how long it takes them to respond. A wide range of IRT literature deals with developing ways of incorporating response time data into measurement models with the aims of, for example, improving the precision of ability estimation, identifying aberrant response behaviour, or explaining response processes (for an overview see Lee & Chen, 2001; van der Linden, 2009).

When jointly modeling the response times and the response accuracy it is often assumed that they can be modelled by a set of latent variables, usually called ability and speed. It is also often assumed that the relationship between the response accuracy and the response time can be fully explained by the relationship between the latent variables. That is, conditional independence between the response time and accuracy is assumed given the latent variables. In other words, although correct responses might be on average slower or faster than incorrect responses, when the latent variables are taken into account there are no differences between the distributions of the response times of correct and incorrect responses.

The conditional independence assumption is important both from the statistical and the substantive point of view (van der Linden, 2009). If one also assumes that response accuracy depends only on the ability latent variable and response time depends only on the speed latent variable, as is done in the hierarchical framework for modeling response time accuracy (van der Linden, 2007), then one obtains a model with a simple structure. Under this model very clear interpretations of speed and ability latent variables would be available and unconfounded measurement of the two would be warranted.

However, conditional independence is an assumption which may be violated in practice, and testing it is an important step in modeling response time and accuracy jointly. Chapters 2 and 3 of this dissertation present two methods for testing this assumption. In Chapter 2, we propose to test conditional independence using Kolmogorov-Smirnov tests (Kolmogorov, 1933). The equality of the response time distributions given a correct and an incorrect response is tested in every subgroup of persons with the same value of the sufficient statistic for ability. This test is of a semi-parametric nature: It is parametric with respect to the model for response accuracy (i.e., an exponential family model is required), but non-parametric with respect to the model for response times and the type of violation of conditional independence against which the assumption is tested. The Type I and Type II error rates of the procedure are investigated in simulation studies and

its use is illustrated in an educational measurement application.

In Chapter 3, posterior predictive checks (Rubin, 1984; Meng, 1994; Gelman, Meng, & Stern, 1996) for conditional independence are proposed which focus on different possible consequences of conditional dependence: residual correlations between response time and accuracy given ability and speed, difference between the variances of response times of the correct and incorrect responses, and difference between the item-rest correlations of the slow and the fast responses. Specificity, sensitivity, and robustness of the procedure are evaluated in simulation studies. Furthermore, an example of applying the procedure to an educational test is presented.

If the assumption of conditional independence is retained after being tested, then standard models based on this assumption can be applied. However, in practical educational measurement applications there might be residual dependencies between response times and response accuracies that cannot be explained by the latent variables (Ranger & Ortner, 2012). In Chapter 4, we focus on one application like this and propose an extension of the hierarchical model for response time and accuracy (van der Linden, 2007) which takes the conditional dependencies into account. The effects of the residual response time (i.e., the response being relatively fast or slow compared to what would have been expected for this person on this item) on both the intercept and the slope of the item response function are incorporated in the model.

1.2 Outline of Part II

The chapters in the second part of the dissertation are connected to each other not by the substantive questions which are addressed, but by the statistical framework that they have in common, namely Bayesian statistical inference. For a general introduction to this framework see, for example, Gelman, Carlin, Stern and Rubin (1995) and Gill (2008). Two important properties of the Bayesian framework which make it very useful in the context of item response theory are that it enables one to estimate very complex models using its powerful simulation-based techniques and that it allows one to include background information beyond the observed data in the analysis.

From the Bayesian perspective the parameters in the statistical model are considered to be random variables. The full specification of a statistical model includes the specification of the data generating distribution (density of the data) and the prior distribution of the model parameters. The latter includes possible knowledge about the plausible values of the parameters before observing the data, for example collected from experts. The density of the data and the prior distribution are combined in the posterior distribution which represents the knowledge about which values the model parameters are likely to have given the observed data.

Inferences are made based on this posterior distribution. The so called Markov chain Monte Carlo methods [MCMC] can be used to sample from the posterior distribution even if it is highly multidimensional and does not have a closed form (Hastings, 1970; Gelfand & Smith, 1990; Tierney, 1994; Robert & Casella, 2004; Gamerman & Lopes, 2006). This provides a relatively simple and straightforward way of estimating models with a large number of parameters which are much more difficult to estimate using classical frequentist methods.

In Chapter 5 the usefulness of the Bayesian approach for estimating complex multidimensional models is highlighted. In this research project we provide a solution to the problem of choosing a scoring rule for an educational test. We argue for the use of scoring rules which are simple and easy to interpret but contain all the information about the person's ability. The simplest scoring rule like that is the using the sumscore (i.e., number of correct responses), which follows from the Rasch model (Rasch, 1960). However, this model is often too restrictive to fit real data. Therefore, a new extension of the Rasch model is proposed which relaxes some of its assumptions but still has a rather simple scoring rule. We call this new model a multi-scale Rasch model, since it assumes that the test consists of a set of scales each following a Rasch model, but the scale memberships of the items are not known a priori and need to be estimated. Once the scales are identified, the test can be scored with a set of sumscores in each of the scales, since these sumscores contain all the information about persons' abilities measured by the scales. An MCMC algorithm for identifying the Rasch scales is developed and evaluated.

In Chapter 6 we illustrate the second advantage of the Bayesian statistical framework introduced in the beginning of this section, namely the possibility of including prior knowledge in the analysis. The substantive problem that we focus on is making the results of different test versions comparable using linking and equating procedures (for a comprehensive overview of linking and equating see Kolen and Brennan, 2004). If a new test version is used each year, then the new test results are not directly comparable to the results of the reference test version due to possible differences in difficulty of the two tests and possible differences in the ability distribution in the new and the reference population. In high-stakes testing the amount of data that are available to link the reference and the new test is often very limited due to security reasons, which leads to low precision of the linking results. We propose and evaluate two methods for the elicitation of prior knowledge about the difference in difficulty of the two tests from subject-matter experts. This prior knowledge is included in test linking procedures in the form of prior distributions to improve the quality of linking. Two empirical elicitation studies for the primary school arithmetics test are presented.

In Chapter 7 the following attractive feature of the Bayesian estimation techniques is used: It allows one to take different sources of uncertainly about the model parameters into account. If some model parameters are not fully identified

it is still possible to sample from their posterior distribution, and their posterior variance would include both the uncertainty due to sampling variability in the data as well as the uncertainty due to the non-identifiability. Using a simulation-based approach it is possible to attempt to separate these two types of uncertainty and hence assess the impact that non-identifiability has on the model inferences. In this chapter, this is investigated in the context of the distribution of the missing data in incomplete non-equivalent group testing designs. We show that while the distribution of the unobserved scores of the reference population on the new test is not fully identified, the uncertainty about this score distribution is very small and can be ignored in practice. Furthermore, we demonstrate using simulated and real data examples that ignoring the non-identifiability issue while assuming a normal distribution for ability may lead to bias in test equating.

Part I

Contributions to modeling response time and accuracy

Chapter 2

A test for conditional independence between response time and accuracy

[1] **Abstract.** An important distinction between different models for response time and accuracy is whether conditional independence (CI) between response time and accuracy is assumed. In the present study, a test for CI given an exponential family model for accuracy (for example, the Rasch model or the One Parameter Logistic model) is proposed and evaluated in a simulation study. The procedure is based on the non-parametric Kolmogorov-Smirnov tests. As an illustrative example, the CI test was applied to data of an arithmetics test for secondary education.
Keywords: conditional independence, item response theory, Kolmogorov-Smirnov tests, response times, sufficiency.

2.1 Introduction

Computer based testing makes it simple to record not only the response a student gives to an item, but also the response time. Response times provide an additional source of information about student's performance. There are different approaches to modeling response times within item response theory (IRT). In the last decade the most popular approach is joint modeling of the distribution of response time and accuracy:

$$f(\mathbf{X} = \mathbf{x}, \mathbf{T} = \mathbf{t} \,|\, \Theta = \theta, H = \eta, \boldsymbol{\xi}), \tag{2.1}$$

[1]This chapter has been published as Bolsinova, M., & Maris, G. (2016) A test for conditional independence between response time and accuracy. *British Journal of Mathematical and Statistical Psychology* DOI: 10.1111/bmsp.12059. Author contributions: B.M. and M.G. designed the research, B.M. performed the research, B.M. wrote the paper, M.G. provided feedback on the paper.

where $\mathbf{X}$ is a random vector of responses with realisations $x = 1$ if it is correct and $x = 0$ if it is incorrect for each element $X_i, \forall i \in [1 : n]$, $\mathbf{T}$ is a random vector of response times with realisations $\mathbf{t}$, Θ and H are random latent variables representing ability and speed with realisations θ and η, and $\boldsymbol{\xi}$ are item parameters, including item characteristics related to time. For shorter notation, we will use

$$f(\mathbf{x}, \mathbf{t} \,|\, \theta, \eta) \tag{2.2}$$

instead of (2.1) throughout the paper.

When modeling this joint distribution, one of the main questions is whether for each item i the response X_i and the response time T_i are independent conditional on the latent variables; that is

$$X_i \perp\!\!\!\perp T_i \,|\, \Theta, H. \tag{2.3}$$

For example, in the hierarchical framework (van der Linden, 2007) a multilevel model for $\mathbf{X}$ and $\mathbf{T}$ is built on the assumption of conditional independence [CI] with all the dependence between responses and response times absorbed in the higher-level correlation between the latent variables Θ and H. This assumption is also crucial for the drift diffusion model (Ratcliff, 1978; Tuerlinckx & De Boeck, 2005; Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006).

At the same time, there are models that allow conditional dependence between response time and accuracy (Ranger & Ortner, 2012). Moreover, some models assume a certain sort of dependence between X_i and T_i, like some of the race models (Pike, 1973; Townsend & Ashby, 1983) and the Signed residual time [SRT] model (Maris & van der Maas, 2012). According to the latter model, the highly able students tend to give fast correct responses and slow incorrect responses, whereas the students with low ability give slow correct responses and fast incorrect responses.

We have to emphasise that independence is assumed conditional on the particular set of latent variables. If it is violated, it does not mean that there may not exist another set of latent variables conditional on which X_i and T_i are independent.

The assumption of CI between X_i and T_i is attractive from a statistical point of view, but in real applications there might be sources of violation of this assumption. For example, the speed with which a student answers test items can fluctuate unsystematically, or attention may fluctuate randomly (Pieters & van der Ven, 1982). Another possible explanation for the dependence between X_i and T_i is that fast and slow responses are governed by different cognitive processes, such that different IRT models hold for the fast and the slow responses (Partchev & De Boeck, 2012).

Testing the assumption of CI is very important because if this assumption does

not hold, then such models as the diffusion models and the hierarchical models cannot be true. On the other hand, if it does hold, then an equally large number of models (such as most race models) cannot be true, because they imply some kind of dependence.

The CI assumption as stated in (2.3) cannot be directly tested because θ and η are unobservable. A testable statement implied by the CI assumption must be formulated. We will show in the next section that, if the marginal model for $\mathbf{X}$ is an exponential family IRT model, CI implies that given a sufficient statistic $S(\mathbf{X})$ the distribution of response times should not differ between the groups of respondents with correct and incorrect responses. In this study we propose a procedure for testing the CI assumption in this form:

$$f(t_i|X_i = 1, S(\mathbf{X}) = s) = f(t_i|X_i = 0, S(\mathbf{X}) = s). \tag{2.4}$$

The approach we take in this paper is of a semi-parametric nature. The purpose is to test the assumption of CI in isolation from other assumptions about the joint distribution of $\mathbf{X}$ and $\mathbf{T}$, but under the assumption of sufficiency of $S(\mathbf{X})$. We are interested not in formulating a specific detailed model for the joint distribution of responses and response times, but in testing general assumptions to rule out models that cannot be true if these assumptions do not hold. Because of the assumption of sufficiency of $S(\mathbf{X})$, our approach is parametric with respect to modeling the probability of a correct response, but it is non-parametric with respect to modeling the response times and the joint distribution of $\mathbf{X}$ and $\mathbf{T}$.

The paper is organised as following. In Section 2.2, we discuss how the non-parametric Kolmogorov-Smirnov test can be used to test the assumption of CI between response time and accuracy. In Section 2.3, a simulation study is presented. In Section 2.4, we present an illustrative example. The paper ends with a discussion.

2.2 Testing CI between response time and accuracy

Van der Linden and Glas (2010) used the Lagrange Multiplier [LM] test for testing the null hypothesis of the CI between responses and response times against a parametric alternative. Under the null hypothesis, responses and response times to item i are assumed to be independent given ability θ and speed η. Hence, the joint distribution of the responses and response times is the product of their marginal distributions:

$$f(x_i, t_i|\theta, \eta) = f(x_i|\theta)f(t_i|\eta). \tag{2.5}$$

Note, that Equation 2.5 is derived using additional assumptions:

$$f(x_i \,|\, \theta, \eta) = f(x_i \,|\, \theta), \tag{2.6}$$

$$f(t_i \,|\, \theta, \eta) = f(t_i \,|\, \eta). \tag{2.7}$$

Under the alternative, response times are distributed differently for $X_i = 0$ and $X_i = 1$, and this difference is modelled by adding an extra parameter λ_i to the response times model. This parameter is the shift in the distribution of T_i for the correct responses relative to the incorrect responses. Hence, the CI assumption is tested against a specific parametric alternative; that is, one has to have appropriate extra parameters for the test to work. A problem with this approach is that after adding a new parameter to the model, the marginal distributions of $\mathbf{X}$ and $\mathbf{T}$ are no longer explained by the same IRT model and response time models.

Another disadvantage of the LM test is that although it is meant as an item level test, a test for one item is dependent on whether CI holds for all other items. The null hypothesis is:

$$H_{0i} : \lambda_i = 0, \lambda_j = 0, \forall j \neq i, \tag{2.8}$$

and the alternative is:

$$H_{ai} : \lambda_i \neq 0, \lambda_j = 0, \forall j \neq i. \tag{2.9}$$

If CI is violated for some of the items $j \neq i$, then H_{0i} may have a high probability of being rejected due to misspecification of these λ_js even if λ_i is actually 0.

Unlike the parametric approach described above, we propose to test the CI assumption without having to specify a parametric model for this dependence. The non-parametric test is constructed in such a way that the null hypothesis is tested against the alternative of any kind of dependence. The LM test would have better results in the cases of dependence of the same form as it is modelled by the additional parameter λ_i. But it is not appropriate and it is likely to perform worse if the dependence between X_i and T_i is of a different kind.

Because we want to avoid being committed a specific model for response times, it is important to note, that CI between response time and accuracy given θ and η (as in Equation 2.5) together with the assumptions in (2.6) and (2.7) implies that response time and accuracy are independent given θ only:

$$f(x_i, t_i|\theta) = \int_{\mathbb{R}} f(x_i|\theta) f(t_i|\eta) f(\eta|\theta) d\eta = f(x_i|\theta) \int_{\mathbb{R}} f(t_i|\eta) f(\eta|\theta) d\eta = \\ = f(x_i|\theta) f(t_i|\theta) \Longleftrightarrow T_i \perp\!\!\!\perp X_i|\Theta. \tag{2.10}$$

Thus, CI can be tested by considering the distribution of X_i and T_i given θ and we do not need to condition on η.

In many cases, the response accuracy alone is modelled using simple IRT models like the Rasch model [RM] (Rasch, 1960), in which the probability of a correct response of person p to item i is predicted from the difficulty of the item (β_i) and the ability of the person (θ_p):

$$\Pr(X_{pi} = 1) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)}. \tag{2.11}$$

For example, the SRT model (Maris & van der Maas, 2012) implies the RM for response accuracy. The probability of a correct response for an unbiased diffusion model (Ratcliff, 1978) can also be represented as a RM if the boundary separation is determined by an experimental condition and does not vary across the items (Tuerlinckx & De Boeck, 2005). While these models amount to having a RM for the marginal distribution of $\mathbf{X}$, they differ in whether they assume the CI assumption when modeling a joint distribution of $\mathbf{X}$ and $\mathbf{T}$.

An important property of the RM is that it belongs to the exponential family, meaning that there exists a sufficient statistic $S(\mathbf{X})$ summarising all the information about the ability contained in the response vector. Having a sufficient statistic is equivalent to the following CI assumption (Dawid, 1979):

$$\mathbf{X} \perp\!\!\!\perp \Theta \,|\, S(\mathbf{X}). \tag{2.12}$$

In the RM the number of items answered correctly is a sufficient statistic. A more flexible model allowing items to have different weights in the estimation of ability is the One parameter logistic model [OPLM] (Verhelst & Glas, 1995) which has the weighted sumscore $\sum_i a_i X_{pi}$ as the sufficient statistic, where a_i is a prespecified item weight of item i which is restricted to positive integers. Another example of an IRT model with the number of correct responses as the sufficient statistic for θ is the interaction model (Haberman, 2007). Furthermore, even more flexible models can be considered with the number of correct responses as the sufficient statistic, for example a model with a separate item parameter for each score group which is used for testing the fit of the Rasch model (Andersen, 1973).

We shall prove next, that if the IRT model holds and if CI between X_i and T_i holds, then the distributions of T_i for a correct and an incorrect response are the same given the value of the sufficient statistic.

Theorem. If $\mathbf{X} \perp\!\!\!\perp \Theta | S(\mathbf{X})$ and $\mathbf{X} \perp\!\!\!\perp \mathbf{T} | \Theta$, then

$$(T_i | X_i = 1, S(\mathbf{X}) = s) \sim (T_i | X_i = 0, S(\mathbf{X}) = s)$$

Proof. Let us denote by $S(\mathbf{X}^{(i)})$ the sufficient statistic computed for the vector $\mathbf{X}$ excluding item i.

$$
\begin{aligned}
f(t_i|X_i{=}1, S(\mathbf{X}){=}s) &= \int_{\mathbb{R}} f(t_i|\theta, X_i{=}1, S(\mathbf{X}){=}s) f(\theta|X_i{=}1, S(\mathbf{X}){=}s) d\theta \\
&\downarrow \quad \mathbf{X} \perp\!\!\!\perp \mathbf{T}|\Theta \\
&= \int_{\mathbb{R}} f(t_i \,|\, \theta) f(\theta|X_i = 1, S(\mathbf{X}) = s) d\theta \\
&\downarrow \quad \mathbf{X} \perp\!\!\!\perp \Theta|S(\mathbf{X}) \\
&= \int_{\mathbb{R}} f(t_i \,|\, \theta) f(\theta|S(\mathbf{X}) = s) d\theta \\
&\downarrow \quad \mathbf{X} \perp\!\!\!\perp \Theta|S(\mathbf{X}) \\
&= \int_{\mathbb{R}} f(t_i \,|\, \theta) f(\theta|X_i = 0, S(\mathbf{X}^{(i)}) = s) d\theta \\
&= f(t_i \,|\, X_i = 0, S(\mathbf{X}) = s)
\end{aligned}
$$

□

A violation of the equality of the distributions in (2.4) can be attributed either to a violation of the sufficiency of $S(\mathbf{X})$, such that the IRT model does not hold, or to a violation of CI. But if the exponential family IRT model holds for the response accuracy taken alone, the decision to take the response times into account should not change it. Thus, a difference between the distributions in (2.4) implies a violation of the CI assumption, given the sufficiency of $S(\mathbf{X})$ for response accuracy.

Having a sufficient statistic is important because it allows one to match the distributions of ability given different response patterns:

$$
f(\theta \,|\, X_i = 1, S(\mathbf{X}^{(i)}) = s - \Delta_i) = f(\theta \,|\, X_i = 0, S(\mathbf{X}^{(i)}) = s), \tag{2.13}
$$

where Δ_i is the difference between the values of the sufficient statistic for two response vectors which differ only in the value of X_i, which is equal to a_i in the OPLM and to 1 in the RM.

More generally the same approach to testing CI can be used in non-exponential family models if for a target item i there exist different response patterns $\mathbf{X}^{(i)} = \mathbf{y}$ and $\mathbf{X}^{(i)} = \mathbf{z}$ such that the conditional distributions of the ability are equal:

$$
f(\theta \,|\, X_i = 1, \mathbf{X}^{(i)} = \mathbf{y}) = f(\theta \,|\, X_i = 0, \mathbf{X}^{(i)} = \mathbf{z}). \tag{2.14}
$$

The response patterns that provide matching distributions of ability can be in principle empirically determined with large enough sample size. Then CI can be

tested by testing the equality of the distributions:

$$f(t_i \mid X_i = 1, \mathbf{X}^{(i)} = \mathbf{y}) = f(t_i \mid X_i = 0, \mathbf{X}^{(i)} = \mathbf{z}). \tag{2.15}$$

To test the equality of distributions in (2.4), we can use any statistic based on empirical distribution functions. A particular statistic of this kind is the Kolmogorov-Smirnov statistic, which is equal to the maximum of the absolute value of the discrepancy between the empirical distribution functions of the random variables Y_1 and Y_2:

$$D = \max_y |F_n(y) - G_m(y)|, \tag{2.16}$$

where $F_n(y)$ is the proportion of observations of Y_1 smaller than or equal to y, and $G_m(y)$ is the proportion of observations of Y_2 smaller than or equal to y. This statistic is consistent and has known exact and asymptotic distributions under $H_0 : Y_1 \sim Y_2$ (Kolmogorov, 1933).

We can divide the persons into groups conditional on their sufficient statistics and test whether within each group response times for correct responses have the same distribution as response times for incorrect responses. For each item i we test for all values of s whether (2.4) holds. Let us by n_{is}^+ and n_{is}^- denote the number of respondents with a sufficient statistic equal to s giving a correct and an incorrect response to item i, respectively. If (2.4) holds, then the maximum of the discrepancy between the empirical distributions of $T_i \mid X_i = 1, S(\mathbf{X}) = s$ and $T_i \mid X_i = 0, S(\mathbf{X}) = s$, denoted by D_{is}, has the following distribution:

$$\Pr\{D_{is} \leq d\} = 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1} \exp\left(-2k^2\left(d\sqrt{\frac{n_{is}^+ n_{is}^-}{n_{is}^+ + n_{is}^-}}\right)^2\right). \tag{2.17}$$

If the probability $\Pr\{D_{is} \geq d\}$ is smaller than a chosen significance level α for the observed value d of the discrepancy between the empirical distributions in (2.4), then the hypothesis of the equality of these distributions is rejected. The equality of the empirical distributions in (2.4) can be tested only if there is at least one correct response and one incorrect response to item i among the persons with the sufficient statistic equal to s.

The CI test consists of multiple comparisons: a Kolmogorov-Smirnov test resulting in a p-value p_{is} is performed for each item i and each value of the sufficient statistic s testing the null-hypothesis

$$H_0^{(is)} : T_i \mid X_i = 1, S(\mathbf{X}) = s \sim T_i \mid X_i = 0, S(\mathbf{X}) = s. \tag{2.18}$$

First, for each item a Kolmogorov-Smirnov test is performed for several s.

Based on these tests for different s, we test (2.4) for an item across all the values of the sufficient statistic:

$$H_0^{(i)} : H_0^{(is)} \text{ is true for all } s. \tag{2.19}$$

Each of these tests is performed on a separate part of the data: a group of students with a certain value of the sufficient statistic; therefore, these tests are independent. If a null-hypothesis holds for k independent tests, then the p-values of those tests are independently and identically uniformly distributed. Therefore, to test (2.19) we can test whether the p-values from testing separate H_0^{is} are uniformly distributed:

$$H_0^{(i)} \text{ is true} \quad \Longrightarrow \quad p_{is} \sim^{iid} U(0,1), \tag{2.20}$$

Using the one-sample Kolmogorov-Smirnov test, the $H_0^{(i)}$ is tested against the one-sided alternative $H_1^{(i)} : p_{is} <_{st} U(0,1)$, meaning that the uniform distribution is stochastically greater than the distribution function of p_{is}. This alternative is used because if $H_0^{(is)}$ is not true, then p_{is} tends to be small. The p-value of this test, denoted by p_i, will indicate whether CI holds for item i. The Kolmogorov-Smirnov test is chosen here over other alternatives, such as the Fisher's method (Fisher, 1925), because exploratory analysis indicated that it seems to be a relatively more powerful procedure in this context.

We prefer testing the uniformity of p_{is} over using a multiple testing correction of the p-values for two reasons. First, we are more interested in directly testing the hypothesis $H_0^{(i)}$ itself, rather than in deciding which of the p_{is} are significant. Second, we expect that in realistic applications if the CI is violated then $H_0^{(is)}$ would not be true for a range of s and not just for one or two values of the sufficient statistic. In this case testing uniformity would have a higher power than a multiple testing correction, since $H_0^{(i)}$ will be rejected if none of the individual p-values is smaller than .05, but many of them tend to be small.

The proposed test should work with both short and long tests. For long tests, the power of the separate tests of $H_0^{(is)}$ may be relatively low, since there may be few persons with a particular value of the sufficient statistic. However, at the level of testing $H_0^{(i)}$ power is re-gained because information from many individual tests is aggregated. On the other hand, if there are few items on a test and, hence, few p_{is} per item, then $H_0^{(i)}$ is tested using a relatively small number of p_{is}s. However, this does not need to result in a low power for testing $H_0^{(i)}$ because the p_{is}s are the results of relatively powerful tests.

In addition to tests for individual items, we want to know whether CI holds for all the items in the test:

$$H_0 : H_0^{(i)} \text{ is true for all } i. \tag{2.21}$$

The uniformity of the p-values p_{is} is tested for n items. Unlike the tests of $H_0^{(is)}$, these n tests are performed on the same data. Since the tests are not independent, the joint distribution of their p-values is not known. Therefore, in this case we employ a multiple testing correction. We can use, for instance, the minimum and the maximum values of p_i to determine whether p_i have marginal uniform distributions without specifying the dependence between them.

$$\Pr\big(\min(p_i) \leq p\big) = \Pr\big(\bigcup_i (p_i \leq p)\big); \tag{2.22}$$

$$\Pr\big(\max(p_i) \leq p\big) = \Pr\big(\bigcap_i (p_i \leq p)\big). \tag{2.23}$$

The right-hand sides of the equations can be bounded using Fréchet inequalities with the probability under perfect positive dependence as the upper bound and the probability under perfect negative dependence as the lower bound (Fréchet, 1951):

$$\begin{gathered}
\max\big(\Pr(p_i \leq p)\big) \leq \Pr\big(\bigcup_i (p_i \leq p)\big) \leq \min\big(1, \sum_i \Pr(p_i \leq p)\big) \\
p \leq \Pr\{\min(p_i) \leq p\} \leq \min(1, np) \\
\max\big(0, 1 - n + \sum_i \Pr(p_i \leq p)\big) \leq \Pr\big(\bigcap_i (p_i \leq p)\big) \leq \min\big(\Pr(p_i \leq p)\big) \\
\max\big(0, 1 - n + np\big) \leq \Pr\big(\max(p_i) \leq p\big) \leq p
\end{gathered} \tag{2.24}$$

If $p_{min} \leq \alpha/n$ or/and $p_{max} \leq \alpha$, then it is highly unlikely that p_i have uniform marginal distribution, where α is a significance level. Then, the CI assumption should be rejected for the set of items as a whole.

If one further wants to decide for which of the items $H_0^{(i)}$ should be rejected, then the Holm-Bonferroni method can be used (Holm, 1979). First, order the p-values p_i from lowest to highest, denoted by $p_{(1)}, \ldots, p_{(n)}$, and let the corresponding hypotheses be $H_{0(1)}, \ldots, H_{0(n)}$. For a given significance level $\alpha = .05$, select the minimal index k such that

$$p_{(k)} > \frac{.05}{n + 1 - k}. \tag{2.25}$$

Reject the null hypotheses $H_{0(1)}, \ldots, H_{0(k-1)}$ and do not reject the null hypotheses $H_{0(k)}, \ldots, H_{0(n)}$. If $k = 1$, none of the null hypotheses $H_0^{(i)}$ should be rejected. If there is no k satisfying (2.25), then all $H_0^{(i)}$ should be rejected.

2.3 Simulation study

A simulation study was carried out to evaluate the proposed CI test. The study included empirical computation of the type I error probabilities and the power against different violations of CI. The procedure was applied to simulated data with the response accuracy following the RM or the Two-parameter logistic model [2PL] (Lord & Novick, 1968). In the second case the true model does not have a sufficient statistic and the procedure is an approximation.

For examining the type I error probabilities, the data were simulated according to the hierarchical model for response time and accuracy (van der Linden, 2007) with a RM or a 2PL for response accuracy and a lognormal model for response times

$$f(\mathbf{x},\mathbf{t}) = \prod_p \prod_i \frac{\exp(x_{pi}a_i(\theta_p - b_i))}{1+\exp(a_i(\theta_p - b_i))} \frac{\alpha_i}{\sqrt{2\pi}t_{pi}} \exp\left(\frac{-\alpha_i^2(\ln t_{pi} - (\beta_i - \eta_p))^2}{2}\right), \quad (2.26)$$

where a_i, b_i, α_i and β_i are item discrimination, item difficulty, item time discrimination, item time intensity parameters, respectively. In the case of the RM, the item discrimination parameters a_i were equal to 1 for all the items. Both for the case of the RM and for the case of the 2PL, the distribution of the response times was:

$$t_{pi} \sim \ln\mathcal{N}\left(\beta_i - \eta_p, \frac{1}{\alpha_i^2}\right). \quad (2.27)$$

The following conditions were varied: sample size (N = 2,000, 5,000, 10,000), number of items ($n = 20, 40$), correlation between speed and accuracy ($\rho = 0, 0.5$).

Three types of violation of the CI were considered:

- Type 1: The location of the lognormal distribution of the response time T_{pi} depends on the response X_{pi}:

$$t_{pi} \sim \ln\mathcal{N}\left(\beta_i - \eta_p + \frac{\lambda x_{pi}}{\alpha_i}, \frac{1}{\alpha_i^2}\right). \quad (2.28)$$

  This is a type of violation for which the LM test has been developed. The following values of λ were considered: 0.2, 0.5, 0.8, -0.2, -0.5, -0.8, representing small, medium and large effect sizes. For the conditions where $\rho = 0$, only the positive values of λ were used, since the procedure is symmetric for correct and incorrect responses.

- Type 2: The difference between the locations of the lognormal distribution given a correct and an incorrect response depends on the ability level of a person

$$t_{pi} \sim \ln\mathcal{N}\left(\beta_i - \eta_p - \frac{\kappa\theta_p(2x_{pi}-1)}{\alpha_i}, \frac{1}{\alpha_i^2}\right). \quad (2.29)$$

The difference between the locations of the distributions of T_i given an incorrect and a correct response is equal to $\frac{2\kappa\theta_p}{\alpha_i}$. If $\kappa > 0$, then correct responses are faster for students with high ability, whereas incorrect responses are faster for students with low ability. Moreover, the difference between the two distributions becomes larger when θ_p takes more extreme values. This type of violation corresponds to what the SRT model predicts (Maris & van der Maas, 2012). The values of κ (0.125, 0.313 and 0.501) were chosen for the simulation such that on average the absolute difference between the values of the location parameters of the response times distributions for the correct and the incorrect responses was equal to 0.2, 0.5 and 0.8, respectively.

- Type 3: The distributions of T_{pi} given a correct and an incorrect response differ not in the location but in the variance:

$$t_{pi} \sim \ln\mathcal{N}\left(\beta_i - \eta_p, \left(\frac{\nu^{(1-x_{pi})}}{\alpha_i}\right)^2\right). \tag{2.30}$$

The extra parameter ν represents the ratio between the standard deviations of the logarithm of the response times for the correct and the incorrect responses. If $\nu > 1$, then the variance of the response times is larger for incorrect responses than for correct responses. In the simulation study we used $\nu = 1.5$ and $\nu = 2$.

The data were simulated in the following way: First, for each person p the ability and the speed parameters were simulated:

$$\{\theta_p, \eta_p\} \sim \mathcal{MVN}\left(\{0,0\}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right). \tag{2.31}$$

Second, for each item i item parameters were sampled: b_i and β_i from $\mathcal{N}(0, 0.5)$, α_i from $\ln\mathcal{N}(0, 0.5)$; $a_i = 1$ for the RM, and $a_i \sim \ln\mathcal{N}(0, 0.5)$ for the 2PL. The distribution chosen for the discrimination parameters in the 2PL allows for quite large differences between the item discriminations within a test: the 1st and the 3rd quartiles of the chosen distribution of a_i are equal to 0.64 and 1.60, respectively. And the expected minimum and maximum values for the discrimination parameters in a test of 20 items are equal to 0.28 and 4.04, respectively.

Third, response accuracy data were simulated:

$$X_{pi} \sim \text{Bernoulli}\left(\frac{1}{1+\exp(a_i(b_i - \theta_i))}\right), \forall p \in [1:N], \forall i \in [1:n]. \tag{2.32}$$

Finally, response times data were simulated according to (2.27) for computing the type I error probabilities, and according to (2.28), (2.29) or (2.30) for computing the power against the different violations of CI.

Table 2.1: Type I error of the CI test

ρ	N	n	RM		2PL	
			H_0	$H_0^{(i)}$	H_0	$H_0^{(i)}$
0	2,000	20	.031	.002	.040	.002
		40	.020	.000	.020	.001
	5,000	20	.026	.001	.026	.001
		40	.031	.001	.019	.001
	10,000	20	.027	.002	.028	.001
		40	.025	.001	.021	.001
0.5	2,000	20	.042	.002	.051	.003
		40	.026	.001	.019	.000
	5,000	20	.026	.001	.114	.007
		40	.026	.001	.040	.001
	10,000	20	.023	.001	.301	.019
		40	.017	.000	.044	.001

In each condition 1,000 data sets were replicated. The CI test was carried out for each item conditioning on the sumscore. The CI test was performed using the `ks.test` R-function (R Core Team, 2014). If $n_{is}^{+}n_{is}^{-}$ <10,000 then the exact p-values were computed, otherwise the approximate p-values were used which may be slightly conservative.

The percentage of data sets for which H_0 was rejected was calculated as the test-level type I error rate or power. Moreover, the percentage of items for which $H_0^{(i)}$ was rejected was calculated as the item-level Type I error or power. The results are presented in Tables 2.1, 2.2, 2.3, and 2.4.

Due to the multiple testing corrections both within an item and across items, the CI test is rather conservative. For the data sets simulated under the RM all type I error probabilities were below .05. When the procedure was an approximation (responses were generated under the 2PL), the number of items was small ($n = 20$), the sample size was large (N=5,000 and 10,000) and $\rho = .5$, the type I error rate was inflated. This happened because for the items with relatively high or relatively low discrimination (on average in the simulated data sets the discrimination parameters ranged from 0.25 to 4) the conditional distributions in (2.13) can be rather different from each other, which causes differences between the distributions in (2.4) due to the correlation between θ and η . These small differences were detected when the sample size was large. When the number of items increases then the differences between the distributions in (2.13), and consequently between the distributions in (2.4) decrease and the type I error inflation is not observed.

Table 2.2: Power of the CI test against the violation of the CI of Type 1

ρ	N	n	RM						2PL					
			$\lambda = 0.2$		$\lambda = 0.5$		$\lambda = 0.8$		$\lambda = 0.2$		$\lambda = 0.5$		$\lambda = 0.8$	
			H_0	$H_0^{(i)}$	H_0	$H_0^{(i)}$	H_0	$H_0^{(i)}$	H_0	$H_0^{(i)}$	H_0	$H_0^{(i)}$	H_0	$H_0^{(i)}$
0	2,000	20	.28	.02	1	.72	1	.92	.25	.02	1	.68	1	.91
		40	.13	.00	1	.57	1	.88	.14	.00	1	.55	1	.87
	5,000	20	.83	.16	1	.91	1	.98	.82	.15	1	.90	1	.98
		40	.73	.06	1	.88	1	.97	.75	.07	1	.86	1	.97
	10,000	20	1	.57	1	.97	1	.99	1	.53	1	.96	1	.99
		40	1	.40	1	.95	1	.99	1	.38	1	.95	1	.99
.5	2,000	20	.31	.02	1	.76	1	.94	.39	.03	1	.69	1	.90
		40	.12	.00	1	.63	1	.91	.16	.01	1	.60	1	.88
	5,000	20	.87	.19	1	.93	1	.99	.97	.27	1	.87	1	.96
		40	.79	.08	1	.91	1	.98	.86	.11	1	.87	1	.97
	10,000	20	1	.63	1	.98	1	1	1	.58	1	.93	1	.98
		40	1	.47	1	.97	1	.99	1	.45	1	.94	1	.98
			$\lambda = -0.2$		$\lambda = -0.5$		$\lambda = -0.8$		$\lambda = -0.2$		$\lambda = -0.5$		$\lambda = -0.8$	
			H_0	$H_0^{(i)}$	H_0	$H_0^{(i)}$	H_0	$H_0^{(i)}$	H_0	$H_0^{(i)}$	H_0	$H_0^{(i)}$	H_0	$H_0^{(i)}$
.5	2,000	20	.28	.02	1	.76	1	.94	.40	.03	1	.76	1	.94
		40	.14	.00	1	.63	1	.91	.16	.01	1	.63	1	.91
	5,000	20	.87	.18	1	.93	1	.99	.92	.24	1	.93	1	.98
		40	.79	.08	1	.91	1	.98	.82	.09	1	.91	1	.98
	10,000	20	1	.63	1	.98	1	1	1	.65	1	.97	1	.99
		40	1	.46	1	.97	1	.99	1	.49	1	.97	1	.99

Table 2.3: Power of the CI test against the violation of the CI of Type 2

ρ	N	n	RM						2PL					
			$\kappa = .125$		$\kappa = .313$		$\kappa = .501$		$\kappa = .125$		$\kappa = .313$		$\kappa = .501$	
			H_0	$H_0^{(i)}$	H_0	$H_0^{(i)}$	H_0	$H_0^{(i)}$	H_0	$H_0^{(i)}$	H_0	$H_0^{(i)}$	H_0	$H_0^{(i)}$
0	2,000	20	.20	.01	1	.62	1	.89	.22	.01	1	.56	1	.85
		40	.11	.00	1	.52	1	.87	.14	.00	1	.48	1	.82
	5,000	20	.66	.08	1	.88	1	.97	.72	.10	1	.84	1	.95
		40	.70	.05	1	.88	1	.97	.75	.07	1	.82	1	.95
	10,000	20	.99	.45	1	.96	1	.99	1	.42	1	.93	1	.98
		40	1	.38	1	.95	1	.99	1	.35	1	.92	1	.98
.5	2,000	20	.23	.02	1	.68	1	.92	.28	.02	1	.61	1	.88
		40	.12	.00	1	.59	1	.90	.13	.00	1	.55	1	.86
	5,000	20	.70	.10	1	.91	1	.98	.84	.15	1	.87	1	.97
		40	.73	.07	1	.90	1	.98	.83	.09	1	.86	1	.96
	10,000	20	1	.52	1	.97	1	.99	1	.55	1	.95	1	.99
		40	1	.44	1	.96	1	.99	1	.43	1	.94	1	.99

Table 2.4: Power of the CI test against the violation of the CI of Type 3

ρ	N	n	RM				2PL			
			$\nu = 1.5$		$\nu = 2$		$\nu = 1.5$		$\nu = 2$	
			H_0	$H_0^{(i)}$	H_0	$H_0^{(i)}$	H_0	$H_0^{(i)}$	H_0	$H_0^{(i)}$
0	2,000	20	.69	.10	1	.66	.68	.09	1	.62
		40	.39	.02	1	.51	.46	.02	1	.48
	5,000	20	1	.49	1	.85	1	.46	1	.83
		40	1	.40	1	.81	1	.38	1	.80
	10,000	20	1	.71	1	.91	1	.69	1	.91
		40	1	.64	1	.89	1	.61	1	.88
.5	2,000	20	.73	.12	1	.72	.78	.13	1	.68
		40	.48	.03	1	.58	.49	.03	1	.55
	5,000	20	1	.56	1	.88	1	.56	1	.88
		40	1	.47	1	.85	1	.46	1	.85
	10,000	20	1	.77	1	.94	1	.80	1	.95
		40	1	.71	1	.92	1	.70	1	.92

The procedure also works better if the variance of the discrimination parameters is smaller: When two extra conditions were added to the simulation study: $n = 20$, $\rho = 0.5$, $a_i \sim \ln\mathcal{N}(0, 0.1)$ (on average the discrimination parameters ranged from 0.56 to 1.83), N=5,000 and N=10,000, the test level type I error probabilities were .026 and .041, respectively.

The test-level power was larger than .8 for the smaller effects ($\lambda = 0.2, -0.2$, $\kappa = .125$ and $\nu = 1.5$) when N=10,000, and for all conditions with larger effects. The item-level power against the violation of the CI of Type 1 and Type 2 was larger than .8 for all conditions when the effect was large ($\lambda = 0.8, -0.8$ or $\kappa = 0.501$), for all conditions except those with $N = 2,000$ when the effect was medium ($\lambda = 0.5, -0.5$ or $\kappa = 0.313$). None of the conditions had an adequate item-level power when the effect was small ($\lambda = 0.2, -0.2$ or $\kappa = 0.125$). The item-level power against the violation of CI of Type 3 was below .8 for all conditions with $\nu = 1.5$, but it was larger than .8 for all conditions except those with smaller samples ($N = 2,000$) when $\nu = 2$.

The power in the conditions simulated under the RM was generally larger than the power in the conditions simulated under the 2PL, but the differences were very small. Therefore, for the conditions in which the Type I error probabilities are adequate our procedure can serve as a useful approximation for testing the CI assumption.

2.4 Example

The CI test was applied to the data of the computerised Arithmetics test which was a part of the central exams in the Netherlands at the end of secondary education in 2013. A test version with a large number of respondents (N=10,369) was used. The respondents, whose response times were not recorded, were deleted from the data together with those respondents who had more than 10% of missing responses. The final sample size was 10,267.

The test consisted of 60 items. However, some items were excluded for substantive reasons. 16 items were removed to make the test more homogeneous in terms of the cognitive processes involved in the item responses. These were the last ten items because due to the restrictive time limit they were not reached by many of the students, and the response process under this time pressure may be different. And to obtain a scale with only open-ended items, all six multiple-choice items were removed, because different cognitive processes may be involved when choosing from alternatives compared to generating a response. Furthermore, eight low quality items were removed due to extreme easiness (with proportions of correct responses $> .85$) or low item-rest correlations ($< .25$). The RM and the OPLM model were fitted to the data with responses to the remaining 36 items using the OPCML software (Verhelst & Glas, 1995). The fit of the models was

Table 2.5: Empirical example: model fit statistics

Model	R_{1c}	df	p
Rasch model	2951.83	105	.000
OPLM (36 items)	284.97	105	.000
OPLM (30 items)	180.41	87	.000

tested with the R_{1c} statistic (Glas, 1988), see Table 2.5. The RM fitted considerably worse, therefore we continued with the OPLM. After deleting six misfitting items, the OPLM had a reasonable fit in the scale of remaining 30 items. Since the sample size was very large, the model was rejected by a formal test, however as can been seen in Figures 2.1a and 2.1b, the discrepancy between the observed and the expected proportion of students giving a correct response given the sum-score was very small even for the items with the highest misfit (Items 13 and 21). In Table 2.6 the item parameters and the item-fit statistics are presented. As the simulation results indicated that the CI test is robust against small violations from the exponential family model, the validity of the procedure should not be threatened in this empirical example.

The CI test was applied to the remaining 30 items (see results in Table 2.6). Since the response times were recorded only up to a second, there were a lot of exactly equal data points in the response times data. Therefore, the exact p-values of the Kolmogorov-Smirnov test could not be computed due to the presence of ties in the empirical distributions. Since in small samples approximate p-values can be very inaccurate, a bootstrap version of the Kolmogorov-Smirnov two-sample test allowing for ties in the data was used, which is can be performed using the `ks.boot` R-function in the `Matching` package (Sekhon, 2011) based on the results from Abadie (2002).

The CI test detected violations of the CI assumption for 25 out of 30 items. For most of these items the correct responses were slower than the incorrect responses. As an illustration, the empirical cumulative distribution functions of the response times to item 30 given four different values of the sufficient statistic are presented in Figure 2.2. There were also two items for which the incorrect responses were slower than the correct responses. Figure 2.3 shows item 3 as an example. Finally, for five items the test did not detect a violation of CI, which is illustrated in Figure 2.4 for item 15.

2.5 Discussion

Determining whether CI between X_i and T_i holds is crucial for modeling their joint

distribution correctly. In the present study, we constructed a statistical procedure for testing this assumption and showed an example of its application.

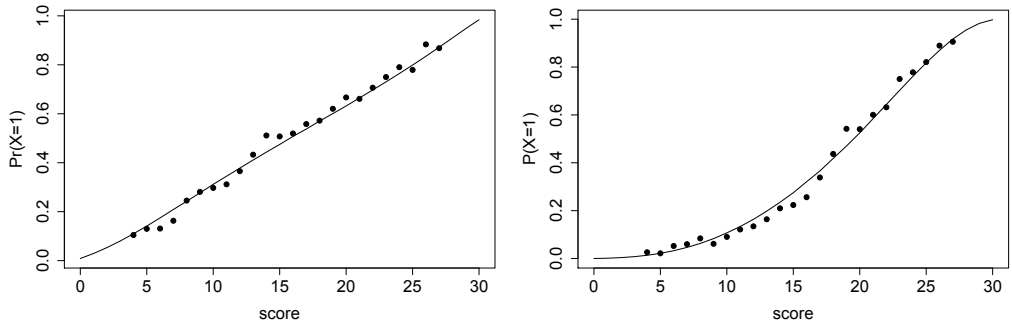
The advantages of the presented test are that the results of the CI test for a particular item do not depend on whether the CI holds for all other items, and that the test does not require a specification of the kind of violation against which CI is tested. One of the limitations of the procedure is that it requires large sample sizes, but this is not a problem for large scale computerised examinations. Another limitation is that the test is based on having an exponential family model for $\mathbf{X}$. However, we have shown how in principle the procedure can be extended (see Equations 2.14 and 2.15), for example it is sufficient to have subscales each fitting a RM. Moreover, in the simulation study we have shown that under certain conditions the procedure is robust to minor violations of the RM. Furthermore, not only the restrictive Rasch model but more flexible models have the number of the correct responses as the sufficient statistic (Andersen, 1973; Haberman, 2007).

In the context of measurement response times are often used to provide extra information about ability by borrowing strength from the measurement of speed for the measurement of ability, which can be done using the hierarchical model for response time and accuracy (van der Linden, 2007). However, the hierarchical model may be applied only if CI holds. Testing this assumption is crucial to decide whether the information from response times can be used for increasing the precision of the estimates of the ability by taking the correlation between θ and η into account. If a violation of the assumption is detected for most of the test items, then it is not justifiable to use response times for improving the precision of the estimates of θ in the way that is done by the hierarchical model. If the assumption is violated for only few items one might consider to remove these items and proceed with applying the hierarchical model, or retain the full set of items but apply a different model which allows for conditional dependencies.

As we have mentioned in the introduction, the CI test is meant for testing independence conditional on a particular set of latent variables. If CI given θ and η is violated, then one of the ways to proceed is introducing other latent variables that would explain the dependence between X_i and T_i, such that conditional on the new set of latent variables X_i and T_i are independent.

Our study raises questions for further research both in the parametric and the non-parametric approaches to the response times modeling. In the parametric approach, the question is what new interpretable parameters can be added to the model to allow for dependence between X_i and T_i. Another question is to what extent the models are robust to violations of the CI assumption.

In the non-parametric approach, the research can go in the direction of testing less restrictive assumptions than the CI assumption. For example, one could consider developing tests for the following assumptions: 1) the distributions in (2.4) differ only in their means but not in the variance; 2) the difference between the distributions in (2.4) is constant across ability levels.



a) Item 13 b) Item 21

Figure 2.1: Expected (solid line) and observed (dots) proportion of correct responses to an item given the number of items answered correctly

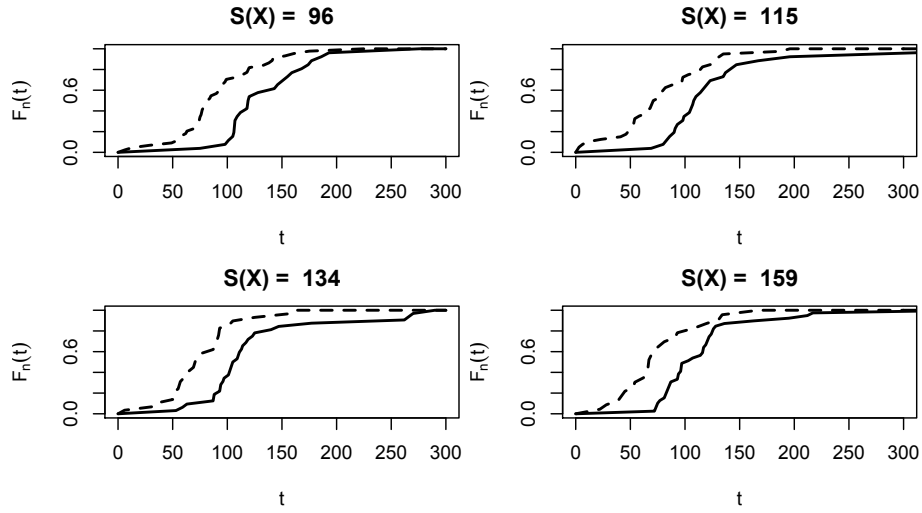


Figure 2.2: Item 30: Correct responses (solid line) are slower than incorrect responses (dashed line)

Table 2.6: Item parameters, item fit statistics and item level results of the CI test

Item	Item parameters		Item fit		CI test
	a_i	$\beta_i(SE)$	S-test (df)	p-value	p_i
Item 1	6	-0.27 (0.004)	12.88 (7)	.075	.014*
Item 2	11	0.07 (0.002)	11.98 (7)	.101	.002*
Item 3	6	-0.23 (0.004)	5.01 (7)	.658	.000*
Item 4	6	0.04 (0.004)	4.40 (7)	.733	.000*
Item 5	6	0.16 (0.004)	18.28 (7)	.011	.000*
Item 6	6	-0.31 (0.005)	13.17 (7)	.068	.250
Item 7	6	-0.15 (0.004)	4.85 (7)	.679	.000*
Item 8	5	0.22 (0.005)	12.92 (7)	.074	.000*
Item 9	5	0.27 (0.005)	5.58 (7)	.590	.000*
Item 10	6	0.06 (0.004)	7.15 (7)	.413	.000*
Item 11	10	-0.06 (0.002)	21.10 (7)	.004	.001*
Item 12	6	-0.30 (0.005)	8.62 (7)	.281	.000*
Item 13	6	0.03 (0.004)	20.42 (7)	.005	.591
Item 14	6	-0.12 (0.004)	5.92 (7)	.549	.000*
Item 15	5	-0.28 (0.005)	15.29 (7)	.033	.050
Item 16	9	0.14 (0.003)	13.69 (7)	.057	.182
Item 17	11	0.10 (0.002)	17.47 (7)	.015	.000*
Item 18	12	0.02 (0.002)	6.17 (7)	.521	.000*
Item 19	7	-0.06 (0.003)	7.57 (7)	.372	.000*
Item 20	6	-0.05 (0.004)	6.63 (7)	.468	.059
Item 21	10	0.11 (0.002)	35.34 (7)	.000	.000*
Item 22	8	0.24 (0.003)	15.36 (7)	.032	.000*
Item 23	12	0.06 (0.002)	5.56 (7)	.592	.000*
Item 24	8	-0.01 (0.003)	6.77 (7)	.453	.013*
Item 25	10	-0.01 (0.002)	2.66 (7)	.915	.000*
Item 26	7	-0.08 (0.003)	17.62 (7)	.014	.009*
Item 27	5	-0.05 (0.004)	16.76 (7)	.019	.000*
Item 28	5	0.24 (0.005)	11.86 (7)	.105	.002*
Item 29	6	0.15 (0.004)	11.17 (7)	.132	.000*
Item 30	9	0.09 (0.003)	8.81 (7)	.266	.000*

Note: * - $H_0^{(i)}$ is rejected.

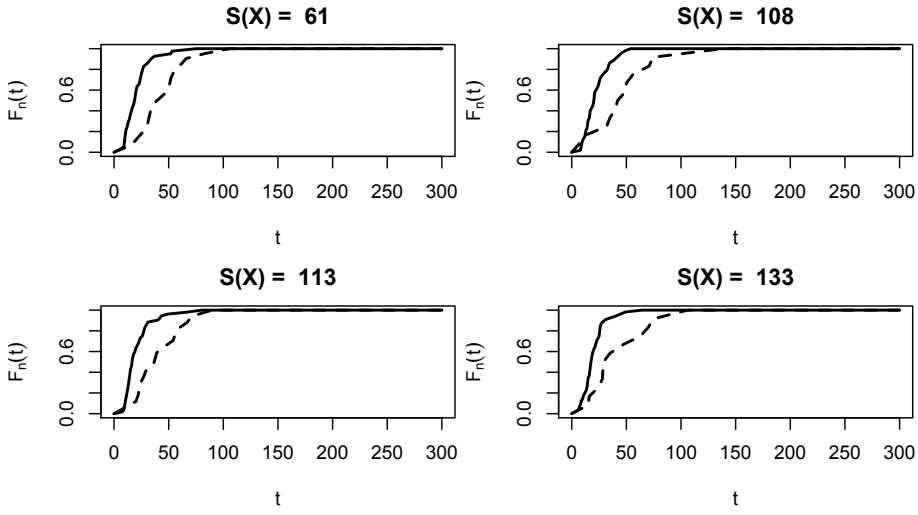


Figure 2.3: Item 3: Incorrect responses (dashed line) are slower than correct responses (solid line)

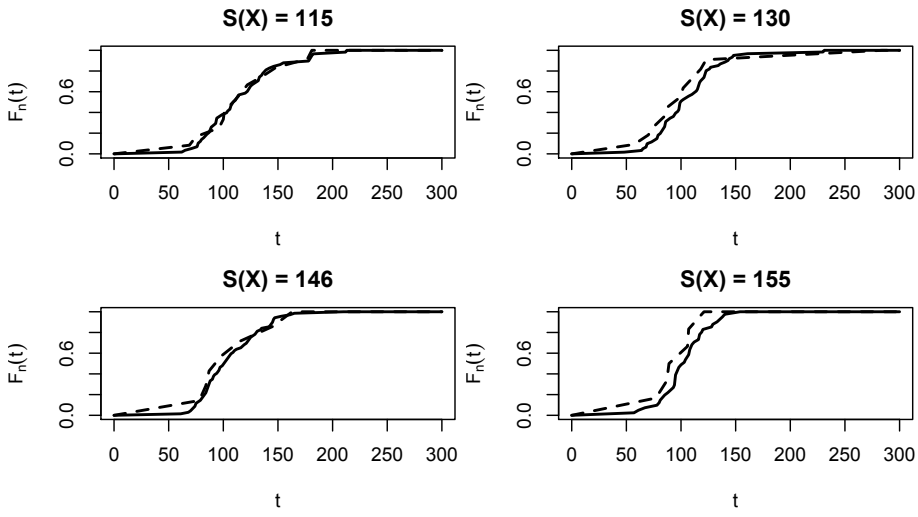


Figure 2.4: Item 15: No difference between correct responses (solid line) and incorrect responses (dashed line)

Chapter 3

Posterior predictive checks for conditional independence between response time and accuracy

[1] **Abstract.** Conditional independence (CI) between response time and response accuracy is a fundamental assumption of many joint models for time and accuracy used in educational measurement. In this study posterior predictive checks (PPC) are proposed for testing this assumption. These PPC are based on three discrepancy measures reflecting different observable consequences of different types of violations of CI. Simulation studies are performed to evaluate the specificity of the procedure, its robustness and its sensitivity to detect different types of conditional dependence, and to compare it to existing methods. The new procedure outperforms the existing methods in most of the simulation conditions. The use of the PPC is illustrated using arithmetics test data.
Keywords: conditional independence, hierarchical model, posterior predictive checks, response times.

3.1 Introduction

When modeling response accuracy [RA] (i.e., a response being correct or incorrect) and response time [RT] in educational and cognitive tests conditional independence [CI] between RA and RT to the same item is often assumed given the speed and

[1]This chapter has been published as Bolsinova, M. & Tijmstra, J. Posterior predictive checks for conditional independence between response time and accuracy. *Journal of Educational and Behavioral Statistics, 41*, 123-145. Author contributions: B.M. and T.J. designed the research, B.M. performed the research, B.M. wrote the paper, T.J. provided feedback on the manuscript.

the ability parameters (van der Linden, 2007, 2009). The relationship between the RAs and the RTs is assumed to be fully explained by the higher-level covariance between speed and ability, such that there is no residual dependence left. The CI assumption can be represented in the following way:

$$f(x_{pi}, t_{pi} \,|\, \theta_p, \tau_p, \boldsymbol{\psi}_i) = f(x_{pi} \,|\, \theta_p, \tau_p, \boldsymbol{\psi}_i) f(t_{pi} \,|\, \theta_p, \tau_p, \boldsymbol{\psi}_i), \tag{3.1}$$

where x_{pi} is the coded binary response of person p to item i having a value of 1 if the response is correct and a value of 0 otherwise, t_{pi} is the RT of person p on item i; θ_p and τ_p are the ability and the speed of person p, respectively; and $\boldsymbol{\psi}_i$ is a vector of item parameters of item i. For the full specification of the joint model for the RT and RA, the models for x_{pi} and t_{pi} have to be specified and the higher-level models for the person parameters and for the item parameters need to be specified.

In the hierarchical framework for modeling RT and RA (van der Linden, 2007), it is also assumed that the RA depends only on the ability of the person and the item parameters related to RA, denoted by $\boldsymbol{\phi}_i$ (e.g., difficulty and discrimination) and that the RT depends only on the speed of the person and the item parameters related to speed, denoted by $\boldsymbol{\varphi}_i$ (e.g., item time intensity):

$$f(x_{pi}, t_{pi} \,|\, \theta_p, \tau_p, \boldsymbol{\phi}_i, \boldsymbol{\varphi}_i) = f(x_{pi} \,|\, \theta_p, \boldsymbol{\phi}_i) f(t_{pi} \,|\, \tau_p, \boldsymbol{\varphi}_i), \tag{3.2}$$

where $\boldsymbol{\psi}_i = \{\boldsymbol{\phi}_i, \boldsymbol{\varphi}_i\}$. These extra assumptions together with the CI assumption result in a simple structure which makes interpretations of speed and ability relatively straightforward, but this interpretation is warranted only if the assumption of CI holds. When CI is violated, then this does not just result in inaccurate estimates but in a misrepresentation of the underlying constructs. Therefore, it is important to investigate whether CI can be maintained or whether more complex models are needed to capture the remaining dependence between RT and RA.

CI can be violated in different ways. For example, a residual correlation between the vector of RAs of all persons to item i and the vector of RTs of all persons answering item i may remain after taking the relationship between speed and ability into account. Such residual correlations can for example be modeled using the joint model for RTs and RAs of Ranger and Ortner (2012). These residual correlations may be more than just measurement artefacts, since for example the sign of the residual correlate may depend on the item difficulty (Bolsinova, de Boeck, Tijmstra, 2015).

Violations of CI might not always show up in the form of residual correlation across all persons, since the residual relationship between RT and RA may differ across persons. For example, if there is a negative residual correlation between RT and RA for persons with low ability and a positive residual correlation for persons with high ability, then these correlations may cancel out at the population level.

However, this type of violation of CI might show up as heterogeneity of variances (i.e., as differences between the variances of RTs given a correct and given an incorrect response), and/or as an interaction effect between RT and ability on RA.

An interaction effect between RT and ability may not only be a result of different residual correlations at different ability levels, but also arise from time heterogeneity of response processes, that is fast and slow responses being qualitatively different, as suggested by the results of Partchev and De Boeck (2012) and of Goldhammer, Naumann, Stelter, Toth, and Rölke (2014). These examples are not meant to be exhaustive, but rather to illustrate that CI can be violated in many different ways, which may threaten the validity of the model. This means that it is important for tests of CI to be able to pick up on these different types of violations.

There are two procedures available in the literature for testing CI. One of them has been proposed from the perspective of the hierarchical model (van der Linden & Glas, 2010). Here, the RTs under CI are modeled using a lognormal distribution with the mean parameter depending on the item time intensity ξ_i and the person speed and the item variance parameter, denoted by σ_i^2:

$$T_{pi} \sim \ln \mathcal{N}(\xi_i - \tau_p, \sigma_i^2). \tag{3.3}$$

This model is tested against a parametric alternative:

$$T_{pi} \sim \ln \mathcal{N}(\xi_i - \tau_p + \lambda_i x_{pi}, \sigma_i^2), \tag{3.4}$$

where the added parameter λ_i captures the difference in the location of the distribution of RTs for correct and for incorrect responses. The hypothesis: $H_0 : \lambda_i = 0$ is tested against $H_a : \lambda_i \neq 0$ using a Lagrange Multiplier [LM] test. While this test is able to detect differences in the location of the two RT distributions, it is not suited for other types of violations of CI. Moreover, the LM-test requires the parametric shape of the distribution of RTs to be correctly specified.

A different approach has been proposed by Bolsinova and Maris (2016). Their test of CI requires an exponential family model, for example the Rasch model (Rasch, 1960), to hold for the RA, which makes it possible to test the following hypothesis:

$$\forall i, \forall s : f(t_i \,|\, X_i = 1, S(\mathbf{X}) = s) = f(t_i \,|\, X_i = 0, S(\mathbf{X}) = s), \tag{3.5}$$

meaning that for each item within each group of persons with the same value of the sufficient statistic $S(\mathbf{X})$, the distribution of the RTs of the correct responses is the same as the distribution of the RTs of the incorrect responses. Equality of the RT distributions is tested using Kolmogorov-Smirnov [KS] tests. The advantage of this approach is that it does not require a specification of the kind of violation

against which CI is tested. However, the exponential family model for the RA is required or at least this model should not be strongly violated. Moreover, due to the fact that the equality of the RT distributions is tested separately for each group of persons with different values of the sufficient statistic, a large overall sample is required to achieve reasonable power (Bolsinova & Maris, 2016).

While the number of available tests for CI is limited, a wide range of methodologies have been developed for evaluating the assumption of local independence (i.e., CI between item scores) in the context of item-response theory (IRT) models that do not take RT into account. These methodologies propose measures that capture residual dependencies within item pairs that remain after conditioning on the latent variable(s) explaining the item scores. Some of these measures are based on the observed and expected frequencies in contingency tables, such as χ^2, G^2, and the standardized log odds ratio residual (Chen & Thissen, 1997). The Mantel-Haenszel statistic is also based on the observed contingency table, but considers this table separately for each value of the restscore (Ip, 2001; Sinharay, Johnson, & Stern, 2006). Other measures are based on associations between item residuals in some form or other, such as the Q_3 (Yen, 1993), the model-based covariance (MBC; Reckase, 1997), and the standardized model-based covariance (SMBC; Levy, Xu, Yel, & Svetina, 2015).

There are a number of relevant differences between the evaluation of local independence and that of CI between RA and RT. For example, while the assessment of local independence focuses on item-pairs, the evaluation of CI between RA and RT is done for individual items. Furthermore, RT is a continuous variable, which prevents a direct application of many of the existing measures to the context of assessing CI between RA and RT, especially those based on contingency tables. However, some of the methods of detecting local dependence may provide valuable starting points for the development of new ways of evaluating CI between RA and RT. Research on detecting violations of local independence (Levy et al., 2009; see also Sinharay et al., 2006) suggests that among others the Q_3 and SMBC were found to have relatively high power to detect violations of local independence (Levy et al., 2009; 2015). This finding provides motivation for considering a similar discrepancy measure in the context of evaluating CI between RA and RT, as will be proposed in the subsequent section.

However, because these methods for assessing local independence solely consider the item scores, they are not tailored towards detecting the different types of violations of CI that may be relevant and realistic in the context of jointly modeling RA and RT that were discussed before. For example, they are not designed to be able to detect differences in the variance of one variable (in our case, RT) for different values of the other variable (RA). Neither are they aimed at detecting differences in discrimination conditional on RT. Therefore, we will not only consider measures of residual correlation (similar to detecting local dependence), but also measures for other consequences of violations of CI.

In the present paper we present a new procedure for detecting violations of CI between RT and RA, which aims to overcome the limitations of the existing methods (LM and KS-tests). Based on the general framework of posterior predictive checks [PPC] (Rubin, 1984; Meng, 1994; Gelman, Meng, & Stern, 1996), we developed a PPC procedure using three discrepancy measures targeted at different ways in which CI can be violated.

The paper is organised as follows. In Section 3.2, the hierarchical model for RTs and RAs is elaborated and PPC are introduced. In Section 3.3, three item-level discrepancy measures of conditional dependence are introduced and a test-level decision criterion for either rejecting or retaining the CI assumption is described. Section 3.4 presents simulation studies focused on the specificity, robustness and sensitivity of PPC for CI. The performance of the new procedure is compared to that of the existing methods for testing CI. In Section 3.5 the use of PPC is illustrated for an empirical example. The paper is concluded with a discussion.

3.2 Model specification, estimation and PPC

In this paper we consider a version of the hierarchical model with a log-normal model for the RTs (van der Linden, 2006) as presented in Equation 3.3; and a two-parameter logistic model for the RA (Birnbaum, 1968):

$$\Pr(X_{pi} = 1) = \frac{\exp(\alpha_i(\theta_p - \beta_i))}{1 + \exp(\alpha_i(\theta_p - \beta_i))}, \tag{3.6}$$

where α_i and β_i are the discrimination and the difficulty parameters of item i. We consider these particular models in this paper, but in general the PPC method can be used for a different specification of the model for the RAs and for the RTs, provided that one can sample from the posterior distribution of the model parameters.

At the item level, the dependence between the item parameters is modeled by a multivariate normal distribution:

$$[\xi_i, \ln \sigma_i^2, \ln \alpha_i, -\alpha_i\beta_i]^T \sim \mathcal{MVN}(\boldsymbol{\mu}_\mathcal{I}, \boldsymbol{\Sigma}_\mathcal{I}), \tag{3.7}$$

where $\boldsymbol{\mu}_\mathcal{I}$ is the item mean vector and $\boldsymbol{\Sigma}_\mathcal{I}$ is the item covariance matrix. The logs of the variance of RT and of the discrimination parameter are modeled because these parameters are restricted to positive values. The intercept of the item response function ($-\alpha_i\beta_i$) is modeled instead of the difficulty (β_i) because it makes estimation of the model more stable.

At the person level, a bivariate normal distribution is used to model the de-

pendence between speed and ability:

$$[\theta_p, \tau_p]^T \sim \mathcal{MVN}\left([0,0]^T, \mathbf{\Sigma}_{\mathcal{P}}\right), \tag{3.8}$$

where $\mathbf{\Sigma}_{\mathcal{P}}$ is the person covariance matrix. To ensure the identification of the model, the mean vector of the person parameters is constrained to a zero vector and the variance of θ is constrained to one.

The model can be estimated using a Gibbs Sampler algorithm (see Appendix A for the details). At each iteration g of the Gibbs Sampler after the burn-in a sample from the posterior distribution of the model parameters given the data is obtained. Using these values a new replicated data set with RAs and RTs is generated under the model:

$$X_{pi}^{(g)} \sim Bernoulli\left(\frac{\exp(\alpha_i^{(g)}(\theta_p^{(g)} - \beta_i^{(g)}))}{1 + \exp(\alpha_i^{(g)}(\theta_p^{(g)} - \beta_i^{(g)}))}\right), \tag{3.9}$$

$$T_{pi}^{(g)} \sim \ln\mathcal{N}\left(\xi_i^{(g)} - \tau_p^{(g)}, \sigma_i^{2(g)}\right). \tag{3.10}$$

The discrepancy measures of interest are computed at each iteration g for both the observed data $D^{(g)}$ and the replicated data $D_{rep}^{(g)}$. The posterior predictive p-value [PPP-value] is the probability of obtaining a replicated discrepancy measure larger than the observed measure (Rubin, 1984; Meng, 1994; Gelman et al., 1996). This probability can be approximated by the number of iterations in which $D^{(g)} \leq D_{rep}^{(g)}$. In the next section we will discuss which discrepancy measures can be indicative of different violations of CI.

3.3 Discrepancy measures

In the Introduction we discussed three possible consequences of violations of CI: residual correlation between RT and RA, heterogeneity of variances of RT between the correct and the incorrect responses, and interaction effect of RT and ability on RA. Here we describe three discrepancy measures that address these consequences of violations of CI.

The first discrepancy measure considers the partial correlation between the observed RAs and the observed RTs to item i given the persons' ability and speed parameters.

$$D_{1i}^{(g)} = Cor(X_i, T_i \,|\, \boldsymbol{\theta}^{(g)}, \boldsymbol{\tau}^{(g)}) \tag{3.11}$$

This correlation can be computed using `ppcor` R-package. This discrepancy measure directly captures residual correlations between X_i and T_i which cannot be explained by the latent variables. If CI holds then the expectation of this correlation is equal to zero for each item. This measure is similar in spirit to discrepancy

measures that have been proposed in the context of evaluating local independence, such as the Q_3-statistic (Yen, 1984) and the SMBC-statistic (Levy et al., 2015). These existing measures capture residual correlations for item-pairs that remain after conditioning on the latent variable(s) explaining RAs. The proposed discrepancy measure differs from these existing measures in that here only a single residual correlation is considered per item (rather than $n-1$ item-pair comparisons per item), and this residual correlation is obtained for two variables that are of a different type (RA is discrete, while RT is continuous).

The second discrepancy measure considers the difference between the log of the variance of the observed RTs of the correct responses and the log of the variance of the observed RTs of the incorrect responses:

$$D_{2i} = \ln\left(Var(T_i \,|\, X_i = 1)\right) - \ln\left(Var(T_i \,|\, X_i = 0)\right) \tag{3.12}$$

This discrepancy measure is aimed at the kind of violation of CI in which there is not necessarily a residual correlation between the RAs and the RTs, but where the two distributions of RTs differ in their variances. For example, for some items correct responses might generally tend to be less variable in terms of RT, because the underlying response processes may be more similar to each other than those leading to incorrect responses.

The final discrepancy measure considers the difference between the item-rest correlation of the item for fast responses and the item-rest correlation of the item for slow responses:

$$D_{3i} = Cor(X_i, \sum_{j \neq i} X_j \,|\, T_i < T_{i,med}) - Cor(X_i, \sum_{j \neq i} X_j \,|\, T_i > T_{i,med}) \tag{3.13}$$

where the slow and the fast responses are defined as the observed responses with a RT longer or shorter than the sample median RT to the item ($T_{i,med}$), respectively. This measure is aimed at a type of violation of the CI where slow and fast responses do not necessarily differ in the probability of a correct response but where they do differ in the strength of the relationship between the item and the measured ability. As has been found in empirical data and as predicted by measurement models (Coyle, 2003; Maris & van der Maas, 2012; Bolsinova, de Boeck, & Tijmstra, 2015) slow responses may sometimes be more or less informative about a person's ability than the fast responses. The item-rest correlation is used since it is a simple classical test theory statistic which roughly captures the discriminatory power of the item.

The last two measures are test statistics, meaning that they do not depend on the values of the model parameters. For the observed data they have to be computed only once. This is not the case for the first measure, since it conditions on the values of $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$, and hence this measure needs to be computed for the

observed data at each iteration.

At each iteration three replicated discrepancy measures are computed per item in the same way as the observed measures in Equations 3.11, 3.12, and 3.13 but using the replicated data $\mathbf{X}^{(g)}$ and $\mathbf{T}^{(g)}$ instead of the observed data $\mathbf{X}$ and $\mathbf{T}$. Each of the three PPP-values per item, denoted by p_{1i}, p_{2i}, and p_{3i}, can be approximated as the number of iterations is which the replicated discrepancy is *larger* than the observed discrepancy.

In the case of these three discrepancy measures both PPP-values close to zero and those close to one indicate that the model does not adequately capture the aspects of the data summarised by these discrepancy measures, since both highly positive and highly negative residual correlations, differences between log-variances, and differences between item-rest correlations are indicative of conditional dependence.

3.3.1 Test level decision criterion

Based on the observed distribution for each of the three PPP-values of the items in a test, a researcher has to decide whether there are too many extreme values to retain the assumption of CI. This means that some criterion has to be chosen that determines whether a PPP-value should be considered extreme. We suggest to symmetrically consider PPP-values below $\frac{\pi}{2}$ and above $1-\frac{\pi}{2}$ to be extreme, where π is some small value, for example .05. It may be noted that since PPP-values are not necessarily uniformly distributed under CI (Meng, 1994), π is not necessarily equal to the false positive rate. We suggest for each discrepancy measure to compare the number of items with an extreme PPP-value, denoted by $n_{extreme}$, with the distribution of the number of items with an extreme PPP-value that would be obtained if they were independently and identically uniformly distributed on the interval [0,1]. This amounts to using a binomial distribution for the number of successes with a probability of π out of n trials. A decision about CI being violated is taken if for at least one of the measures the number of extreme PPP-values satisfies the condition

$$\left(1 - \sum_{k=0}^{n_{extreme}} \binom{n}{k} \pi^k (1-\pi)^{n-k}\right) < p_{crit}, \tag{3.14}$$

where p_{crit} is a chosen value on the interval from 0 to 1 which is supposed to be acceptably low, for example .05. If the distribution of the PPP-values is not uniform but more concentrated around .5, then this makes the criterion more conservative, because the probability of an extreme value will be smaller than under uniformity. Because uniformity and independence of the PPP-values are not guaranteed the proposed criterion should not be taken to imply that the false positive rate is fixed at a specific known value p_{crit}. Rather, the theoretical binomial distribution pro-

vides a mathematically convenient starting point from which to derive a criterion that may be useful but which performance needs to be assessed, as will be done in the simulation study. When using multiple PPC, one might choose to use a lower value for p_{crit} to prevent the inflation of the overall chance of a misclassification of the data set as having a violation of the CI due to multiple testing. Here we use $.05/m$, where m is the number of PPC that are used to assess CI.

3.4 Simulation studies

3.4.1 Specificity of the PPC

Specificity when the lower-level models are correctly specified

Methods

The data were generated under CI using the hierarchical model for the RT and RA, using a 2PL model for the RA (see Equation 3.6) and a lognormal model for the RTs (see Equation 3.3). The item and the person parameters were simulated in the following way:

$$[\theta_p, \tau_p]^T \sim \mathcal{MVN}\left(\boldsymbol{\mu}_{\mathcal{P}} = \mathbf{0}, \sigma_\theta^2 = \sigma_\tau^2 = 1, \rho_{\theta\tau}\right); \tag{3.15}$$

$$[\xi_i, \ln(\sigma_i^2), \ln(\alpha_i), -\alpha_i\beta_i]^T \sim \mathcal{MVN}\left(\boldsymbol{\mu}_{\mathcal{I}} = \mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{I}} = \mathbf{I}_4[1, 0.5, 0.5, 1]^T\right). \tag{3.16}$$

For each combination of sample size (500, 1000 or 2000), test length (20 and 40) and correlation between ability and speed (0 and .5) 500 replicated data sets were simulated. The hierarchical model was fitted to each of the replicated data sets using the Gibbs Sampler (see Appendix A) with 3000 iterations (including 1000 of burn-in), and each second iteration after the burn-in was used for the PPC with the three discrepancy measures described in the previous section.

First, the performance of each of the discrepancy measures was evaluated separately as if only one measure were used for testing CI, where $p_{crit} = .05$ was used. Following that criterion and the guidelines in Equation 3.14, a conclusion about CI being violated for the test was drawn if $n_{extreme} > 3$ for $n = 20$ and if $n_{extreme} > 4$ for $n = 40$. Second, the performance of the combination of the three discrepancy measures was evaluated, where $p_{crit} = .05/3$ was used for each of the PPC. CI was considered to be violated if for at least for one of the three measures $n_{extreme} > 3$ for $n = 20$ or $n_{extreme} > 5$ for $n = 40$.

The type I error rates of the existing procedures (the LM-test and the KS-tests) were also evaluated in this simulation study. For the LM-test Bonferroni correction was used to control for the effect of multiple comparisons. For the KS-tests, the equality of the RT distributions was tested after conditioning on the number of items correct and CI was rejected if either the minimum of the

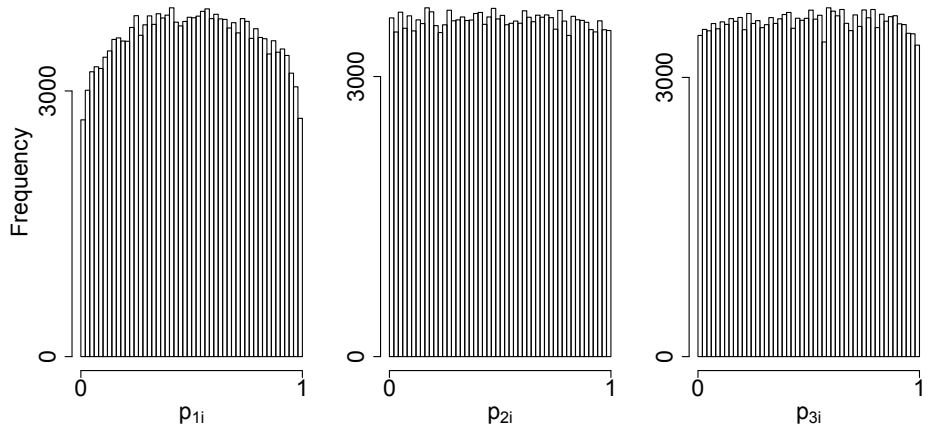


Figure 3.1: Distributions of the PPP-values for the three discrepancy measures under CI for all 12 conditions combined. Each histogram is based on 180,000 PPP-values.

item-level p-values was smaller than $.05/n$ or their maximum was smaller than .05 (Bolsinova & Maris, 2016).

Results

Figure 3.1 shows the histograms of the PPP-values for the three discrepancy measures for all the simulation conditions combined. The three observed distributions closely resemble a uniform distribution. However, for p_{1i} the extreme values are slightly under-represented compared to a uniform distribution.

Table 3.1 shows for each condition the proportion of data sets where a violation of CI was falsely detected, based on each of the PPC individually and based on the three checks combined. For p_{1i}, the false discovery rate was generally lower than for the other two checks, which is in line with the observation made in the previous paragraph that the distribution of p_{1i} (Figure 3.1) shows some deviation from uniformity. For p_{2i} and p_{3i}, the false discovery rate was close to the proportion of rejections that would be expected for PPP-values with an i.i.d. uniform distribution (.016 for $n = 20$ and .048 for $n = 40$). For the combination of the three checks, the proportion of false positives is somewhat lower than expected for i.i.d. uniformly distributed PPP-values (.048 for $n = 20$ and .042 for $n = 40$).

Table 3.1 also includes the observed type I error rates of the LM-test and the KS-tests. The type I error rate of the LM-test is in most of the conditions slightly lower than .05. The KS-tests are even more conservative, since correction for multiple testing is performed both within an item and between the items (Bolsinova

Table 3.1: Proportion of data sets generated under CI where a violation of CI was detected based on each of the three PPC individually and combined, the LM-test and the KS-tests (500 replications).

n	N	$\rho_{\theta\tau}$	PPC				LM-test	KS-tests
			p_{1i}	p_{2i}	p_{3i}	$\{p_{1i}, p_{2i}, p_{3i}\}$		
20	500	0	.000	.014	.012	.026	.030	.016
		.5	.004	.020	.010	.034	.056	.026
	1000	0	.004	.008	.006	.018	.046	.012
		.5	.012	.018	.008	.038	.036	.034
	2000	0	.004	.014	.012	.030	.044	.034
		.5	.012	.014	.016	.040	.038	.068
40	500	0	.028	.038	.032	.020	.040	.000
		.5	.024	.032	.030	.018	.056	.004
	1000	0	.030	.036	.044	.030	.036	.016
		.5	.040	.056	.028	.022	.042	.012
	2000	0	.018	.034	.040	.018	.030	.028
		.5	.010	.048	.050	.032	.036	.026

& Maris, 2016).

Specificity when the lower-level models are misspecified

The first simulation results showed that the PPC rarely classify data sets generated under CI as having violations of the assumption. Next, we evaluate the robustness of the PPC to the misspecification of the RT model and the RA model. These lower level misspecifications do not affect the relationship between the RAs and the RTs and do not influence CI. However, it could be that the performance of the procedure is affected because the posterior predictive distribution is obtained using the wrong model. We investigated whether the specificity of the PPC suffers from these misspecifications.

Methods

Two misspecifications of the lower-level models were considered. First, a possibility is that the model for the RTs is misspecified. Here, we consider a situation where the data generating model for the RTs includes an extra item parameter:

$$t_{pi} \sim \ln N(a_i(\xi_i - \tau_p), \sigma_i^2), \tag{3.17}$$

where a_i is an extra item parameter similar to the item discrimination in the model for the RA (Fox & van der Linden, 2007; Klein Entink, Fox, & van der Linden, 2009) which reflects that items might differ with respect to the decrease in RT as speed increases. Second, the model for RA can be misspecified. Here, we consider the case where the data generating model is not a 2PL but a three-parameter logistic model (Birnbaum, 1968):

$$\Pr(x_{pi} = 1) = c_i + (1 - c_i)\frac{\exp(\alpha_i(\theta_p - \beta_i))}{1 + \exp(\alpha_i(\theta_p - \beta_i))}, \tag{3.18}$$

where c_i is the guessing parameter of item i. The extra item parameters were simulated as follows: $a_i \sim \ln\mathcal{N}(0, 0.5)$ and $c_i \sim \mathcal{U}(.2, .3)$.

First, the robustness of the PPC under a baseline condition ($N = 1000$, $n = 20$, $\rho_{\theta\tau} = .5$) was analysed. Then, the effect of changing one of these design parameters was investigated, resulting in 5 simulation conditions for each type of misspecification (note that to investigate the effect of sample size both a smaller and a larger sample size were considered).

For each condition 500 data sets were simulated under CI. The hierarchical model with lower level models defined in (3.3) and (3.6) was fitted to the replicated data sets and the PPC were performed. Robustness of the PPC was compared to that of the LM-test (see Appendix B for details) and the KS-tests.

Results

Table 3.2 shows that when the RT model was misspecified, the specificity of the PPC did not appear to be effected: The proportions of the data sets in which a violation of CI was falsely detected were similar to those when the lower-level models were correctly specified (see Table 3.1). The type I error rate of the KS test was not inflated. The type I error rate of the LM-test was strongly inflated (up to 1) in the conditions with correlated speed and ability. This means that unlike the other two tests the LM-test is very much dependent on the correct specification of the RT distribution.

When the RA model was misspecified, the specificity of p_{1i} and p_{2i} was hardly affected. However, the specificity of the p_{3i} suffered from this misspecification when $\rho_{\theta\tau} = .5$, in the sense that the false discovery rate was considerably larger than when the RA model was correctly specified (see Table 3.1). This happens because when θ and τ are correlated the ability of persons with slow responses is on average lower than the ability of persons with fast responses, and therefore the item-rest correlation for the slow responses (corresponding to on average lower ability) decreases due to guessing, which makes the distribution of p_{3i} not uniform but skewed with a large number of values close to 0. The performance of the combination of the three PPC was affected due to the problem with p_{3i}. If guessing

Table 3.2: Robustness of the tests of CI against the misspecifications of the lower level models (500 replications). The data were generated under CI, but during estimation either the RT or the RA model was misspecified.

N	n	$\rho_{\theta\tau}$	PPC				LM-test	KS-tests
			p_{1i}	p_{2i}	p_{3i}	$\{p_{1i}, p_{2i}, p_{3i}\}$		
RT model misspecified								
1000	20	0.5	.012	.012	.026	.050	.984	.046
500	20	0.5	.000	.018	.024	.042	.872	.012
2000	20	0.5	.008	.018	.016	.042	1	.056
1000	**40**	0.5	.038	.020	.042	.038	1	.016
1000	20	**0**	.000	.014	.006	.020	.006	.030
RA model misspecified								
1000	20	0.5	.020	.018	.034	.072	.052	.024
500	20	0.5	.006	.018	.028	.048	.048	.020
2000	20	0.5	.012	.008	.114	.130	.040	.040
1000	**40**	0.5	.024	.046	.106	.056	.054	.006
1000	20	**0**	.006	.008	.028	.040	.056	.026

is an important factor in a test then it may be advisable either not to use p_{3i} or to use a model that accounts for the guessing behaviour. The type I error rates of the LM-test and of the KS-tests were not inflated when the RA model was misspecified

3.4.2 Sensitivity of the PPC

Methods

To evaluate how well the PPC detect violations of CI, we simulated data under different models with five different types of violations of CI. The exact specification of each violation can be found in Tables 3.3 and 3.4, and the choice of these conditions is motivated below.

Types 1 and **Type 2** both specify that the distribution of t_{pi} depends on whether the response x_{pi} is correct or incorrect. The violation of **Type 1** is the kind of violation for which the LM-test for CI (van der Linden & Glas, 2010) was designed: The location parameters of the lognormal distribution of RTs differ for the correct responses ($\xi_i + \lambda_i\sigma_i$) compared to the incorrect responses (ξ_i). **Type 2** captures the idea that the dependence between RT and RA is not necessarily constant across persons and could depend on the person's ability, as is predicted by the Signed Residual Time model (Maris & van der Maas, 2012) and as has

Table 3.3: Specification of the models for RT and RA for the different types of violations of CI

	Distribution of t_{pi}	$\Pr(x_{pi}=1)$	Distribution of extra parameters
1	$\ln\mathcal{N}\left(\xi_i+\lambda_i\sigma_i x_{pi}-\tau_p,\sigma_i^2\right)$	see Equation 3.6	$\lambda_i\sim\mathcal{N}(0,\sigma_\lambda^2)$
2	$\ln\mathcal{N}\left(\xi_i-\frac{(-1)^{x_{pi}}}{2}\eta\theta_p\sigma_i-\tau_p,\sigma_i^2\right)$	see Equation 3.6	-
3	$\ln\mathcal{N}(\xi_i-\tau_p,\sigma_i^2+\sigma_p^2)$	see Equation 3.6	$\sigma_p^2\sim\ln\mathcal{N}(\rho_\sigma\theta_p,(1-\rho_\sigma^2))$
4	see Equation 3.3	$\frac{e^{(\alpha_i(\theta_p-\beta_i-\delta_i t^*_{pi}))}}{1+e^{(\alpha_i(\theta_p-\beta_i-\delta_i t^*_{pi}))}}$	$\delta_i\sim N(0,\sigma_\delta^2)$
5	see Equation 3.3	$\frac{e^{\left(\alpha_i\gamma_i^{t^*_{pi}}(\theta_p-\beta_i)\right)}}{1+e^{\left(\alpha_i\gamma_i^{t^*_{pi}}(\theta_p-\beta_i)\right)}}$	$\gamma_i\sim\ln N(0,\sigma_\gamma^2)$

Table 3.4: Specification of medium and small violations of CI of each type.

	Medium violation	Small violation	Comments
1	$\sigma_\lambda^2=0.627$	$\sigma_\lambda^2=0.251$	Across items on average the absolute standardised difference in the means of the log RTs for correct and incorrect responses is 0.5 or 0.2, respectively.
2	$\eta=-0.627$	$\eta=-0.251$	Across persons on average the absolute standardised ldifference in the expected log RTs for correct and incorrect responses is 0.5 or 0.2, respectively.
3	$\rho_\sigma=-.5$	$\rho_\sigma=-.2$	-
4	$\sigma_\delta^2=0.627$	$\sigma_\delta^2=0.125$	Across items on average the absolute difference in the difficulties given fast and slow responses is 0.5 or 0.2, respectively.
5	$\sigma_\gamma^2=0.627$	$\sigma_\gamma^2=0.125$	Across items on average the absolute difference in the logs of discriminations given fast and slow responses is 0.5 or 0.2 (i.e, the ratio of the discriminations is 1.65 or 1.22), respectively.

been addressed by the KS-tests of Bolsinova and Maris (2016). Here, $\eta\theta_p$ is the standardized difference between the expectation of the log RTs for the correct responses and for the incorrect responses for person p. A negative value of η has been used, meaning that for persons with a high ability correct responses are faster than incorrect responses, while for persons with a low ability incorrect responses are faster than correct responses.

Type 3 focuses on the variances of log RTs. While CI predicts equal residual variances of the log RTs for correct and incorrect responses, this does not have to hold in practice. It could be the case that the RTs of low-ability test takers are more varied than those of high-ability test takers, for example because the former may sometimes skip or guess on a question. Because high-ability test takers will have a higher proportion of correct responses, this will result in a lower residual variance for correct responses than for incorrect responses, meaning that CI (given θ and τ) is violated. We specified this condition by setting the variance of log RT to be person- as well as item-dependent, where the person component of the variance is negatively correlated with ability.

Types 4 and 5 capture violations of CI due to differences between fast and slow response processes, a possibility that has been discussed in the literature (Partchev & De Boeck, 2012; Goldhammer et al., 2014; Bolsinova et al., 2015). Different item response functions were specified depending on a response being relatively fast or slow, compared to what would be expected under the log-normal RT model. In Table 3.3 we use a dummy variable t^*_{pi} to indicate whether a response is relatively slow for person p on item i, obtained through

$$t^*_{pi} = \mathcal{I}\left(t_{pi} > \exp\left(\xi_i - \tau_p + \sigma_i^2/2\right)\right). \tag{3.19}$$

In **Type 4** the difficulties differ for slow responses ($\beta_i + \delta_i$) compared to fast responses (β_i). In **Type 5** the discriminations are different: $\alpha_i\gamma_i$ and α_i for slow and fast responses, respectively, capturing the idea that the amount of information that a response provides is different for slow and fast responses (Coyle, 2003; Maris & van der Maas, 2012).

First, for each of the types of violations the assumption of CI was tested in a baseline condition: $N = 1000$, $n = 20$, $\rho_{\theta\tau} = .5$ and a medium violation of CI. Second, the effect of changing one of the following parameters compared to the baseline condition was evaluated: sample size (500 or 2000 instead of 1000), test length (40 items instead of 20), correlation between ability and speed (.0 instead of .5) and size of violation (small instead of medium), resulting in five extra conditions per type of violation. Finally, one extra condition was used: with $N = 2000$ and a small violation, because we expected that a sample size of 1000 might not be enough for detecting small violations. In each of the conditions, PPC with the combination of $\{p_{1i}, p_{2i}, p_{3i}\}$ (see Section 3.4.1 for details), the LM-test and the KS-tests were performed in each of 100 simulated data sets.

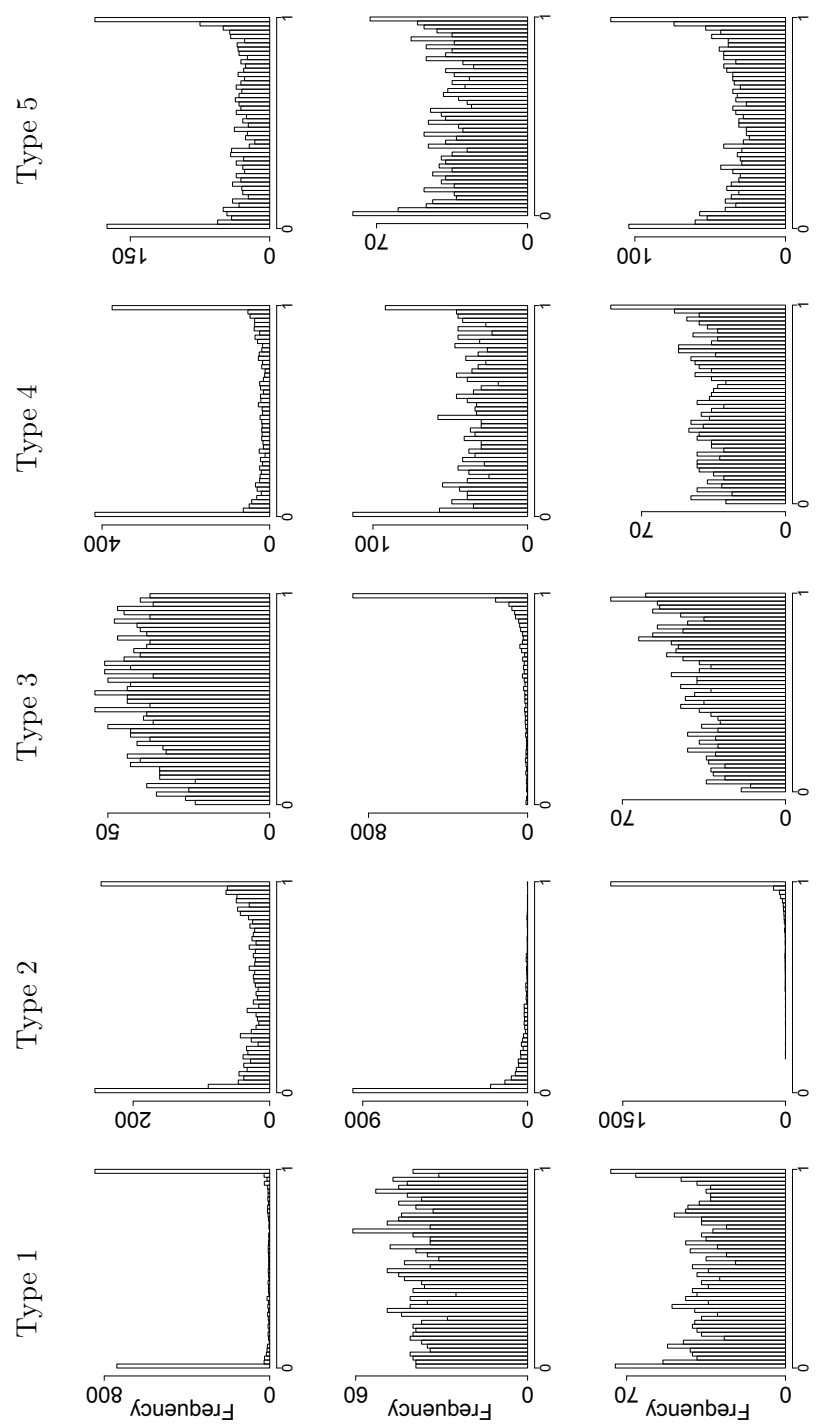


Figure 3.2: Distributions of the PPP-values for the three discrepancy measures for the five types of violations of CI in the baseline condition (based on 2,000 PPP-values)

Results

Figure 3.2 shows the distribution of the PPP-values for the baseline condition for each of the types of violations of CI. Violations of Type 1 and 4 resulted in a large number of extreme p_{1i}s, meaning that the observed residual correlations between RAs and RTs are generally stronger than expected under CI. Violations of Type 2 resulted in a large number of small p_{2i}s, meaning that differences between the observed variances of the RTs of the correct responses and those of the incorrect responses are generally higher (i.e., more positive) than expected under CI. Additionally, violations of Type 2 resulted in an even larger number of large p_{3i}, meaning that the differences between the observed item-rest correlations for the slow responses and for the fast responses are generally lower (i.e., more negative) than expected under CI. Violations of Type 3 resulted in a large number of large p_{2i}s, meaning that differences between the observed variances of the RTs of the correct responses and those of the incorrect responses are generally lower (i.e., more negative) than expected under CI. Violations of Type 5 resulted in a large number of extreme p_{1i}s and also a rather large number of extreme p_{3i}s.

Table 3.5 shows the proportion of data sets in which violations of CI were detected in each of the conditions by each of the three procedures (PPC, LM-test and KS-tests). Because in practice we recommend to use the combination of discrepancy measures in order to be able to detect a variety of ways in which CI might be violated, we present only the results based on the combination of the three measures in Table 3.5 (see Section 3.3.1 for details). The results for the individual discrepancy measures can be found in Appendix C.

The PPC and the LM-test detected violations of Type 1 in all simulated data sets in all conditions. The KS-tests did not have adequate power (>.8) in the condition with $N = 1000$ and a small violation, but did have adequate power in all other conditions. Only the PPC had adequate power in all the conditions of Type 2. The LM-test lacked power to detect violations of Type 2 when the sample size is small ($N = 500$) and when the size of violation is small. The KS-tests had lower power than the LM-test and the PPC in all conditions of Type 2. For violations of Type 3 the PPC had adequate power in every condition, except for the combination of small violation and $N = 1000$. The KS-tests were however unable to detect this type of violation adequately. The LM-test had in most conditions lower power to detect violation of Type 3 than the PPC, but performed better than the PPC when the violations were small and $N = 1000$. However, using the LM-test results in a misrepresentation of the kind of violation present in the data: The locations of the two distributions are actually the same while the variances are different. The PPC outperformed the other two procedures in all conditions of Type 4, but had relatively low sensitivity when the violation was small (.48 and .74 when $N = 1000$ and $N = 2000$, respectively). The most difficult type of violation to detect was Type 5. Only with the PPC power above .8 was achieved

Table 3.5: Proportion of correctly detected violations of CI using PPC - posterior predictive checks using the combination $\{p_{1i}, p_{2i}, p_{3i}\}$, LM - Lagrange Multiplier test (van der Linden & Glas, 2010), KS - Kolmogorov-Smirnov tests (Bolsinova & Maris, 2016). For the specifications of medium (m) and small (s) sizes of violation see Table 4. Based on 100 replications.

N	n	$\rho_{\theta\tau}$	Size	Type 1			Type 2			Type 3		
				PPC	LM	KS	PPC	LM	KS	PPC	LM	KS
1000	20	.5	m	1	1	1	1	.95	.87	1	.74	.06
500	20	.5	m	1	1	.91	1	.66	.19	.91	.75	.02
2000	20	.5	m	1	1	1	1	.99	1	1	.72	.05
1000	**40**	.5	m	1	1	1	1	.99	.69	1	.87	.01
1000	20	**.0**	m	1	1	.99	1	.94	.80	.97	.64	.04
1000	20	.5	**s**	1	1	.45	.87	.27	.09	.59	.76	.04
2000	20	.5	**s**	1	1	.85	.99	.44	.23	.78	.79	.05

N	n	$\rho_{\theta\tau}$	Size	Type 4			Type 5		
				PPC	LM	KS	PPC	LM	KS
1000	20	.5	m	.99	.99	.54	.68	.52	.12
500	20	.5	m	.96	.84	.24	.34	.37	.03
2000	20	.5	m	1	.99	.86	.96	.76	.38
1000	**40**	.5	m	1	1	.59	.92	.77	.10
1000	20	**.0**	m	.99	.92	.43	.65	.63	.11
1000	20	.5	**s**	.48	.30	.01	.09	.11	.04
2000	20	.5	**s**	.74	.53	.21	.24	.27	.12

in two conditions: When either the sample size was large or the number of items was large. In all other conditions the PPC performed similarly to the LM-test, while both outperformed the KS-tests.

3.5 Empirical example

To illustrate the performance of the PPC for CI, the procedure was applied to data of an arithmetics test which is a part of the exit examination in Dutch secondary education. The data of the students from a common educational background (preparatory higher vocational education) to one of the test versions (consisting of 52 items) was used. Only the items with proportions of correct responses between .2 and .8 were used, resulting in a final test length of 38 items. Data from one person were deleted because the total time on the test was 197 seconds while all other students spent more than 1000 seconds on the test. The final sample size

was 610.

A hierarchical model with a 2PL model for RA (3.6) and a log-normal model for RT (3.3) was fitted to the data using a Gibbs Sampler (see Appendix A). Two chains with 11,000 iterations each (including 1,000 iterations of burn-in) were used. Convergence was evaluated using $\hat{R}$-statistic (Gelman & Rubin, 1992) for all item parameters and for the higher level item and person parameters, and overall with the multivariate scale reduction factor (Brooks & Gelman, 1998). All univariate $\hat{R}$-statistics and the multivariate scale reduction factor were below 1.1, indicating reasonable convergence. After the burn-in in every second iteration a replicated data set was simulated and discrepancy measures were computed (i.e., in total 10,000 samples from the posterior predictive distribution obtained using the two chains were used).

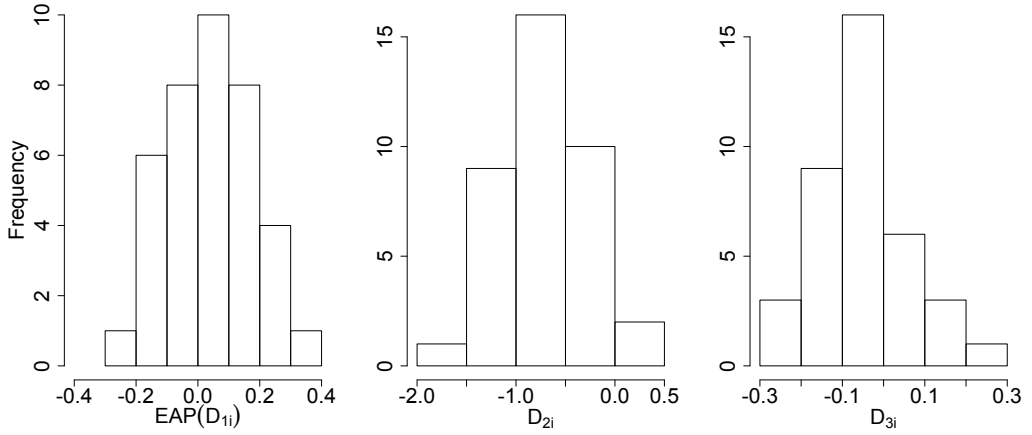


Figure 3.3: Histograms of the discrepancy measures between the observed data and the hierarchical model assuming CI for the arithmetics test data.

The PPC resulted in 22 extreme p_{1i}s (either above .975 or below .025), 32 extreme p_{2i}s and 8 extreme p_{3i}s. Based on these results we conclude that CI is violated for the data set at hand, since for at least one of the discrepancy measures (and in this case for all) the number of extreme PPP-values exceeded 5 (see Equation 3.14). Therefore, the hierarchical model does not seem to hold for these data. Among the items with an extreme p_{1i}, RA had a positive residual correlation with RT for 14 items and a negative residual correlation with RT for 8 items. Among the items with an extreme p_{2i}, the RT distribution of correct responses had a higher variance than the RT distribution of incorrect responses for 2 items and a lower variance than the RT distribution of incorrect responses for 30 items. Among the items with an extreme p_{3i}, observed item-rest correlation was higher for the fast responses for 6 items and was higher for the slow responses

for 2 items. Figure 3.3 shows the histograms of the discrepancy measures of the observed data. Since the first discrepancy measure is computed at each iteration, its expected a posteriori estimate is calculated for each item, denoted by EAP(D_{1i}). The values of the observed discrepancy measures give some indication of the size of violations of CI. For example, for 6 items the EAP of the residual correlations exceeded the benchmark of small effect size, and for 7 items the variance of the log RTs of incorrect responses was at least three times as large as the variance of the log RTs of correct responses.

3.6 Discussion

PPC presented in this paper offer a powerful, robust and flexible approach to testing CI. In most conditions of the simulation study, the PPC detected violations of CI more often or at least as often as the existing tests of CI. Results strongly indicate that the proposed PPC method can be useful in detecting different types of violations of CI, but further research may be needed to determine the performance of the procedure in a wider range of realistic scenarios.

The three proposed discrepancy measures capture different ways in which CI may be violated. D_{1i} measures the residual correlation that remains between RA and RT after taking speed and ability into account.

Positive residual correlations which were the most common in the empirical example could be explained by persons varying their effective speed as a consequence of changing their position on the speed-accuracy trade-off (van der Linden, 2009). However, negative residual correlations cannot be explained in this way. A possible explanation for the negative correlations, which were also found in the empirical example, is that respondents finish working on an item if they have found a correct response, but may continue working on the item if they have not yet found the correct answer. In that case it could be that slow responses are less often correct, resulting in a negative residual correlation. Further research is needed to reveal the different possible causes of residual correlations for different types of items.

D_{2i} captures differences in the variance of RTs depending on whether a response is correct or incorrect. High values of p_{2i}, as were observed for many of the items in the empirical example, indicate that RT is more variable for incorrect responses and may reflect that the underlying response processes are more heterogeneous than those resulting in correct responses. This could also be of substantive interest, because it may be relevant to distinguish between different processes that lead to an incorrect response (e.g., guessing, anxiety, or misapplying a response strategy).

D_{3i} was designed to measure difference between fast and slow responses in terms of the item discrimination, as has been suggested in the literature (Partchev & De Boeck, 2012; Goldhammer et al., 2014). A large number of extreme p_{3i}

values might indicate that fast and slow responses have underlying processes that are qualitatively different and as a result certain type of responses based on one process (e.g, careful deliberation) might be more informative about ability than response based on other processes (e.g., guessing). It may be relevant to take this into consideration when modeling the data, as has been suggested by Bolsinova et al. (2015).

In our treatment of the PPC method we proposed to use a combination of discrepancy measures, rather than base the decision to reject CI on a single discrepancy measure. The motivation for this choice is that in practice it may be difficult to anticipate which types of violations are likely to occur in the data. Furthermore, it could very well be that for different items different types of violations are present, making it useful to look at the set of discrepancy measures that cover a range of potential consequences of these violations. Additionally, having information about these different types of discrepancies provides a more informative picture of the likely sources of the violation and may provide a user with suggestions of where it makes sense to extend the model to incorporate those sources. However, as mentioned before care should be taken when inferring the source of conditional dependence, because the observed discrepancies may be due to a variety of different sources. Also there may be other types of violations that are not addressed by the discrepancy measures presented in this paper. However, we offer a flexible framework that can be extended with new discrepancy measures if particular violations of CI are suspected.

As we have shown in the simulation study the PPC procedure is rather robust to violations in the lower level models for RT and RA unlike the LM-test, which is important because a model used for analysis is always a simplification of reality. Moreover, the procedure can be easily extended to deal with more complex lower-level models, for example including a time discrimination parameter or a guessing parameter or including a different RT distribution, as long as a Bayesian estimation algorithm for these models is available.

The method requires a choice of a higher-level model, but is flexible with respect to which model is chosen. In this paper the hierarchical model for RT and RA (van der Linden, 2007) which is prevalent in educational measurement was used, but the general method can readily be adapted to be applicable for other models that assume CI, for example the diffusion model (Ratcliff, 1978) or the independent race model (Tuerlinckx & De Boeck, 2005), provided methods are available for sampling from the posterior distribution of the model parameters and for simulating data under the model.

Whereas the PPP-values may help users determine if conditional dependence is present, the proposed discrepancy measures may provide insight into the severity of those violations. These measures can function as indicators of effect sizes of the different ways in which CI can be violated (partial correlation, ratio of variances, difference between item-rest correlations), as illustrated in the empirical example.

As such they may be useful in determining the likely impact of observed violations of CI on model-based inferences. However, further research into the robustness of models assuming CI is needed for such an assessment to be realistic.

Using PPC CI is tested for a particular set of latent variables (in the case of the hierarchical model, these are θ and τ). If CI is violated for this particular set, it does not mean that CI cannot hold for a different set of latent variables. If an alternative CI model with a new set of latent variables is formulated, then the CI assumption can be again tested with PPC using samples from the posterior distribution of the parameters of the new model. This also means that the PPC can be used in the context of modeling conditional dependence to check whether an attempt to model conditional dependence has been successful.

The aim of this paper has been to provide new powerful methods of detecting a variety of violations of CI. Evaluating the practical impact of particular types of violations of CI on inferences that are made based on the model that make this assumption is beyond the scope of this paper. However, violations of CI may not only be relevant with regard to their consequences for the accuracy of the model inferences, but may also reveal substantively relevant aspects of the response process that are not accounted for in the model that is used. Investigating the ways in which CI is violated may therefore be of substantive interest. The proposed PPC framework and possible extensions of it may prove useful in addressing this substantive questions in future research.

3.7 Appendices

Appendix A

Although the variance of θ is constrained to 1, to improve the convergence of the model at each iteration of the Gibbs Sampler the full covariance matrix $\mathbf{\Sigma}_{\mathcal{P}}$ is sampled and at the end of each iteration all parameters are transformed to fit the scale defined by $\sigma_\theta^2 = 1$.

For the Bayesian estimation of the model, prior distributions for the item and the person hyper parameters ($\boldsymbol{\mu}_{\mathcal{I}}$, $\mathbf{\Sigma}_{\mathcal{I}}$ and $\mathbf{\Sigma}_{\mathcal{P}}$) have to be chosen. We choose vague prior distributions: independent normal distributions with a zero mean and a large variance (100) for the elements of $\boldsymbol{\mu}_{\mathcal{I}}$, half-$t$ distributions with $\nu = 2$ degrees of freedom and a scale parameter $A = 2$ for the standard deviations of the item parameters (Gelman, 2006), marginally uniform joint distribution for the correlations between the item parameters (Huang & Wand, 2013) and an inverse-Wishart distribution with 4 degrees of freedom and the identity matrix $\mathbf{I}_2$ as the scale parameter for $\mathbf{\Sigma}_{\mathcal{P}}$ (Hoff, 2009). Note, that the results are not sensitive to the choice of the scale parameter, because in the posterior distribution the prior is dominated by the data when $N \gg 4$ (Hoff, 2009, p.110). Independent prior distributions are assumed

Here we describe a Gibbs Sampler for sampling from the joint distribution of the model parameters, which is proportional to the product of the prior distribution and the density of the data:

$$p(\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\xi}, \boldsymbol{\sigma}^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\mathcal{P}}, \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}} \mid \mathbf{X}, \mathbf{T}) \propto p(\boldsymbol{\Sigma}_{\mathcal{P}}) p(\boldsymbol{\mu}_{\mathcal{I}}) p(\boldsymbol{\Sigma}_{\mathcal{I}}) \times \prod_p \mathcal{MVN}(\theta_p, \tau_p; \boldsymbol{\Sigma}_{\mathcal{P}}) \prod_i \frac{1}{\sigma_i^2 \alpha_i} \mathcal{MVN}(\xi_i, \ln \sigma_i^2, \ln \alpha_i, -\alpha_i \beta_i; \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}) \times \prod_p \prod_i \frac{1}{t_{pi} \sigma_i} \exp\left(-\frac{(\ln t_{pi} - (\xi_i - \tau_p))^2}{2\sigma_i^2} \right) \frac{\exp(x_{pi}(\alpha_i(\theta_p - \beta_i)))}{1 + \exp(\alpha_i(\theta_p - \beta_i))}. \quad (3.20)$$

To ensure better convergence we re-parameterise the 2PL model for RA in terms of the slope (α_i) and the intercept ($\beta_i^* = -\alpha_i \beta_i$) of the item response function:

$$\Pr(X_{pi} = 1) = \frac{\exp(\alpha_i \theta_p + \beta_i^*)}{1 + \exp(\alpha_i \theta_p + \beta_i^*)}. \quad (3.21)$$

In the Gibbs Sampler the model parameters are subsequently sampled from their conditional posterior distributions given the current values of all the other parameters (Geman & Geman, 1984; Casella & George, 1992). Below is the description of the steps of the algorithm.

Step 1: For each item i sample the time intensity parameter from

$$p(\xi_i \mid \dots) \propto \exp\left(-\frac{(\xi_i - \mu_\xi^*)^2}{2\sigma_\xi^{*2}} \right) \prod_p \exp\left(-\frac{(\ln t_{pi} - (\xi_i - \tau_p))^2}{2\sigma_i^2} \right), \quad (3.22)$$

where μ_ξ^* and σ_ξ^{*2} are the conditional mean and the conditional variance of ξ_i given the other item parameters of item i:

$$\mu_\xi^* = \mu_1 + \boldsymbol{\sigma}_{1,-1} \boldsymbol{\Sigma}_{-1}^{-1} \left((\ln \sigma_i^2, \ln \alpha_i, \beta_i)^T - \boldsymbol{\mu}_{-1} \right) \quad (3.23)$$

$$\sigma_\xi^{*2} = \sigma_{\mathcal{I},1}^2 - \boldsymbol{\sigma}_{1,-1} \boldsymbol{\Sigma}_{\mathcal{I},-1}^{-1} \boldsymbol{\sigma}_{1,-1}^T, \quad (3.24)$$

where μ_{-1} is a vector of the means of other item parameters except time intensity, $\boldsymbol{\Sigma}_{\mathcal{I},-1}$ is a covariance matrix of the three remaining item parameters and $\boldsymbol{\sigma}_{1,-1}$ is a vector of covariances between item time intensity and other item parameters.

The distribution in (3.22) is a normal distribution:

$$(\xi_i \mid \dots) \sim \mathcal{N}\left(\frac{\frac{\sum_p (\ln t_{pi} + \tau_p)}{\sigma_i^2} + \frac{\mu_\xi^*}{\sigma_\xi^{*2}}}{\frac{N}{\sigma_i^2} + \frac{1}{\sigma_\xi^{*2}}}, \frac{1}{\frac{N}{\sigma_i^2} + \frac{1}{\sigma_\xi^{*2}}} \right) \quad (3.25)$$

Step 2: For each item i sample the variance of log RT from

$$p(\sigma_i^2 | \dots) \propto \left(\frac{1}{\sigma_i^2}\right)^{\frac{N}{2}+1} \exp\left(-\frac{(\ln\sigma_i^2 - \mu^*_{\ln\sigma^2})^2}{2\sigma^{*2}_{\ln\sigma^2}}\right) \prod_p \exp\left(-\frac{(\ln t_{pi} - (\xi_i - \tau_p))^2}{2\sigma_i^2}\right), \tag{3.26}$$

where $\mu^*_{\ln\sigma^2}$ and $\sigma^{*2}_{\ln\sigma^2}$ are defined similarly to μ^*_ξ and σ^{*2}_ξ in Equations 3.23 and 3.24. To sample from (3.26) Metropolis-Hastings algorithm is used with a proposal distribution

$$\sigma^{*2} \sim \mathcal{IG}\left(\frac{N}{2}, \frac{\sum_p (\ln t_{pi} - (\xi_i - \tau_p))^2}{2}\right), \tag{3.27}$$

and an acceptance probability

$$\Pr(\sigma_i^2 \to \sigma^{*2}) = \min\left(1, \exp\left(\frac{-(\ln\sigma^{*2} - \mu^*_{\ln\sigma^2})^2 + (\ln\sigma_i^2 - \mu^*_{\ln\sigma^2})^2}{2\sigma^{*2}_{\ln\sigma^2}}\right)\right). \tag{3.28}$$

Step 3: For each person p sample the ability parameter from

$$p(\theta_p \mid \dots) \propto \mathcal{N}\left(\theta_p; \frac{\rho_{\theta\tau}}{\sigma_\tau}\tau_p, (1-\rho^2_{\theta\tau})\right) \prod_i \frac{\exp(\alpha_i\theta_p + \beta_i^*)}{1+\exp(\alpha_i\theta_p + \beta_i^*)}, \tag{3.29}$$

where $\frac{\rho_{\theta\tau}}{\sigma_\tau}\tau_p$ and $(1-\rho^2_{\theta\tau})$ are the mean and the variance of the conditional normal distribution of θ_p given τ_p.

To sample from this distribution the single variable exchange algorithm (Marsman, Maris, Bechger, & Glas, 2015) is used. First, sample a candidate value

$$\theta^* \sim \mathcal{N}\left(\frac{\rho_{\theta\tau}}{\sigma_\tau}\tau_p, (1-\rho^2_{\theta\tau})\right); \tag{3.30}$$

then using this value simulate a response vector $\mathbf{x}^*$:

$$x_i^* \sim Bernoulli\left(\frac{\exp(\alpha_i\theta^* + \beta_i^*)}{1+\exp(\alpha_i\theta^* + \beta_i^*)}\right), \forall i \in [1:n]. \tag{3.31}$$

The probability of accepting θ^* as a new value of θ_p is:

$$\Pr(\theta_p \to \theta^*) = \min\left(1, \exp\left((\theta^* - \theta_p)\left(\sum_i \alpha_i x_{pi} - \sum_i \alpha_i x_i^*\right)\right)\right). \tag{3.32}$$

Step 4: For each person p sample the speed parameter from

$$p(\tau_p \mid \dots) \propto \exp\left(-\frac{(\tau_p - \sigma_\tau \rho_{\theta\tau}\theta_p)^2}{2(1-\rho_{\theta\tau}^2)\sigma_\tau^2}\right) \prod_i \exp\left(-\frac{(\ln t_{pi} - (\xi_i - \tau_p))^2}{2\sigma_i^2}\right), \quad (3.33)$$

where $\sigma_\tau \rho_{\theta\tau}\theta_p$ and $(1-\rho_{\theta\tau}^2)\sigma_\tau^2$ are the mean and the variance of the conditional normal distribution of τ_p given θ_p. The conditional posterior in (3.33) is normal:

$$(\tau_p \mid \theta_p, \dots) \sim \mathcal{N}\left(\frac{\sum_i \frac{(\xi_i - \ln t_{pi})}{\sigma_i^2} + \frac{\sigma_\tau \rho_{\theta\tau}\theta_p}{(1-\rho_{\theta\tau}^2)\sigma_\tau^2}}{\sum_i \frac{1}{\sigma_i^2} + \frac{1}{(1-\rho_{\theta\tau}^2)\sigma_\tau^2}}, \frac{1}{\sum_i \frac{1}{\sigma_i^2} + \frac{1}{(1-\rho_{\theta\tau}^2)\sigma_\tau^2}}\right). \quad (3.34)$$

Step 5: For each item i sample the item discrimination parameter from

$$p(\alpha_i \mid \dots) \propto \frac{1}{\alpha_i} \exp\left(-\frac{(\ln \alpha_i - \mu^*_{\ln\alpha})^2}{2\sigma^{*2}_{\ln\alpha}}\right) \prod_p \frac{\exp(\alpha_i\theta_p + \beta_i^*)}{1+\exp(\alpha_i\theta_p + \beta_i^*)}. \quad (3.35)$$

where $\mu^*_{\ln\alpha}$ and $\sigma^{*2}_{\ln\alpha}$ are defined similarly to μ^*_ξ and σ^{*2}_ξ in Equations 3.23 and 3.24. To sample from (3.35) Metropolis-Hastings algorithm is used with a lognormal distribution with the location parameter equal to the log of the current value of the discrimination parameter.

Step 6: For each item i sample the item intercept parameter from

$$p(\beta_i^* \mid \dots) \propto \exp\left(-\frac{(\beta_i^* - \mu^*_{\beta^*})^2}{2\sigma^{*2}_{\beta^*}}\right) \prod_p \frac{\exp(\alpha_i\theta_p + \beta_i^*)}{1+\exp(\alpha_i\theta_p + \beta_i^*)}. \quad (3.36)$$

where $\mu^*_{\beta^*}$ and $\sigma^{*2}_{\beta^*}$ are defined similarly to μ^*_ξ and σ^{*2}_ξ in Equations 3.23 and 3.24. To sample from (3.36) Metropolis-Hastings algorithm is used with a normal distribution with the mean equal to the current value of the intercept parameter.

Step 7: Sample the covariance matrix of the item parameters from:

$$p(\boldsymbol{\Sigma}_\mathcal{I} \mid \boldsymbol{\xi}, \boldsymbol{\sigma}^2, \boldsymbol{\alpha}, \boldsymbol{\beta}^*, \boldsymbol{\mu}_\mathcal{I}) \propto p(\boldsymbol{\xi}, \boldsymbol{\sigma}^2, \boldsymbol{\alpha}, \boldsymbol{\beta}^* | \boldsymbol{\mu}_\mathcal{I}, \boldsymbol{\Sigma}_\mathcal{I}) |\boldsymbol{\Sigma}_\mathcal{I}|^{-\frac{\nu+8}{2}} \prod_{k=1}^{4} \left(\nu\,(\boldsymbol{\Sigma}_\mathcal{I})_{kk} + \frac{1}{A^2}\right)^{-\frac{\nu+4}{2}}. \quad (3.37)$$

A Metropolis-Hastings algorithm is used to sample from this conditional posterior with the candidate value $\boldsymbol{\Sigma}^*$ sampled from the Inverse-Wishart distribution with n degrees of freedom and the scale matrix equal to

$$\sum_i \left((\xi_i, \ln\sigma_i^2, \ln\alpha_i, \beta_i^*)^T - \boldsymbol{\mu}_\mathcal{I}\right)\left((\xi_i, \ln\sigma_i^2, \ln\alpha_i, \beta_i^*)^T - \boldsymbol{\mu}_\mathcal{I}\right)^T, \quad (3.38)$$

such that the acceptance probability is equal to:

$$\Pr(\boldsymbol{\Sigma}_{\mathcal{I}} \rightarrow \boldsymbol{\Sigma}^*) = \min\left(1, \frac{|\boldsymbol{\Sigma}^*|^{-\frac{\nu+3}{2}} \prod_{k=1}^{4} \left(\nu\left(\boldsymbol{\Sigma}^*\right)_{kk} + \frac{1}{A^2}\right)^{-\frac{\nu+4}{2}}}{|\boldsymbol{\Sigma}_{\mathcal{I}}|^{-\frac{\nu+3}{2}} \prod_{k=1}^{4} \left(\nu\left(\boldsymbol{\Sigma}_{\mathcal{I}}\right)_{kk} + \frac{1}{A^2}\right)^{-\frac{\nu+4}{2}}}\right). \tag{3.39}$$

Step 8: Sample the mean vector of the item parameters from

$$p(\boldsymbol{\mu}_{\mathcal{I}} \,|\, \boldsymbol{\xi}, \boldsymbol{\sigma}^2, \boldsymbol{\alpha}, \boldsymbol{\beta}^*, \boldsymbol{\Sigma}_{\mathcal{I}}) \propto p(\boldsymbol{\xi}, \boldsymbol{\sigma}^2, \boldsymbol{\alpha}, \boldsymbol{\beta}^* \,|\, \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}) p(\boldsymbol{\mu}_{\mathcal{I}}). \tag{3.40}$$

With a multivariate normal prior for $\boldsymbol{\mu}_{\mathcal{I}}$, this conditional posterior is also a multivariate normal with a mean vector equal to

$$\left((100\mathbf{I}_4)^{-1} + n\boldsymbol{\Sigma}_{\mathcal{I}}^{-1}\right)^{-1} \left(\boldsymbol{\Sigma}_{\mathcal{I}}^{-1} \left(\sum_i \xi_i, \sum_i \ln \sigma_i^2, \sum_i \ln \alpha_i, \sum_i \beta_i^*\right)^T\right) \tag{3.41}$$

and the covariance matrix equal to $\left((100\mathbf{I}_4)^{-1} + n\boldsymbol{\Sigma}_{\mathcal{I}}^{-1}\right)^{-1}$.

Step 9: Sample the covariance matrix of person parameters from

$$p(\boldsymbol{\Sigma}_{\mathcal{P}} \,|\, \boldsymbol{\theta}, \boldsymbol{\tau}) \propto p(\boldsymbol{\theta}, \boldsymbol{\tau} \,|\, \boldsymbol{\Sigma}_{\mathcal{P}}) p(\boldsymbol{\Sigma}_{\mathcal{P}}). \tag{3.42}$$

This is the conditional posterior of the covariance matrix of a multivariate normal distribution, which given the Inverse-Wishart prior is known to be an inverse-Wishart distribution (see for example, Hoff (2009)):

$$(\boldsymbol{\Sigma}_{\mathcal{P}} \,|\, \boldsymbol{\theta}, \boldsymbol{\tau}) \sim \text{Inv-Wishart}\left(4 + N, \mathbf{I}_2 + \sum_p (\theta_p, \tau_p)(\theta_p, \tau_p)^T\right) \tag{3.43}$$

Step 10: Re-scale model parameters to equate the variance of ability to 1:

$$\begin{aligned}
\theta_p &\rightarrow \tfrac{\theta_p}{\sigma_\theta}, && \forall p \in [1:N] \\
\alpha_i &\rightarrow \alpha_i \sigma_\theta, && \forall i \in [1:n] \\
\mu_{\ln \alpha} &\rightarrow \mu_{\ln \alpha} + \ln \sigma_\theta \\
\boldsymbol{\Sigma}_{\mathcal{P}} &\rightarrow \begin{bmatrix} 1 & \rho_{\theta\tau} \\ \rho_{\theta\tau} & \sigma_\tau^2 \end{bmatrix}
\end{aligned} \tag{3.44}$$

Appendix B

Here, we briefly describe how the LM-statistics were computed in the simulation studies. When investigating the robustness of the LM-test, for the data sets simulated under the two misspecifications of the lower level models, the item time intensities and the item variances for the lognormal model in Equation 3.3 were estimated using a Gibbs Sampler and the person speed parameters were computed

as follows (van der Linden & Glas, 2010):

$$\hat{\tau}_p = (\sum_i \frac{1}{\hat{\sigma}_i^2}(\hat{\xi}_i - \ln t_{pi}))/\sum_i \frac{1}{\hat{\sigma}_i^2}. \tag{3.45}$$

RT residuals were calculated:

$$\hat{z}_{pi} = (\ln t_{pi} - (\hat{\xi}_i - \hat{\tau}_p))/\hat{\sigma}_i, \tag{3.46}$$

and the LM-test statistic was computed for each item:

$$LM_i = \left(\sum_p \frac{x_{pi}\hat{z}_{pi}}{\hat{\sigma}_i}\right)^2 / \sum_p \left(\frac{x_{pi}}{\hat{\sigma}_i^2} - \frac{\left(\frac{x_{pi}}{\hat{\sigma}_i^2}\right)^2}{\sum_i \frac{1}{\hat{\sigma}_i^2}}\right). \tag{3.47}$$

Under CI LM_i has a χ^2-distribution with one degree of freedom.

Since the LM-test was designed for situations where the item parameters are known (van der Linden & Glas, 2010), in the first and the third simulation studies the true values of $\boldsymbol{\xi}$ and $\boldsymbol{\sigma}^2$ were used instead of the estimates when computing the persons parameters (see Equation 3.45), the RT residuals (see Equation 3.46) and the LM-statistics (see Equation 3.47).

Appendix C

Table 3.6: Proportion of correctly detected violations of CI with three discrepancy measures (100 replications)

N	n	$\rho_{\theta\tau}$	Size	Type 1			Type 2			Type 3			Type 4			Type 5		
				p_{1i}	p_{2i}	p_{3i}	p_{1i}	p_{2i}	p_{3i}	p_{1i}	p_{2i}	p_{3i}	p_{1i}	p_{2i}	p_{3i}	p_{1i}	p_{2i}	p_{3i}
1000	20	.5	m	1	.00	.11	.81	1	1	.00	1	.01	.99	.14	.04	.58	.09	.25
500	20	.5	m	1	.01	.07	.43	.91	1	.00	.91	.02	.96	.10	.03	.28	.05	.01
2000	20	.5	m	1	.04	.33	1	1	1	.02	1	.04	1	.28	.11	.81	.30	.61
1000	**40**	.5	m	1	.05	.52	1	1	1	.01	1	.05	1	.37	.18	.96	.25	.67
1000	20	**.0**	m	1	.01	.14	.81	.01	1	.00	.97	.00	.99	.08	.08	.54	.04	.28
1000	20	.5	**s**	1	.00	.03	.11	.26	.83	.01	.58	.02	.48	.01	.00	.06	.02	.01
2000	20	.5	**s**	1	.02	.04	.27	.62	.99	.01	.77	.01	.74	.05	.00	.20	.02	.02

Chapter 4

Modeling conditional dependence between response time and accuracy

[1] **Abstract.** The assumption of conditional independence between response time and accuracy given speed and ability is commonly made in response time modeling. However, this assumption might be violated in some cases, meaning that the relationship between the response time and the response accuracy of the same item cannot be fully explained by the correlation between the overall speed and ability. We propose to explicitly model the residual dependence between time and accuracy by incorporating the effects of the residual response time on the intercept and the slope parameter of the IRT model for response accuracy. We present an empirical example of a violation of conditional independence from a low-stakes educational test and show that our new model reveals interesting phenomena about the dependence of the item properties on whether the response is relatively fast or slow. For more difficult items responding slowly is associated with a higher probability of a correct response, whereas for the easier items responding slower is associated with a lower probability of a correct response. Moreover, for many of the items slower responses were less informative for the ability because their discrimination parameters decrease with residual response time.

Keywords: conditional independence, hierarchical model, item response theory, residual dependence, response times.

[1]This chapter is conditionally accepted for publication in Psychometrika as Bolsinova, M., de Boeck, P. & Tijmstra, J. Modeling conditional dependence between response time and accuracy. Author contributions: B.M., deB.P. and T.J. designed the research, B.M. performed the research, B.M. wrote the paper, deB.P. and T.J provided feedback on the manuscript.

4.1 Introduction

Educational tests are often taken in a computerised form, which allows one to not only collect the students' responses, but also the response times. This can be useful since response times can be an important source of information about the students' performance (Luce, 1986; van der Linden, 2009). One of the most popular approaches for the joint modeling of item response accuracies and their response times in educational measurement is the hierarchical framework (van der Linden, 2007). In this framework the dependence between response time and accuracy of an item is taken to be fully explained by a correlation between a person's overall ability and overall speed, such that conditional on the latent variables speed and ability, for each item i the response time T_i and the response accuracy X_i are assumed to be independent.

The hierarchical framework has been successfully used in several applications in educational and psychological testing (van der Linden, 2008; van der Linden & Guo, 2008; Klein Entink, Kuhn, Hornke, & Fox, 2009; Goldhammer & Klein Entink, 2011; Loeys, Rossel, & Baten, 2011; Petscher, Mitchell, & Foorman, 2014; Scherer, Greiff, & Hautamäki, 2015). Although the hierarchical model assuming conditional independence is convenient from a statistical point of view and provides clear interpretations of the individual differences in speed and accuracy and the relations between them, in some cases the fundamental assumption of conditional independence is violated, implying that the higher-level dependencies between the speed and the ability parameters do not fully explain the dependence between the response time and the response accuracy (Partchev & De Boeck, 2012; Ranger & Ortner, 2012; Chen & De Boeck, 2014; Bolsinova & Maris, 2016). Conditional dependence between time and accuracy may for example arise from respondents varying their speed or using different strategies to solve the items throughout the test. In this paper we propose to explicitly model the residual dependence between time and accuracy within each item after the higher-level correlation between overall speed and ability has been taken into account.

In the hierarchical framework (van der Linden, 2007) the joint distribution of response accuracy and response time is modeled as a product of the marginal distributions of accuracy and time, which are obtained using standard IRT and response time models, respectively (e.g., a two-parameter logistic model for response accuracy and a log-normal model for response time). A more general way of modeling the joint distribution of response time and accuracy that does not require conditional independence is to decompose their joint distribution into a product of a marginal and a conditional distribution in one of two ways. One possibility is to have a standard IRT model for the marginal dsitribution of response accuracy (e.g., a two-parameter logistic model) and multiply it with the conditional distribution of response time given a response being correct or incorrect, as has been suggested by Bloxom (1985). Van der Linden and Glas (2010) have pur-

sued such an approach, but with the goal of developing a test for the assumption of conditional independence rather than obtaining a substantively interpretable joint model. A second possibility is to have a standard model for the marginal distribution of response time (e.g., a log-normal model) and multiply it with the conditional distribution of response accuracy given response time.

In this study we consider the latter approach: Letting the parameters of the response accuracy model depend on whether the response is relatively fast or slow. We consider this second approach, because this aims at improving the model for response accuracy, which is the model that is often most important for practical applications. This choice is in line with the idea that response accuracy could be affected when a respondent provides a response to a particular item that is faster or slower than would be expected based on that person's overall speed. Extending the model for response accuracy by incorporating response time allows one to study in more detail the impact that the relative speed of the response has on the response accuracy. Research by Partchev and De Boeck (2012) indicates that there likely are important differences in the response processes of fast versus slow responses. Working with the conditional distribution of response accuracy given response time makes it possible to study these differences.

We consider an extension of the two-parameter model for response accuracy, in which both the intercept (item easiness) and the slope (strength of relationship between the probability of a correct response and the measured ability) of the item characteristic curve (ICC) are dependent on whether the response is relatively fast or slow. Including response time in the model for response accuracy has a long tradition in response time modeling (Roskam, 1987; van Breukelen & Roskam, 1991; Verhest, Verstralen, & Jansen, 1997; van Breukelen, 2005; Wang and Hanson, 2005; Wang, 2006). One important aspect that differentiates our approach from these existing approaches is that not only the main effect of time on accuracy (i.e., effect on the intercept), but also the interaction effect between time and ability (i.e., effect on the slope) is included in the model for response accuracy.

It may be noted that one could also choose to model conditional dependence at the level of the joint distribution by specifying a bivariate distribution for response time and response accuracy that includes an item-level residual correlation, as is done by Ranger and Ortner (2012). However, this approach does not provide a direct translation of how the probability of a correct response changes as a function of response time, and corresponds to only allowing possible effects of response time on the intercept of the item response function, and not the slope. For these reasons, we propose to introduce time-dependent parameters in the response accuracy model which allows for more versatility and results in additional item parameters that have a straightforward interpretation.

The paper is organised as follows. In Section 4.2 we describe the specification of the hierarchical model for response time and accuracy which is extended in

this paper. In Section 4.3 we introduce a motivating empirical example which we will be using throughout the paper. We show that for this data set conditional independence between response time and accuracy assumed by the hierarchical model is violated. In Section 4.4 we propose a modification of the hierarchical model that explicitly models the effects of the relative speed of a response on the parameters of the ICC. Instead of dichotomising the response times into fast versus slow, as in the IRTree approach of Partchev and De Boeck (2012), we consider the relative speed of the response as a continuous measure that serves as a covariate for the parameters of the ICC. When using a continuous measure of speed no arbitrary dichotomisation of response time is required, which may have the advantage of avoiding a loss of information. The full model and its constrained versions are described and a Bayesian estimation method is proposed. In Section 4.5 we return to the empirical data set. Different models for the conditional dependence between time and accuracy are fitted to these data. The best model is chosen based on the DIC, and its goodness-of-fit is investigated with posterior predictive checks. Substantive interpretations are given to the estimated model parameters. In order to provide evidence for the stability of the conclusions drawn from the data set of interest, we present a small scale simulation study in Section 4.6 that investigates the parameter recovery when data are simulated using the estimates from the empirical data set as true values. The paper concludes with a discussion.

4.2 Specification of the hierarchical model

Let us by $\mathbf{X}$ denote an $N \times n$ matrix of responses of N persons to n items taking values of 1 if the response is correct and 0 otherwise, and by $\mathbf{T}$ an $N \times n$ matrix of the corresponding response times. The hierarchical model for response times and accuracy is (van der Linden, 2007):

$$f(\mathbf{X}, \mathbf{T}) = \prod_p \prod_i f(t_{pi} \mid \tau_p, \boldsymbol{\gamma}_i) f(x_{pi} \mid \theta_p, \boldsymbol{\delta}_i), \tag{4.1}$$

where t_{pi} and x_{pi} are the response time and accuracy of person p on item i, τ_p and θ_p are the speed and the ability parameters of person p, and $\boldsymbol{\gamma}_i$ and $\boldsymbol{\delta}_i$ are the vectors of item parameters of item i related to time and accuracy, respectively. At the lower level of the hierarchical model both the model for response accuracy and the model of response time need to be specified. At the higher level the models for the relationship between the person parameters (θ_p and τ_p) and for the relationship between the item parameters ($\boldsymbol{\delta}_i$ and $\boldsymbol{\gamma}_i$) need to be specified. Below we describe the full specification of the hierarchical model considered in this study.

The model for the response accuracy is the following (Birnbaum, 1968)

$$\Pr(X_{pi} = 1 \,|\, \theta_p) = \frac{\exp(\alpha_i \theta_p + \beta_i)}{1 + \exp(\alpha_i \theta_p + \beta_i)}, \tag{4.2}$$

that is, $\boldsymbol{\delta}_i = \{\alpha_i, \beta_i\}$, where $\alpha_i > 0$ and β_i are the slope and the intercept of the ICC of item i which relates the ability of the person to the probability of a correct response to the item. The slope α_i reflects the discriminative power of the item, since it specifies the strength of the relationship between the latent ability and the response to the item, and the intercept β_i reflects item easiness.

The model for the response times is

$$T_{pi} \,|\, \tau_p \sim \ln \mathcal{N}(\xi_i - \tau_p, \sigma_i^2), \tag{4.3}$$

that is $\boldsymbol{\gamma}_i = \{\xi_i, \sigma_i^2\}$. The mean of the logarithm of response time of person p to item i depends on the item time intensity (ξ_i) and the person speed, and the variance parameter depends on the item. The residual variance of response time σ_i^2 can be interpreted as the inverse of item time discrimination (van der Linden, 2006), that is, the larger $\frac{1}{\sigma_i^2}$ is, the larger the proportion of variance of response time explained by the variation of speed across persons is.

At the higher level, a multivariate normal distribution for the item parameters $\{\xi_i, \ln(\sigma_i^2), \ln(\alpha_i), \beta_i\}$ and a multivariate normal distribution for the persons parameters $\{\theta_p, \tau_p\}$ with the identifiability restrictions of $\mu_\theta = \mu_\tau = 0$ and $\sigma_\theta^2 = 1$ are assumed.

4.3 Motivating example: violation of conditional independence

We present an analysis of a data set of the Major Field Test for the Bachelor's Degree in Business[2] which is a low-stakes educational test. This test is used to assess mastery of concepts, principles and knowledge of graduating bachelor students in business. The test is not used for making individual-level decisions, but for evaluating educational programmes. The test consists of 120 multiple-choice items separated into two part. Only the first part of the test was analysed in this study. The time limit for this part was one hour, while the average time used by the respondents was 42 minutes. Some items in the test are based on diagrams, charts and data tables. The test items cover a wide range of difficulties and are aimed at the evaluation of both depth and breadth of business knowledge. From the original sample with the responses of 1000 persons to 60 items, 11 items

[2]Source: Derived from data provided by Educational Testing Service Copyright ©201x ETS. www.ets.org. The opinions set forth in this publication are those of the author(s) and not ETS.

were removed due to low item rest correlations ($< .1$). The responses for which response times were equal to 0 were treated as missing values.

To test the assumption of conditional independence between response times and accuracy given speed and ability we used the Lagrange Multiplier test of van der Linden and Glas (2010). In this test for each item the hierarchical model assuming conditional independence (van der Linden, 2007) is tested against a model which allows for differences in the expected log-response time for correct and incorrect responses by including an extra item parameter in the model for the response times:

$$t_{pi} \sim \ln \mathcal{N}\left(\xi_i + \lambda_i(1 - x_{pi}) - \tau_p, \sigma_i^2\right). \tag{4.4}$$

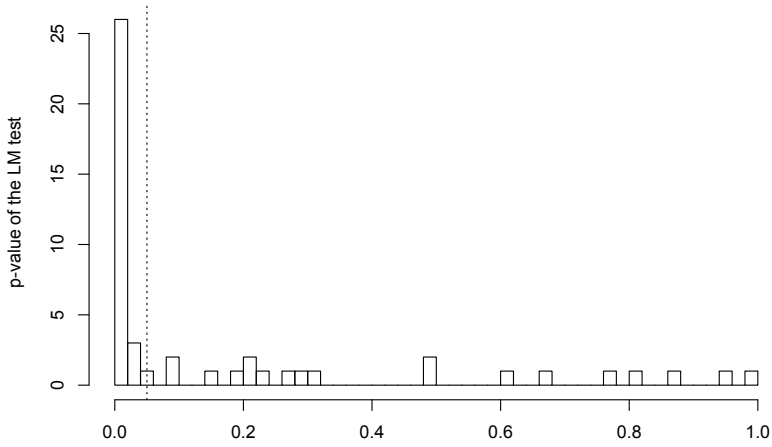


Figure 4.1: Distribution of the p-values of the Lagrange Multiplier test for conditional independence between response time and accuracy. Most of the p-values are below .05, indicating that conditional independence is violated.

Figure 4.1 shows the distribution of the p-values for the item-level conditional independence test. For more than half of the items, conditional independence is violated ($\alpha = .05$ for each test). These results indicate that the assumption of conditional independence cannot be maintained for these data. However, as has been demonstrated in simulation studies (Bolsinova & Tijmstra, 2016), the Lagrange multiplier test of van der Linden and Glas (2010) also may pick up violations of conditional independence that are of a different type than what is specified in Equation 4.4. Therefore, this test does not yet tell us in what way exactly the hierarchical model for response time and accuracy is violated.

To investigate in which way conditional independence is violated we performed two posterior predictive checks (Meng, 1994; Gelman et al., 1996; Sinharay et al.,

2006) that focus on differences in the behaviour of the items with respect to response accuracy between slow and fast responses, following the approach proposed by Bolsinova and Tijmstra (2016). Here, we defined slow and fast responses by a median split by defining a transformed time variable:

$$t^*_{pi} = \begin{cases} 1 \text{ if } t_{pi} \geq t_{med,i} \\ 0 \text{ if } t_{pi} < t_{med,i} \end{cases}, \tag{4.5}$$

where $t_{med,i}$ is the median response time for item i. Our goal is to investigate whether observed differences between the slow and the fast responses to the difficulty of the items and their discriminatory power are unlikely to be observed under the hierarchical model assuming conditional independence.

When posterior predictive checks are implemented the measures of interest have to be repeatedly computed for replicated data sets. Therefore, for reasons of computational convenience we decided not to estimate IRT difficulty and discrimination parameters separately for the slow and for the fast responses, but to compute simple classical test theory statistics which can be viewed as proxies for the difficulty and discrimination, namely the proportion of correct responses and the item-rest correlation. For each item two discrepancy measures were computed: the difference between the proportion of correct responses to the item among slow responses and among fast responses,

$$D_{1i} = \frac{\sum_p x_{pi} t^*_{pi}}{\sum_p t^*_{pi}} - \frac{\sum_p x_{pi}(1 - t^*_{pi})}{\sum_p (1 - t^*_{pi})}, \tag{4.6}$$

and the difference between item-rest correlations of the item among slow and fast responses,

$$D_{2i} = Cor\left(\mathbf{x}_{i,slow}, \mathbf{x}^{(i)}_{+,slow}\right) - Cor\left(\mathbf{x}_{i,fast}, \mathbf{x}^{(i)}_{+,fast}\right), \tag{4.7}$$

where $\mathbf{x}_{i,slow}$ and $\mathbf{x}_{i,fast}$ are vectors of responses of all persons such that $t^*_{pi} = 1$ or $t^*_{pi} = 0$, respectively; $\mathbf{x}^{(i)}_{+,slow}$ and $\mathbf{x}^{(i)}_{+,fast}$ are vectors of the numbers of correct responses to all the items excluding item i of all persons such that $t^*_{pi} = 1$ or $t^*_{pi} = 0$, respectively.

To assess whether the observed values D_{1i} and D_{2i} are plausible under the hierarchical model, they can be compared to values drawn from the posterior predictive distribution of these measures given the data and the model. This posterior predictive distribution can be obtained using draws from the posterior distribution of the model parameters. The conditional independence model was estimated using a Gibbs Sampler. The prior distributions and the sampling procedure were specified following the specification of Bolsinova and Tijmstra (2016), which means that independent vague priors were used for the hyper-parameters (mean vector

and covariance matrix of the item parameters, variance of speed, and correlation between speed and accuracy). Two independent chains with 10000 iterations each were used. The first 5000 iterations in each chain were treated as burn-in and were discarded. To reduce the risk of autocorrelation every 5-th sample after the burn-in was used, resulting in 1000 samples from each chain. Using each of these 2000 samples a new replicated data set was simulated according to the hierarchical model: $\mathbf{X}_{rep}^{(g)}$, $\mathbf{T}_{rep}^{(g)}$, where superscript g denotes g-th sample from the posterior distribution. In each replicated data set $D_{1i}^{(g)}$ and $D_{2i}^{(g)}$ were computed for each item. For each item two posterior predictive p-values were computed:

$$p_{1i} = \frac{\sum_g \mathcal{I}(D_{1i} > D_{1i}^{(g)})}{2000}, \tag{4.8}$$

$$p_{2i} = \frac{\sum_g \mathcal{I}(D_{2i} > D_{2i}^{(g)})}{2000}. \tag{4.9}$$

If p_{1i} is close to 0 or close to 1, it means that the difference between the proportion of correct responses among the slow and the fast responses observed in the empirical data are not likely under the model. Similarly, if p_{2i} is close to 0 or is close to 1, the observed difference between the item-rest correlations is unlikely under the model. Figure 4.2 shows the histograms of the posterior predictive p-values of the items for the model assuming conditional independence. The large number of extreme p-values indicate that the model does not capture an important aspect of the data, namely that the items behave differently for the slow and the fast responses. We will modify the model such that it will better explain the behaviour of the items.

Next, we investigated whether the observed deviation from conditional independence in the data can be explained by the extended hierarchical model in Equation 4.4, in which the model for response time is extended, while the same model for response accuracy (see Equation 4.2) is used as in the hierarchical conditional independence model. In this model, conditional independence as specified in Equation 4.1 is violated, and this violation is taken to be fully explained using the additional parameter λ_i. This way of modelling conditional dependence is in line with the first approach described in the Introduction, that is, the distribution of response time is modeled conditional on whether the response is correct or not.

To determine whether this extension of the hierarchical model is able to fully explain the observed deviation from conditional independence in the data, we analysed to what extent the observed D_{1i} and D_{2i} would be plausible under this model. Figure 4.3 shows the histograms of the posterior predictive p-values for the extended hierarchical model. The situation has improved with respect to the difference in the proportion of correct responses, but not with respect to the difference in the item-rest correlations. The finding that there was only an improvement

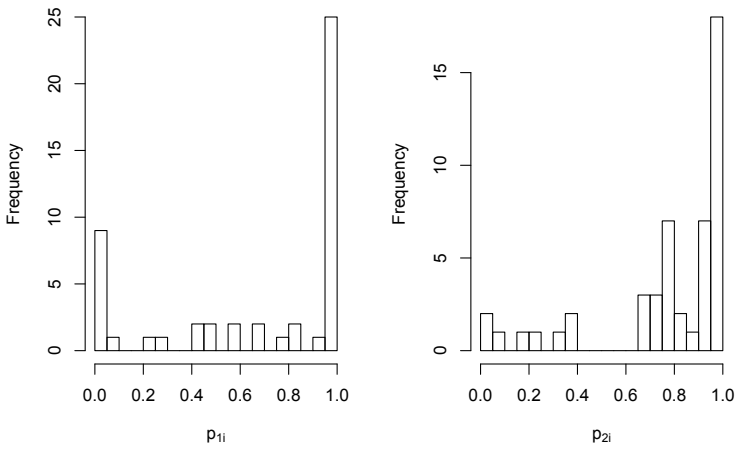


Figure 4.2: Posterior predictive p-values for the hierarchical model assuming conditional independence: a) difference between the proportion of correct responses to an item for slow and fast responses; b) difference between the item rest correlations for slow and fast responses

with respect to the differences in the proportion of correct responses could be expected, since the extended hierarchical model only allows for a shift in the mean conditional response time for correct an incorrect responses. That is, such a shift cannot account for any observed variation in the discriminative power of an item as a function of response time, as reflected in differences in the item-rest correlations.

Based on these results, we suggest extending the hierarchical model such that it takes into account that both the proportion of correct responses to the items and their discriminative power might change as a function of response time. In the next section we propose to consider extending the model for response accuracy using residual log response time as a covariate for the parameters of the IRT model (both the intercept and the slope).

4.4 Residual log response time as a covariate for the parameters of the ICC

4.4.1 Model specification

When considering the possible effects of having relatively fast or slow responses, we want to disentangle the particular response time from the overall speed of a person and the overall time intensity of an item. That is, it is not t_{pi} in isolation

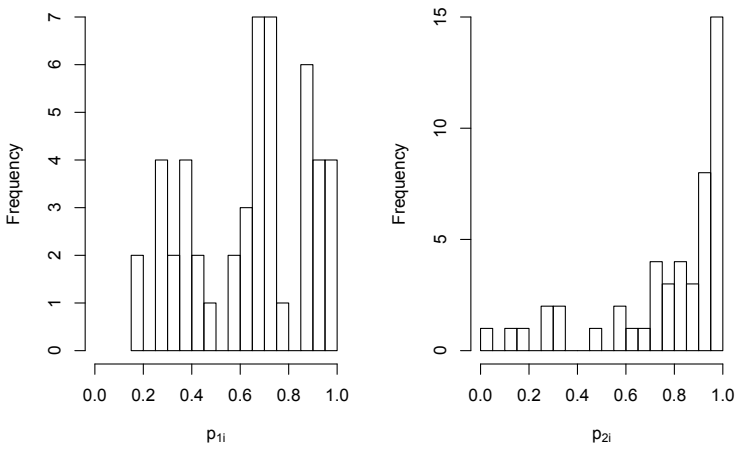


Figure 4.3: Posterior predictive p-values for the model with an extra parameter for the difference in response times distributions of the correct and incorrect responses: a) difference between the proportion of correct responses to an item for slow and fast responses; b) difference between the item rest correlations for slow and fast responses

that informs us whether a response is relatively slow or fast, but rather the difference between t_{pi} and the expected response time for person p on item i. Two identical response times might differ with regard to whether they are fast or slow, depending on the speed of the corresponding persons and the time intensity of the items. Because of this, it may be reasonable to use a standardised residual of the log response time (derived from Equation 4.3) to capture the extent to which a response should be considered to be fast or slow. Let us denote the standardised residual log response time of person p to item i by z_{pi}:

$$z_{pi} = \frac{\ln t_{pi} - (\xi_i - \tau_p)}{\sigma_i}. \tag{4.10}$$

If $z_{pi} > 0$, it means that the response of person p to item i is relatively slow, while if $z_{pi} < 0$ it is relatively fast. The residuals are standardised in order to make the regression coefficients specifying the effects of residual time on accuracy comparable across items by taking into account the differences in σ_i. If conditional independence between response times and accuracy given ability and speed holds, then the probability of a correct response to item i does not depend on whether the response is relatively slow or relatively fast, given ability and speed.

We suggest to use a time-related covariate both for the intercept and for the

slope of the ICC, such that it can also capture the difference between the discriminative power of the items for the slow and fast responses as observed in the empirical data. Let the slope and the intercept in Equation 4.2 depend on the standardised residual of the log response time:

$$\begin{array}{rcll} \alpha_i & = & \alpha_{0i}\alpha_{1i}^{z_{pi}}, & \text{or equivalently} \\ \ln(\alpha_i) & = & \ln(\alpha_{0i}) + \ln(\alpha_{1i})z_{pi}, & \text{and} \end{array} \tag{4.11}$$

$$\beta_i = \beta_{0i} + \beta_{1i}z_{pi}. \tag{4.12}$$

Since the slope parameter in the two-parameter logistic model is restricted to positive values, a linear model for $\ln(\alpha_{pi})$ rather than for α_i is used. Another reason for using a multiplicative effect for the slope instead of the linear effect is that the slope parameter is itself a multiplicative parameter. The parameters α_{0i} and β_{0i} are the baseline slope and the baseline intercept of the item response function of item i, which refer to the responses x_{pi} which are answered as fast as it is expected for person p on item i (i.e., $z_{pi} = 0$). The parameters α_{1i} and β_{1i} are the effects of z_{pi} on the slope and the intercept of the ICC. If $\alpha_{1i} = 1$ or $\beta_{1i} = 0$ it means that there is no effect of residual log response time of the slope or on the intercept, respectively.

If one would assume that persons keep a constant speed across items (which is usually assumed within the hierarchical modeling framework), then β_{1i} would be closely related to the conditional accuracy function (van Breukelen, 2005), since the residual response time does not reflect a change in effective speed. However, if the effective speed does vary across items (e.g., faster at the end of the test because of a strict time limit), then this would be partly reflected in the residual response time (i.e., part of the residual is the deviation of the effective speed on item i from the average effective speed τ). In that case the effect $\beta_{1i} > 0$ relates to the speed-accuracy trade-off (Luce, 1986), that is, investing less time in solving the item decreases the probability of a correct response.

The full model allows the effects of the covariate to vary across the items. However, one might assume that the effect of responding relatively fast or slow is the same for all the items, choosing one of the constrained models: equal α_{1i} and β_{1i}, equal α_{1i} but varying β_{1i}, or equal β_{1i} but varying α_{1i}. It may be noted that if one chooses to model only the effects on the intercept (i.e., varying β_{1i} and $\alpha_1 = 1$ for all items) then the model is similar in structure to the model of Ranger and Ortner (2012) with the exception that we consider a logistic model for response accuracy instead of a normal ogive model.

As in the hierarchical model that assumes conditional independence, we need to specify the joint distribution for the item parameters and the joint distribution of the person parameters. The dependence between the item parameters of individual items is modeled by a multivariate normal distribution for the vector $\{\xi_i, \ln(\sigma_i^2), \ln(\alpha_{0i}), \ln(\alpha_{1i}), \beta_{0i}, \beta_{1i}\}$ with a mean vector $\boldsymbol{\mu}_{\mathcal{I}}$ and a covariance matrix

$\boldsymbol{\Sigma}_{\mathcal{I}}$. Logarithmic transformation is used for the parameters which are restricted to positive numbers. A multivariate normal distribution is used to model the dependence between speed and ability. The mean vector of the person population distribution is constrained to zero and the variance of θ is constrained to one to ensure the identification of the model, similar to the hierarchical conditional independence model. The two person population parameters of interest are the correlation between speed and ability (denoted by $\rho_{\theta\tau}$) and the variance of speed (denoted by σ_τ^2).

4.4.2 Estimation

For the estimation of the model we developed a Gibbs Sampler (Geman & Geman, 1984; Casella & George, 1992) implemented in `R` programming language (R Core Team, 2014) to obtain samples from the joint distribution of the model parameters:

$$f\left(\boldsymbol{\alpha_0}, \boldsymbol{\alpha}_1, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\sigma}^2, \boldsymbol{\tau}, \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}, \sigma_\tau^2, \rho_{\theta\tau} \,|\, \mathbf{X}, \mathbf{T}\right), \tag{4.13}$$

which includes both the parameters of the individual persons and items and the hyper-parameters of the person population distribution and the item population distribution.

Although the variance of θ is constrained to 1, to improve the convergence of the model at each iteration of the Gibbs Sampler the full covariance matrix $\boldsymbol{\Sigma}_{\mathcal{P}}$ is sampled and at the end of each iteration all parameters are transformed to fit the scale defined by $\sigma_\theta^2 = 1$ (see Appendix for details). Prior distributions for the item and the person hyper-parameters have to be specified. We choose vague prior distributions: normal distributions with a zero mean and a large variance (100) for the means of the item parameters, half t-distributions with $\nu = 2$ degrees of freedom and a scale parameter $A = 2$ for the standard deviations of the items parameters, marginally uniform joint distribution for the correlations between the item parameters (Huang & Wand, 2013) and an inverse-Wishart distribution with 4 degrees of freedom and identity matrix $\mathbf{I}_2$ as the scale parameter for $\boldsymbol{\Sigma}_{\mathcal{P}}$ (Hoff, 2009). Results are not sensitive to the specification of the prior scale parameter, because the posterior distribution is dominated by the data when the sample size is large (Hoff, 2009, p.110). Prior distributions are assumed to be independent.

The estimation algorithm includes Metropolis-Hastings steps (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) and a modification of the composition algorithm by Marsman et al. (2015). In the Gibbs Sampler, the model parameters are subsequently sampled from their full conditional posterior distributions given the current values of all other parameter. The details about how to sample from each of the conditional posteriors are described in the Appendix.

For model comparison purposes, modifications of the algorithm have also been developed to estimate the constrained models (equal α_{1i} and β_{1i}, equal α_{1i} but

varying β_{1i}, or equal β_{1i} but varying α_{1i}), models with different time related covariates (one might be interested in the effect of t_{pi} and $\ln t_{pi}$ on the IRT parameters instead of z_{pi}), the hierarchical model assuming conditional independence, and the modified hierarchical model with an extra parameter for the difference in the location parameters of the distribution of the response times given a correct and an incorrect response (see Equation 4.4).

4.4.3 Model selection and Goodness-of-fit

To select the best model the deviance information criterion [DIC] can be used, because it adequately takes the complexity of hierarchical models into account (Spiegelhalter, Best, Carlin, & van der Linden, 2002). The DIC can be computed using the output of the Gibbs Sampler. First, at each iteration (after discarding the burn-in and thinning) the deviance is computed. For example for the full model the deviance is:

$$D^{(g)} = -2 \ln \left(f \left(\mathbf{X}, \mathbf{T} \,|\, \boldsymbol{\alpha}_0^{(g)}, \boldsymbol{\alpha}_1^{(g)}, \boldsymbol{\beta}_0^{(g)}, \boldsymbol{\beta}_1^{(g)}, \boldsymbol{\theta}^{(g)}, \boldsymbol{\xi}^{(g)}, \boldsymbol{\sigma}^{2(g)}, \boldsymbol{\tau}^{(g)} \right) \right), \tag{4.14}$$

where the superscript $^{(g)}$ denotes the g-th iteration of the Gibbs Sampler; this expression does not include the hyper-parameters σ_τ^2, $\rho_{\theta\tau}$, $\boldsymbol{\mu}_\mathcal{I}$ and $\boldsymbol{\Sigma}_\mathcal{I}$ because the distribution of the data is independent of the hyper-parameters given the individual item and person parameters. Second, the deviance is computed for the posterior means of the model parameters:

$$\hat{D} = -2 \ln \left(f \left(\mathbf{X}, \mathbf{T} \,|\, \hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\sigma}}^2, \hat{\boldsymbol{\tau}} \right) \right). \tag{4.15}$$

The DIC is equal to:

$$DIC = 2 \frac{\sum_g D^{(g)}}{G} - \hat{D}, \tag{4.16}$$

where G is the total number of iterations which are taken into account when computing the DIC.

To evaluate the absolute fit of the best fitting model, posterior predictive checks for a global discrepancy measure between the data and the model can be used, for example the log-likelihood of the data under the model. For each g-th sample from the posterior distribution of the model parameters given the observed data, a replicated data set $(\mathbf{X}_{rep}^{(g)}, \mathbf{T}_{rep}^{(g)})$ is simulated under the model and the log-likelihood is computed both for the observed data and the replicated data:

$$LL_{obs}^{(g)} = \ln \left(f \left(\mathbf{X}, \mathbf{T} | \boldsymbol{\alpha}_0^{(g)}, \boldsymbol{\alpha}_1^{(g)}, \boldsymbol{\beta}_0^{(g)}, \boldsymbol{\beta}_1^{(g)}, \boldsymbol{\theta}^{(g)}, \boldsymbol{\xi}^{(g)}, \boldsymbol{\sigma}^{2(g)}, \boldsymbol{\tau}^{(g)} \right) \right), \tag{4.17}$$

$$LL_{rep}^{(g)} = \ln \left(f \left(\mathbf{X}_{rep}^{(g)}, \mathbf{T}_{rep}^{(g)} | \boldsymbol{\alpha}_0^{(g)}, \boldsymbol{\alpha}_1^{(g)}, \boldsymbol{\beta}_0^{(g)}, \boldsymbol{\beta}_1^{(g)}, \boldsymbol{\theta}^{(g)}, \boldsymbol{\xi}^{(g)}, \boldsymbol{\sigma}^{2(g)}, \boldsymbol{\tau}^{(g)} \right) \right). \tag{4.18}$$

The posterior predictive p-value is the proportion of samples in which the observed data is less likely under the model than the replicated data. If the posterior predictive p-value is small, then the data are unlikely under the model.

The goodness-of-fit can be further evaluated using posterior predictive checks based on D_{1i} and D_{2i} statistics (see Equations 4.6 and 4.7).

4.5 Results

4.5.1 Fitted models

Eight different models for response time and accuracy were fitted to the data set of interest. First, the hierarchical model assuming conditional independence was estimated. Second, the modification of the hierarchical model with an extra parameter for the difference between the log-normal distributions of the response times given a correct and an incorrect response (see Equation 4.4) was fitted. Third, four models with residual log-response time (z_{pi}) as a covariate for the parameters of the ICC were estimated: the full model and its three constrained versions (equal α_{1i} and β_{1i}, equal α_{1i} but varying β_{1i}, equal β_{1i} but varying α_{1i}). Finally, two models with alternative time related covariates (t_{pi} and $\ln t_{pi}$) for the parameters of the ICC were fitted.

4.5.2 Convergence

Convergence was assessed using $\hat{R}$-statistic (Gelman & Rubin, 1992) for all the hyper-parameters individually and overall with the multivariate scale reduction factor (Brooks & Gelman, 1998). For all eights fitted models all multivariate $\hat{R}$ and the multivariate scale reduction factor were smaller than 1.1, indicating that convergence was not an issue.

4.5.3 Model selection

The values of the DIC of the different models are presented in Table 4.1. As expected based on the results of the test for conditional independence, the hierarchical model assuming conditional independence fits worse than the models taking conditional dependence between time and accuracy into account. When models for response accuracy are considered that include z_{pi} as a covariate for the item slope and intercept, allowing both these effects to vary across items improves the model, as evidenced by the fact that the full model has the lowest DIC, while the model with fixed effects has the highest DIC of the four models. It can also be observed that the full model outperforms the extension of the hierarchical model that includes a shift parameter (λ_i) for the model for response time. Finally, the residual log-response time is a better predictor of the parameters of the ICC than

Table 4.1: DIC of the fitted models

Model		DIC
CI model		466624.5
Model with extra λ_i		465498.5
z_{pi} as a covariate	equal α_1 and β_1	466280.4
	equal α_1	465550.5
	equal β_1	466100.3
	full model	465452.7
$\ln(t_{pi})$ as a covariate	full model	465605.9
t_{pi} as a covariate	full model	465853.2

the response time or the log-response time, as can be seen from the comparison of the three full models with different specifications of the time related covariates. Since the full model with z_{pi} as a covariate is the best performing model, this is the model that will be the focus in the remaining of the paper.

4.5.4 Posterior predictive checks

In the previous subsection we concluded that the full model with z_{pi} as a covariate fits the best of the fitted models. Now, we will further investigate its goodness-of-fit using posterior predictive checks. First, we performed a posterior predictive check for the global discrepancy measure. The posterior predictive p-value is equal to .35 (i.e., the proportion of iterations in which the log-likelihood of the observed data was lower than the log-likelihood of the data replicated under the model), which means that the observed data is not much more unlikely under the model than the data simulated under the model.

Second, we performed the same posterior predictive check as for the model assuming conditional independence in Section 4.3. Figure 4.4 shows the histogram of the posterior predictive p-values for the difference between the proportion of correct responses to the items among the slow and the fast responses (see Equation 4.6) and for the difference between the item-rest correlations among the slow and the fast responses (see Equation 4.7). Neither of the two measures resulted in a disproportionate amount of extreme posterior predictive p-values, which indicates that the model adequately captures these aspects of the data. These results are in line with what could be expected, since the posterior predictive checks focus on exactly the kind of dependencies that are meant to be captured by the added parameters β_{1i} and α_{1i}.

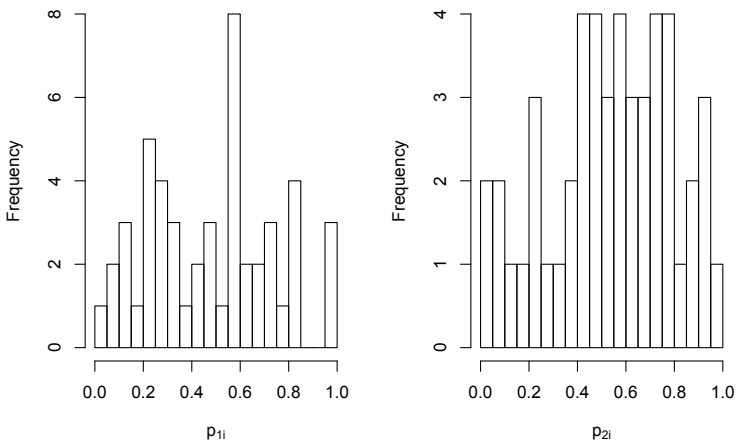


Figure 4.4: Posterior predictive p-values for the full model with residual response time as a covariate for item parameters: a) difference between the proportion of correct responses to an item for slow and fast responses; b) difference between the item rest correlations for slow and fast responses.

4.5.5 Effect of residual time on the ICC

Figure 4.5 displays the estimates of the effect of z_{pi} on the intercept and the slope of the ICC. For many of the items the credible intervals for these effects exclude 0 and 1, respectively, which indicates that the residual log response time does have an effect on the behaviour of the items. The estimates of α_{1i} differ across items. However, for most of them, these estimates are below 1, which means that these items discriminate worse if the response is slower. The effect of z_{pi} on the intercept (β_{1i}) is more variable across items compared to the effects on the slope (α_{1i}). For most of the items the effects on the intercept are negative, that is, the probability of correct responses among the relatively slow responses ($z_{pi} > 0$) was lower than among the relatively fast responses ($z_{pi} < 0$). For some of the items β_{1i} are positive, that is, the easiness of the item is higher for the relatively slow responses.

To zoom in on the differences between relatively fast responses ($z_{pi} = -1$) and relatively slow responses ($z_{pi} = 1$), we present the scatterplots of the predicted slopes and intercept of the items given two values of z_{pi} (see Figure 4.6). The predicted values for the intercepts (see Figure 4.6a) lie both above and below the diagonal, meaning that for some items the probability of a correct response is higher for slow responses and for other items the probability of a correct response is higher for fast responses. The predicted values of the slopes (see Figure 4.6b) lie

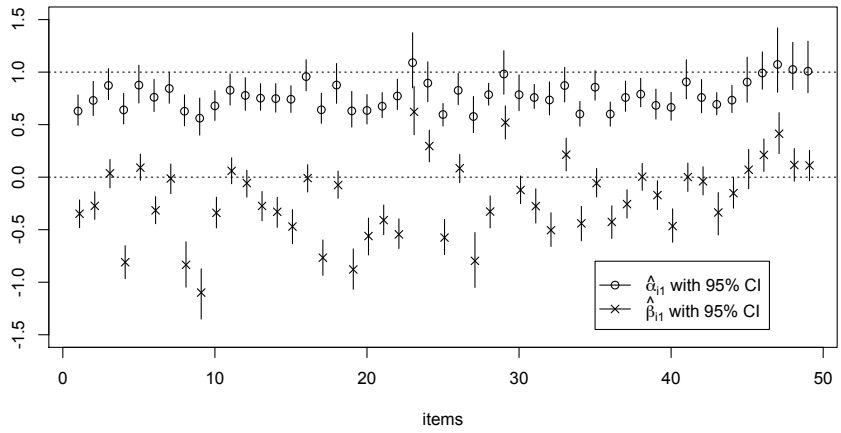
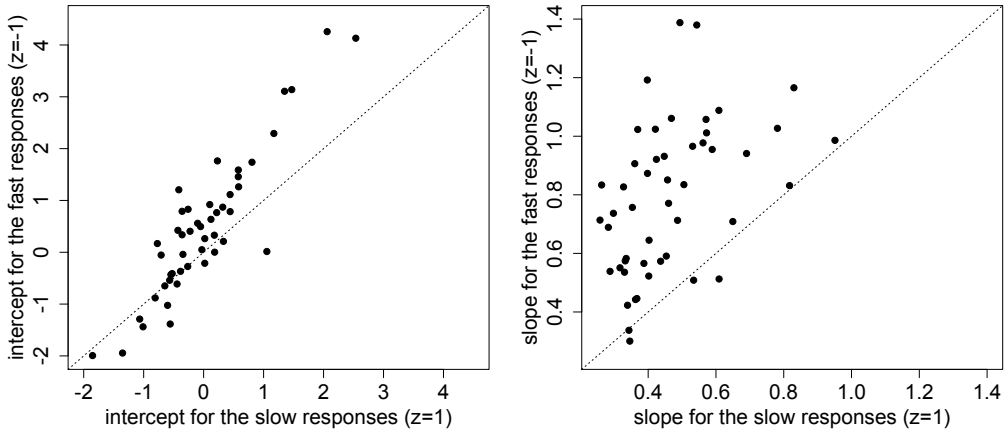


Figure 4.5: Estimated effects of residual response time on the slope and the intercept of the ICC

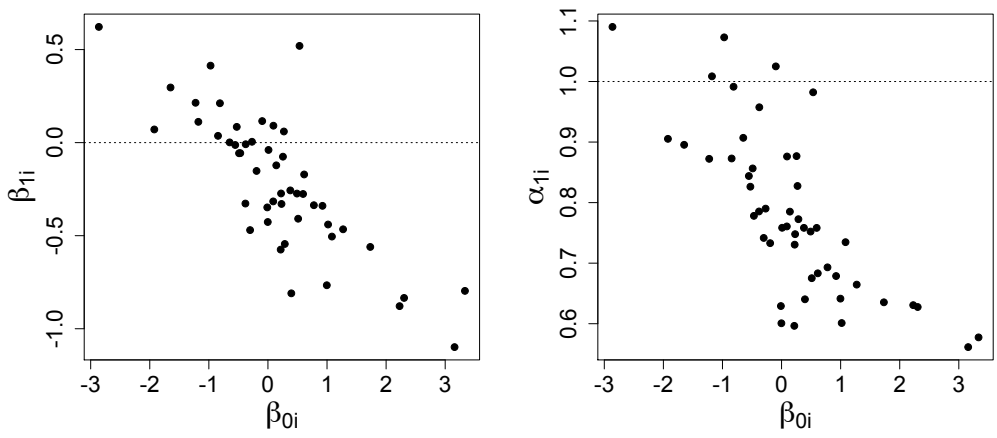
mainly above the diagonal, meaning that for these items the relationship between item response and the ability θ is stronger for fast responses.

Table 4.2 presents the posterior means and the 95% credible intervals for the means and the variances of the item parameters, and the correlations between them. On average the items have a low baseline discrimination (-0.57 on the log scale, corresponding to a baseline discrimination parameter of 0.57), and are relatively easy (the mean of the baseline intercept is 0.17). The effects of residual log response time on the intercept and on the logarithm of the slope are on average negative (-0.21 and -0.27, respectively) but the variance of the effect is larger for the intercepts than for the slopes (0.15 and 0.04, respectively).



a) b)

Figure 4.6: Predicted intercepts (a)) and slopes (b)) of the ICC given a slow response ($z_{pi} = 1$) on the x-axis and given a fast response ($z_{pi} = -1$) on the y-axis



a) b)

Figure 4.7: The effects of the residual log-response time on the intercept (a) and on the slope (b) of the ICC on the y-axis against the baseline intercept of the ICC on the x-axis.

Table 4.2: Between-item variances of the item parameters (on the diagonal), correlations between the item parameters (off-diagonal), and the mean vector of the item parameters

	ξ_i	$\ln(\sigma_i^2)$	$\ln(\alpha_{0i})$	$\ln(\alpha_{1i})$	β_{0i}	β_{1i}
ξ_i	0.20					
	[0.13,0.30]					
$\ln(\sigma_i^2)$	.40	0.12				
	[.14,.61]	[0.08,0.18]				
$\ln(\alpha_{0i})$	.11	- .05	0.12			
	[-.20,.41]	[-.34,.24]	[0.07,0.20]			
$\ln(\alpha_{1i})$	.44	.33	-.05	0.04		
	[.14,.69]	[.02,.59]	[-.43,.34]	[0.02,0.06]		
β_{0i}	-.44	-.29	.13	-.62	1.39	
	[-.64,-.20]	[-.53,-.02]	[-.21,.43]	[-.83,-.34]	[0.93,2.09]	
β_{1i}	.52	.32	-.10	.73	-.75	0.15
	[.30,.70]	[.05,.55]	[-.42,.23]	[.48,.90]	[-.85,-.60]	[0.10,0.22]
$\boldsymbol{\mu}_{\mathcal{I}}$	3.50	-1.49	-0.57	-0.27	0.17	-0.21
	[3.37,3.63]	[-1.59,-1.39]	[-0.69,-0.45]	[-0.34,-0.2]	[-0.15,0.51]	[-0.33,-0.11]

The baseline intercept of the ICC is negatively correlated with the effects on the intercept (-.75) and on the log of the slope (-.62). Figure 4.7 shows the scatterplots of the effects of z_{pi} on the intercept (a) and on the slope (b) of the ICC against the baseline intercept of the ICC. For very difficult items the effect of being slow is positive and for easier items the effect of being slow is more and more negative. In other words, for very difficult items being slow increases the probability of a correct response, whereas for very easy items being slow decreases the probability of a correct response. Moreover, slow responses are less informative (have lower discrimination) than the fast responses for the easy items, and are either more informative or equally informative as the fast responses for the difficult items.

As can be observed in Table 4.2, the effect of z_{pi} on the intercept and the effect on the log of the slope are strongly correlated (.73). Part of this correlation is explained by considering the baseline intercept, which is negatively correlated with both effects. However, after conditioning on the baseline intercept a positive correlation of .47 remains. This can be taken to indicate that items differ in the extent to which differences between fast and slow responses are present. That is, some items show both a strong effect on the slope and the intercept, whereas for other items both effects are weaker, indicating that there may not be any large differences between fast and slow responses for those items.

Item time intensity is negatively correlated with item baseline intercept ($-.44$), that is more difficult items require more time. This negative correlation between item baseline intercept and time intensity is in line with expectations of van der Linden (2007). Furthermore, time intensity is positively correlated with the effects of residual time on the item intercept (.52) and the item slope (.44). The first of these two correlations means that in our example spending relatively more time on the time intensive items increases the probability of a correct response while it decreases the probability of a correct response on items that do not require a lot of time. The second correlation implies that in this test time intensive items become more informative if answered relatively slowly, whereas items with low time intensity discriminate better if they are answered fast relative to what is expected for the combination of the person and the item.

4.5.6 Sensitivity analysis: robustness to outliers

For the original analysis none of the response time outliers were removed. However, it is important to check if the presence of outliers with respect to response time affects the estimates of the model parameters. To do that, we fitted the full model with z_{pi} as a covariate to the data set without the responses that were considered to be possible outliers. Responses with the item-wise z-scores of the log-response times below the 0.1-th quantile or above the 99.9-th quantile of the standard normal distribution were identified as outliers, resulting in the removal of 514 responses out of the total of 49,000 responses.

Table 4.3: Difference between the estimates of the hyper-parameters of the items after the removal of the outliers compared to the original estimates

	ξ_i	$\ln(\sigma_i^2)$	$\ln(\alpha_{0i})$	$\ln(\alpha_{1i})$	β_{0i}	β_{1i}
ξ_i	0.01					
$\ln(\sigma_i^2)$	-0.10	-0.03				
$\ln(\alpha_{0i})$	-0.02	0.04	0.00			
$\ln(\alpha_{1i})$	-0.06	-0.10	-0.01	0.00		
β_{0i}	0.00	0.02	-0.02	0.03	0.05	
β_{1i}	0.03	-0.05	0.01	0.03	-0.02	0.01
$\boldsymbol{\mu}_{\mathcal{I}}$	0.02	-0.13	0.01	0.00	0.03	-0.01

Removing the outliers resulted in the decrease of standard deviation of speed from 0.33 [0.31,0.34] in the original data set to 0.28 [0.27,0.29] in the data set without the outliers, and in the decrease of the correlation between speed and accuracy from -.09 [-.16,-.02] to -.02 [-.09,.05]. These effects of the removal of the outliers are not very influential for the overall conclusions, since σ_τ is not the primary parameter of interest, and $\rho_{\theta\tau}$ was already too low to assign any substantive relevance to it.

With respect to the estimates of the item hyper-parameters, removing outliers mostly effected the estimates related to the $\ln(\sigma_i^2)$, as could have been expected from the fact that removing extreme values from the sample decreases the estimated variance. Its mean and variance decreased and the correlations with other item parameters became less strong. The 95% credible intervals for the correlations between $\ln(\sigma_i^2)$ and other three parameters $(\ln(\alpha_{1i}), \beta_{0i}, \beta_{1i})$ included zero after the removal of the outliers. For this reason, we do not give any substantive interpretations to these correlations. Table 4.3 summarises the differences between the estimates of the item hyper-parameters after and before the removal of the outliers.

4.6 Simulation study

To assess parameter recovery of the model a simulation study based on the empirical example was performed. In this applied paper we are not aiming at showing the performance of the model for various combinations of item hyper-parameters, person hyper-parameters, test and sample sizes, but rather mainly at the specific combination of those factors from the empirical data that we are dealing with. Therefore, in the simulation study we used the estimates of the item and the person hyper-parameters to simulate replicated data sets of the same sample size

(1000 persons) and the same number of items (49). To investigate how parameter recovery is effected by decrease in sample size and number of items and to evaluate the applicability of the model in a wider range in conditions, three more conditions were considered: $N = 500, n = 49$, $N = 1000, n = 25$, and $N = 500, n = 25$. For each condition, 100 data sets were simulated under the full model with z_{pi} as a covariate for the parameters of the ICC. In each replication the model was fitted using the Gibbs Sampler with one chain of 10000 iterations (including 5000 iterations of burn-in).

Table 4.4 shows the simulation results: the average EAP estimates of the hyper-parameters and the number of the replications (out of 100) in which the true value was within the 95% credible interval. First, let us consider the results obtained when the same sample size and the number of items as in the empirical example were used. The mean vector of the item parameters, the standard deviation of speed and the correlation between speed and ability were correctly recovered. The correlations between the item parameters are estimated to be closer to zero than the true values, and the variances of the item parameters are slightly overestimated. However, this bias is relatively small and does not influence the substantive interpretation of the relations between the item parameters.

When the sample size was reduced (500 instead of 1000), the results were not seriously effected. However, when the number of items was reduced (25 compared to 49), the bias of the variances of item parameters and of the correlations between them increased. This is likely due to the fact that these hyper-parameters were estimated based on a relatively small sample of items. In the condition with $N = 1000$ and $n = 25$ the number of 95% credible intervals which contained the true value is smaller than when the sample size was smaller. This can be explained by the fact that the posterior variance is smaller when the sample size is larger. Overall these results indicate that for accurate recovery of the item hyper-parameters test size should not be too small.

4.7 Discussion

In the paper we provide empirical evidence that a higher level dependence between the persons' speed and ability cannot always fully explain the dependence between response time and accuracy. For cases in which CI is violated, we propose an approach to modeling the conditional dependence by introducing an effect of residual response time on the intercept and on the slope of the ICC. In the applied example the proposed model accounts for differences in item properties between the fast and the slow responses.

The conclusions drawn from the fitted model in the empirical example are interesting from a substantive point of view. The negative correlation between the baseline item intercept and the effect of the residual response time on the

Table 4.4: Results of the simulation study: the EAP estimates of the hyper-parameters averaged across 100 replications and the number of replications in each the true value was within the 95% credible interval

	True value	Average EAP				Coverage rate (%)			
N		1000		500		1000		500	
n		49	25	49	25	49	25	49	25
μ_ξ	3.50	3.51	3.50	3.50	3.48	96	95	95	95
$\mu_{\ln(\sigma^2)}$	-1.49	-1.48	-1.48	-1.48	-1.49	95	97	96	98
$\mu_{\ln(\alpha_0)}$	-0.57	-0.57	-0.59	-0.59	-0.59	96	90	91	95
$\mu_{\ln(\alpha_1)}$	-0.27	-0.26	-0.27	-0.27	-0.28	97	96	93	94
μ_{β_0}	0.17	0.17	0.16	0.20	0.18	97	95	96	93
μ_{β_1}	-0.21	-0.21	-0.22	-0.22	-0.22	97	94	94	91
σ^2_ξ	0.20	0.22	0.24	0.24	0.24	97	81	94	94
$\sigma^2_{\ln(\sigma^2)}$	0.12	0.14	0.17	0.16	0.19	94	77	77	79
$\sigma^2_{\ln(\alpha_0)}$	0.12	0.15	0.17	0.14	0.16	95	85	90	88
$\sigma^2_{\ln(\alpha_1)}$	0.04	0.05	0.10	0.05	0.06	92	83	92	91
$\sigma^2_{\beta_0}$	1.39	1.50	1.53	1.50	1.55	95	87	92	92
$\sigma^2_{\beta_1}$	0.15	0.16	0.20	0.16	0.18	95	88	97	93
$\sigma_{\xi,\ln(\sigma^2)}$	0.40	0.33	0.33	0.33	0.29	91	88	97	96
$\sigma_{\xi,\ln(\alpha_0)}$	0.11	0.10	0.02	0.12	0.06	96	84	95	94
$\sigma_{\xi,\ln(\alpha_1)}$	0.44	0.38	0.33	0.36	0.27	93	88	96	94
σ_{ξ,β_0}	-0.44	-0.40	-0.37	-0.39	-0.36	99	84	96	95
σ_{ξ,β_1}	0.53	0.48	0.40	0.47	0.41	96	86	95	93
$\sigma_{\ln(\sigma^2),\ln(\alpha_0)}$	-0.05	-0.05	-0.03	-0.01	-0.02	98	88	94	98
$\sigma_{\ln(\sigma^2),\ln(\alpha_1)}$	0.33	0.27	0.21	0.23	0.16	96	84	96	96
$\sigma_{\ln(\sigma^2),\beta_0}$	-0.30	-0.24	-0.23	-0.24	-0.23	97	89	94	98
$\sigma_{\ln(\sigma^2),\beta_1}$	0.32	0.25	0.24	0.24	0.22	92	87	94	95
$\sigma_{\ln(\alpha_0),\ln(\alpha_1)}$	-0.05	-0.06	-0.02	-0.04	-0.02	98	90	96	98
$\sigma_{\ln(\alpha_0),\beta_0}$	0.13	0.12	0.08	0.09	0.09	94	88	94	96
$\sigma_{\ln(\alpha_0),\beta_1}$	-0.10	-0.10	-0.08	-0.07	-0.11	96	87	95	96
$\sigma_{\ln(\alpha_1),\beta_0}$	-0.62	-0.55	-0.43	-0.52	-0.39	94	86	96	91
$\sigma_{\ln(\alpha_1),\beta_1}$	0.73	0.63	0.49	0.58	0.42	94	82	92	83
σ_{β_0,β_1}	-0.75	-0.69	-0.59	-0.67	-0.63	90	84	90	91
$\rho_{\theta\tau}$	-0.09	-0.09	-0.10	-0.09	-0.09	95	97	93	95
σ_τ	0.33	0.33	0.33	0.33	0.33	96	96	93	94

intercept is consistent with the results of Goldhammer et. al (2014), who also provided evidence for the increase of the probability of a correct response for difficult items and the decrease of the probability of the correct response for easy items for slow responses. It is important to note that since most of the effects of residual log-response time on the item intercept are negative, this kind of conditional dependence cannot be explained by the speed-accuracy trade-off.

The average negative effect of the residual response time on the item slope is contradicting the findings regarding the 'worst performance rule' (Coyle, 2003), which predict that slow responses contain the most information about persons' ability. One possible explanation could be that the 'worst performance rule' applies to the difficult items but not to the easy items (see Figure 4.7), which are perhaps better answered using fast automated strategies. Another possible explanation for the decrease of the item discriminative power if a person takes more time on the item than expected, is that if responses are fast, then all persons are using the same strategy, whereas the more time persons take the more diverse strategies they may use, hence making the relationship between the measured ability and the probability of a correct response weaker. However, care should be taken with any of these interpretations, since a long response time might also simply be a product of the respondent not having spent all of the recorded time on solving the item.

Modeling conditional dependence between response time and accuracy allows one to reveal more about the relationship between time and accuracy than just one overall correlation between ability and speed. In the presented example, we were able to detect interesting patterns of positive and negative relationships between time and accuracy, while overall the correlation between ability and speed was close to zero. It would be interesting to investigate whether similar conditional dependence phenomena would be observed in other data sets in which the correlation between the two latent traits would be strong and negative or strong and positive.

4.8 Appendix

Here, we describe a Gibbs Sampler for sampling from the joint posterior distribution of the model parameters, which is proportional to the product of the prior distribution and the density of the data:

$$p(\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\xi}, \boldsymbol{\sigma}^2, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\Sigma}_{\mathcal{P}}, \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}} \,|\, \mathbf{X}, \mathbf{T}) \propto p(\boldsymbol{\Sigma}_{\mathcal{P}}) p(\boldsymbol{\mu}_{\mathcal{I}}) p(\boldsymbol{\Sigma}_{\mathcal{I}}) \times$$
$$\prod_p \mathrm{MVN}(\theta_p, \tau_p; \boldsymbol{\Sigma}_{\mathcal{P}}) \prod_i \frac{1}{\sigma_i^2 \alpha_{0i} \alpha_{1i}} \mathrm{MVN}(\xi_i, \ln \sigma_i^2, \ln \alpha_{0i}, \ln \alpha_{1i}, \beta_{0i}, \beta_{1i}; \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}) \times$$
$$\prod_p \prod_i \frac{1}{t_{pi} \sigma_i} \exp\left(-\frac{(\ln t_{pi} - (\xi_i - \tau_p))^2}{2\sigma_i^2} \right) \frac{\exp(x_{pi}(\alpha_{0i} \alpha_{1i}^{z_{pi}} \theta_p + \beta_{0i} + \beta_{1i} z_{pi}))}{1 + \exp(\alpha_{0i} \alpha_{1i}^{z_{pi}} \theta_p + \beta_{0i} \beta_{1i} z_{pi}))}. \quad (4.19)$$

Before the algorithm can be started initial values have to be specified. Identity matrices are used for $\boldsymbol{\Sigma}_{\mathcal{P}}$ and $\boldsymbol{\Sigma}_{\mathcal{I}}$; zero mean vector is used for $\boldsymbol{\mu}_{\mathcal{I}}$. It is important for the initial values of $\boldsymbol{\xi}$, $\boldsymbol{\sigma}^2$ and τ to be chosen close to where the posterior density is concentrated since these parameters determine the values of the residuals of log response times. First, random values are chosen for these parameters: $\xi_{i0} \sim \mathcal{N}(0,1), \sigma_{i0}^2 \sim \ln\mathcal{N}(0,1), \forall i \in [1:n]$ and $\tau_{p0} \sim \mathcal{N}(0,1), \forall p \in [1:N]$. Second, for 20 iterations values are drawn from the conditional posterior distributions of each of these parameters given the response time data only and an improper prior $p(\boldsymbol{\xi}, \boldsymbol{\sigma}, \boldsymbol{\tau}) \propto \prod_i \frac{1}{\sigma_i^2}$. Random initial values are chosen for the parameters in the response accuracy models: $\alpha_{0i0} \sim \ln\mathcal{N}(0, 0.2), \alpha_{1i0} \sim \ln\mathcal{N}(0, 0.2), \beta_{0i0} \sim \mathcal{N}(0, 0.5)$, $\beta_{1i0} \sim \mathcal{N}(0, 0.5), \forall i \in [1:n]$, and $\theta_p \sim \mathcal{N}(0,1), \forall p \in [1:N]$.

After initialisation the algorithm goes through the steps described below, in which the parameters are sampled from their full conditional posterior distributions.

Step 1: For each person p sample the person speed parameter τ_p from:

$$p(\tau_p \,|\, \dots) \propto p(\tau_p \,|\, \boldsymbol{\Sigma}_{\mathcal{P}}, \theta_p) f(\mathbf{T}_p \,|\, \tau_p, \dots) f(\mathbf{X}_p \,|\, \tau_p, \dots). \quad (4.20)$$

Sampling is done using Metropolis-Hastings algorithm with a candidate value drawn from the proposal density:

$$\tau^* \sim \mathcal{N}\left(\frac{\sum_i \frac{(\xi_i - \ln t_{pi})}{\sigma_i^2} + \frac{\sigma_\tau \rho_{\theta\tau} \theta_p}{(1-\rho_{\theta\tau}^2)\sigma_\tau^2}}{\sum_i \frac{1}{\sigma_i^2} + \frac{1}{(1-\rho_{\theta\tau}^2)\sigma_\tau^2}}, \frac{1}{\sum_i \frac{1}{\sigma_i^2} + \frac{1}{(1-\rho_{\theta\tau}^2)\sigma_\tau^2}} \right). \quad (4.21)$$

which is proportional to the product $p(\tau_p \,|\, \boldsymbol{\Sigma}_{\mathcal{P}}, \theta_p) f(\mathbf{T}_p \,|\, \tau_p, \dots)$. The acceptance ratio is equal to:

$$\Pr(\tau_p \to \tau^*) = \min\left(1, \frac{f(\mathbf{X}_p \,|\, \tau^*, \dots)}{f(\mathbf{X}_p \,|\, \tau_p, \dots)}\right). \quad (4.22)$$

Step 2: For each item i sample the time intensity parameter from

$$p(\xi_i \mid \dots) \propto p(\xi_i \mid \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}, \sigma_i^2, \alpha_{0i}, \alpha_{1i}, \beta_{0i}, \beta_{1i}) f(\mathbf{T}_i \mid \xi_i, \dots) f(\mathbf{X}_i \mid \xi_i, \dots). \quad (4.23)$$

Sampling is done using Metropolis-Hastings algorithm with a candidate value drawn from the proposal density

$$\xi^* \sim \mathcal{N}\left(\frac{\frac{\sum_p (\ln t_{pi} + \tau_p)}{\sigma_i^2} + \frac{\mu_\xi^*}{\sigma_\xi^{*2}}}{\frac{N}{\sigma_i^2} + \frac{1}{\sigma_\xi^{*2}}}, \frac{1}{\frac{N}{\sigma_i^2} + \frac{1}{\sigma_\xi^{*2}}} \right), \quad (4.24)$$

where μ_ξ^* and σ_ξ^{*2} are the conditional mean and the conditional variance of ξ_i given the other item parameters of item i. This proposal is proportional to the product of the density of the RT data and the density of ξ_i given other item parameters of item i and the item hyper parameters. The acceptance probability is:

$$\Pr(\xi_i \to \xi^*) = \min\left(1, \frac{f(\mathbf{X}_i \mid \xi^*, \dots)}{f(\mathbf{X}_i \mid \xi_i, \dots)}\right). \quad (4.25)$$

Step 3: For each item i sample the variance of log RT from

$$p(\sigma_i^2 \mid \dots) \propto p(\sigma_i^2 \mid \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}, \xi_i, \alpha_{0i}, \alpha_{1i}, \beta_{0i}, \beta_{1i}) f(\mathbf{T}_i \mid \sigma_i^2, \dots) f(\mathbf{X}_i \mid \sigma_i^2, \dots) \quad (4.26)$$

Metropolis-Hastings algorithm is used with a proposal distribution

$$\sigma^* \sim \mathcal{IG}\left(\frac{N}{2}, \frac{\sum_p (\ln t_{pi} - (\xi_i - \tau_p))^2}{2}\right), \quad (4.27)$$

and an acceptance probability

$$\Pr(\sigma_i^2 \to \sigma^*) = \min\left(1, \frac{f(\mathbf{X}_i \mid \sigma^*, \dots)}{f(\mathbf{X}_i \mid \sigma_i^2, \dots)} \frac{\exp\left(\frac{-(\ln \sigma^* - \mu^*_{\ln \sigma^2})^2}{2\sigma^{2*}_{\ln \sigma^2}}\right)}{\exp\left(\frac{-(\ln \sigma_i^2 - \mu^*_{\ln \sigma^2})^2}{2\sigma^{2*}_{\ln \sigma^2}}\right)}\right), \quad (4.28)$$

where $\mu^*_{\ln \sigma^2}$ and $\sigma^{*2}_{\ln \sigma^2}$ are the conditional mean and the conditional variance of $\ln \sigma_i^2$ given the other item parameters of item i.

Step 4: For each person p sample person ability parameter from:

$$p(\theta_p \mid \dots) \propto p(\theta_p \mid \boldsymbol{\Sigma}_{\mathcal{P}}, \tau_p) f(\mathbf{X}_p \mid \theta_p, \dots) \quad (4.29)$$

To sample from this distribution the single variable exchange algorithm (Marsman

et al., 2015) is used. First, sample a candidate value from

$$\theta^* \sim \mathcal{N}\left(\frac{\rho_{\theta\tau}}{\sigma_\tau}, 1-\rho^2_{\theta\tau}\right); \tag{4.30}$$

then using this value simulate a response vector $\mathbf{x}^*$:

$$x_i^* \sim Bernoulli\left(\frac{\exp(\alpha_{0i}\alpha_{1i}^{z_{pi}}\theta^* + \beta_{0i} + \beta_{1i}z_{pi})}{1+\exp(\alpha_{0i}\alpha_{1i}^{z_{pi}}\theta^* + \beta_{0i} + \beta_{1i}z_{pi})}\right), \forall i \in [1:n]. \tag{4.31}$$

The probability of accepting θ^* as a new value of θ_p is:

$$\Pr(\theta_p \to \theta^*) = \min\left(1, \exp\left((\theta^*-\theta_p)\left(\sum_i \alpha_{0i}\alpha_{1i}^{z_{pi}}x_{pi} - \sum_i \alpha_{0i}\alpha_{1i}^{z_{pi}}x_i^*\right)\right)\right). \tag{4.32}$$

Step 5: For each item i sample item parameters $\{\alpha_{0i},\alpha_{1i},\beta_{0i},\beta_{1i}\}$ from

$$p(\alpha_{0i},\alpha_{1i},\beta_{0i},\beta_{1i} \mid \dots) \propto p(\alpha_{0i},\alpha_{1i},\beta_{0i},\beta_{1i} \mid \boldsymbol{\mu}_\mathcal{I}, \boldsymbol{\Sigma}_\mathcal{I}, \xi_i, \sigma_i^2) \\ f(\mathbf{X}_i \mid \alpha_{0i},\alpha_{1i},\beta_{0i},\beta_{1i},\dots). \tag{4.33}$$

Metropolis-Hastings algorithm is used with a multivariate normal distribution with a mean vector equal to the current values of the parameters, all variances equal to 0.01 and all correlations equal to 0 as a proposal density.

Step 6: Sample the covariance matrix of person parameters from

$$p(\boldsymbol{\Sigma}_\mathcal{P} \mid \boldsymbol{\theta}, \boldsymbol{\tau}) \propto p(\boldsymbol{\theta}, \boldsymbol{\tau} \mid \boldsymbol{\Sigma}_\mathcal{P}) p(\boldsymbol{\Sigma}_\mathcal{P}). \tag{4.34}$$

This is the conditional posterior of the covariance matrix of a multivariate normal distribution, which given the Inverse-Wishart prior is known to be an inverse-Wishart distribution (see for example, Hoff (2009)):

$$(\boldsymbol{\Sigma}_\mathcal{P} \mid \boldsymbol{\theta}, \boldsymbol{\tau}) \sim \text{Inv-Wishart}\left(4+N, \mathbf{I}_2 + \sum_p (\theta_p, \tau_p)(\theta_p, \tau_p)^T\right). \tag{4.35}$$

Step 7: Sample the mean vector of the item parameters from

$$p(\boldsymbol{\mu}_\mathcal{I} \mid \dots) \propto p(\boldsymbol{\xi}, \boldsymbol{\sigma}^2, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1 \mid \boldsymbol{\mu}_\mathcal{I}, \boldsymbol{\Sigma}_\mathcal{I}) p(\boldsymbol{\mu}_\mathcal{I}). \tag{4.36}$$

With a multivariate normal prior for $\boldsymbol{\mu}_\mathcal{I}$, this conditional posterior is also a mul-

tivariate normal with a mean vector equal to

$$\left((100\mathbf{I}_6)^{-1}+n\boldsymbol{\Sigma}_{\mathcal{I}}^{-1}\right)^{-1}\times \left(\boldsymbol{\Sigma}_{\mathcal{I}}^{-1}\left(\sum_i \xi_i, \sum_i \ln\sigma_i^2, \sum_i \ln\alpha_{0i}, \sum_i \ln\alpha_{1i}, \sum_i \beta_{0i}, \sum_i \beta_{1i}\right)^T\right), \quad (4.37)$$

and the covariance matrix equal to $\left((100\mathbf{I}_6)^{-1} + n\boldsymbol{\Sigma}_{\mathcal{I}}^{-1}\right)^{-1}$.

Step 8: Sample the covariance matrix of the item parameters from:

$$p(\boldsymbol{\Sigma}_{\mathcal{I}} \mid \dots) \propto p(\boldsymbol{\xi}, \boldsymbol{\sigma}^2, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1 \mid \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}})|\boldsymbol{\Sigma}_{\mathcal{I}}|^{-\frac{\nu+12}{2}} \times \prod_{k=1}^{6}\left(\nu\left(\boldsymbol{\Sigma}_{\mathcal{I}}\right)_{kk} + \frac{1}{A^2}\right)^{-\frac{\nu+6}{2}}. \quad (4.38)$$

A Metropolis-Hastings algorithm is used to sample from this conditional posterior with the candidate value $\boldsymbol{\Sigma}^*$ sampled from the Inverse-Wishart distribution with n degrees of freedom and the scale matrix equal to

$$\sum_i \left((\xi_i, \ln\sigma_i^2, \ln\alpha_{0i}, \ln\alpha_{1i}, \beta_{0i}, \beta_{1i})^T - \boldsymbol{\mu}_{\mathcal{I}}\right)\left((\xi_i, \ln\sigma_i^2, \ln\alpha_{0i}, \ln\alpha_{1i}, \beta_{0i}, \beta_{1i})^T - \boldsymbol{\mu}_{\mathcal{I}}\right)^T, \quad (4.39)$$

such that the acceptance probability is equal to:

$$\Pr(\boldsymbol{\Sigma}_{\mathcal{I}} \to \boldsymbol{\Sigma}^*) = \min\left(1, \frac{|\boldsymbol{\Sigma}^*|^{-\frac{\nu+5}{2}}\prod_{k=1}^{6}\left(\nu\left(\boldsymbol{\Sigma}^*\right)_{kk} + \frac{1}{A^2}\right)^{-\frac{\nu+6}{2}}}{|\boldsymbol{\Sigma}_{\mathcal{I}}|^{-\frac{\nu+5}{2}}\prod_{k=1}^{6}\left(\nu\left(\boldsymbol{\Sigma}_{\mathcal{I}}\right)_{kk} + \frac{1}{A^2}\right)^{-\frac{\nu+6}{2}}}\right). \quad (4.40)$$

Step 9: Re-scale model parameters to equate the variance of ability to 1:

$$\begin{aligned} \theta_p &\to \tfrac{\theta_p}{\sigma_\theta}, && \forall p \in [1:N]; \\ \alpha_{0i} &\to \alpha_{0i}\sigma_\theta, && \forall i \in [1:n]; \\ \mu_{\ln\alpha_0} &\to \mu_{\ln\alpha_0} + \ln\sigma_\theta; \\ \boldsymbol{\Sigma}_{\mathcal{P}} &\to \begin{bmatrix} 1 & \rho_{\theta\tau} \\ \rho_{\theta\tau} & \sigma_\tau^2 \end{bmatrix}. \end{aligned} \quad (4.41)$$

Part II

Bayesian contributions to item response theory

Chapter 5

Unmixing Rasch scales: How to score an educational test

[1] **Abstract.**One of the important questions in the practice of educational testing is how a particular test should be scored. In this paper we consider what an appropriate simple scoring rule should be for the Dutch as a second language test consisting of listening and reading items. As in many other applications, here the Rasch model which allows to score the test with a simple sumscore is too restrictive to adequately represent the data. In this study we propose an exploratory algorithm which clusters the items into subscales each fitting a Rasch model and thus provides a scoring rule based on observed data. The scoring rule produces either a weighted sumscore based on equal weights within each subscale, or a set of sumscores (one for each of the subscales). An MCMC algorithm which enables to determine the number of Rasch scales constituting the test and unmix these scales is introduced and evaluated in simulations. Using the results of unmixing we conclude that the Dutch language test can be scored with a weighted sumscore with three different weights.
Keywords: educational testing, Markov chain Monte Carlo,
mixture model, multidimensional IRT, One Parameter Logistic model,
Rasch model, scoring rule.

5.1 Introduction

Consider a test measuring language ability. One of the important practical questions when using this test is how it should be scored. This includes the following

[1]This chapter has been accepted for publication in Annals of Applied Statistics as Bolsinova, M., Maris, G. & Hoijtink, H. Unmixing Rasch scales: How to score an educational test. Author contributions: B.M., M.G. and H.H. designed the research, B.M. performed the research, B.M. wrote the paper, M.G. and H.H. provided feedback on the manuscript

subquestions: Should the results be summarised in a single score or in multiple scores? Should all items have the same weight or different weights when computing the score or the subscores? If subscores are used, how to determine which items belong to which subscale? If different weights are used, how to restrict the number of possible weights such that not every response pattern results in a unique weighted score? And how to determine which items should have the same weight? In this paper, we argue for an empirical approach for choosing a scoring rule. We want the data to tell us what is an appropriate score to use for grading this language test: the sumscore (the number of correct responses to all the items), two sumscores (the number of correct responses to the listening items and the number of correct responses to the reading items), a set of multiple sumscores with an alternative division of the items into subscales, a weighted sumscore, or a set of weighted sumscores. The most appropriate choice will often require a thorough investigation of the structure of the test data.

The aim of this article is to choose a simple scoring rule for the state exam of Dutch as foreign language (Staatsexamen NT2). By passing this test non-native speakers show sufficient mastery of the Dutch language to work and study in the Netherlands. We consider the multiple-choice part of the test consisting of reading and listening items. The reading and the listening subtests consist of multiple texts or audio fragments followed by multiple choice questions.

Having a measurement model providing an explicit scoring rule is very important and convenient in the context of educational measurement. A scoring rule based on a sufficient statistic is favourable because no information about the ability is lost by summarising a vector of responses in one or more scores. One of the simplest IRT models - the Rasch model [RM] (Rasch, 1960) - has the number of correct responses as a sufficient statistic for ability (Andersen, 1977; Fischer, 1995). However, the RM very often does not fit the empirical data due to the strict assumptions of unidimensionality and equal discrimination of the items. It is not uncommon that an educational test measures more than one ability. Moreover, some of the test items are more closely related to the latent trait than others (i.e., have a steeper item characteristic curve) and should have a bigger weight in the estimation of a person's ability. In our case of the Dutch as foreign language test, it is unlikely that a diverse pool of items (with both reading and listening items) would constitute a single Rasch scale. In this study we propose a new model which relaxes the assumptions of the Rasch model, but still gives an explicit scoring rule for the test summarising all the information about the student's ability (or abilities). This scoring rule can be more complicated than simply summing the number of correct responses, but should still result in one or more scores that are easy to use and interpret, for example, a set of sumscores or a weighted sumscore with a limited number of different weights.

The paper is organised as follows. First, in Section 5.2 the state examination of Dutch as second language is introduced in more detail and the problem of choosing

a scoring rule for it is discussed. In the following four sections we introduce our solution to the problem. In Section 5.3, we discuss how the assumptions of the RM can be relaxed without losing the important property of sufficiency of the sumscores. This results in the multi-scale RM which is a mixture of Rasch scales. Note, that throughout the paper when we use the term 'mixture of scales', we are referring to a mixture of *item clusters*, each with different properties, and not to the more common type of mixture models with different *groups of persons*, such as present in the mixture Rasch model (Rost, 1990). In Section 5.4, the presented model is discussed in more detail in relation to the problem of choosing the scoring rule for the Dutch language test. In Section 5.5, the estimation of the model is discussed. In Section 5.6, we evaluate the estimation procedure in a simulation study. After introducing the methodology and showing its usability in simulations, in Section 5.7 we return to the application and address the practical questions raised in the beginning of this section concerning the NT2 exam. The paper is concluded with a discussion.

5.2 State examination of Dutch as a second language

We consider the version of the NT2 exam called Program II which is meant for those who have gained higher education in their home country and wish to continue their education in Dutch or work at the level of university education in the Netherlands. This version of the NT2 exam corresponds to the B2 language level within the Common European Framework of Reference for languages (Council of Europe, 2011). The exam is taken in a computerised form. Test-takers are given 100 minutes to complete the reading part of the exam and the listening part takes about two hours. A short article from a newspaper or scientific journal, or an information brochure can be examples of reading texts. In some reading items participants are asked about some particular detail from a certain part of the text, while other items require understanding the text as a whole. A common example of an audio fragment in the listening part is a radio interview.

Test scores which are easy to understand and interpret need to be communicated to test-takers and policy makers. The easiest way to score the test would be with the number of correct responses, such that all persons with the same number of correct responses receives the same score and it does not matter which items are answered correctly. This scoring rule implies the Rasch model for the data. The RM models the probability of answering an item correctly using only two parameters (one for the item and one for the person):

$$\Pr(X_{pi} = 1 \,|\, \delta_i, \theta_p) = \frac{\exp(\theta_p - \delta_i)}{1 + \exp(\theta_p - \delta_i)}, \tag{5.1}$$

where X_{pi} is the item response which can be scored 1 if it is correct or 0 if it is

incorrect, δ_i is the difficulty parameter of item $i \in [1 : n]$ and θ_p is the ability parameter of person $p \in [1 : N]$. However, the RM being rather restrictive rarely fits the data. To evaluate whether a simple sumscore is appropriate as the score for the NT2 exam, we tested the fit of the RM to the data set from this examination - responses of 2398 persons to 74 items (40 reading and 34 listening).

The fit of the RM to the data was tested using Anderson's Likelihood-ratio [LR] test (Andersen, 1973). The idea of the test is the following: The sample is split into H groups based on the number of correct responses. If the RM holds then there are no differences between the estimates of the item parameters obtained in separate groups. The likelihood ratio is computed using the likelihood based on the estimates of the item parameters in each group and the likelihood based on the overall estimates of the item parameters. The logarithm of this ratio follows the χ^2-distribution with $(n-1)(H-1)$ degrees of freedom under the RM, where n is the number of items.

For the data set from NT2 exam the LR-statistic for a median-split ($H = 2$) was equal to 1165.54 ($df = 73$), $p < 0.0005$. Hence, the RM does not fit the data of the NT2 exam. An alternative to using one sumscore, could be using two scores: the number of correct responses to the reading items and the number of correct responses to the listening items. To evaluate whether this scoring rule is appropriate for the NT2 exam, we tested the fit of the RM separately to the reading items and to the listening items. The LR-statistics for the reading and the listening subscales were equal to 551.93 ($df = 39, p < 0.0005$) and 473.54 ($df = 33, p < 0.0005$), respectively. Hence, the RM does not hold in the two parts of the test taken separately. Therefore, a different scoring rule has to be chosen. We argue for a data-driven exploratory approach which identifies scales fitting the RM within the full set of items, such that the test can be scored with a set of sumscore in this subscales, or with a weighted sumscore with equal weights within each scale, which is easy to interpret and to communicate to the test-takers.

In this study we do not consider the two-parameter logistic model (2PL; Lord & Novick, 1968) which includes for each item not only a difficulty parameter, but also a discrimination parameter, usually denoted by α_i, such that the probability of a correct response depends on $\alpha_i(\theta_p - \delta_i)$ instead of the simple difference between ability and difficulty. The reason for not fitting this model to the data is that in the 2PL each item has a unique weight and each response pattern corresponds to a unique score which makes interpretation and communication of the results more difficult and less transparent. Another reason for not considering models like the 2PL or the three-parameter logistic model (Birnbaum, 1968) is that these models do not allow for multidimensionality in the data, while we are aiming at relaxing not only the assumption of the equal discriminations but also the unidimensionality assumption of the RM.

5.3 Relaxing the assumptions of the RM

Simple Rasch model

As we mentioned in the previous section, the main advantage of the RM is that it has a sufficient statistic for person parameters (the number of items correct) and a sufficient statistic for item parameters (the number of correct responses to the item). This is important both for the estimation of the parameters, because it makes the RM an exponential family model, and for the interpretation of test results, because all persons answering the same number of items correctly have the same estimate of the ability parameter. From a student's perspective, it is desirable that students who answer the same number of items correctly get the same grade. Although the RM has these important advantages, a disadvantage is that it makes restrictive assumptions, often resulting in misfit to empirical data.

General multidimensional IRT model

Let us consider how some of the assumptions of the RM can be relaxed. If we relax the assumptions of unidimensionality and equal discriminations, a general model allowing for multidimensionality and different discriminations of items can be obtained (Reckase, 2008):

$$\Pr(X_{pi} = 1 \,|\, \delta_i, \boldsymbol{\alpha}_i, \boldsymbol{\theta}_{p\cdot}) = \frac{\exp(\sum_{k=1}^{M} \alpha_{ik}\theta_{pk} - \delta_i)}{1 + \exp(\sum_{k=1}^{M} \alpha_{ik}\theta_{pk} - \delta_i)}, \tag{5.2}$$

where M is the number of scales, δ_i is the difficulty parameter of item i and $\boldsymbol{\alpha}_i = \{\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iM}\}$ are the discrimination parameters of item i corresponding to the dimensions $\{1, 2, \ldots M\}$, and $\boldsymbol{\theta}_{p\cdot} = \{\theta_{p1}, \theta_{p2}, \ldots, \theta_{pM}\}$ is the vector of abilities of person p. This is a very flexible model, but its flexibility comes with some statistical and interpretational problems. For example, with respect to the model in (5.2) only $\sum_k \alpha_{ik}\theta_{pk}$ is identifiable, but not the individual parameters. Like in factor analysis, the problem of rotation has to be addressed to obtain estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$. Moreover, the model does not have sufficient statistics. We will restrict the model in such a way that it retains some of its flexibility while also regaining some of the important properties of the RM.

Simple structure multidimensional model

If $\boldsymbol{\alpha}$ is restricted to have a simple structure, that is each vector $\boldsymbol{\alpha}_i$ (see Equation 5.2) has only one non-zero element, then the model becomes a mixture of unidimensional scales, each fitting the 2PL. The simple structure of $\boldsymbol{\alpha}$ clarifies the

interpretation of the abilities $\boldsymbol{\theta}_{\cdot k} = \{\theta_{1k}, \theta_{2k}, \ldots, \theta_{Nk}\}$, since each item measures only one ability. However, since the 2PL is not an exponential family model, persons having the same number of correct responses to the items measuring ability $\boldsymbol{\theta}_{\cdot k}$ but different response patterns do not have the same estimates of the ability, and hence the sumscore on that scale is not a sufficient statistic.

Multi-scale Rasch model

If we further restrict the non-zero element of $\boldsymbol{\alpha}_i$ to be equal to one, then the model is a mixture of Rasch scales and $\sum_i \alpha_{ik} X_{pi}$ contains all information about ability θ_{pk}. This gives a rather convenient scoring rule where all information about student's abilities is summarised in a vector of subscores

$$\left\{ \sum_i \alpha_{i1} X_{pi}, \sum_i \alpha_{i2} X_{pi}, \ldots, \sum_i \alpha_{iM} X_{pi} \right\}. \tag{5.3}$$

We call the mixture of Rasch scales a multi-scale Rasch model. It assumes that a test consists of a number of Rasch homogeneous subscales which have to be unmixed. The model has the same form as in Equation 5.2, but with the constraints $\alpha_{ik} \in \{0, 1\}$ and $\sum_k \alpha_{ik} = 1$. Thus, $\boldsymbol{\alpha}_i = \{\alpha_{i1}, ..., \alpha_{iM}\}$ is a vector of item scale memberships specifying to which scale item i belongs: $\alpha_{ik} = 1$ if item i belongs to dimension k and 0 otherwise.

One-parameter logistic model as a multi-scale RM

It might seem that with a multi-scale RM we are still restricted to items with the same discrimination. However, we will now show that in such a model we can allow items referring to the same ability to have different discriminations. To do that we present the one-parameter logistic model (OPLM; Verhelst & Glas, 1995) as a special case of a mixture of Rasch scales. The usual way of considering the OPLM is as a special case of the 2PL in which items have known integer-valued discrimination indices a_i instead of the discrimination parameters that are estimated freely. We propose an alternative perspective. We consider it as a special case of the multi-scale RM in which the scales differ only in the item discriminations.

Since in the OPLM the discrimination indices are constrained to be integer-valued, there will be a limited number of possible values for the discrimination indices in a test, denoted by $\sigma_1, \sigma_2, \ldots, \sigma_M$. Instead of having one person parameter θ_p per person, we introduce several person parameters $\theta_{pk} = \sigma_k \theta_p$, one for each group of items with a common discrimination index equal to σ_k - referred to as item set discrimination by Humphry (2012). Furthermore, let us denote by $\boldsymbol{\alpha}$ a simple structure matrix where $\alpha_{ik} = 1$ if $a_i = \sigma_k$ and $\alpha_{ik} = 0$ otherwise. Finally, we re-parameterise the difficulty parameter as $\delta_i^* = a_i \delta_i$. Then, within each set of

items $\{i \mid a_i = \sigma_k\}$ a RM with person parameter $\boldsymbol{\theta}_{\cdot k}$ holds (Humphry, 2011), and the whole test is modelled as a mixture of Rasch scales with a fixed matrix $\boldsymbol{\alpha}$ and person parameters in different scales k and l being proportional to each other:

$$\theta_{pk} = \frac{\sigma_k}{\sigma_l}\theta_{pl}. \tag{5.4}$$

These scales measure the same latent variable but represent different frames of reference and have different units of measurement (Humphry & Andrich, 2008). Thus, we have shown that a multi-scale RM can allow items measuring the same ability to have different discriminative power, if they belong to different scales with perfectly correlated person parameters. In this case, not only a vector of sumscores, but a weighted score $\sum_i \sum_k \alpha_{ik}\sigma_k X_{pi}$ contains all information about the original person parameter θ_p.

The problem of unmixing Rasch scales

The purpose of the present study is to develop a Bayesian algorithm for selecting the best partitioning of items into scales each fitting a RM, that is to estimate the item scale membership matrix $\boldsymbol{\alpha}$. This is done by sampling from the posterior distribution of item scale memberships (parameters of interest) given the data: $p(\boldsymbol{\alpha} \mid \mathbf{X})$. All other parameters of the multi-scale RM are nuisance parameters which are also sampled to simplify the computations. The item scale memberships are identified, because for each pair of items it can be determined from the data whether they belong to the same Rasch scale or to different scales (see Appendix A). Since the parameters are identified, they can also be consistently estimated.

The multi-scale RM is related to the between-item multidimensional Rasch model (Adams, Wilson, & Wang, 1997), which also assumes a RM for subsets of items in the test. However, while in the between-item multidimensional RM and in the OPLM the subscales or the groups of items with the same discrimination indices, respectively, have to be pre-specified, in the new model the item memberships are parameters which can be estimated. Therefore, our method provides an objective statistical tool that researchers can use to select an optimal partitioning of items into Rasch scales, instead of having to specify the scales or the item discrimination indices in advance using only background information.

There have been other attempts to solve the problem of selecting groups of items fitting the RM. Hardouin and Mesbah (2004) proposed a method that is based on the AIC. Debelak and Arendasy (2012) identified item clusters fitting the RM using hierarchical clustering. Both approaches are not model-based and instead provide heuristics for building scales bottom-up. Simulation results from both studies show that the procedures do not work very well when the person parameters are highly correlated, when the sample sizes are small and when the item pools are large. Moreover, the procedures are not at all suited for deter-

mining scales differing only in the discriminative power of the items, due to the perfect correlation of the person parameters. A simulation study comparing the performance of our model-based approach algorithm with that of the hierarchical clustering algorithm can be found in the Appendix C.

5.4 Model specification

5.4.1 Mixture of Rasch scales

As we stated in the introduction, the purpose of the algorithm which we developed is to obtain estimates of the item memberships in the multi-scale RM by sampling from their posterior distribution.

We consider a marginal model, in which individual person parameters are treated as random-effects with a multivariate normal distribution with a zero mean vector and covariance matrix $\boldsymbol{\Sigma}$. Constraining the mean vector of ability to zero ensures the identification of the model.

Let us by $\mathbf{X}_{p\cdot}$ denote a random vector of responses to n items from person p randomly sampled from the population, and its realisation by $\mathbf{x}$ with $x_i = 1$ if a response to item i is correct and $x_i = 0$ otherwise. The probability of $\mathbf{X}_{p\cdot}$ being equal to $\mathbf{x}$ is the following:

$$\Pr(\mathbf{X}_{p\cdot} = \mathbf{x} \,|\, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) = \int_{\mathbb{R}} \prod_{i=1}^{n} \frac{\left(\exp\left(\sum_{k=1}^{M} \alpha_{ik}\theta_k - \delta_i\right)\right)^{x_i}}{1 + \exp\left(\sum_{k=1}^{M} \alpha_{ik}\theta_k - \delta_i\right)} p(\boldsymbol{\theta} \,|\, \boldsymbol{\Sigma}) d\boldsymbol{\theta}, \qquad (5.5)$$

where $\boldsymbol{\delta} = \{\delta_1, \delta_2, \ldots, \delta_n\}$ is a vector of item difficulties, $\boldsymbol{\alpha}$ is an $n \times M$ matrix of item membership parameters, and $p(\boldsymbol{\theta} \,|\, \boldsymbol{\Sigma})$ denotes the population distribution. In the multi-scale RM the probability of observing a correct response to item i given the vector of ability parameters is the same as in the general multidimensional IRT model in Equation 5.2, but the vector $\boldsymbol{\alpha}_i$ is constrained to have one element equal to one and all other elements equal to zero. As can be seen from Equation 5.5, the multi-scale RM assumes local independence, meaning that the item responses are independent given the vector of ability parameters.

In Section 5.3, using the OPLM as an example, we have shown that multiple Rasch scales might not only represent different abilities as in the most general multidimensional model in (5.2), but may also differ in the discriminative power of the items. We will now elaborate more on the different types of scenarios in which the Rasch scales could be unmixed:

Type 1. The test measures several substantively different abilities, and each of the Rasch scales refers to a separate ability. For example, in an arithmetic test the items can group into substantively different scales: addition, subtraction, division and multiplication. Within each of the subscales the discriminations of

the items are equal and each of the abilities can be summarised in a subscore. For a model with four item clusters the covariance matrix has the from:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & \sigma_{1,4} \\ \sigma_{1,2} & \sigma_2^2 & \sigma_{2,3} & \sigma_{2,4} \\ \sigma_{1,3} & \sigma_{2,3} & \sigma_3^2 & \sigma_{3,4} \\ \sigma_{1,4} & \sigma_{2,4} & \sigma_{3,4} & \sigma_4^2 \end{bmatrix} \tag{5.6}$$

with all covariances $\sigma_{k,l}$ being free parameters. In the NT2 exam, the items might cluster in subsets measuring reading ability and listening ability with items within a dimension having equal discriminations. In that case the appropriate scoring rule would be to use a set of two subscores: $\{\sum_i \alpha_{i1} X_{pi}, \sum_i \alpha_{i2} X_{pi}\}$.

Type 2. The test measures several abilities, but not each scale represent a separate ability. Some of the abilities are represented by one or more scales with different discriminations. Such scales can occur, for example, due to different response formats of the items, or because some of the items are more relevant for the measured ability and, therefore, should have a bigger weight. For a model with four item clusters the covariance matrix can have the form:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2 & \rho\sigma_1\sigma_3 & \rho\sigma_1\sigma_4 \\ \sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \rho\sigma_2\sigma_4 \\ \rho\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 & \sigma_3\sigma_4 \\ \rho\sigma_1\sigma_4 & \rho\sigma_2\sigma_4 & \sigma_3\sigma_4 & \sigma_4^2 \end{bmatrix}, \tag{5.7}$$

that is the correlations between θ_1 and θ_2, and between θ_3 and θ_4 are constrained to one, and there is only one correlation parameter to be freely estimated. This model is equivalent to a two-dimensional IRT model with

$$\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right) \tag{5.8}$$

and four item clusters with discrimination parameters equal to σ_1 and σ_2 in the first dimension and equal to σ_3 and σ_4 in the second dimension. In the case of the NT2 exam, it might the two distinct abilities (reading and listening) each measured by several scales with different discrimination parameters. Then the appropriate scoring rule would be to use a set of two weighted scores, one for the reading ability and one for the listening ability. $\{\sum_i (\alpha_{i1}\sigma_1 + \alpha_{i2}\sigma_2) X_{pi}, \sum_i (\alpha_{i3}\sigma_3 + \alpha_{i4}\sigma_4) X_{pi}\}$.

Type 3. The test measures a single ability, but the different Rasch scales represent groups of items with different discriminations between the groups. For example, a model with four item clusters the covariance matrix could have the

form:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2 & \sigma_1\sigma_3 & \sigma_1\sigma_4 \\ \sigma_1\sigma_2 & \sigma_2^2 & \sigma_2\sigma_3 & \sigma_2\sigma_4 \\ \sigma_1\sigma_3 & \sigma_2\sigma_3 & \sigma_3^2 & \sigma_3\sigma_4 \\ \sigma_1\sigma_4 & \sigma_2\sigma_4 & \sigma_3\sigma_4 & \sigma_4^2 \end{bmatrix}, \tag{5.9}$$

that is all correlation parameters are constrained to one, This variant of the model would be equivalent to a unidimensional IRT model with $\theta \sim \mathcal{N}(0, 1)$ and four item clusters with discrimination parameters equal to $\sigma_1, \sigma_2, \sigma_3$, and σ_4, respectively. In our case of the NT2 exam, it might turn out that the reading and listening items together measure the same passive language ability, but some of them turn out to have higher discriminations than others, for example depending on the length of the reading passage or the audio fragment to which it refers. Then the appropriate scoring rule for the test would be to use a weighted sumscore: $\sum_i \sum_k \alpha_{ik}\sigma_k X_{pi}$. Our algorithm makes it possible to identify the scales within which the items would have the same weight and estimate these weights.

The algorithm presented in this paper is exploratory, therefore, it need not be pre-specified which of the scenarios we expect, and the covariance matrix is freely estimated. Once the unmixing results for the Dutch language test are obtained, we can formulate hypotheses about the nature and the interrelations of the scales. If the estimate of the correlation between the scales is close to one, then through cross-validation we would test a hypothesis that these scales measure, in fact, the same ability, which would lead to the conclusion that for the scoring rule we could not only use a set of subscores, but also a weighted sumscore. Hypotheses like this can be evaluated by comparing the fit of models of Type 1, Type 2 (if only some of the scales are perfectly correlated) and Type 3 (if all the scales are perfectly correlated) in cross-validation.

The number of scales also does not need to be pre-specified beforehand, but can be decided upon based on the estimation results (see Section 5.5.2). This is an important feature of our procedure, because while the researcher can have some idea about how many substantively different abilities are measured, it can hardly be known based only on the theoretical insight how many different weights are needed for the items measuring each of these abilities. Moreover, the expectation about the number of substantively different abilities could be wrong.

5.4.2 Density of the data, prior and posterior distributions

The density of the data is:

$$f(\mathbf{X} \mid \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) = \prod_{p=1}^{N} \Pr(\mathbf{X}_{p\cdot} = \mathbf{x} \mid \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}), \tag{5.10}$$

where $\mathbf{X}$ is an $N \times n$ matrix of persons with each row $\mathbf{X}_{p\cdot}$ representing responses of person $p \in [1:N]$.

A priori the parameters of the model are assumed to be independent:

$$p(\boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} p(\delta_i) \prod_{i=1}^{n} p(\boldsymbol{\alpha}_i) p(\boldsymbol{\Sigma}). \tag{5.11}$$

We use non-informative priors here, because prior knowledge is not needed to make the model estimable. For the item difficulties a uniform prior distribution $U(-\infty, +\infty)$ is used. This is an improper prior, but the resulting posterior is proper if for every item there is at least one person giving a correct response and at least one person giving an incorrect response (Ghosh, Ghosh, Chen, & Agresti, 2000). For the item memberships a multinomial prior is used:

$$\Pr(\alpha_{ik} = 1, \alpha_{il} = 0, \forall l \neq k) = \frac{1}{M}, \forall k \in [1:M], \forall i \in [1:n], \tag{5.12}$$

where the choice of $\frac{1}{M}$ implies that a priori all item scale memberships are considered equally likely. We choose a semi-conjugate prior for the covariance matrix which is an inverse-Wishart distribution with degrees of freedom $\nu_0 = M + 2$ and a scale parameter $\boldsymbol{\Lambda}_0 = \mathbf{I}_M$ (i.e., an M-dimensional identity matrix). With this choice of ν_0 the results are not sensitive to the choice of $\boldsymbol{\Lambda}_0$ because in the posterior distribution the data dominates the prior when $N \gg (M + 2)$ (Hoff, 2009, p.110).

In order to unmix Rasch scales we need to obtain samples from the joint posterior distribution:

$$p(\boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma} \,|\, \mathbf{X}) \propto f(\mathbf{X} \,|\, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) p(\boldsymbol{\delta}) p(\boldsymbol{\alpha}) p(\boldsymbol{\Sigma}). \tag{5.13}$$

In the next section we discuss how these samples can be obtained using a data augmented MCMC algorithm.

5.5 Estimation

5.5.1 Algorithm for unmixing Rasch scales

In this subsection we discuss how Rasch scales can be unmixed using a Markov chain Monte Carlo algorithm (Gamerman & Lopes, 2006) when the number of scales is pre-specified. This algorithm makes it possible to obtain samples from the posterior distribution in Equation 5.13. In the next section a procedure to determine the number of scales is described.

To start the MCMC algorithm, initial values for the model parameters are specified: samples from $U(-2, 2)$ for the item difficulties, samples from a multinomial distribution with a probability $\frac{1}{M}$ for every scale for the item memberships,

and $\mathbf{I}_M$ for $\boldsymbol{\Sigma}$. After initialisation, in every iteration of the MCMC algorithm the parameters are subsequently sampled from their full conditional posterior distributions given the current values of all other parameters (Geman & Geman, 1984; Casella & George, 1992). Data augmentation is implemented (Tanner & Wong, 1987; Zeger & Karim, 1991). That is, every iteration starts with sampling from the posterior distribution of individual person parameters, which results in a set of conditional posterior distributions that are relatively easy to sample from. Each iteration of the algorithm consists of four steps, described below:

Step 1. For every scale $k \in [1:M]$ for every person $p \in [1:N]$, sample from the full conditional posterior distribution of θ_{pk}:

$$p(\theta_{pk} \,|\, \mathbf{X}, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\theta}_{p(k)}, \boldsymbol{\theta}_{(p)\cdot}, \boldsymbol{\Sigma}) = p(\theta_{pk} \,|\, \mathbf{X}_{p\cdot}, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\theta}_{p(k)}, \boldsymbol{\Sigma}), \tag{5.14}$$

where $\boldsymbol{\theta}_{(p)\cdot}$ are person parameters of all persons except p, and $\boldsymbol{\theta}_{p(k)}$ are person parameters of person p in all scales except k. This conditional posterior depends on the data only through the value X_{p+k} - the number of correct responses of person p to the set of items $\{i \,|\, \alpha_{ik} = 1\}$, because the RM holds in each scale:

$$p(\theta_{pk}|\mathbf{X}_{p\cdot}, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\theta}_{p(k)}, \boldsymbol{\Sigma}) = p(\theta_{pk}|X_{p+k}, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\theta}_{p(k)}, \boldsymbol{\Sigma}) \propto f(X_{p+k}|\boldsymbol{\delta}, \boldsymbol{\alpha}, \theta_{pk}) p(\theta_{pk}|\boldsymbol{\theta}_{p(k)}, \boldsymbol{\Sigma}). \tag{5.15}$$

The conditional composition algorithm (Marsman et al., 2015) is used to sample from this distribution:

a. Sample a candidate value $\theta^* \sim p(\theta_{pk} \,|\, \theta_{p(k)}, \boldsymbol{\Sigma})$, which is a conditional distribution of a multivariate normal distribution, thus itself a normal distribution with known parameters (see for example, Gelman et al. (1995)).

b. Simulate a vector of responses $\mathbf{X}_k^*$ to the set of item $\{i \,|\, \alpha_{ik} = 1\}$ according to the RM.

c. Compute $X_k^{+*} = \sum_i X_i^*$. If $X_k^{+*} = X_{pk}^+$ then θ^* is taken as a new sample from (5.14). Otherwise, Steps a, b, and c are repeated.

Step 2. Sample from the full conditional posterior distribution of $\boldsymbol{\Sigma}$ which depends only on the person parameters $\boldsymbol{\theta}$:

$$p(\boldsymbol{\Sigma} \,|\, \mathbf{X}, \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = p(\boldsymbol{\Sigma}|\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\boldsymbol{\Sigma}) p(\boldsymbol{\Sigma}) = \mathcal{IW}\left(\nu_0 + N, \boldsymbol{\Lambda}_0 + \sum_{p=1}^{N} \boldsymbol{\theta}_{p\cdot} \boldsymbol{\theta}_{p\cdot}^T\right), \tag{5.16}$$

which is an inverse-Wishart distribution with the posterior degrees of freedom equal to the sum of the prior degrees of freedom and the sample size, and the

posterior scale parameter equal to the prior scale parameter plus the sum of squares of the person parameters (Hoff, 2009, p.111).

Step 3. For every item $i \in [1:n]$, sample item difficulty δ_i from its full conditional posterior distribution:

$$p(\delta_i|\mathbf{X}, \boldsymbol{\delta}_{(i)}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\Sigma}) = p(\delta_i \,|\, \mathbf{X}_{\cdot i}, \boldsymbol{\alpha}_i, \boldsymbol{\theta}) \propto \prod_p \frac{\exp(X_{pi}(\sum_{k=1}^{M} \alpha_{ik}\theta_{pk} - \delta_i))}{1 + \exp(\sum_{k=1}^{M} \alpha_{ik}\theta_{pk} - \delta_i)}, \quad (5.17)$$

where $\boldsymbol{\delta}_{(i)}$ denotes a vector of item difficulties of all items except item i and $\mathbf{X}_{\cdot i}$ is a vector of responses of all persons to item i. For this distribution the normalising constant is very difficult to compute, therefore a Metropolis step within the Gibbs sampler is used to sample from the posterior (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). A normal distribution with the mean equal to the current value of the parameter and the variance τ^2 (equal to 0.25 for the first 500 iterations and 0.01 for the remaining iterations) is used as the proposal distribution.

Step 4. For every item $i \in [1:n]$, sample item scale membership $\boldsymbol{\alpha}_i$ from the full conditional posterior distribution:

$$p(\boldsymbol{\alpha}_i \,|\, \mathbf{X}, \boldsymbol{\delta}, \boldsymbol{\alpha}_{(i)}, \boldsymbol{\theta}, \boldsymbol{\Sigma}) = p(\boldsymbol{\alpha}_i \,|\, \mathbf{X}_{\cdot i}, \delta_i, \boldsymbol{\theta}) \propto \prod_p \frac{\exp(X_{pi}(\sum_{k=1}^{M} \alpha_{ik}\theta_{pk} - \delta_i))}{1 + \exp(\sum_{k=1}^{M} \alpha_{ik}\theta_{pk} - \delta_i)}, \quad (5.18)$$

where $\boldsymbol{\alpha}_{(i)}$ are item scale memberships of all items except i. This amounts to sampling from a Multinomial $(1, \{p_{i1}, \ldots, p_{iM}\})$, with parameters

$$p_{ik} = \Pr(\alpha_{ik} = 1, \alpha_{il} = 0, \forall l \neq k) = \frac{\prod_p \frac{\exp(X_{pi}(\theta_{pk} - \delta_i))}{1+\exp(\theta_{pk} - \delta_i)}}{\sum_{j=1}^{M} \prod_p \frac{\exp(X_{pi}(\theta_{pj} - \delta_i))}{1+\exp(\theta_{pj} - \delta_i)}}. \quad (5.19)$$

As is the case with most finite mixture models, the posterior distribution of the parameters of the multi-scale RM has a complex structure (Diebolt & Robert, 1994; Frühwirth-Schnatter, 2006). It has multiple modes corresponding to every partition of items into scales. Among the modes there are $M!$ modes of equal height representing the same partition of items into scales due to the possible permutations of the scale labels. However, the problem of label switching usually does not occur within one chain, because it is not likely for the chain to leave the mode corresponding to a particular set of labels once it has been reached.

In practice it is impossible for the Markov chain to visit all the modes in

a reasonable number of iterations (Celeux, Hurn, & Robert, 2000). It is more likely that the chain will stay in the neighbourhood of one of the strongest modes. Consequently, the initial values influence to which mode the sampler is directed. Multiple chains from random initial values are, therefore, used to explore whether there are many strong modes representing different partitions of items into scales and what the relative likelihood of these modes is. The procedure goes as follows:

a) Run ten independent chains from random starting values for a chosen number of iterations and discard the first half of the iterations in each chain (burn-in) to remove the influence of the initial values. The number of iterations depends on: a) the number of items, b) the number of scales, c) the correlation between the scales, d) the ratio of the variances of the person parameters in different scales. Simulations have shown, that for 20 items in two scales with a moderate correlation between them 2000 iterations per chain are usually enough.

b) Order the chains based on:

$$\bar{L}_c = \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n}\sum_{p=1}^{N}\ln\left(\frac{\left(\exp\left(\sum_{k=1}^{M}\alpha_{ik}^{gc}\theta_{pk}^{gc}-\delta_i^{gc}\right)\right)^{X_{pi}}}{1+\exp\left(\sum_{k=1}^{M}\alpha_{ik}^{gc}\theta_{pk}^{gc}-\delta_i^{gc}\right)}\right), \tag{5.20}$$

where G denotes the number of iterations after the burn-in and superscripts g and c denote the value of a parameter at the g-th iteration in the c-th chain.

c) Select the best chain with the highest value of $\bar{L}_c$. This quantity is used to select the best chain because it allows one to choose the chain corresponding to the strongest mode among the chains.

d) Try to re-label the scales in the second best chain in such a way that the scales become almost the same as in the best chain. By "almost the same" we mean the following: in each scale the number of mismatching items (i.e., items which are assigned to this scale in the best chain, but to a different chain in the scale under consideration) cannot exceed 20% of the number of items in this scale. Continue with all other chains, until you arrive at a chain in which the scales cannot be re-labelled in such a way that the item partition into scales is almost the same as in the best chain. The results from the selected and re-labelled chains can be combined. For each item i and each scale k compute the posterior probability of this item to belong to this scale:

$$\hat{\pi}_{ik} = \frac{\sum_{c\in\mathbf{C}}\sum_g \alpha_{ik}^{g,c}}{|\mathbf{C}|G}, \tag{5.21}$$

where $\{\mathbf{C}\}$ denotes a set of selected chains. If for item i for neither of the scales $\hat{\pi}_{ik}$ is larger than 0.65, one can conclude that this item does not fit well in any

of the Rasch scales.

e) If there are no chains with the same partition of items into scales as in the best chain, then more chains with more iterations should be used. If consistent results are not obtained after running more chains, then either the algorithm can not handle this combination of parameters (N, n, M, $\mathbf{\Sigma}$), or it is a sign of model misfit: the test cannot be well modelled as a mixture of M Rasch scales. Note, that if an $(M-1)$-scale RM is a true model, then if M scales are used, it will be hardly possible to have a consistent partition of items into M scales.

5.5.2 Determining the number of scales

The MCMC algorithm described in the previous section requires the number of scales in the item set to be known. However, the value of M is generally not known and has to be chosen. Choosing the appropriate number of mixture components or the number of clusters (scales) is a complicated problem that is not yet fully solved (Frühwirth-Schnatter, 2006; McLachlan & Peel, 2000). In this article, we use two information criteria for choosing the model with an appropriate number of dimensions.

Once unmixing with M scales is finished, the item scale memberships are fixed to be equal to their posterior mode, denoted by $\hat{\boldsymbol{\alpha}}$. Given $\hat{\boldsymbol{\alpha}}$ the item difficulties and the covariance matrix are re-estimated using a data augmented Gibbs Sampler: First, initial values for the item difficulties (samples from $U(-2,2)$) and the covariance matrix (identity matrix) are specified. Second, for G iterations the individual person parameters, the covariance matrix and the item difficulties are subsequently sampled from their full conditional posterior distributions (see Steps 1-3 in Section 5.5.1). Since the item memberships are fixed, the posterior distribution is not multimodal and using one chain with a large number of dimensions is sufficient. Third, after discarding the first half of the iterations (burn-in), compute the expected a posteriori [EAP] estimates of the item difficulties and the covariance matrix:

$$\hat{\mathbf{\Sigma}} = \frac{1}{G/2} \sum_{g=G/2+1}^{G} \mathbf{\Sigma}^g, \tag{5.22}$$

$$\hat{\delta}_i = \frac{1}{G/2} \sum_{g=G/2+1}^{G} \delta_i^g, \forall i \in [1:n]. \tag{5.23}$$

The modified AIC (Akaike, 1974), and the BIC (Schwarz, 1978) are computed

as follows[2]:

$$\text{AIC} = -2\ln f(\mathbf{X} \mid \hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Sigma}}) + 2\left(\frac{M(M+1)}{2} + n + (M-1)n\right), \text{ and} \tag{5.24}$$

$$\text{BIC} = -2\ln f(\mathbf{X} \mid \hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Sigma}}) + \ln N\left(\frac{M(M+1)}{2} + n + (M-1)n\right). \tag{5.25}$$

The estimates $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\delta}}$ are used instead of the estimates based on the posterior in Equation 5.13, since if throughout the iterations the items move frequently across the scales, the EAP estimates based on the draws from Equation 5.13 would be less optimal and give a lower likelihood than $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\Sigma}}$. In the expression for the number of parameters, the first element is the number of freely estimated elements of $\boldsymbol{\Sigma}$, the second is the number of difficulty parameters and the third one is the number of freely estimated elements of $\boldsymbol{\alpha}$. With each extra scale there are an extra n elements to estimate for the items, since for each item it has to be decided whether it should be re-assigned to a new scale or not. The evaluation of the log-likelihood in Equations 5.24 and 5.25 involves integration over multidimensional space, which is done here through numerical integration with Gauss-Hermite quadrature (for details see Appendix B).

When choosing the number of scales, one should not only follow the above described procedure, but also consider the possible interpretations of the scales. Once a number of scales $\hat{M}$ is chosen using the information criteria, one should evaluate the solutions with $\hat{M}-1$, $\hat{M}$ and $\hat{M}+1$ scales from the substantive point of view. For example, given the context of the test, it might be reasonable to choose a smaller number of scales if it improves the interpretability of the scales, or choose a larger number of scales if they contain substantially different items.

5.6 Evaluation of the MCMC algorithm

In this section by means of a simulation study we show how well Rasch homogeneous subscales can be reconstructed using the MCMC algorithm and evaluate the performance of the modified AIC and BIC for selecting the appropriate number of scales[3]. The scales are correctly reconstructed, if for every item the posterior

[2]These are modifications because the original AIC and BIC are based on the maximum likelihood estimates. However, in our case the EAP estimates are very close to the maximum likelihood estimates since vague priors are used.

[3]In Appendix C three more simulation studies are presented, in which the performance of the MCMC algorithm is evaluated in more detail. The first simulation study deals with unmixing the scales representing substantively different abilities (multi-scale RM of Type 1). We also compare the performance of the MCMC algorithm with the method of hierarchical cluster analysis (Debelak & Arendasy, 2012), which also aims at constructing a set of scales that each fitting a RM. The second study illustrates how the algorithm performs when the scales measure the same ability and differ only in the discrimination of the items (multi-scale RM of Type 3).

Table 5.1: Results of choosing the number of scales: % of data sets in which the number of scales was chosen correctly ($\hat{M} = M$), was overestimated ($\hat{M} = M+1$), and underestimated ($\hat{M} = M-1$); % of data sets in which all items were classified correctly ($\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}$) given that $\hat{M} = M$.

Method		Condition					
		1	2a	2b	3	4	5
AIC	$\hat{M} = M$	98	100	100	100	100	100
	$\hat{M} = M+1$	2	0	0	0	0	0
	$\hat{M} = M-1$	-	0	0	0	0	0
BIC	$\hat{M} = M$	100	100	100	100	52	0
	$\hat{M} = M+1$	0	0	0	0	0	0
	$\hat{M} = M-1$	-	0	0	0	48	100
$\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}$		-	99	100	99	100	100

mode of its item membership is equal to the true item membership.

Data were simulated under a 1-, 2-, 3-, 4- and 5-scale RM. For the 2-scale RM, we considered two cases: one with different abilities measured (multi-scale RM of Type 1) and another with the two scales only differing in the discrimination parameter (multi-scale RM of Type 3). When $M > 2$, the simulated tests consisted both of scales differing only in the discriminative power, and of scales representing different abilities with a moderate correlation between them (multi-scale RM of Type 2). For every M, responses of 1000 persons to $10 \times M$ items (10 per scale) were simulated. Item difficulties were sampled from $U(-2\sum_k \alpha_{ik}\sigma_k, 2\sum_k \alpha_{ik}\sigma_k)$. The specification of each condition was the following:
1) $M = 1 : \sigma_1 = 1$;
2a) $M = 2 : \sigma_1 = \sigma_2 = 1, \rho_{1,2} = 0.5$;
2b) $M = 2 : \sigma_1 = 1, \boldsymbol{\theta}_{\cdot 2} = 2\boldsymbol{\theta}_{\cdot 1}$(implying that $\sigma_2 = 2$ and $\rho_{1,2} = 1$);
3) $M = 3 : \sigma_1 = \sigma_2 = 1, \rho_{1,2} = 0.5, \boldsymbol{\theta}_{\cdot 3} = 2\boldsymbol{\theta}_{\cdot 1}$;
4) $M = 4 : \sigma_1 = \sigma_2 = 1, \rho_{1,2} = 0.5, \boldsymbol{\theta}_{\cdot 3} = 2\boldsymbol{\theta}_{\cdot 1}, \boldsymbol{\theta}_{\cdot 4} = 2\boldsymbol{\theta}_{\cdot 2}$;
5) $M = 5 : \sigma_1 = \sigma_2 = \sigma_3 = 1, \rho_{1,2} = \rho_{1,3} = \rho_{2,3} = 0.5, \boldsymbol{\theta}_{\cdot 4} = 2\boldsymbol{\theta}_{\cdot 1}, \boldsymbol{\theta}_{\cdot 5} = 2\boldsymbol{\theta}_{\cdot 2}$;

In each condition, the MCMC algorithm was applied to 100 simulated data sets. The number of iterations per chain depended on the number of scales that were fitted and was equal to $M \times 500, \forall M \in [2:6]$. The modified AIC and the modified BIC (see Equations 5.24 and 5.25)were used for choosing the model with an appropriate number of scales out of the $(M-1)$-, M- and $(M+1)$-scale RM.

The results are presented in Table 5.1. The AIC showed very good performance, choosing the true number of scales in almost all data sets. The BIC underestimated the number of scales, when the tests were long (40 and 50 items)

The third study evaluates the autocorrelations in the Markov chain.

and the true number of scales was large (4 and 5). Therefore, we use the AIC in determining the number of scales in the NT2 exam. When the procedure selected the correct number of scales, then those scales were correctly reconstructed in more than 95% of the cases as can be seen from the last line of Table 5.1.

5.7 Choosing a scoring rule for the NT2 exam

5.7.1 Data

Data from the state exam of Dutch as a second language collected in July 2006 was used. The reading and listening parts of the NT2 exam consisted of 40 items each. However, six of the items were not taken for analysis because they were too easy (with proportions of correct responses larger than 0.85). The test was taken by 2425 persons. Responses of persons having more than 20% missing responses in one of the subtests were discarded (27 persons in total). The remaining missing values were considered as incorrect responses. The resulting sample size was $N = 2398$ and the test length was $n = 74$ (40 reading items and 34 listening items). The average proportion of correct responses to the items was equal to 0.67. The distribution of the number of correct responses had a mean of 49.74, a standard deviation of 12.24, a maximum of 74 and a minimum of 17.

The data set was randomly divided into two parts: a training set ($N = 1500$) on which the exploratory unmixing using the MCMC algorithm was carried out as was discussed in Section 5.5, and a testing set ($N = 898$) which was used for testing whether the scales identified in exploratory part are indeed Rasch scales, and testing hypotheses about the relations between the unmixed scales.

5.7.2 Unmixing Rasch scales

Three multi-scale RMs were fitted to the data: with two, three, and four scales, respectively. In each case, ten chains with $M \times 2000$ iterations each were used. The results of the unmixing are summarised in Table 5.2. While for the 2-scale and the 3-scale RMs all chains converged to the same partition of items into scales, in the case of the 4-scale RM only four chains converged to the same solution. The 3-scale RM had the lowest AIC value, therefore it was chosen as the best model.

In the three-scale RM 24 items were assigned to scale 1, 34 items were assigned to Scale 2, 13 items were assigned to Scale 3, and three items were not assigned to any scale, because for none of the scales the posterior probability of belonging to this scale ($\hat{\pi}_{ik}$) was above 0.65. All three scales included both reading and listening items: in Scale 1 there were 10 reading and 14 listening items, in Scale 2 there were 22 reading items and 12 listening items, and in Scale 3 there were 6 reading and 7 listening items.

Table 5.2: Results of unmixing Rasch scales in the Dutch as a foreign language test: the scales are ordered based on the value of $\hat{\sigma}_k^2$ from the largest to the smallest, the last number shows the number of items which did not belong to any of the scales ($\hat{\pi}_{ik} < 0.65, \forall k$)

Model	# items per scale	AIC	ΔAIC from the best model
2-scale RM	39/33/2	122494.2	41.5
3-scale RM	24/34/13/3	122452.7	0
4-scale RM	22/9/34/7/2	122608.7	156.0

The estimated covariance matrix was

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 1.67[1.48, 1.88] & 1.13[1.01, 1.25] & 0.71[0.63, 0.80] \\ \textit{0.96 [0.95,0.97]} & 0.83[0.74, 0.93] & 0.49[0.44, 0.55] \\ \textit{0.92 [0.88,0.95]} & \textit{0.90 [0.86,0.94]} & 0.36[0.31, 0.42] \end{bmatrix}, \tag{5.26}$$

where the elements below the diagonal (italicised) are the correlation coefficients, and the elements above the diagonal are the covariances, the 95% credible intervals for the estimates are given between brackets. The estimates of the correlations between the person parameters in the three scales were very high, therefore, a hypothesis about the relationship between the scales was formulated, namely that the three scales, in fact, measure the same ability, and the test can be scored with a weighted sumscore instead of a set of subscores. This hypothesis was tested on the second part of the data by selecting the best model out of the Type 1 model and the Type 3 model. Since three items did not belong to any of the three scales, in the following analysis only 71 items were used.

5.7.3 Cross-validation of the unmixed scales

Does the RM fit in the unmixed scales?

Identification of the three scales provided a hypothesis that we tested on the remaining part of the data, namely that the three scales are Rasch scales (without yet specifying whether these scale measure a single ability). We also tested a different hypothesis, which was formulated based on the background information: "the reading and the listening parts of the test form Rasch scales". Both hypotheses were tested by testing the fit of the RM in the subscales: 1) in the three subscales which resulted from the unmixing; 2) in the reading and the listening subscales. The fit of the RM model was tested using the LR-statistic. The RM was fitted to the three identified scales and to the listening and the reading scales using R package `eRm` (Mair & Hatzinger, 2007).

We did not expect a perfect fit of the RM to the complete scales (see lines 1,

Table 5.3: Fit of the RM in the three unmixed scales (before and after removing misfitting items), and in the reading and listening scales (in these two scales removing less than 10 items did not result in a reasonable fit) in the testing data set: LR-statistic

Scale	LR	*df*	*p*-value
Scale 1 (full scale: 24 items)	57.43	23	<0.005
Scale 1 (misfitting items removed: 21 items)	32.29	20	0.04
Scale 2 (full scale: 34 items)	49.59	33	0.03
Scale 2 (misfitting items removed: 32 items)	44.77	31	0.05
Scale 3 (full scale: 13 items)	39.91	12	<0.005
Scale 3 (misfitting item removed: 11 items)	14.54	10	0.15
Reading scale (38 items)	200.42	37	<0.005
Listening scale (33 items)	181.67	32	<0.005

3, and 5 of Table 5.3), because if there were some misfitting items among the 74 items used in the exploratory unmixing they would have been assigned to one of the scales, where they fit relatively better, but still badly in absolute terms. That is why, for example, we go from 24 to 21 items in scale 1. The analysis presented in Table 5.3 helped to identify these misfitting items. If one would discard three misfitting items in the first scale, two in the second and two in the third, the RM would have a reasonable fit in all three scales. However, when the reading and the listening scales were considered, discarding of a small number (less than ten) of misfitting items would not result in a reasonable fit of the RM.

Three different abilities or one?

In cross-validation, we tested whether a multi-scale RM of Type 1 or of Type 3 fitted the test consisting of 71 items best. First, the two models with fixed scales were fitted to the training data with 71 items. For the model of Type 1, the estimates of the item difficulties and the covariance matrix were obtained (denoted by $\hat{\boldsymbol{\delta}}_{type1}$ and $\hat{\boldsymbol{\Sigma}}_{type1}$). As has been mentioned in Section 5.4.1, the model of Type 3 is equivalent to a unidimensional model with a standard normal distribution of ability and three item clusters with discriminations equal to σ_1, σ_2, and σ_3. Therefore, this unidimensional model with fixed scales has been fitted to the data (see Appendix D) and estimates of the item difficulties and the three discrimination parameters were obtained (denoted by $\hat{\boldsymbol{\delta}}_{type3}$ and $\hat{\boldsymbol{\sigma}}_{type3}$).

Second, the fit of the models of Type 1 and Type 3 to the testing data set (denoted by $\mathbf{X}_{test}$) with the parameters fixed at the estimates obtained in the training data was evaluated. Log-likelihood of both models was computed (see

Table 5.4: Two scoring rules (based on two unweighted subscores and based on one weighted sumscore) for six persons

p	$\sum_{i\in\{R\}} X_{pi}$	$\sum_{i\in\{L\}} X_{pi}$	Decision	$\sum_i \sum_k \alpha_{ik}\sigma_k X_{pi}$	Decision
1	27	27	pass	53.42	pass
2	31	29	pass	59.40	pass
3	16	24	fail	38.66	fail
4	20	9	fail	27.95	fail
5	23	27	fail	50.04	pass
6	25	20	pass	42.85	fail

Appendix B):

$$\ln\left(f\left(\mathbf{X}_{test}\,|\,\hat{\boldsymbol{\delta}}_{type1},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\Sigma}}_{type1}\right)\right) = -34776.93 \tag{5.27}$$

$$\ln\left(f\left(\mathbf{X}_{test}\,|\,\hat{\boldsymbol{\delta}}_{type3},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\sigma}}_{type3}\right)\right) = -34767.83 \tag{5.28}$$

The Type 3 model had better fit, which suggested that all three scales measure the same dimension and that a weighted sumscore is the best scoring rule for this particular Dutch language ability test. The estimated weights were equal to 1.30, 0.89, and 0.56 in the three scales, respectively.

Does it make a difference?

Finally, we investigated whether using the chosen scoring rule

$$\sum_i (1.30\alpha_{i1} + 0.89\alpha_{i2} + 0.56\alpha_{i3})X_{pi} \tag{5.29}$$

leads to different decisions about the persons passing or failing the test compared to the decision based on unweighted sumscores on the set of reading items, denoted by $\{R\}$, and on the set of listening items, denotes by $\{L\}$.

Suppose, the original pass-fail criterion is that a person passes the test if he/she has at least 25 correct responses on the reading test and at least 20 correct responses on the listening test. This decision criterion results in 412 persons from the testing set passing the test. A cut-off value for the weighted sumscore leading to the same number of students passing the test is 48.21. Table 5.4 shows the application of the two scoring rules to six persons from the testing set. It can be seen that for some persons the decisions based on two scoring rule match each other, while for others they do not. In we consider a two-by-two classification table for the pass/fail decision according to the original rule and the pass/fail

decision according to the new scoring rule, then we discover that 31 persons who fail the test according to the original rule would pass it according to the new rule and vice versa. Hence, we have shown that a scoring rule chosen based on the empirical data and therefore representing the data structure better leads to a different pass/fail decision for 62 persons (7% of the testing data set) compared to non-compensatory scoring rule based on the two unweighted subscores.

5.8 Discussion

In this article we presented a novel solution to the problem of choosing a scoring rule for the test. Using the exploratory unmixing algorithm in the state examination of Dutch as a foreign language three Rasch scales were identified. Each of these scales consisted both of reading and listening items. Further analysis showed that the scale represent the same substantive dimension and the scales differ only in the discriminative power of the items. That is the test can be scored with a weighted score with three different weights. The fact that the reading and the listening items were not classified in separate scales is not surprising if the kind of tasks that these items represent are considered: Both the reading and the listening items require understanding of information that is communicated through language (i.e., passive language skills).

The scoring rule that has been chosen for the NT2 exam is not a conjunction of reading and listening but a compensatory rule based on a longer test which makes the score more reliable. Hence, the confirmatory part of our method can be used to evaluate whether using weighted sumscore instead of the set of scores does not threaten the validity of the measurement while improving the reliability of the scores. In the NT2 exam application it turned out that using the weighted sumscore as the scoring rule better represents the structure in the data than the set of unweighted sumscores for the reading and the listening parts, and it makes a difference for 7% of the sample.

Identification of scales with different levels of discrimination can give start to further studying of the item characteristics that make them discriminate worse, and might lead to item revisions. In this way, out method may serve as a diagnostic instrument for detecting poorly performing items and improving them.

As has been observed in our example of the NT2 exam, in practice some of the items might not be assigned to any of the scales, because they are assigned to different scales in different chains, as we have seen in the example. This means that for these items the model does not fit very well. This can be caused by within-item multidimensionality of these items, that is, when $\boldsymbol{\alpha}_i$ of the item does not have a simple structure. It is possible to test this hypothesis by comparing the constrained multi-scale RM with the multidimensional model (see Equation 5.2) in which some of the $\boldsymbol{\alpha}_i$ are freely estimated. Thus, considering the model as

a constrained version of a general multidimensional model makes it possible to further investigate in which ways the model can be improved by allowing some items to load on more than one dimension.

Theoretically, OPLM is an elegant and attractive model. From a practical point of view, however, the assumption that researchers can cluster items together on the basis of their discriminatory power is quite often unrealistic, as is the assumption that clusters of items only differ with respect to discriminatory power. The new model retains the theoretical elegance of the OPLM model, but provides substantive researchers with a tool for the automatic clustering of items. At the same time, with the new model we can relax the stringent assumption in the OPLM model that item clusters only differ with respect to their discriminatory power. The new model provides the researcher with important information that can be used to uncover in what respect Rasch homogeneous scales differ from one another.

5.9 Appendices

Appendix A: Identification of the multi-scale RM

The main question about the identification of the multi-scale RM is whether the partition of the items into scales ($\boldsymbol{\alpha}$) is unique if the model holds and the set of items indeed consists of a number of Rasch scales. The trivial different item membership matrix $\boldsymbol{\alpha}^*$ giving the same distribution of the data $p(\mathbf{X})$ as $\boldsymbol{\alpha}$ can be achieved by label switching. This can be easily recognised and should not be treated as a different solution. Therefore, to prove that the model is identifies one should prove that there is no set of parameters $\boldsymbol{\alpha}^*$, such that there exist at least two items i and j, such that $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_j$ and $\boldsymbol{\alpha}_i^* \neq \boldsymbol{\alpha}_j^*$.

For simplicity, let us consider a model with only two scales. Let $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_j = \boldsymbol{\alpha}_i^* = \{1, 0\}$ and $\boldsymbol{\alpha}_j^* = \{0, 1\}$. Consider the ratio of probabilities: $\frac{p(X_i=1,X_j=0,\mathbf{X}^{(i,j)})}{p(X_i=0,X_j=1,\mathbf{X}^{(i,j)})}$, where $\mathbf{X}^{(i,j)}$ is a vector of responses to all items except i and j. According to the multi-scale RM:

$$\frac{p(X_i=1, X_j=0, \mathbf{X}^{(i,j)})}{p(X_i=0, X_j=1, \mathbf{X}^{(i,j)})} = \frac{\int \frac{\exp(\theta_1-\delta_i)}{(1+\exp(\theta_1-\delta_i))(1+\exp(\theta_1-\delta_j))} p(\mathbf{X}^{(i,j)}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}{\int \frac{\exp(\theta_1-\delta_j)}{(1+\exp(\theta_1-\delta_i))(1+\exp(\theta_1-\delta_j))} p(\mathbf{X}^{(i,j)}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}} = \exp(\delta_j - \delta_i), \quad (5.30)$$

that is, it is constant across all response vectors $\mathbf{X}^{(i,j)}$. Now we will consider whether this ratio can be constant if the items i and j do not belong to the same

scale:

$$\frac{p(X_i=1,X_j=0,\mathbf{X}^{(i,j)})}{p(X_i=0,X_j=1,\mathbf{X}^{(i,j)})} = \int \frac{p(X_i=1,X_j=0,\mathbf{X}^{(i,j)}\,|\,\boldsymbol{\theta}^*)}{p(X_i=0,X_j=1,\mathbf{X}^{(i,j)})} f(\boldsymbol{\theta}^*)\,\mathrm{d}\boldsymbol{\theta}^* =$$
$$= \int \frac{p(X_i=1,X_j=0,\mathbf{X}^{(i,j)}\,|\,\boldsymbol{\theta}^*)}{p(X_i=0,X_j=1,\mathbf{X}^{(i,j)}\,|\,\boldsymbol{\theta}^*)} f(\boldsymbol{\theta}^*\,|\,X_i=0,X_j=1,\mathbf{X}^{(i,j)})\,\mathrm{d}\boldsymbol{\theta}^* =$$
$$= \int \frac{\exp(\theta_1^*-\delta_i^*)}{\exp(\theta_2^*-\delta_j^*)} f(\boldsymbol{\theta}^*\,|\,X_i=0,X_j=1,\mathbf{X}^{(i,j)})\,\mathrm{d}\boldsymbol{\theta}^* =$$
$$= \exp(\delta_j^*-\delta_i^*)\int (\exp(\theta_1^*-\theta_2^*) f(\boldsymbol{\theta}^*\,|\,X_i=0,X_j=1,\mathbf{X}^{(i,j)})\,\mathrm{d}\boldsymbol{\theta}^* =$$
$$= \exp(\delta_j^*-\delta_i^*)\mathcal{E}\left(\exp(\Theta_1^*-\Theta_2^*)\,|\,X_i=0,X_j=1,\mathbf{X}^{(i,j)}\right). \quad (5.31)$$

The conditional expectation of $\exp(\theta_1^*-\theta_2^*)$ has to take the same values for different $\mathbf{X}^{(i,j)}$. Let us consider two response vectors with $\sum\limits_{k\neq i,j}\alpha_{k1}^*X_k = m_1$ and $\sum\limits_{k\neq i,j}\alpha_{k1}^*X_k = m_1+1$, both vectors have $\sum\limits_{k\neq i,j}\alpha_{k2}^*X_k = m_2$. Since, a vector of subscores $\{\sum_i \alpha_{i1}^*X_i, \sum_i \alpha_{i2}^*X_i\}$ is a sufficient statistic, the posterior distributions of $\boldsymbol{\theta}^*$ are stochastically ordered (Migrom, 1981):

$$f(\boldsymbol{\theta}^*\,|\,X_i=0,X_j=1,\sum_{k\neq i,j}\alpha_{k1}^*X_k=m_1,\sum_{k\neq i,j}\alpha_{k2}^*=m_2) <_{st}$$
$$f(\boldsymbol{\theta}^*\,|\,X_i=0,X_j=1,\sum_{k\neq i,j}\alpha_{k1}^*X_k=m_1+1,\sum_{k\neq i,j}\alpha_{k2}^*=m_2). \quad (5.32)$$

The function $\exp(\theta_1^*-\theta_2^*)$ is a non-decreasing function in θ_1^*. If this function is also finite, then

$$\mathcal{E}\left(\exp(\Theta_1^*-\Theta_2^*)\,|\,X_i=0,X_j=0,\sum_{k\neq i,j}\alpha_{k1}^*X_k=m_1,\sum_{k\neq i,j}\alpha_{k2}^*=m_2\right) <$$
$$\mathcal{E}\left(\exp(\Theta_1^*-\Theta_2^*)\,|\,X_i=0,X_j=1,\sum_{k\neq i,j}\alpha_{k1}^*X_k=m_1+1,\sum_{k\neq i,j}\alpha_{k2}^*=m_2\right). \quad (5.33)$$

Therefore, the ratio of probabilities in (5.30) can not be constant across different $\mathbf{X}^{(i,j)}$, when items i and j do not belong to the same scales. Hence, $\boldsymbol{\alpha}$ is the only possible set of item memberships, meaning that the partition of items into Rasch scales can be identified from the data. Note, that with infinite data the ratios of probabilities in (5.30) can be directly observed. For every pair of items it is then possible to say whether they belong to the same scale, if the ratios are constant

across $\mathbf{X}^{(i,j)}$, or not, which gives a unique partition of items into scales.

If the item memberships are identified, then the item difficulties and the variances of abilities can identified (San Martin & Rolin, 2013), as you can treat each Rasch scale as a separate marginal RM, and consequently the covariances between the abilities are also identified.

Appendix B: Approximating the likelihood of the multi-scale RM

To compute the likelihood of the multi-scale RM we decompose it in the following way:

$$f(\mathbf{X} \mid \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) = \prod_p f(\mathbf{X}_p \mid \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) =$$
$$\prod_p \int \prod_{i \in \{i \mid \alpha_{i1}=1\}} f(x_{pi} \mid \delta_i, \theta_1) p(\theta_1 \mid \boldsymbol{\Sigma}) \times \cdots \times$$
$$\int \prod_{i \in \{i \mid \alpha_{ik}=1\}} f(x_{pi} \mid \delta_i, \theta_k) p(\theta_k \mid \boldsymbol{\Sigma}, \theta_1, \ldots, \theta_{k-1}) \times \cdots \times$$
$$\int \prod_{i \in \{i \mid \alpha_{iM}=1\}} f(x_{pi} \mid \delta_i, \theta_M) p(\theta_M \mid \boldsymbol{\Sigma}, \theta_1, \ldots, \theta_{M-1}) d\theta_M \ldots d\theta_k \ldots d\theta_1. \quad (5.34)$$

The conditional distributions of person parameters in scales $k > 1$ are normal distributions with the mean μ_k^* and the variance σ_k^{*2}. The likelihood is approximately equal to

$$f(\mathbf{X}_p|\boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \approx (\pi)^{-\frac{M}{2}} \prod_{k=1}^{M} \sum_{j_k=1}^{J_k} w_{j_k} \prod_{i \in \{i \mid \alpha_{ik}=1\}} f(x_{pi}|\delta_i, \theta_k = \sqrt{2}\sigma_k^* y_{j_k} + \mu_k^*), \quad (5.35)$$

where $\mu_1^* = 0$, $\sigma_1^* = \sigma_1$; $\mathbf{y}$ and $\mathbf{w}$ are the Gauss-Hermite nodes and weights, and the number of weights per scale is determined by the variance of the conditional distribution: $J_k = \lfloor 20\sigma_k^* \rfloor$.

Appendix C: Additional simulation studies

Simulation study 1

In the first simulation study we compare the performance of the new algorithm with the method of hierarchical cluster analysis (Debelak & Arendasy, 2012), which also aims at constructing a set of scales that each fitting a RM. Here, we briefly describe their method. At the initial step, all combinations of three items $\mathcal{O}_3$ from the item set $\mathcal{O}$ are considered. The fit of the model to each set is evaluated using the R_{1c} statistic (Glas, 1988), and the item set that fits the RM

best is selected. The R_{1c} is based on the discrepancy between the observed and expected probabilities of a correct response to the items within groups of persons separated based on the number of correct responses.

After that, the selected set is expanded. At every step, if $\mathcal{A}_m$ is the set of already selected items, from all combinations of $m+1$ items, such that $\mathcal{A}_m \subset \mathcal{O}_{m+1}$, the best fitting set A_{m+1} is selected. The procedure continues until none of the possible sets would fit the RM or until there are no items left. It is then repeated for the items that have not been selected in the first scale to construct the second scale, and so on.

In order to obtain a direct comparison of the two methods, we applied our algorithm to data sets with the same parameters as those that were used in the study of Debelak and Arendasy. Their data sets consisted of two equal sized subscales each fitting a RM. Item difficulties were sampled from $N(0, \tau_1^2)$ in scale 1 and $N(0, \tau_2^2)$ in scale 2, and person parameters were sampled from a multivariate normal distribution with a zero mean vector and a covariance matrix $\mathbf{\Sigma}$. From the total of 180 conditions used by Debelak and Arendasy we chose only the conditions with extreme values for the varied parameters, since that should be sufficient for a comprehensive comparison of the two methods. We included only the conditions with non-zero correlation between the person parameters ($\rho = .5$), since if it is highly unlikely for an educational test to measure uncorrelated abilities. The selected extreme values were: 1) test length: short ($n = 10$) or long ($n = 50$), 2) sample size: small ($N = 250$) or large ($N = 1000$), 3) standard deviations of the person parameters and the item parameters: small in both scales: $\sigma_1 = \sigma_2 = 1$ and $\tau_1 = \tau_2 = 0.5$ (condition AA), large in both scales: $\sigma_1 = \sigma_2 = 2.5$ and $\tau_1 = \tau_2 = 1.5$ (condition DD), or small in one dimension and large in the other: $\sigma_=1$ and $\sigma_2 = 2.5$, $\tau_1 = 0.5$ and $\tau_2 = 1.5$ (condition AD).

For each of the 12 conditions, the algorithm with 10 chains of 2000 iterations each (including 1000 iterations of the burn-in) was applied to 100 simulated data sets. In Table 5.5, the percentage of correct reconstructions of the scales by the MCMC algorithm and the hierarchical clustering algorithm are shown. The results of the hierarchical cluster analysis are taken from the original publication.

The new MCMC algorithm performs better in all 12 conditions, with the difference being dramatic in many of the cells. 100% correct results are obtained in 10 conditions. The percentage of correct results is the lowest when the sample size is small, and the variances of the person parameters are equal and low (77% for the short scales and 92% for the long scales). The results for these conditions did not improve after increasing the number of iterations. Thus, for optimal performance of the algorithm the scales should not be too short and/or the sample size should not be too small.

Table 5.5: Comparison of results of the MCMC algorithm and the partial hierarchical clustering algorithm (HCA): percentage of correct scale reconstructions; the results for the HCA are cited from Debelak and Arendasy (2012)

		HCA			MCMC		
N	n	AA	AD	DD	AA	AD	DD
250	10	0.42	10.0	65.34	77	100	100
	50	0.02	9.56	87.01	87	100	100
1000	10	29.17	10.15	92.26	100	100	100
	50	10.84	10.15	82.99	100	100	100

Simulation study 2

In the Introduction we described the OPLM as a special case of the multi-scale RM. If the item memberships are not pre-specified, then our approach enables the identification of scales fitting the RM, that differ only in the discriminative power of the items in them. In the second simulation study, we show the effectiveness of the algorithm for unmixing Rasch scales when the scales are perfectly correlated ($\rho = 1$) and the relations between standard deviations of person parameters are the same as in the OPLM. The following parameters were varied in the simulations:

1. Sample size: $N = 500, 1000, 2000, 5000$;
2. Test length: $n = 2 \times 5, 2 \times 10$;
3. Ratio between standard deviations of person parameters:
   - $\frac{\sigma_1}{\sigma_2} = \frac{1}{2}$ corresponding to $a_1 = 1$, $a_2 = 2$ in the OPLM;
   - $\frac{\sigma_1}{\sigma_2} = \frac{2}{3}$ corresponding to $a_1 = 2$, $a_2 = 3$ in the OPLM;
   - $\frac{\sigma_1}{\sigma_2} = \frac{3}{4}$ corresponding to $a_1 = 3$, $a_2 = 4$ in the OPLM.

In all conditions $\sigma_1 = 1$. Item parameters were sampled from a uniform distribution: $U(-2\sigma_1, 2\sigma_1)$ for items in scale 1 and $U(-2\sigma_2, 2\sigma_2)$ for items in scale 2. Person parameters were sampled from a multivariate normal distribution. For every condition the algorithm was applied to 100 simulated data sets.

One can expect that the closer to 1 the ratio between σ_1 and σ_2 becomes, the more difficult it is to separate the scales and the more iterations per chain are needed. With this in mind, the following numbers of iterations were chosen for the simulation conditions: 2000 iterations (including 1000 iterations burn-in) for the data sets with $\frac{\sigma_1}{\sigma_2} = \frac{1}{2}$, 3000 iterations (including 1500 iterations burn-in) for the data sets with $\frac{\sigma_1}{\sigma_2} = \frac{2}{3}$ and 4000 iterations (including 2000 iterations burn-in) for the data sets with $\frac{\sigma_1}{\sigma_2} = \frac{3}{4}$.

Table 5.6: Results of unmixing the scales which are perfectly correlated

	$n = 2 \times 5$			$n = 2 \times 10$		
N	$\frac{\sigma_1}{\sigma_2} = \frac{1}{2}$	$\frac{\sigma_1}{\sigma_2} = \frac{2}{3}$	$\frac{\sigma_1}{\sigma_2} = \frac{3}{4}$	$\frac{\sigma_1}{\sigma_2} = \frac{1}{2}$	$\frac{\sigma_1}{\sigma_2} = \frac{2}{3}$	$\frac{\sigma_1}{\sigma_2} = \frac{3}{4}$
Percentage of correct scale reconstruction						
500	80	36	23	84	26	16
1000	96	40	29	98	54	19
2000	98	70	41	99	94	50
5000	90	56	40	99	99	72
Percentage of correctly classified items						
500	96.50	76.00	64.50	97.15	83.35	64.75
1000	99.60	80.60	66.80	99.90	91.50	74.35
2000	99.30	91.40	76.70	99.95	98.75	91.55
5000	96.60	85.80	68.80	99.95	99.95	96.50

The results, both on the scale level (percentage of correct scale reconstruction) and the item level (percentage of correctly classified items in all data sets), are presented in Table 5.6.

As expected, for every sample size and test length the larger the difference between variances of person parameters, the more correct results are obtained. For the short scales, increasing the sample size at first improves scale reconstruction and then makes it worse because of slow mixing of the Markov chain. If both scales consisted of 10 items, then good results (more than 80% of correct scale reconstruction and more than 95% of correctly classified items) were obtained for all sample sizes when the ratio between standard deviations was equal to $1/2$, and for large sample sizes ($N = 2000, 5000$) when the ratio between standard deviations was equal to $2/3$. The algorithm performs worse when the scales are short: good results were obtained only for the condition with $\sigma_1/\sigma_2 = 1/2$.

If the scales are perfectly correlated, the algorithm performs optimally, if the scales are not too short (about 10 items), and the variances of person parameters are not too close to each other ($\frac{\sigma_1}{\sigma_2} \in [\frac{2}{3} : \frac{1}{2}]$).

Operational characteristics of the algorithm

We did a small simulation to evaluate the autocorrelation of the Markov chain. We examined how operational characteristics change

1. for a fixed test length ($n = 20$) and the number of scales ($M = 2$) if the sample size changes ($N = 250, 500, 1000, 2000, 4000$),

2. for a fixed sample size ($N = 1000$) and number of scales ($M = 2$) if the

Table 5.7: Autocorrelation of the Markov chain at lag 1 and 20 for different sample sizes (N), number of items (n), and number of scales (M) averaged for item difficulties (δ) and elements of the covariance matrix ($\mathbf{\Sigma}$)

N	n	M	δ		$\mathbf{\Sigma}$	
			lag 1	lag 20	lag 1	lag 20
250	20	2	.90	.23	.68	.06
500	20	2	.83	.09	.65	.01
2000	20	2	.70	.02	.66	.04
4000	20	2	.69	.04	.68	.07
1000	10	2	.78	.03	.84	.06
1000	20	2	.77	.05	.67	-.10
1000	40	2	.75	.05	.47	.02
1000	60	2	.75	.05	.33	-.03
1000	60	3	.75	.03	.42	.00
1000	60	4	.75	.02	.47	.00
1000	60	5	.75	.03	.57	-.01

number of items per scale and therefore the total test length changes ($n = 10, 20, 40, 60$),

3. for a fixed sample size ($N = 1000$) and test length ($n = 60$) if the number of scales and, therefore, the number of items per scale changes ($M = 2, 3, 4, 5$).

These conditions are summarised in Table 5.7. For illustrative purposes, one data set was simulated and analysed in each condition. The person parameters were sampled from a multivariate normal distribution with a zero mean vector, all variances were equal to 1 and all covariances were equal to 0.5. Item difficulties of items in each scale were equally distanced between -2 and 2. One chain with $M \times 1000$ iterations was used. Autocorrelation was computed for each parameter after first half of the iterations was discarded and it was averaged for item difficulties and elements of the covariance matrix. In Table 5.7 autocorrelations at lag 1 and 20 are presented.

Autocorrelation of the variance and covariance parameters decreases if the test length increases and if the number of items per scale increases. For obtaining stable estimates of the covariance matrix very long chains are needed. For example, when for a short test ($n = 10$) with 2 scales and medium sample size ($N = 1000$), an independent sample can be obtained only at lag 50, therefore 10 chains with 5000 iterations each are needed to obtain an independent sample of size 1000. For all other conditions the length of the chains that is needed is roughly twice shorter, because autocorrelation is almost zero at lag 20.

Autocorrelation of the item difficulty parameters decreases when the sample size increases. If test length is fixed but the number of scales is varied or if the number of items in the test increases, autocorrelation of item difficulties hardly changes. At lag 1 the autocorrelation is always high ($> .70$), but it is usually satisfactory ($< .1$) after lag 20.

The larger the sample size, the higher is the global mode of the posterior distribution relative to all other local modes and the higher is the probability of an item to belong to the correct scale than to an incorrect one. Therefore, increasing the sample size until some point helps to recover item membership parameters. But at the same time each mode becomes more peaked which makes mixing of the Markov chain slower, because the probability of an item to ever leave the scale to which it was assigned in one of the iterations after all other parameters were optimised for this partition of items into scales is close to zero. Therefore, for very large samples ($N > 5000$) unmixing becomes hardly possible in a reasonable number of iterations. This effect is present more, when the variances of person parameters in two scales are almost the same. Therefore, applications of the algorithm have to restricted to be moderate sample sizes.

Appendix D: Estimation of the model with fixed correlation parameters

Here we describe in detail the estimation of the multi-scale RM of Type 2 and Type 3 with fixed item scale memberships. We will discuss the estimation of the multi-scale RM of Type 2 on the example a model with $M > 2$ scales two of which are perfectly correlated. The algorithm can be extended to cases when there are more perfectly correlated scales.

As has been mentioned in Section 5.4.1. the multi-scale RM with M scales two of which are perfectly correlated with the covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2 & \sigma_1\rho_3\sigma_3 & \dots & \sigma_1\rho_M\sigma_M \\ \sigma_1\sigma_2 & \sigma_2^2 & \sigma_2\rho_3\sigma_3 & \dots & \sigma_2\rho_M\sigma_M \\ \sigma_1\rho_3\sigma_3 & \sigma_2\rho_3\sigma_3 & \sigma_3^2 & \dots & \sigma_{3M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_1\rho_M\sigma_M & \sigma_2\rho_M\sigma_M & \sigma_{3M} & \dots & \sigma_M^2 \end{bmatrix} \tag{5.36}$$

is equivalent to a $(M-1)$-dimensional model with the covariance matrix:

$$\boldsymbol{\Sigma}^* = \begin{bmatrix} \sigma_1^2 & \sigma_1\rho_3\sigma_3 & \dots & \sigma_1\rho_M\sigma_M \\ \sigma_1\rho_3\sigma_3 & \sigma_3^2 & \dots & \sigma_{3M} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1\rho_M\sigma_M & \sigma_{3M} & \dots & \sigma_M^2 \end{bmatrix}. \tag{5.37}$$

In the latter model Rasch scales are in $M-2$ dimensions and one dimension has two item clusters with different discriminations, equal to 1 in the first cluster and to $\sigma_2^* = \frac{\sigma_2}{\sigma_1}$ in the second cluster. Below we describe how to sample from the posterior distribution of the parameters of the collapsed model:

$$p(\boldsymbol{\delta}, \boldsymbol{\Sigma}^*, \sigma_2^* \,|\, \mathbf{X}, \hat{\boldsymbol{\alpha}}). \tag{5.38}$$

The same as in the original algorithm described in Section 5.5.1 at the beginning of each iteration the individual person parameters are sampled to simplify the full conditional posteriors of other parameters. The inverse-Wishart prior with $M+1$ degrees of freedom and a scale parameter $\mathbf{I}_{M-1}$. Improper priors are used for the difficulty parameters $p(\delta_i) \propto 1$; and for the discrimination in the second cluster of dimension 1: $p(\sigma_2^*) \propto \frac{1}{\sigma_2^*}$. Initial values are specified in the same way as in the original algorithm.

Step 1. For each person p sample $\theta_{p3}, \ldots, \theta_{pM}$ from their full conditional posterior distributions, which is done in the same way as in the original algorithm.

Step 2. For each person p sample θ_{p1} from

$$p(\theta_{p1} \,|\, \mathbf{X}, \boldsymbol{\delta}, \hat{\boldsymbol{\alpha}}, \sigma_2^*, \boldsymbol{\Sigma}^*, \theta_{p3}, \ldots, \theta_{pM}). \tag{5.39}$$

Similarly to sampling the parameters in the other dimensions we draw a candidate value from:

$$\theta^* = p(\theta_{p1} \,|\, \boldsymbol{\Sigma}^*, \theta_{p3}, \ldots, \theta_{pM}), \tag{5.40}$$

generate responses to the items in dimension 1 using:

$$\Pr(X_i^* = 1) = \begin{cases} \frac{\exp(\theta^* - \delta_i)}{1+\exp(\theta^* - \delta_i)}, & \text{if } \hat{\alpha}_{i1} = 1, \\ \frac{\exp(\sigma_2^* \theta^* - \delta_i)}{1+\exp(\sigma_2^* \theta^* - \delta_i)}, & \text{if } \hat{\alpha}_{i2} = 1, \end{cases} \tag{5.41}$$

and compute the sumscore $X_+^* = \sum_i (\hat{\alpha}_{i1} + \hat{\alpha}_{i2}\sigma_2^*) X_i^*$; but unlike the original algorithm values are not sampled until the simulated sumscore is equal to the observed sumscore, because the number of possible values of the weighted sumscore increases dramatically when different weights are used. Instead, exchange variable algorithm is used (Marsman et al. 2015), that is the candidate value if accepted with a probability:

$$\Pr(\theta_{p1} \to \theta^*) = \min\left(1, \exp\left((\theta^* - \theta_{p1})\left(\sum_i (\hat{\alpha}_{i1} + \hat{\alpha}_{i2}\sigma_2^*) X_{pi} - X_+^*\right)\right)\right), \tag{5.42}$$

otherwise the current value is retained.

Step 3. Sample the covariance matrix $\boldsymbol{\Sigma}^*$ given the person parameters in all dimensions, which is done in the same way as in the original algorithm.

Step 4. For each item i sample δ_i from its full conditional posterior, which is

done analogously to the Step 3 in the original algorithm.

Step 5. Sample σ_2^* from

$$p(\sigma_2^* \mid \mathbf{X}, \boldsymbol{\delta}, \hat{\boldsymbol{\alpha}}, \boldsymbol{\theta}, \boldsymbol{\Sigma}^*) \propto \frac{1}{\sigma_2^*} \prod_p \prod_{i \in \{i:\alpha_{i2}=1\}} \frac{\exp(\sigma_2^* \theta_{p1} - \delta_i)}{1 + \exp(\sigma_2^* \theta_{p1} - \delta_i)}, \tag{5.43}$$

which is done using a Metropolis-Hastings algorithm (Metropolis et al., 1953) with a log-normal distribution with the logarithm of the current value as the mean and the standard deviation equal to 0.01.

In cases when there are more perfectly correlated scales than two, the covariance matrix of the collapsed model has less dimensions than $M - 1$ and there are more discrimination parameters (like σ_2^*) to be sampled. In the case of the model of Type 3, it collapses to a unidimensional model with a standard normal distribution for the person parameters and discriminations $\sigma_1, \sigma_2, \ldots, \sigma_M$. The sampling scheme is then the following:

Step 1: For each person p sample θ_p from $p(\theta_p \mid \mathbf{X}_p, \boldsymbol{\delta}, \hat{\boldsymbol{\alpha}}, \sigma_1, \ldots, \sigma_M)$: Sample a candidate value from $N(0, 1)$ and generate a response vector using

$$\Pr(X_i^* = 1) = \frac{\exp\left(\sum_{k=1}^M \hat{\alpha}_{ik} \sigma_k \theta^* - \delta_i\right)}{1 + \exp\left(\sum_{k=1}^M \hat{\alpha}_{ik} \sigma_k \theta^* - \delta_i\right)}. \tag{5.44}$$

The candidate value is accepted with the probability:

$$\Pr(\theta_p \to \theta^*) = \min\left(1, \exp\left((\theta^* - \theta_p)\left(\sum_i \sum_k \alpha_{ik} \sigma_k (X_{pi} - X_i^*)\right)\right)\right). \tag{5.45}$$

Step 2: For each item i sample item difficulty δ_i from its full conditional posterior distribution using Metropolis-Hastings algorithm.

Step 3: For each scale k sample the item discrimination σ_k from its full conditional posterior, analogously to sampling σ_2^* in Equation 5.43.

Chapter 6

Using expert knowledge for test linking

[1] **Abstract.** Linking and equating procedures are used to make the results of different test forms comparable. In the cases where no assumption of random equivalent groups can be made some form of linking design is used. In practice the amount of data available to link the two tests is often very limited due to logistic and security reasons, which affects the precision of linking procedures. This study proposes to enhance the quality of linking procedures based on sparse data by using Bayesian methods which combine the information in the linking data with background information captured in informative prior distributions. We propose two methods for the elicitation of prior knowledge about the difference in difficulty of two tests from subject-matter experts and explain how these results can be used in the specification of priors. To illustrate the proposed methods and evaluate the quality of linking with and without informative priors, an empirical example of linking primary school mathematics tests is presented. The results suggest that informative priors can increase the precision of linking without decreasing the accuracy.

Keywords: elicitation, expert knowledge, informative priors, test equating, test linking.

[1]This chapter is under review with Psychological Methods as Bolsinova, M., Hoijtink, H., Vermeulen, J.A., & Béguin, A. Using expert knowledge for test linking. Author contributions: B.M., B.A., and H.H. designed the research, B.M performed the research, B.M. and V.J.A. collected expert data, B.M. wrote the paper, H.H., B.A. and V.J.A. provided feedback on the manuscript.

6.1 Introduction

If different test forms of an educational test measuring the same ability are administered to different populations of students (e.g., from different years), their results are not directly comparable because of the differences in the difficulty of the tests and the differences in the ability in the populations. Linking and equating techniques are ways to make the scores on the tests comparable. For linking a new test form to scores of the reference test form different linking designs can be used (Angoff, 1971; Wright & Stone, 1979; Lord, 1980; Petersen, Kolen, & Hoover, 1989; Kolen & Brennan, 2004).

In high-stakes testing (e.g., examinations) different test forms often do not have items in common due to security reasons. If the forms are administered under the assumption of non-equivalent groups it is necessary to collect additional data to link the different test forms (Mittelhaëuser, Béguin, & Sijtsma, 2015). Most commonly a type of anchor test is used, but the administration of anchor tests under appropriate conditions is challenging and expensive (Keizer-Mittelhaëuser, 2014). In this article we consider a situation in which two test forms can be connected through the so called linking groups in a pre-test non-equivalent group design (Béguin, 2000), because that is common for high-stakes examinations in the Netherlands, but the methodology developed in this paper can be also used with different linking designs.

When item response theory (IRT) is used for linking, item parameters of the items in the current and the reference tests have to be placed on the same scale (Kolen & Brennan, 2004; von Davier, 2011). This can be done either by estimating the IRT parameters in the two tests separately and then placing them on the same scale using scale transformation methods (Marco, 1977; Loyd & Hoover, 1980; Haebara, 1980; Stocking & Lord, 1983), or by estimating the parameters of the two test forms together in concurrent calibration (Wingersky & Lord, 1984). Once the item parameters are put on the same scale, the predicted score distribution of the reference population on the current test can be estimated. A cut-score (i.e., a minimum number of correct responses needed to pass the test) for the current test can be determined using IRT observed score equating using equipercentile linking (Lord & Wingersky, 1984).

Unlike examination data, which usually consist of responses of thousands of students to test items, the linking data are often not sufficiently large, such that the uncertainty about the difference between the difficulties of the two tests and, hence, about the new cut-score is rather large. These data are often collected in non-examination conditions (with different levels of stress and motivation) and not from the populations of interest. Thus, the linking data often do not provide a high enough quality of linking (in terms of uncertainty and bias of the cut-scores).

From the Bayesian perspective, the data are not the only source of information about the item parameters. A second source is the background information which

can be captured in the prior distributions. It has been advocated that using informative priors is useful in practical applications (Goldstein, 2006; Vanpaemel, 2011). The purpose of this study is to develop methods to improve the quality of linking by combining the information from the linking data and the informative priors. We explore different ways of eliciting the prior distributions about the difference in difficulty of the two tests from subject-matter experts and using them to link the two tests.

There have been studies with a focus on specification of informative priors for the parameters of IRT models. Item features (e.g., operations involved in solving the item, number of response alternatives, or use of negation in the formulation of the item) can be used to predict the item parameters (Fisher, 1973; Tatsuoka, 1987), which can be included as prior knowledge for Bayesian estimation of the item parameters (Mislevy, 1988). This source of prior information has been also used in the context of test equating and linking (Mislevy, Sheehan, & Wingersky, 1993). However, information about item features that are good predictors of the difficulty is not always available. Other authors include judgemental information from subject-matter experts in the estimation of the item parameters (Bejar, 1983; Swaminathan, Hambleton, Sireci, Xing, & Rivazi, 2003; Ozaki & Toyoda, 2006; Wauters, Desmet, & van der Noordgate, 2012). The latter has not been done in the context of test linking, and the expert judgements were only used to improve the estimation of the individual item parameters. Judgemental information about the items difficulties is also collected in the context of standard setting (Shepard, 1980; Geisinger, 1991; Cizek & Bunch, 2007). In that case, the cut-score for the test is selected based on the results of standard setting procedures and not by including the information from the experts in the form of the prior distributions and combining it with data in a Bayesian estimation procedure.

Experts' judgements of individual items are often not very accurate and reliable, however when combined on the test level they can provide valuable information about the relations between two tests. Therefore, we argue that the expert knowledge about the item difficulties is especially useful for test equating and linking. Another reason for a special interest in using experts' judgements in the context of test linking is that from the examination data we can estimate the differences between the item difficulties within the reference test and within the current test with high precision and the only thing that is missing is the information about the relations between the tests. Therefore, the information available from the examination data can be used to help in obtaining more valid and reliable judgements with respect to the relations between the two tests.

This paper is structured as follows. First, the measurement model and the equating design used throughout the paper are discussed. Then, in Section 6.3 we propose two methods for elicitation of the prior knowledge about the test difficulty from experts. The first one is an adaptation of the Angoff standard setting procedure (Angoff, 1971). The second method was designed by us for more direct

elicitation of the experts' knowledge about the differences between the difficulties of the two tests. In Section 6.4, the two elicitation methods are compared in terms of the quality of linking with the elicited priors using an empirical example based on the primary school mathematics test "Entreetoets Groep 7". The paper is concluded with a discussion.

6.2 Measurement model and equating design

In this study the marginal Rasch model (Rasch, 1960) is used assuming a normal distribution for proficiency. It models the probability of a correct response to an item in a population:

$$\Pr(X_i = 1) = \int_{\mathbb{R}} \frac{\exp(\theta - \delta_i)}{1 + \exp(\theta - \delta_i)} \mathcal{N}(\theta; \mu, \sigma^2) d\theta, \tag{6.1}$$

where X_i denotes a binary coded response (1 for correct and 0 for incorrect) to item i with difficulty δ_i of a person randomly sampled from the population with the mean and the variance of proficiency θ equal to μ and σ^2. The Rasch model was chosen because it has a clear interpretation of the item difficulty. If $\delta_i > \delta_j$, then both the conditional (i.e., given a particular value of θ) and the marginal probability (6.1) of a correct response to item i is smaller than to item j. This is important when translating experts' judgements of the type "Item i is more difficult than item j" into statements about the model parameters ($\delta_i > \delta_j$). This is not possible if an item discrimination parameter is added to the model, like is done in the two parameter logistic model (Lord & Novick, 1968). We assume that all the items, both in the current and in the reference test, have the same discriminative power. Although the Rasch model is a rather restrictive model, it has been shown that equating results using the Rasch model are rather robust to model violations (Béguin, 2000).

We consider a pre-test non-equivalent group equating design with G linking groups. This design is visualised in Figure 6.1, where rows represent persons and columns represent items. We denote the data matrix of the reference exam by $\mathbf{X}$, the data of the current exam by $\mathbf{Y}$ and the data of the G linking groups by $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_G$. By $\mathbf{X}^*$ we denote the unobserved responses of the reference population to the current test.

When concurrent Bayesian calibration is used for linking the two tests, samples from the joint posterior of the model parameters need to be obtained:

$$p(\boldsymbol{\delta}_r, \boldsymbol{\delta}_c, \mu_r, \sigma_r^2, \mu_c, \sigma_c^2, \mu_1, \sigma_1^2, \ldots, \mu_G, \sigma_G^2 \,|\, \mathbf{X}, \mathbf{Y}, \mathbf{Z}_1, \ldots, \mathbf{Z}_G), \tag{6.2}$$

where $\boldsymbol{\delta}_r$ and $\boldsymbol{\delta}_c$ are the vectors of item difficulties of the items in the reference and the current tests, respectively; μ_r and μ_c are the means of proficiency of the

reference and the current populations respectively, σ_r^2 and σ_c^2 are the corresponding population variances, and $\mu_1, \sigma_1^2, \ldots, \mu_G, \sigma_G^2$ are the population parameters in the linking groups. A zero point for the IRT scale is fixed by setting the average difficulty of the items in the reference test equal to zero: $\bar{\delta}_r = 0$. Using samples from the posterior in (6.2) the score distribution of the reference population on the current test (score distribution in $\mathbf{X}^*$) is estimated and the new cut-score is determined using equipercentile equating (see Appendix A).

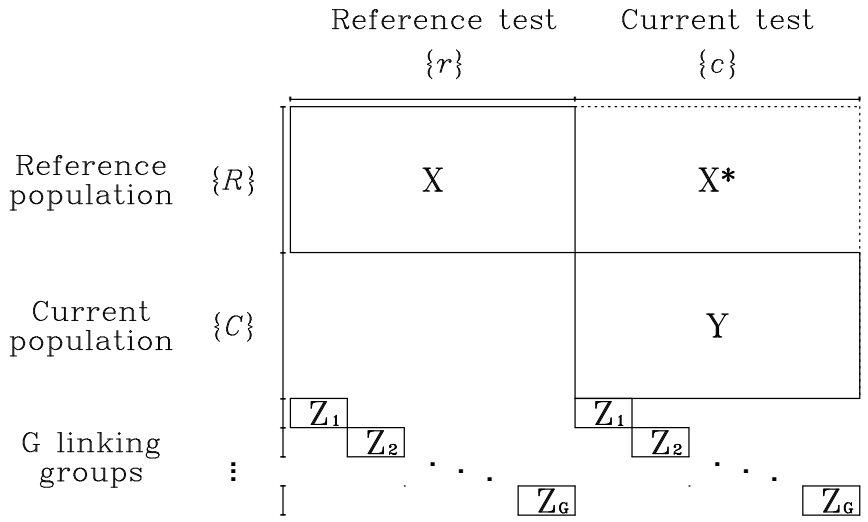


Figure 6.1: Equating design with G linking groups: rows represent persons, columns represent items.

The linking data are the only data that provide information about the difference between the average difficulty of the items in the current test and the average difficulty of the items in the reference test, denoted by $\tau = \bar{\delta}_c - \bar{\delta}_r$. Since the linking data are sparse, the largest part of the uncertainty about what the new cut-score should be is coming from the uncertainty about τ. We aim to increase the precision of the estimate of the new cut-score by including prior information about τ in the estimation. The following re-parametrisation is used throughout the paper (see Figure 6.2):

$$\boldsymbol{\delta}_c^* = \boldsymbol{\delta}_c - \bar{\delta}_c = \boldsymbol{\delta}_c - \tau \tag{6.3}$$

$$\mu_c^* = \mu_c - \bar{\delta}_c = \mu_c - \tau, \tag{6.4}$$

The rest of the paper is focused on the specification of the prior distribution of τ.

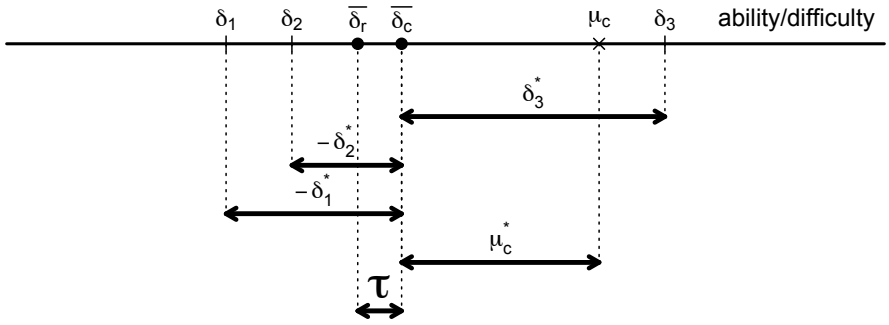


Figure 6.2: Re-parametrisation of the model parameters.

6.3 Elicitation of prior knowledge about the difference between the difficulty of two tests

Information about the difference between the average difficulty of the items in the current test and the average difficulty of the items in the reference test can be collected from subject-matter experts who can judge the difficulty of the items in the two tests. In this section we describe the two methods developed for the elicitation of the prior knowledge about τ. In the following section we compare the performance of these methods in two empirical elicitation studies.

Both methods use item difficulties estimated from the examination data. Since we also used the data to calibrate the two test forms on the same scale, we need to divide the examination data (both from the reference and from the current exams) into two halves: the first half which is used to facilitate the elicitation of experts' knowledge about the mutual order of the items and to construct priors for the item and the population parameters, here called the training data, and the second half which is used for the estimation of the new cut-score, here called the estimation data. See Appendix A for technical details.

6.3.1 Adaptation of the Angoff method for elicitation of the prior knowledge about τ

The first method that we developed is an adapted version of the Angoff method of standard setting (Angoff, 1971). Unlike the regular use of the Angoff method and other standard setting procedures we use the experts' judgements not to set the cut-scores directly but use these judgements for the specification of the informative prior for τ. In this way the cut-scores can be estimated based on both the experts' knowledge and the linking data.

Traditionally in the Angoff method, each expert $e \in \{1 : E\}$ from a panel of E experts is asked for each test item to give the probability that a minimally competent (borderline) candidate will answer this item correctly:

$$p_{ie} = \Pr(X_{pi} = 1 \,|\, \theta_p = \theta^*), \tag{6.5}$$

where θ^* is proficiency of a borderline candidate. These probabilities are then added over items to obtain the expected score of a borderline candidate which is chosen as a cut-score. In our study, we use the experts' evaluations of the probabilities p_{ie} differently. If each expert evaluates all items, then based on the Rasch model her/his estimate of the difference between the average difficulties of the items in the current and the reference test denoted by τ_e can be computed as

$$\tau_e = \frac{\sum_{i\in\{c\}} \ln\left(\frac{1-p_{ie}}{p_{ie}}\right)}{|c|} - \frac{\sum_{j\in\{r\}} \ln\left(\frac{1-p_{je}}{p_{je}}\right)}{|r|}, \tag{6.6}$$

where $\{r\}$ and $\{c\}$ are the sets of items in the reference and the current tests, respectively. It ca be seen that τ_e does not depend on the level of the proficiency of the borderline candidate.

Letting each expert evaluate all items is very time consuming and can lead to experts' judgements being less valid and reliable due to fatigue and loss of motivation. In our adapted procedure only a subset of items $\{r^*\}$ from the reference test and a subset of items $\{c^*\}$ from the current test are used. Then, for each expert we compute the difference between the average difficulties of the items in the subsets $\{c^*\}$ and $\{r^*\}$, denoted by τ_e^* :

$$\tau_e^* = \frac{\sum_{i\in\{c^*\}} \ln\left(\frac{1-p_{ie}}{p_{ie}}\right)}{|c^*|} - \frac{\sum_{j\in\{r^*\}} \ln\left(\frac{1-p_{je}}{p_{je}}\right)}{|r^*|}. \tag{6.7}$$

τ_e^* is not equal to τ_e since the items in the subsets are not fully representative of the full tests:

$$\tau_e = (\hat{d}_r - \hat{d}_c) - \tau_e^*, \tag{6.8}$$

where $\hat{d}_r$ is the difference between the average difficulty in the subset $\{r^*\}$ and the average difficulty in the set $\{r\}$, and $\hat{d}_c$ is the difference between the average difficulty in the subset $\{c^*\}$ and the average difficulty in the set $\{c\}$. These two quantities can be estimated from the training data.

The prior distribution of τ is chosen to be a normal distribution

$$p_1(\tau) = \mathcal{N}\left((\hat{d}_r - \hat{d}_c) - \mu_w, \sigma_w^2\right) \tag{6.9}$$

where $\mu_w = \frac{\sum_e w_e \tau_e^*}{\sum_e w_e}$ is the weighted mean of τ_e^* across the experts and $\sigma_w^2 =$

$\frac{\sum_e w_e(\tau_e^* - \mu_w)^2}{1-\sum_e w_e^2}$ is the weighted variance. The weights are determined by how well the estimated p_{ie} from each expert correlate with the observed proportions of correct responses to the items within each test in the training data:

$$w_e = \frac{1}{2}(Cor(\mathbf{p}_{re}, \mathbf{p}_r) + Cor(\mathbf{p}_{ce}, \mathbf{p}_c))\mathcal{I}_{Cor(\mathbf{p}_{re},\mathbf{p}_r)>0}\mathcal{I}_{Cor(\mathbf{p}_{ce},\mathbf{p}_c)>0}, \tag{6.10}$$

where $\mathbf{p}_{re}$ and $\mathbf{p}_{ce}$ are the vectors of probabilities of length $|r^*|$ and $|c^*|$, respectively, evaluated by expert e, and $\mathbf{p}_c$ and $\mathbf{p}_r$ are the observed proportions of correct responses to the items in the training data. If one of these correlations is negative for a particular expert then this expert gets a weight of zero.

6.3.2 Rulers method for direct elicitation of the prior knowledge about τ

One could imagine two rulers on which the positions of the items within each test are indicated, see Figure 6.3. As the arrows show, the rulers can be arbitrarily shifted to the left or to the right relative to each other, since the examination data do not tell us anything about the relative position of these two rulers. Prior knowledge about how these rulers should be positioned relative to each other can be elicited from experts. The most direct way would be to give an expert the rulers with the empirical item positions within each test and ask her/him to determine the proper mutual position of the two rulers. But comparing two complete tests with a large number of items is a very complicated task which is practically impossible to complete. For that reason, we developed a procedure in which experts are asked to compare smaller sets of items which have to be carefully selected. Our method is different from just asking the experts to specify the mutual order of the items in the two tests, because the empirical within-test items orders and the distances between the item difficulties within each test make many of the orders impossible, for example in Figure 6.3 the order $\delta_1 < \delta_1^* < \delta_2^* < \delta_3^* < \delta_2$ is not possible since the distance between δ_3^* and δ_1^* is larger than the distance between δ_2 and δ_1.

When developing the elicitation procedure, we carried out a pilot study with one expert to figure out what problems experts might experience when comparing sets of items. First, we observed that it was much easier for the expert to compare items of similar content. Another observation was that sometimes the expert did not agree with the empirical order of the item difficulties within a test which made specifying the mutual order of the items in the two sets senseless. Based on these observations, we developed the elicitation procedure consisting of several steps. In the first step, items are selected from the reference and the current test so that they are similar to each other in content and differ from each other in difficulty. In the second step, each expert $e \in \{1 : E\}$ orders the item difficulties within each test and only those items for which the expert's order and the order observed in

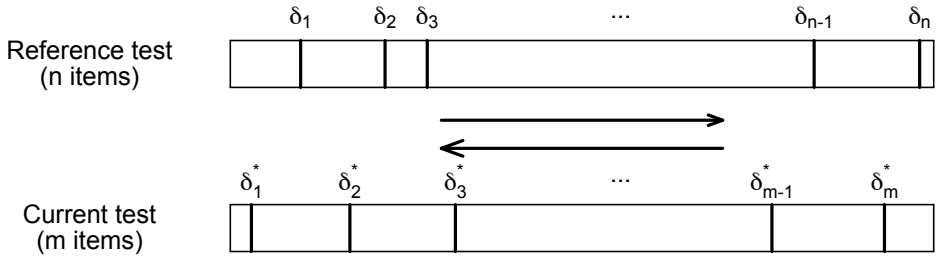


Figure 6.3: Two rulers with item difficulties estimated within each test: the arrows indicate that relative position of the rulers needs to be determined by experts.

the training data match are retained. In the third step, each expert aligns the remaining items from the reference test to the items from the current test. In this way, we make sure that the problems observed in the pilot study will not occur. Below we describe the three steps of the procedure in detail.

1. **Preliminary item selection**.

   (a) Divide the items within each test into homogenous groups based on the content, for example in a language test items may be divided in spelling, grammar, and punctuation subgroups. Often tests consist of clearly defined subdomains. If that is not the case, then experts can be asked to help divide the items in homogeneous subgroups. The subgroups should not be made too small, 6-8 items per subgroup should be sufficient.

   (b) Estimate the item difficulties separately for the items in the reference test and in the current test given the training data.

   (c) Select the largest subset of items from each homogenous group, such that the posterior probability of each pair of items within the subset to have a certain order of item difficulties is larger than 95%[2]. If multiple subsets can be constructed, then select one of them at random. The elicitation procedure cannot be used if these subsets cannot be constructed, that is if either there is not enough variation in item difficulty or if there are not enough data to be certain about the order of the item difficulties.

[2]This posterior probability can be approximated by the proportion of samples from the posterior distribution (given in Equations 6.22 and 6.23 in Appendix A) in which a certain order holds for the sampled values of the item difficulties.

2. **Final item selection** (performed separately for each expert e).

   (a) An expert orders the items within each homogeneous group based on their difficulty, for the two tests separately.

   (b) A subset of items from a set is retained if the expert's order of this set does not contradict the order observed in the training data. For example, if the observed order is:

   $$\hat{\delta}_1 < \hat{\delta}_2 < \hat{\delta}_3 < \hat{\delta}_4, \tag{6.11}$$

   and the expert's order is:

   $$\delta_{1e} < \delta_{3e} < \delta_{2e} < \delta_{4e}, \tag{6.12}$$

   then the expert's order within the subsets $\{1, 2, 4\}$ and $\{1, 3, 4\}$ do not contradict the empirical order. Both subsets can be used in the procedure. To make the selection of items automatic, one of these subsets is chosen randomly.

   (c) If for one of the content groups of items, in one of the tests there is no pair of items to be retained, then this group is discarded from the both tests. Therefore, after the final selection of items different experts might have different number of item groups to compare.

   (d) The quality of the judgements of expert e is quantified by the average proportion of items for which the expert's order and the empirical order were the same, denoted by p_e. In the case of empirical and expert orders in (6.11) and (6.12) this proportion is equal to .75. p_e used to weigh the expert's judgements, such that the effect of the prior distribution elicited from an expert on the combined prior decreases if her/his judgements rarely match the observed data. The weight of expert e is equal to

   $$w_e = \frac{p_e - p_0}{\sum_e (p_e - p_0)}, \tag{6.13}$$

   where p_0 is the expected average proportion of items which would be obtained if the order produced by a hypothetical expert is random.

3. **Shifting two rulers**. This stage is done for all content groups $j \in [1 : J_e]$, where J_e is the number of the content groups retained for expert e and consequently the number of judgements about the parameter τ which are elicited from her/him. Starting from Step 3 the rulers method is illustrated in Figure 6.4.

   (a) The expert is presented with a set of items from the reference test from one of the content groups ordered from the easiest to the most

difficult one, and with an ordered set of the items from the current test (from the same content group). The items are positioned according to the distances between the item difficulties estimated from the training data. The expert can shift these two sets of items relative to each other to the right or to the left, but cannot change the positions of the items within a test, since they are fixed at the estimated difficulties.

(b) The expert places the two sets together on the common scale according to her/his best guess. In each of the content groups we obtain the mutual order of the items and the estimate of the mode of the prior distribution, denoted by $\hat{\tau}_{ej}$. See "Best guess" in Figure 6.4.

(c) To evaluate the expert's uncertainty about the estimate, she/he is asked to place the item sets in the two most extreme positions which the expert still considers plausible: first in which the set of items from the current test has the rightmost position on the scale (resulting in an upper bound $\tau_{ej}^{\max}$) and second in which it has the leftmost position in the scale (resulting in a lower bound $\tau_{ej}^{\min}$). See "Extreme 1" and "Extreme 2" in Figure 6.4.

The above described procedure of experts' judgements collection for each expert e results in multiple sets $\{\hat{\tau}_{ej}, \tau_{ej}^{\min}, \tau_{ej}^{\max}\}, \forall j \in [1 : J_e]$. Next, we describe how this information can be translated into the prior distribution of τ.

Each pair of sets of items to be ordered presents the expert with an opportunity to express her/his beliefs about the relative difficulties of the items in the two sets and her/his uncertainty about the assessment of these relative difficulties. Through our method, each pair of sets produces an estimate of the mode of the expert's prior, as well as an estimate of the lower and the upper bound of the region that that expert still considers to be credible. This credible range was operationalised as the 90% credible interval for that expert's prior, and hence the lower and the upper bound that were specified by the expert are used as an estimate of her/his 5% and 95% percentiles, respectively. Since the extreme positions do not have to be symmetric around the mode, to approximate the prior knowledge elicited from expert e in judgement j we need a distribution which allows for skewness. A skew-normal distribution (Azzalini, 2005) is used:

$$p_{ej}(\tau) = \text{Skew-normal}(\xi_{ej}, \omega_{ej}, \alpha_{ej}), \tag{6.14}$$

where ξ_{ej} specifies the location, ω_{ej} specifies the spread and α_{ej} specifies the degree of skewness of the distribution (see Figure 6.4). The skew-normal distribution includes a normal distribution as its special case when $\alpha_{ej} = 0$. The parameters of the distribution are chosen in such a way that $\tau_{ej}^{\min}$ and $\tau_{ej}^{\max}$ are the 5-th and the 95-th quantiles and $\hat{\tau}_{ej}$ is the mode (see Appendix B for the details).

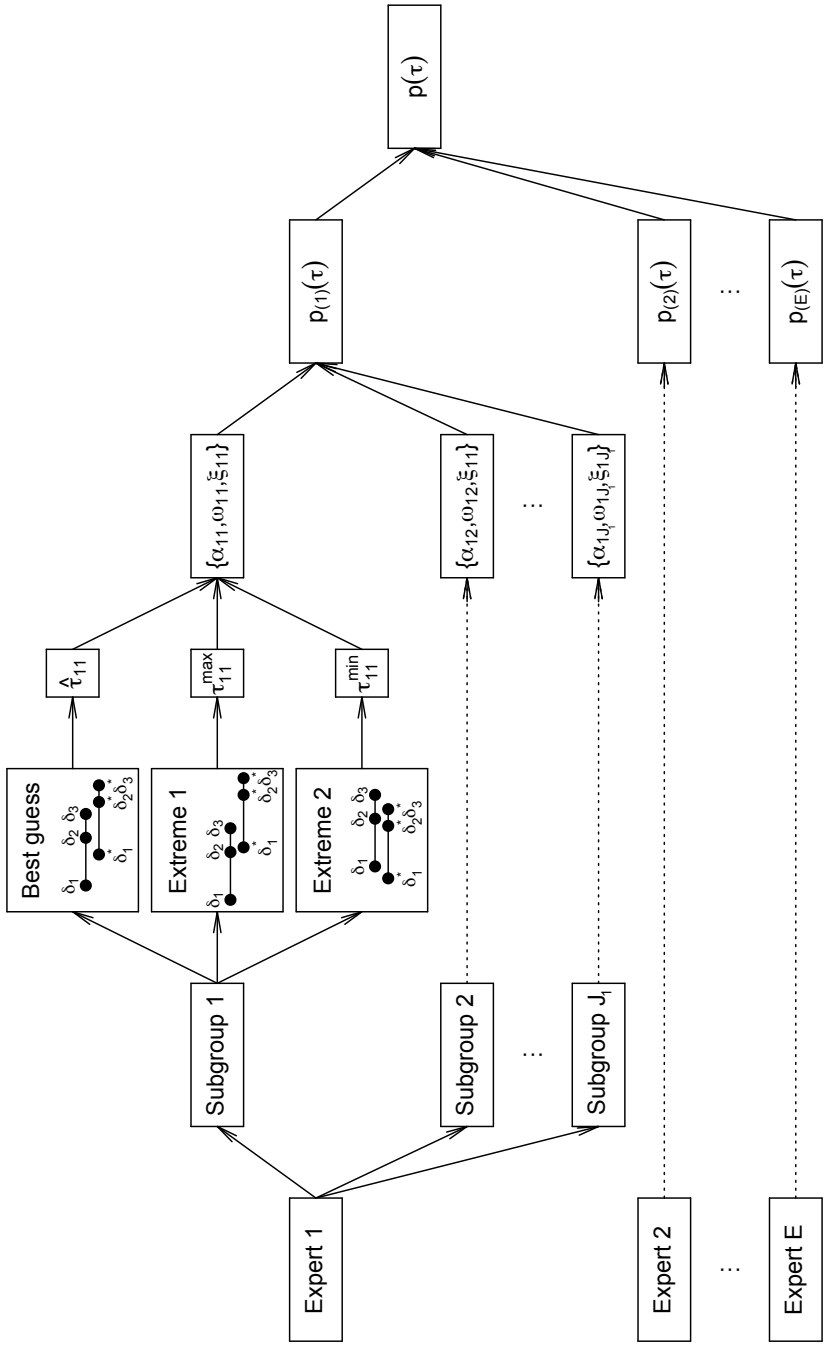


Figure 6.4: Scheme of the rulers method after the final selection of items.

Each judgement of expert e adds extra information about which values of τ are plausible according to this expert. Separate judgements of the same expert are assumed to be independent, therefore to combine the information from several judgements, we use a product of the distributions $p_{ej}(\tau)$:

$$p_e(\tau) = \frac{\prod_j p_{ej}(\tau)}{\int_{-\infty}^{\infty} \prod_j p_{ej}(\tau) d\tau}, \tag{6.15}$$

where the normalising constant in the denominator ensures that $p_e(\tau)$ is a proper distribution. This integral does not have a closed form solution and is therefore approximated here by numerical integration with Gauss-Hermite quadrature (see Equation 6.57 in Appendix B). The motivation for the independence assumption is that each judgement refers to a different set of items with a unique combination of item features influencing the item difficulty, which an expert takes into account.

When combining the information from the different experts, we use linear opinion pooling (Stone, 1961; O'Hagan et al., 2006):

$$p_2(\tau) = \sum_e w_e p_{(e)}(\tau), \tag{6.16}$$

where the weights w_e computed using (6.13) make sure that the results of the experts with higher quality judgements have a larger influence on the prior $p_2(\tau)$. We prefer linear opinion pooling over logarithmic opinion pool because the latter being a geometric mean of individual distributions leads to unrealistically strong aggregated believes. Moreover, linear opinion pool does not rule out the low or the high values of the parameter that are supported by a minority of the experts (O'Hagan et al., 2006, p.184).

6.4 Empirical example

6.4.1 Data

For illustrating and comparing the methods of test linking using prior knowledge, we used the data from the test of mathematics for primary school "Entreetoets Group 7" taken by students in the Netherlands at the end of the 5th grade. The same test consisting of 120 items was administered in 2008 and 2009. The test was divided into 10 groups based on content: 1) mental arithmetics, 2) mental arithmetics - estimation, 3) arithmetic operations, 4) number relations, 5) geometry, 6) measurement of length and surface, 7) measurement of weight and volume, 8) percentages, fractions and ratios, 9) time, 10) money; and then each subgroup was randomly divided into two parts. We treated the first part as the reference test

and the second part as the current test. The populations of 2008 and 2009 were treated as the reference and the current population respectively. Hence, an equating problem was artificially created for the data set in which the responses of the persons from what we labelled as "reference population" to the items from what we labelled "current test" were actually observed (see Figure 6.5). This makes it possible not only to illustrate the procedures introduced in Section 6.3 but also to evaluate them by comparing the estimate of the new cut-score obtained with the different priors based on the *predicted* responses of the reference population to the current test (see Figure 6.5b denoted by "?") with the cut-score, derived from the *observed* responses of the reference population to the current test in the complete data (see Figure 6.5a). Hence, the latter is used as a proxy of the true cut-score.

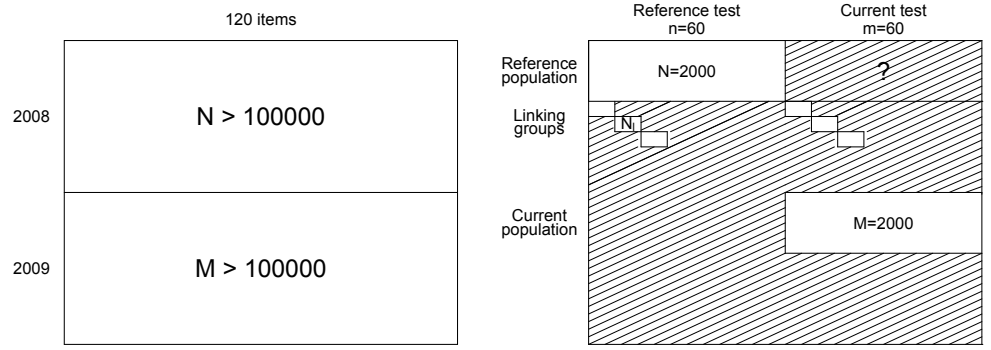


Figure 6.5: Creating an equating problem from a complete design. The completely dashed areas indicate data that were not used in the analysis.

The data set of each year consisted of responses of more than 100,000 students. Since the linking procedures developed in this study are meant for tests administered to smaller samples, responses of 2000 persons from each year were randomly selected as examination data. The data for three linking groups with responses to randomly selected 8 items from the reference test and 8 items from the current test were selected from the data of 2008. Although they were sampled from the same population as the examination data, this fact was ignored in the estimation, assigning separate parameters for the mean and the variance of proficiency in each linking group.

Since based on this particular arithmetics test students are assigned to one of five levels of proficiency (from A to E), four cut-scores need to be estimated. The cut-scores between the levels in the reference test (denoted by s_{ref}) were 49, 42, 35 and 24 correct responses. The corresponding cut-scores for the current test

maintaining the same standard need to be estimated. From the complete data (see Figure 6.5a) the true cut-scores in the current test (denoted by $s_{pass,true}$) were determined to be equal to 50, 43, 36 and 26 correct responses.

6.4.2 Method

To evaluate the linking results based on different priors we analysed how the estimated cut-scores for the current test (s_{pass}) changes depending on the prior specification.

For each cut-score in the reference test we estimated a cut-score in the current test using different priors and compared it to the true cut-score $s_{pass,true}$. To take into account the sampling variability, we re-sampled the data (both the persons and the items) for the linking groups 100 times to examine the distribution of the estimated cut-scores across these re-sampled data sets, to evaluate how often each of the cut-scores is correctly estimated:

$$k_{true} = \sum_{k=1}^{100} \mathcal{I}(s_{pass,k} = s_{pass,true}) \tag{6.17}$$

and how large the mean squared error (MSE) is:

$$MSE = \frac{1}{100} \sum_{k=1}^{100} (s_{pass,k} - s_{pass,true})^2, \tag{6.18}$$

where $s_{pass,k}$ is the estimated cut-score from the k-th re-sampled data set.

We also used the average number of misclassified students across the re-sampled data sets to compare the quality of linking based on different priors. In each re-sampled data set the number of persons from the current population who were assigned to an incorrect level of proficiency due over- or underestimation of the cut-scores was counted. If the estimates of the cut-scores are centred around the true values and do not vary a lot then this number would be small.

To analyse how the influence of the prior distribution changes depending on the size of the linking data, we varied the number of persons per linking group: $N_l = 100, 200, 300$, and 500.

In addition to $p_1(\tau)$ defined in Equation 6.9 and $p_2(\tau)$ defined in Equation 6.16, we also used a vague prior:

$$p_0(\tau) = \mathcal{N}(0, 100) \tag{6.19}$$

to show the added value of using prior knowledge.

6.4.3 Expert elicitation. Study 1

Participants

Nine experts participated in the study (4 females and 5 males). The age of the experts ranged from 25 to 66 with a mean of 51. These experts were members of the construction groups, who develop items for mathematical tests at the Dutch National Institute for educational measurement. The number of years of experience as an item constructor ranged from less than a year to 30 years. Most of the experts also had experience with teaching primary school mathematics and/or teaching at the programmes for primary school teachers.

Preliminary selection of items

From the 10 content groups only 7 groups were selected for the expert elicitation, because the other three groups were too small (number relations, geometry, and money). The items within each test were ordered based on the observed proportions of correct responses in the training data. Within each of the seven content groups a subset was selected such that for each pair of items i and j within this subset the posterior probability of them having a certain order of item difficulties was larger than .95. In total 62 items were selected which were used both for the Angoff and for the rulers methods.

The item difficulties within each test, which were used for specifying the locations of the items in the elicitation procedure, were estimated from the training data.

Procedure

The group of nine experts was randomly divided into two groups: one which started with the Angoff method (four experts) and continued with the rulers method, and the other which had the reverse order of the elicitation methods (five experts). Due to technical problems one of the experts from the second group had to start with the Angoff method. All the assignments were performed by the experts individually.

The rulers method was implemented in a computer application developed for this study. The application had two parts corresponding to Step 2 and Step 3 of the elicitation procedure described in Section 6.3.2. Each expert got a prerecorded audio-instruction accompanied by a power-point presentation illustrating the procedures.

In the first part of the procedure, experts were presented with sets of 3, 4, 5, or 6 items from each content group and each test (see Figure 6.6). The content

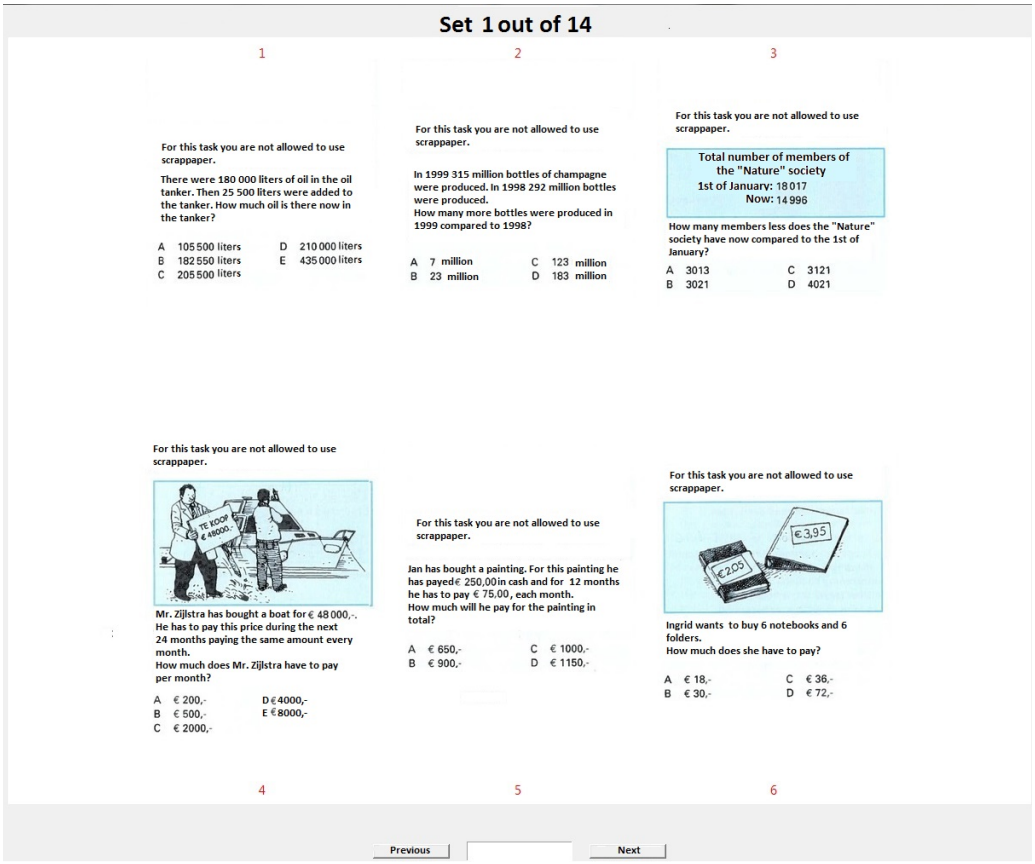


Figure 6.6: Illustration of the first part of the computerised procedure for the rulers method: six items have to be ordered based on their difficulty (translated from Dutch).

groups were presented in the same order to everyone, but the order of the item sets within a content group was randomly chosen for each expert (either reference test - current test, or current test - reference test). Within each set items were presented in a random order. For each set experts had to fill in the order of the items based on their difficulty starting from the easiest item.

After the first part, experts received the instruction for the second part. In the second part of the procedure experts were presented with two sets of items: one at the top and one at the bottom. The items were located according to their estimated item difficulties based on the training data. Each set contained at most four items. If after the first part there were more than four items in a set for which the expert's order matched the empirical order, then four items were randomly selected to be retained. It was not possible for the experts to move the

items within a set, but only sets as a whole relative to each other using two sliders, one at the top and one at the bottom. First, the experts had to place the sets in the most plausible position (see Figure 6.7). Second, they had to specify the first extreme position by moving either the top set to the right or the bottom set to the left away from the most likely position (in Figure 6.7 "Best guess" was substituted by "Extreme 1: top to the right or bottom to the left"). Third, they had to choose the second extreme position by moving the sets away from the most likely position in the opposite direction (in Figure 6.7 "Best guess" was substituted by "Extreme 2: top to the left or bottom to the right").

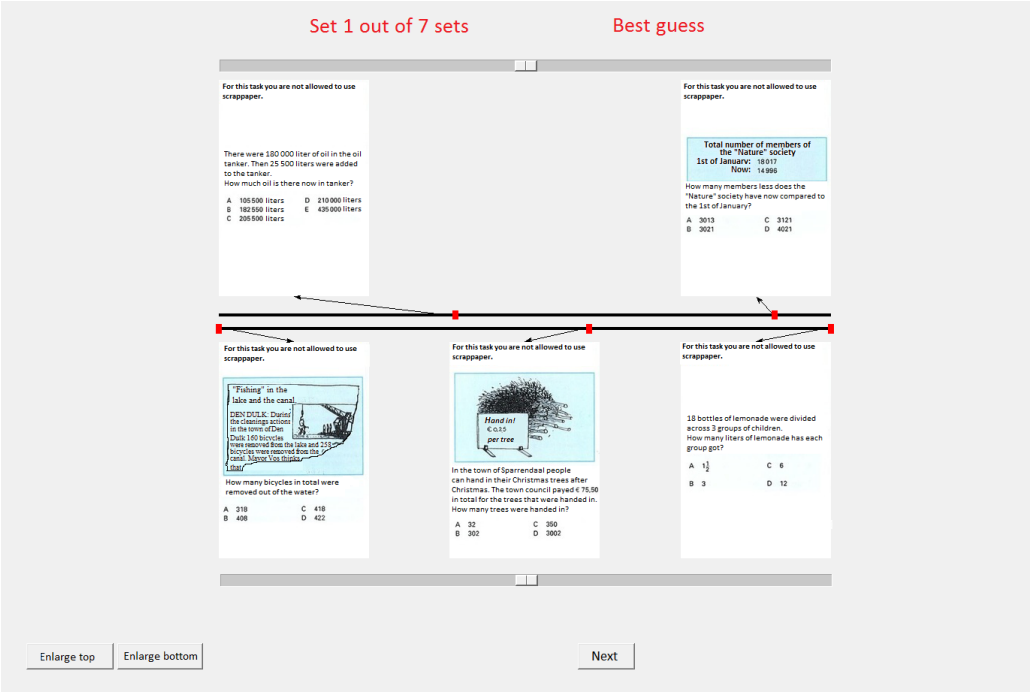


Figure 6.7: Illustration of the second part of the computerised procedure for the rulers method: Best guess (translated from Dutch). The red blocks indicate the item positions on the difficulty scale. Two sliders (at the top and at the bottom) can be used to change the position of the item sets.

Due to an error in coding[3], the item difficulty of one of the items was estimated incorrectly at the initial phase of the study. This was a problem only for one set of items and did not influence the estimates of the difficulties of all other items.

[3]Note that we did not make this error in Study 2.

However, the item position in which this item was presented in the second part of the procedure was incorrect, therefore the experts' judgements from the pair of sets containing this item (mental arithmetics content group) were discarded. The estimation of the cut-scores was based on the corrected difficulty of that one item.

In the Angoff procedure the same 62 items as used in the rulers method were given to each expert in a random order. The experts filled in their responses to a question "Imagine a group of students from the 5th grade with an average level of proficiency (B/C or level 3 of the student monitoring system). How many of these students will answer this item correctly?" in a booklet with one item per page (see Figure 6.8). Three extra items were included in the beginning of the booklet to familiarise the experts with the procedure, such that the total number of items was 65 but only the results of the 62 items were taken into account.

Results

The Angoff method resulted in a prior

$$p_1(\tau) = \mathcal{N}(\mu = 0.097, \sigma^2 = 0.005), \tag{6.20}$$

which is shown in Figure 6.9b.

With respect to the rulers method, whenever the estimated $\hat{\tau}_{ej}$ did not lie within the two extreme estimates $\tau_{ej}^{\min}$ and $\tau_{ej}^{\min}$, the judgement was excluded, since it meant that the instruction was not followed properly. This was the case for half of the experts' judgements[4]. All judgements of Expert 8 were excluded because only once his response was consistent with the instructions. One of the judgements of Expert 7 was considered to be an outlier. The lower bound in this judgement was 1.48, whereas the upper bound in the other three judgements were -0.09, -0.05 and -0.02. Therefore, this judgement was discarded.

Figure 6.9a shows the priors elicited from individual experts in the computerised procedure. In Figure 6.9b the combined prior $p_2(\tau)$ and the prior elicited with the Angoff method $p_1(\tau)$ are shown. Figure 6.9 also includes $\bar{\tau}$ which was estimated from the observed responses of almost 20,000 persons from both 2008 and 2009 to all 120 mathematics items (see Figure 6.5a). A RM was fitted to the complete data set and the difference between the average estimates of the difficulties of the items in the current and the reference test was computed. We used it as a proxy for the true value of τ to evaluate how close to the truth the expert's judgements were. Two of the priors elicited from the individual experts are located close to the proxy of the true values $\bar{\tau}$. Consequently, one of the local modes of

[4]Please note that in Study 2 this problem did not occur.

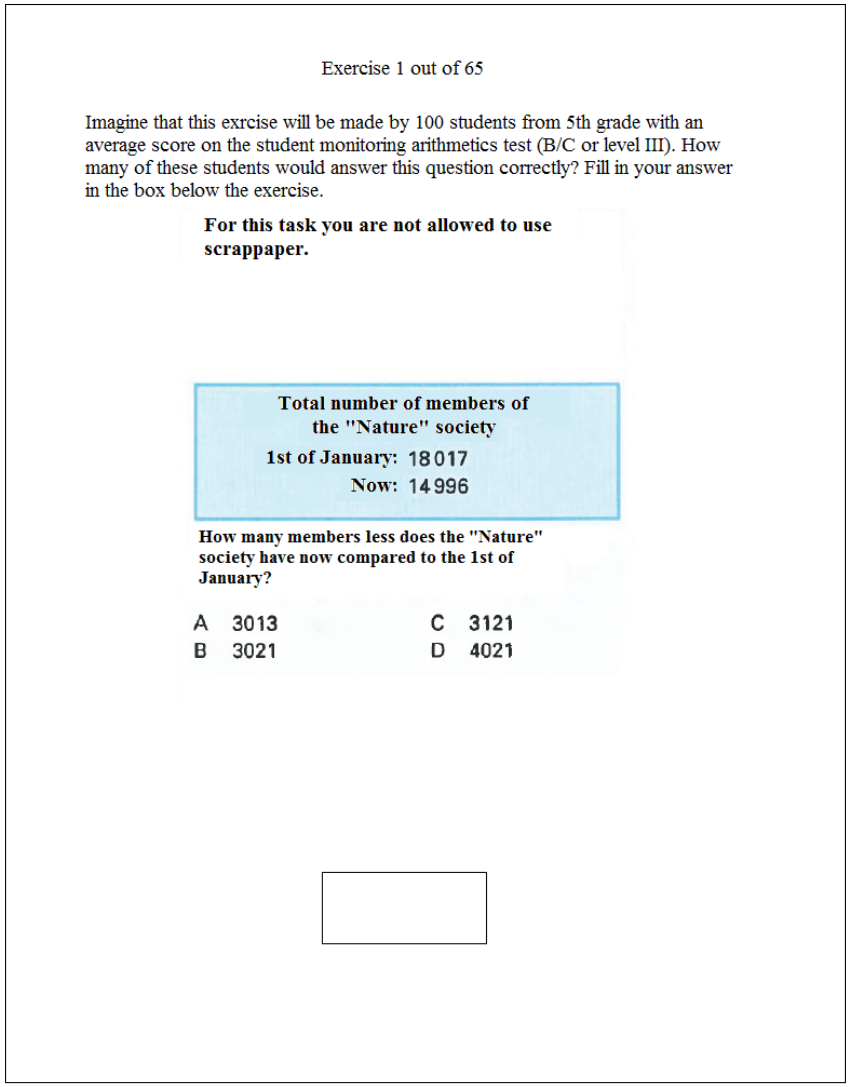


Figure 6.8: Illustration of the Angoff procedure: one item per page is presented (translated from Dutch).

$p_2(\tau)$ is also close to $\bar{\tau}$. On the contrary, the Angoff prior assigned extremely low density to $\bar{\tau}$.

Figure 6.10 shows the distribution of the estimates of the cut-scores across the re-sampled data sets, and Table 6.1 shows how often each of the four cut-scores was

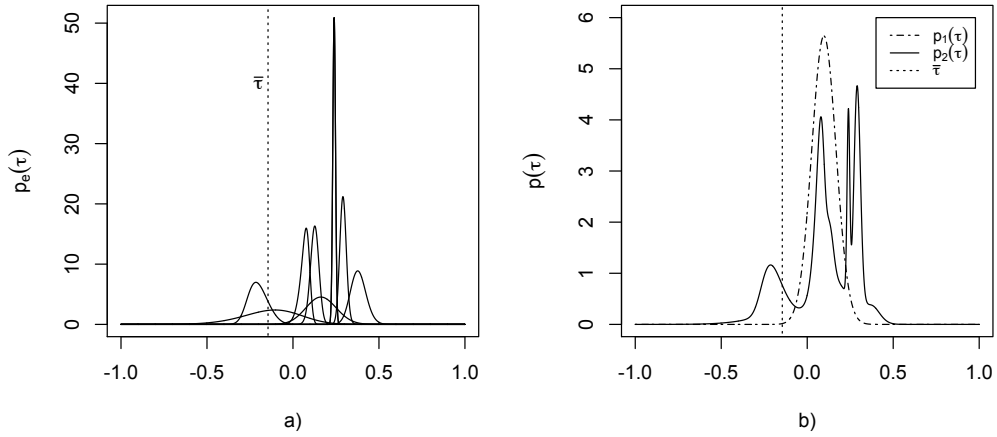


Figure 6.9: Prior distributions elicited from individual experts using the rulers method in the Elicitation Study 1 (a) and combined prior distributions (b).

correctly estimated with different priors. From Figure 6.10 one may see that with the vague prior there was a lot of variation in the estimated cut-scores, especially when $N_l = 100$. With the informative priors, the estimates of the cut-scores across the re-sampled data sets were less spread around the mode. But the Angoff prior gave biased results in most conditions, since the mode of the distribution of the estimated cut-scores was lower than the true cut-score, in some conditions even by two score points. For the prior based on the rulers method only the mode of the distribution of the cut-score B/C was not equal to the true cut-score for one of the sample sizes. In all other conditions k_{true} using the rulers method was either larger than or roughly the same as k_{true} using the vague prior. Moreover, the mean squared error of the estimates based on the rulers method were also either smaller or approximately equal to those based on the vague prior. The MSEs of the estimates based on the Angoff prior were in all conditions larger than those based on the vague prior.

To illustrate the consequences of incorrectly estimating the cut-scores more concretely, we looked at the average number of persons misclassified when different priors were used (see Table 6.2). The worst results were obtained with the Angoff prior. With the smallest sample size ($N_l = 100$), the vague prior resulted in the smallest average number of misclassified persons. With larger sample sizes, the results were better for $p_2(\tau)$. However, the differences between $p_2(\tau)$ and the vague prior decreased with sample size, such that the difference between the two methods is negligible when $N_l = 500$.

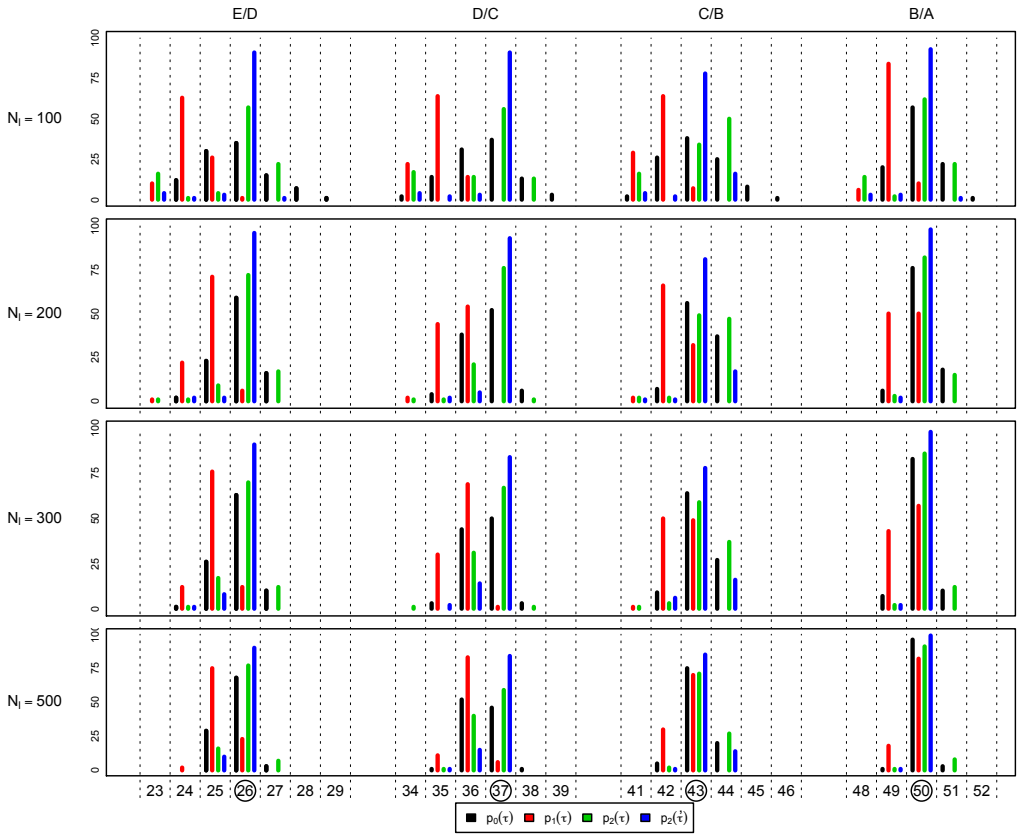


Figure 6.10: Distributions of the estimates of the cut-scores across the re-sampled data sets using different priors: $p_0(\tau)$ - vague prior, $p_1(\tau)$ - Angoff prior, $p_2(\tau)$ - rulers prior Study 1, $p_2'(\tau)$ - ruler prior Study 2. The true cut-scores are marked with a circle.

Table 6.1: Numbers of data sets in which the estimated cut-score was equal to the true cut-score and mean squared error of the estimates of the cut-scores

cut-score	Prior	$N_l = 100$		$N_l = 200$		$N_l = 300$		$N_l = 500$	
		k_{true}	MSE	k_{true}	MSE	k_{true}	MSE	k_{true}	MSE
D/E	$p_0(\tau)$	35*	1.21	59*	0.47	63*	0.40	66*	0.34
	$p_1(\tau)$	1	3.63	6	1.68	12	1.24	23	0.83
	$p_2(\tau)$	57*	1.74	72*	0.39	70*	0.33	77*	0.23
	$p_2'(\tau)$	91*	0.39	96*	0.10	91*	0.12	90*	0.10
C/D	$p_0(\tau)$	36*	1.31	53*	0.59	48*	0.61	46*	0.57
	$p_1(\tau)$	0	4.68	0	2.48	1	1.89	6	1.27
	$p_2(\tau)$	56*	1.80	76*	0.40	67*	0.41	59*	0.44
	$p_2'(\tau)$	91*	0.42	93*	0.13	84*	0.22	84*	0.19
B/C	$p_0(\tau)$	39*	0.99	56*	0.44	64*	0.36	74*	0.26
	$p_1(\tau)$	7	1.80	31	0.75	50*	0.53	69*	0.31
	$p_2(\tau)$	34	1.14	49*	0.57	59*	0.44	71*	0.29
	$p_2'(\tau)$	78*	0.34	81*	0.22	78*	0.22	85*	0.15
A/B	$p_0(\tau)$	57*	0.46	77*	0.23	83*	0.17	96*	0.04
	$p_1(\tau)$	9	1.09	50*	0.50	57*	0.43	82*	0.18
	$p_2(\tau)$	62*	0.80	82*	0.18	86*	0.14	92*	0.08
	$p_2'(\tau)$	93*	0.16	98*	0.02	98*	0.02	99*	0.01

Note. * - the most frequent estimate of the cut-score is equal to the true cut-score.

Table 6.2: Average number of misclassified persons across 100 re-sampled data sets

Prior	$N_l = 100$	$N_l = 200$	$N_l = 300$	$N_l = 500$
$p_0(\tau)$	144.52	82.57	72.13	54.24
$p_1(\tau)$	296.12	184.4	152.96	103.32
$p_2(\tau)$	148.30	68.60	62.02	51.39
$p_2'(\tau)$	37.10	19.63	26.56	21.34

Discussion

Generally, the results of the first study showed that including experts' knowledge in the form of informative priors in test linking is promising. However, the two elicitation methods differed in their utility. The prior based on the rulers method $p_2(\tau)$ is a mixture of experts' opinions, whereas the Angoff prior is simply an average of experts' results. Therefore, $p_2(\tau)$ can better capture the diversity of experts' opinions, and if there are experts in the panel whose judgements are close to the truth then the posterior would get closer to the truth as well. On the contrary with the Angoff method the prior of τ is dominated by the judgements' of the majority of the experts and does not reflect the variation of expert opinions. Another drawback of the Angoff method is that it does not allow to take the uncertainty of the individual experts into account.

Because the prior distribution $p_2(\tau)$ was multimodal and only one of the local modes was close to the truth (see Figure 6.9), we looked into the characteristics of the experts whose opinions were further away or closer to $\bar{\tau}$. From the panel of eight experts, the worst results were obtained from an expert with the shortest experience as an item writer (less than a year) and from teachers with very large experience of teaching mathematics at primary schools. Better results were obtained from those experts who did not necessarily have experience with teaching but had more theoretical knowledge about primary school mathematics and educational measurement.

Our hypothesis is that for teachers having experience with concrete students in class it is very difficult to make judgements about the difficulty of the items in general in the population. On the contrary, researchers are more used to thinking in more abstract terms. Moreover, the idea of the putting items along a single difficulty dimension might be rather difficult for those without theoretical background in educational measurement. To test this hypothesis, we carried out a second study in which we involved primary school mathematics researchers from universities and employees of a testing organisation specialised in testing primary school mathematics.

6.4.4 Expert elicitation. Study 2

Participants

Seven experts (four females and three males) participated in the second elicitation study. These were four primary schools mathematics researchers from two Dutch universities and three employees of the Dutch National Institute for educational measurement working with primary school mathematics tests.

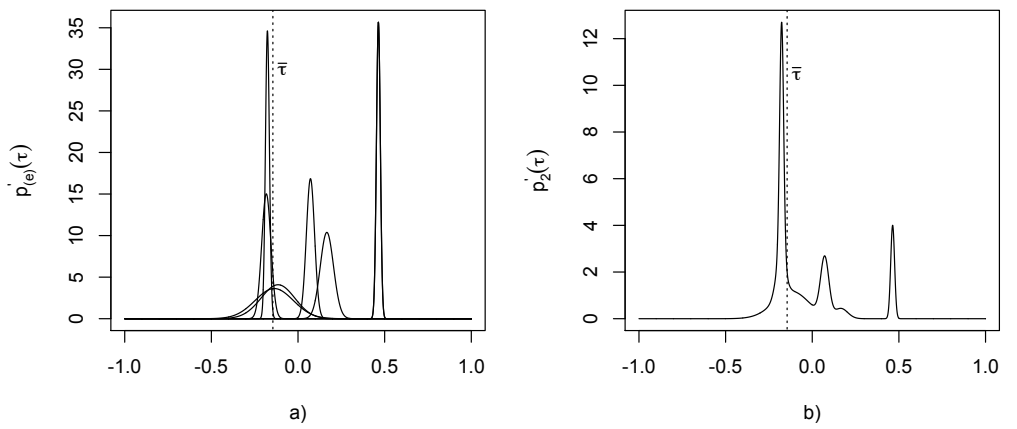


Figure 6.11: Prior distributions elicited from individual experts in the Elicitation Study 2 (a) and combined prior distribution based on experts' judgements (b).

Procedure

In the second elicitation study only the rulers method was used, because it had better results than the Angoff method in the first study.

As has been mentioned when presenting the results of Study 1, instructions were not always properly followed, resulting in the mode $\hat{\tau}_{ej}$ being outside of the credible interval $[\tau_{ej}^{\min}; \tau_{ej}^{\max}]$. To avoid this problem the second part of the computerised procedure was modified. Instead of having two sliders to change the mutual position of the two item sets there was only one slider at the top. Moreover, the instruction was clearly stating that after the most likely position, the top item set had to be first moved to the right and then to the left. The rest of the procedure remained the same. The same 62 items as in the first study were used. There were no problems anymore with $\hat{\tau}_{ej}$ not being within the credible interval.

Results

Figure 6.11a shows the priors elicited from the individual experts. Four of the distributions $p'_e(\tau)$ are concentrated around the proxy of the true value $\bar{\tau}$. One of the experts provided a distribution with a mode far away from the judgements of other experts. This was the expert with the lowest quality of the judgement and the smallest weight $w_e = .04$ (for comparison, $w_e = .14$ would be the weight of each expert if equal weights were assigned). In Figure 6.11b the combined prior $p'_2(\tau)$ is shown. The largest mode of this distribution is very close to $\bar{\tau}$.

For all sample sizes and all cut-scores the informative prior $p_2'(\tau)$ resulted in correctly estimated cut-scores in more re-sampled data sets than the vague prior (see Table 6.1). The variance of the estimated cut-scores was reduced, while they were still concentrated around the same mode as the estimates based on the vague prior (see Figure 6.10). In all conditions and for all levels of proficiency, the most frequent estimate of the cut-score was equal to the true cut-score. The MSEs of the estimated cut-scores were also smaller for $p_2'(\tau)$ than for the vague prior (see Table 6.1). Finally, on average less students were misclassified when $p_2'(\tau)$ was used compared to $p_0(\tau)$, see Table 6.2.

6.5 Conclusions

In this paper we introduced different procedures for elicitation of prior knowledge for test linking from subject-matter experts. The empirical elicitation studies showed promising results for the rulers method of elicitation of the experts' knowledge about the difference in test difficulty. However, there are some challenges and limitations of the presented approaches. The crucial part of the elicitation procedures is the selection of experts. One should be very careful when including someone in the study as an expert. In our study we have noticed that experts with different characteristics provide judgements of different quality. Based on the results, we presume that it is more useful to have experts with more theoretical knowledge about testing particular abilities (in our case primary school mathematics) and experience with analysing test results on a large scale, rather than those with a lot of practical teaching experience.

Evaluating item difficulties is a very difficult task for experts and even in a panel of carefully selected experts the results obtained from different experts might disagree with each other. In our proposed approach, the results of individual experts are combined in a mixture prior, which reflects the differences in the experts' opinions. When a mixture prior is combined with the data, then if there are one or two experts whose judgements disagree with the rest of the experts and do not match the evidence from the linking data, then the results of these outlying experts will hardly influence the posterior.

When using informative priors based on subject-matter experts' judgements in real practice it is important to compare the obtained results with the results based on the vague priors. The latter reflects the information contained in the data only. If the estimated cut-scores obtained with and without taking the expert judgements into account differ dramatically then one should decide whether to trust the linking data or the experts more. An important feature of the proposed elicitation methods is that it provides methods for evaluating the quality of experts judgements, by comparing the experts' judgements about the item difficulties within the reference and the current tests with the observed examination

data (see expert weights defined by Equation 6.10 for the Angoff method and by Equation 6.13 for the rulers method).

The results of the empirical study suggest that using informative priors can improve linking results. Expert judgements collected using the rulers method and included in the Bayesian estimation can increase the precision of linking without introducing a lot of bias.

6.6 Appendices

Appendix A: Gibbs Sampler for estimating the cut-score s_{pass}

Here we provide details about how to estimate the cut-score for the current test s_{pass}. To estimate s_{pass} we need to estimate the score distribution of the reference population on the current test (see $\mathbf{X}^*$ in Figure 6.1) for which samples from the joint posterior distribution

$$p(\boldsymbol{\delta}_c^*, \mu_r, \sigma_r^2, \tau \,|\, \mathbf{X}, \mathbf{Y}, \mathbf{Z}_1, \ldots, \mathbf{Z}_G) \tag{6.21}$$

need to be obtained. To obtain samples from this multivariate distribution we use a Gibbs Sampler algorithm in which at each iteration each parameter of interest is sampled from its conditional posterior distribution given the current values of all other parameters. To simplify the conditional posterior distributions, not only the parameters μ_r, σ_r^2, $\boldsymbol{\delta}_c^*$ and τ are sampled but also the item parameters of the items in the reference test, the population parameters of the current population and the linking groups and the individual ability parameters of all the persons in the reference population (denoted by the vector $\boldsymbol{\theta}_r$), the current population (denoted by the vector $\boldsymbol{\theta}_c^*$) and the linking groups (denoted by the vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G$).

The examination data sets $\mathbf{X}$ and $\mathbf{Y}$ (see Figure 6.1) are both split in two parts: one for constructing prior distributions, denoted by $\mathbf{X}^{(1)}$ and $\mathbf{Y}^{(1)}$, and another for the estimation of s_{pass}, denoted by $\mathbf{X}^{(2)}$ and $\mathbf{Y}^{(2)}$. The subsets of persons from the reference population whose responses are in $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are denoted by $\{R^{(1)}\}$ and $\{R^{(2)}\}$, respectively. The subsets of persons from the current population whose responses are in $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ are denoted by $\{C^{(1)}\}$ and $\{C^{(2)}\}$, respectively. $N^{(1)}$, $N^{(2)}$, $M^{(1)}$, and $M^{(2)}$ denote the number of persons in $\{R^{(1)}\}$, $\{R^{(2)}\}$, $\{C^{(1)}\}$. and $\{C^{(2)}\}$, respectively.

In the following subsections we will first describe how the training data $\mathbf{X}^{(1)}$ and $\mathbf{Y}^{(1)}$ are used for constructing priors, second we will describe how using these priors, the estimation data $\mathbf{X}^{(2)}$ and $\mathbf{Y}^{(2)}$ and the data of the linking groups $\mathbf{Z}_1, \ldots, \mathbf{Z}_G$, samples from the posterior distribution needed to determine the cut-score s_{pass} can be obtained, and finally we will show how to obtain the posterior distribution of s_{pass} and its estimate.

Using training data to construct priors

When constructing priors, we need to obtain samples from the posterior distributions:

$$p(\boldsymbol{\theta}_r, \mu_r, \sigma_r^2, \boldsymbol{\delta}_r \,|\, \mathbf{X}^{(1)}) \propto \prod_{p \in \{R^{(1)}\}} \prod_{i \in \{r\}} \frac{\exp(x_{pi}(\theta_p - \delta_i))}{1 + \exp(\theta_p - \delta_i)} \times$$
$$\times \mathcal{N}(\mu_r; 0, 100) \text{Inv-}\mathcal{G}(\sigma_r^2; .001, .001) \prod_{i \in \{r\}} \mathcal{N}(\delta_i; 0, 100) \prod_{p \in \{R^{(1)}\}} \mathcal{N}(\theta_p; \mu_r, \sigma_r^2) \quad (6.22)$$

and

$$p(\boldsymbol{\theta}_c^*, \mu_c^*, \sigma_c^2, \boldsymbol{\delta}_c^* \,|\, \mathbf{Y}^{(1)}) \propto \prod_{p \in \{C^{(1)}\}} \prod_{i \in \{c\}} \frac{\exp(y_{pi}(\theta_p^* - \delta_i^*))}{1 + \exp(\theta_p^* - \delta_i^*)} \times$$
$$\times \mathcal{N}(\mu_c^*; 0, 100) \text{Inv-}\mathcal{G}(\sigma_c^2; .001, .001) \prod_{i \in \{c\}} \mathcal{N}(\delta_i^*; 0, 100) \prod_{p \in \{C^{(1)}\}} \mathcal{N}(\theta_p^*; \mu_c^*, \sigma_c^2). \quad (6.23)$$

The normal priors for the means and the inverse-gamma priors for the variance are chosen because of the mathematical convenience of conjugacy.

The initial values, denoted by a superscript (0), for all the parameters have to be chosen: $\mu_r^{(0)} = 0; \sigma_r^{2(0)} = 1; \mu_c^{*(0)} = 0, \sigma_c^{2(0)} = 1; \delta_i^{(0)} \sim U(-2, 2), \forall i \in \{r\}; \delta_i^{*(0)} \sim U(-2, 2), \forall i \in \{c\}, \tau^{(0)} = 0$. It is not needed to choose the initial values for the individual person parameters since they are sampled in the first step of the algorithm.

Below we describe how to sample from the posterior distribution in (6.22). Sampling from the posterior distribution in (6.23) is analogous to sampling from (6.22). The algorithm has 5 steps.

$$\textbf{Step 1. } \theta_p \sim p(\theta_p \,|\, \dots) = p(\theta_p \,|\, X_{p+}, \mu_r, \sigma_r^2, \boldsymbol{\delta}_r), \forall p \in \{R^{(1)}\}: \quad (6.24)$$

which depends on the data only through the sumscore $X_{p+} = \sum_i X_{pi}$, since the RM holds. Sampling from this distribution can be done using the conditional composition algorithm (Marsman et al., 2015):

a. Sample a candidate value from the population distribution $\theta \sim \mathcal{N}(\theta_p; \mu_r, \sigma_r^2)$

b. Simulate a vector of responses $\mathbf{X}$ to the items in the reference test

$$\Pr(X_i = 1 \,|\, \delta_i, \theta) = \frac{\exp(\theta - \delta_i)}{1 + \exp(\theta - \delta_i)}. \quad (6.25)$$

c. Compute $X_+ = \sum_i X_i$. If $X_+ = X_{p+}$ then θ is accepted as a sample from (6.24). Otherwise, Steps a, b, and c are repeated.

Step 2. $$\mu_r \sim p(\mu_r \mid \boldsymbol{\theta}_r, \sigma_r^2) = \mathcal{N}\left(\mu_r; \frac{\sum_{p\in\{R^{(1)}\}} \theta_p}{\sigma_r^2}, \frac{1}{\frac{1}{100}+\frac{N^{(1)}}{\sigma_r^2}}\right); \quad (6.26)$$

Step 3. $$\sigma_r^2 \sim p(\sigma_r^2|\boldsymbol{\theta}_r, \mu_r) = \text{Inv-}\mathcal{G}\left(\sigma_r^2; .001+\frac{N^{(1)}}{2}, .001+\frac{\sum_{p\in\{R^{(1)}\}}(\theta_p-\mu_r)^2}{2}\right); \quad (6.27)$$

Step 4. $$\forall i \in \{r\} : \delta_i \sim p(\delta_i \mid \dots) \propto \mathcal{N}(\delta_i; 0, 100) \prod_{p\in\{R^{(1)}\}} \frac{\exp(x_{pi}(\theta_p-\delta_i))}{1+\exp(\theta_p-\delta_i)}. \quad (6.28)$$

The normalising constant for the distribution in Equation 6.28 does not have a closed form solution, therefore we use Metropolis algorithm (Metropolis et al., 1953) to sample from this conditional posterior with a normal proposal density centred around the current value of the parameter.

Step 5. At the end of each iteration, the parameters have to be re-scaled to keep the chosen identification of the scale, namely $\bar{\delta}_r = 0$: $\mu_r = \mu_r - \bar{\delta}_r$, and $\delta_i = \delta_i - \bar{\delta}_r, \forall i \in \{r\}$. The individual person parameters do not have to be re-scaled since their values at the end of iteration t do not influence the values of the parameters in iteration $t+1$.

By repeatedly going through these five steps, samples from the posterior in (6.22) and analogously in (6.23) are obtained. The priors for the parameters used in the next step - estimation of the cut-score s_{pass} - are the following:

$$p(\mu_r,\sigma_r^2,\boldsymbol{\delta}_r,\mu_c^*,\sigma_c^2,\boldsymbol{\delta}_c^*) = \mathcal{N}(\mu_r;\mu_{r0},\sigma_{r0}^2)\text{Inv-}\mathcal{G}(\sigma_r^2;\alpha_{r0},\beta_{r0})\prod_{i\in\{r\}}\mathcal{N}(\delta_i;\mu_{\delta_i},\sigma_{\delta_i}^2)$$
$$\mathcal{N}(\mu_c^*;\mu_{c0}^*,\sigma_{c0}^2)\text{Inv-}\mathcal{G}(\sigma_c^2;\alpha_{c0},\beta_{c0})\prod_{i\in\{c\}}\mathcal{N}(\delta_i^*;\mu_{\delta_i^*},\sigma_{\delta_i^*}^2) \quad (6.29)$$

where the prior means and the prior variances of the population means and the item difficulties are equal to the average values of these parameters across the samples from (6.22) and (6.23) and to the variances of the sampled values of these parameters, respectively. The hyper parameters of the inverse-gamma distributions of σ_r^2 and σ_c^2 are also chosen based on the averages and the variances of the sampled values of these parameters. Since the mean and the variance of a random variable with the inverse-gamma distribution with parameters α and β are equal to $\frac{\beta}{\alpha-1}$ and $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$, respectively, we choose the following values for the hyper parameters:

$$\alpha_{r0} = \frac{(\bar{\sigma}_r^2)^2}{Var(\sigma_r^2)} + 2 \quad (6.30)$$

$$\beta_{r0} = (\bar{\sigma}_r^2)\left(\frac{(\bar{\sigma}_r^2)^2}{Var(\sigma_r^2)} + 1\right) \quad (6.31)$$

where $\bar{\sigma}_r^2$ and $Var\left(\sigma_r^2\right)$ are the average and the variance of the sampled values of σ_r^2. The hyper parameters for the distribution of σ_c^2 are chosen analogously.

The average sampled values of the item difficulties are used to facilitate the collection of expert judgements. In the Angoff method they are used to compute $\hat{\delta}_c$ and $\hat{\delta}_r$ (see Equation 6.8). In the rulers method they are used to select the items and to determine their position on the rulers which experts move in the third stage of the procedure.

Sampling from the posterior distribution needed to determine the cut-score

To estimate the cut-score s_{pass} we need to sample from the posterior:

$$p(\boldsymbol{\theta}_r,\mu_r,\sigma_r^2,\boldsymbol{\theta}_c^*,\mu_c^*,\sigma_c^2,\boldsymbol{\theta}_1,\mu_1,\sigma_1^2,\ldots,\boldsymbol{\theta}_G,\mu_G,\sigma_G^2,\boldsymbol{\delta}_r,\boldsymbol{\delta}_c^*,\tau|\mathbf{X},\mathbf{Y},\mathbf{Z}_1,\ldots,\mathbf{Z}_G) \quad (6.32)$$

which is proportional to the product of the density of the data:

$$\begin{aligned} f(\mathbf{X}^{(2)},\mathbf{Y}^{(2)},\mathbf{Z}_1,\ldots,\mathbf{Z}_G) = &\prod_{p\in\{R^{(2)}\}}\prod_{i\in\{r\}}\frac{\exp(x_{pi}(\theta_p-\delta_i))}{1+\exp(\theta_p-\delta_i)}\prod_{p\in\{C^{(2)}\}}\prod_{i\in\{c\}}\frac{\exp\left(y_{pi}(\theta_p^*-\delta_i^*)\right)}{1+\exp(\theta_p^*-\delta_i^*)}\times \\ &\times\prod_{g=1}^{G}\prod_{p\in\{E_g\}}\prod_{i\in\{e_g\cap r\}}\frac{\exp\left(z_{gpi}(\theta_p-\delta_i)\right)}{1+\exp(\theta_p-\delta_i)}\prod_{i\in\{e_g\cap c\}}\frac{\exp\left(z_{gpi}(\theta_p-\delta_i^*-\tau)\right)}{1+\exp(\theta_p-\delta_i^*-\tau)}, \end{aligned} \quad (6.33)$$

where $\{E_g\}$ denotes the set of persons in linking group G and $\{e_g\}$ denotes the set of items answered by linking group G; and the joint prior distribution

$$\begin{aligned} &p(\boldsymbol{\theta}_r,\mu_r,\sigma_r^2,\boldsymbol{\theta}_c^*,\mu_c^*,\sigma_c^2,\boldsymbol{\theta}_1,\mu_1,\sigma_1^2,\ldots,\boldsymbol{\theta}_G,\mu_G,\sigma_G^2,\boldsymbol{\delta}_r,\boldsymbol{\delta}_c^*,\tau) = \\ &= p(\tau)p(\mu_r,\sigma_r^2,\boldsymbol{\delta}_r,\mu_c^*,\sigma_c^2,\boldsymbol{\delta}_c^*)\prod_{p\in\{R^{(2)}\}}\mathcal{N}(\theta_p;\mu_r,\sigma_r^2)\prod_{p\in\{C^{(2)}\}}\mathcal{N}(\theta_p^*;\mu_c^*,\sigma_c^2))\times \\ &\times\prod_g\left(\mathcal{N}(\mu_g;0,100)\text{Inv-}\mathcal{G}(\sigma_g^2;.001,.001)\prod_{p\in\{E_g\}}\mathcal{N}(\theta_p;\mu_g,\sigma_g^2)\right), \end{aligned} \quad (6.34)$$

where the priors of the population means and variances of the reference and the current population, and the item difficulties are estimated from the training data, see Equation 6.29.

The initial values, denoted by a superscript (0), for all the parameters are chosen in the same way as when sampling from (6.22) and (6.23) with the following initial values for the additional parameters $\mu_g^{(0)}=0,\sigma_g^{2(0)}=1,\forall g\in[1:G]$

Sampling from the conditional posteriors of the parameters in (6.32) is similar to sampling from the conditional posteriors of $\boldsymbol{\theta}_r$, μ_r, σ_r^2, and $\boldsymbol{\delta}_r$ in (6.22). The

following steps are involved:

Step 1a. $\forall p \in \{R^{(2)}\}$:

$$\theta_p \sim p(\theta_p \mid \dots) \propto \mathcal{N}(\theta_p; \mu_r, \sigma_r^2) \prod_{i \in \{r\}} \frac{\exp(x_{pi}(\theta_p - \delta_i))}{1 + \exp(\theta_p - \delta_i)}; \tag{6.35}$$

Step 1b. $\forall p \in \{C^{(2)}\}$:

$$\theta_p^* \sim p(\theta_p^* \mid \dots) \propto \mathcal{N}(\theta_p^*; \mu_c^*, \sigma_c^2) \prod_{i \in \{c\}} \frac{\exp(y_{pi}(\theta_p^* - \delta_i^*))}{1 + \exp(\theta_p^* - \delta_i^*)}; \tag{6.36}$$

Step 1c. $\forall g \in [1:G], \forall p \in \{E_g\}$

$$\theta_p \sim p(\theta_p | \dots) \propto \mathcal{N}(\theta_p; \mu_g, \sigma_g^2) \prod_{i \in \{r \cap e_g\}} \frac{\exp(z_{gpi}(\theta_p - \delta_i))}{1 + \exp(\theta_p - \delta_i)} \prod_{i \in \{c \cap e_g\}} \frac{\exp(z_{gpi}(\theta_p - \delta_i^* - \tau))}{1 + \exp(\theta_p - \delta_i^* - \tau)}; \tag{6.37}$$

Step 2a.

$$\mu_r \sim p(\mu_r \mid \dots) = \mathcal{N}\left(\mu_r; \frac{\frac{\mu_{r0}}{\sigma_{r0}^2} + \frac{\sum_{p \in \{R^{(2)}\}} \theta_p}{\sigma_r^2}}{\frac{1}{\sigma_{r0}^2} + \frac{N^{(2)}}{\sigma_r^2}}, \frac{1}{\frac{1}{\sigma_{r0}^2} + \frac{N^{(2)}}{\sigma_r^2}}\right); \tag{6.38}$$

Step 2b.

$$\mu_c^* \sim p(\mu_c^*, \mid \dots) = \mathcal{N}\left(\mu_c^*; \frac{\frac{\mu_{c0}^*}{\sigma_{c0}^2} + \frac{\sum_{p \in \{C^{(2)}\}} \theta_p^*}{\sigma_c^2}}{\frac{1}{\sigma_{c0}^2} + \frac{M^{(2)}}{\sigma_c^2}}, \frac{1}{\frac{1}{\sigma_{c0}^2} + \frac{M^{(2)}}{\sigma_c^2}}\right); \tag{6.39}$$

Step 2c. $\forall g \in [1:G]$:

$$\mu_g \sim p(\mu_g, \mid \dots) = \mathcal{N}\left(\mu_g; \frac{\frac{\sum_{p \in \{E_g\}} \theta_p}{\sigma_g^2}}{\frac{1}{100} + \frac{N_e}{\sigma_g^2}}, \frac{1}{\frac{1}{100} + \frac{N_e}{\sigma_g^2}}\right); \tag{6.40}$$

Step 3a. $\sigma_r^2 \sim p(\sigma_r^2 \mid \dots) = \text{Inv-}\mathcal{G}\left(\sigma_r^2; \alpha_{r0} + \frac{N^{(2)}}{2}, \beta_{r0} + \frac{\sum_{p \in \{R^{(2)}\}} (\theta_p - \mu_r)^2}{2}\right);$ (6.41)

Step 3b. $\sigma_c^2 \sim p(\sigma_c^2 \mid \dots) = \text{Inv-}\mathcal{G}\left(\sigma_c^2; \alpha_{c0} + \frac{M^{(2)}}{2}, \beta_{c0} + \frac{\sum_{p \in \{C^{(2)}\}} (\theta_p^* - \mu_c^*)^2}{2}\right);$ (6.42)

Step 3c. $\forall g \in [1:G]$:

$$\sigma_g^2 \sim p(\sigma_g^2 \mid \dots) = \text{Inv-}\mathcal{G}\left(\sigma_g^2; .001 + \frac{N_e}{2}, .001 + \frac{\sum_{p\in\{E_g\}}(\theta_p - \mu_g)^2}{2}\right); \quad (6.43)$$

Step 4a. $\forall i \in \{r/\{e_1 \cap \dots \cap e_G\}\}$:

$$\delta_i \sim p(\delta_i \mid \dots) \propto \mathcal{N}(\delta_i; \mu_{\delta_i}, \sigma_{\delta_i}^2) \prod_{p\in\{R^{(2)}\}} \frac{\exp(x_{pi}(\theta_p - \delta_i))}{1 + \exp(\theta_p - \delta_i)}; \quad (6.44)$$

Step 4b. $\forall i \in \{c/\{e_1 \cap \dots \cap e_G\}\}$:

$$\delta_i^* \sim p(\delta_i \mid \dots) \propto \mathcal{N}(\delta_i^*; \mu_{\delta_i^*}, \sigma_{\delta_i^*}^2) \prod_{p\in\{C^{(2)}\}} \frac{\exp(y_{pi}(\theta_p^* - \delta_i^*))}{1 + \exp(\theta_p^* - \delta_i^*)}; \quad (6.45)$$

Step 4c. $\forall g \in [1:G], \forall i \in \{r \cap e_g\}$:

$$\delta_i \sim p(\delta_i \mid \dots) \propto \mathcal{N}(\delta_i; \mu_{\delta_i}, \sigma_{\delta_i}^2) \prod_{p\in\{R^{(2)}\}} \frac{\exp(x_{pi}(\theta_p - \delta_i))}{1 + \exp(\theta_p - \delta_i)} \prod_{p\in\{E_g\}} \frac{\exp(z_{gpi}(\theta_p - \delta_i))}{1 + \exp(\theta_p - \delta_i)}; \quad (6.46)$$

Step 4d. $\forall g \in [1:G], \forall i \in \{c \cap e_g\}$:

$$\delta_i^* \sim p(\delta_i^* \mid \dots) \propto \mathcal{N}(\delta_i^*; \mu_{\delta_i^*}, \sigma_{\delta_i^*}^2) \prod_{p\in\{C^{(2)}\}} \frac{\exp(x_{pi}(\theta_p^* - \delta_i^*))}{1 + \exp(\theta_p^* - \delta_i^*)} \prod_{p\in\{E_g\}} \frac{\exp(z_{gpi}(\theta_p - \delta_i^* - \tau))}{1 + \exp(\theta_p - \delta_i^* - \tau)}. \quad (6.47)$$

Step 5. $\tau \sim p(\tau \mid \dots) \propto p(\tau) \prod_{g=1}^{G} \prod_{p\in E_g} \prod_{i\in\{c\cap e_g\}} \frac{\exp(z_{gpi}(\theta_p - \delta_i^* - \tau))}{1 + \exp(\theta_p - \delta_i^* - \tau)}, \quad (6.48)$

Step 6. The parameters are re-scaled to make sure that $\bar{\delta}_r = 0$ and $\bar{\delta}_c^* = 0$: $\mu_r = \mu_r - \bar{\delta}_r$, $\mu_g = \mu_g - \bar{\delta}_r, \forall g \in [1:G]$, $\delta_i = \delta_i - \bar{\delta}_r, \forall i \in \{r\}$, $\mu_c^* = \mu_c^* - \bar{\delta}_c^*$, and $\delta_i^* = \delta_i^* - \bar{\delta}_c^*, \forall i \in \{c\}$.

Sampling from the conditional posteriors in Steps 1a, 1b, and 1c is analogous to sampling from (6.24). Sampling from the conditional posteriors in Steps 4a, 4b, 4c and 4d is analogous to sampling from (6.28). Sampling from the conditional posterior in Step 5 is similar to sampling from the conditional posterior distributions of the item difficulties, the Metropolis algorithm is used to sample from this distribution.

Estimating the cut-score s_{pass}

After the burn-in, at each iteration t the unobserved responses of the persons from the reference population to the current exam ($\mathbf{X}^*$) are simulated according to the Rasch model using the values of the model parameters at iteration t sampled from (6.32) using the Gibbs Sampler described in the previous subsection:

$$x_{pi}^{*(t)} \sim \text{Bernoulli}\left(\frac{\exp\left(\theta_p^{(t)} - (\delta_i^{*(t)} + \tau^{(t)})\right)}{1 + \exp\left(\theta_p^{(t)} - (\delta_i^{*(t)} + \tau^{(t)})\right)}\right), \forall p \in \{R^{(2)}\}, \forall i \in \{c\}. \quad (6.49)$$

A sample from the posterior distribution of the cut-score, denoted by $s_{pass}^{(t)}$, is such a score that the number of students from the reference population with observed scores on the reference test below s_{ref} is as close as possible to the number of students from the reference population with simulated scores on the current test at iteration t below this score:

$$s_{pass}^{(t)} = \text{argmin}_s \left(\left(\sum_{p \in \{R^{(2)}\}} \left(\mathcal{I}\left(\sum_{i \in \{r\}} x_{pi} < s_{ref}\right) - \mathcal{I}\left(\sum_{i \in \{c\}} x_{pi}^{*(t)} < s\right)\right)\right)^2\right). \quad (6.50)$$

Using a large number of sampled values from the posterior in (6.32), a sequence of values $\{s_{pass}^{(1)}, s_{pass}^{(2)}, \ldots, s_{pass}^{(T)}\}$ is obtained, which is a sample from the posterior distribution of the cut-score is obtained. The maximum a posteriori estimate of s_{pass} is the mode of this sample. The posterior variance of s_{pass} is the variance in this sample.

Appendix B: Approximating experts' judgements with a skew-normal distribution

In this section we describe how to choose the parameters of the skew-normal distribution $p_{ej}(\tau)$, such that its mode, 5-th percentile and 95-th percentile would match the values of $\hat{\tau}_{ej}$, $\tau_{ej}^{\min}$ and $\tau_{ej}^{\max}$, respectively. The skew-normal distribution with parameters α, ξ, ω is given by:

$$f(x) = \frac{2}{\sqrt{2\pi}\omega} \exp\left(\frac{-(x-\xi)^2}{2\omega^2}\right) \int_{-\infty}^{\alpha\frac{x-\xi}{\omega}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt \quad (6.51)$$

First, we determine the value of the skewness parameter α. Let us by $q_p(\alpha, \xi, \omega)$ denote the p-th percentile of the skew-normal distribution with parameters α, ξ and ω and by $m(\alpha, \xi, \omega)$ the mode of this distribution. The larger the skewness

is, the further away from 1 the ratio below is:

$$r(\alpha, \xi, \omega) = \frac{q_{95}(\alpha, \xi, \omega) - m(\alpha, \xi, \omega)}{m(\alpha, \xi, \omega) - q_5(\alpha, \xi, \omega)} \tag{6.52}$$

The values of the parameters ξ and ω do not influence the value of this ratio. For all values of α ranging from -4 to 4, with equal interval steps of .001, we estimated the mode $m(\alpha, 0, 1)$ (with the precision up to .0001, which is sufficient for the application at hand) and computed the ratio $r(\alpha, 0, 1)$, using the 'sn' R-package. And then for each judgement of each expert we chose:

$$\alpha_{ej} = \arg\min_{\alpha} \left| \frac{\tau_{ej}^{\max} - \hat{\tau}_{ej}}{\hat{\tau}_{ej} - \tau_{ej}^{\min}} - r(\alpha, 0, 1) \right|. \tag{6.53}$$

Second, we choose the value of the parameter ω which determines the spread of the distribution:

$$\omega_{ej} = \frac{\tau_{ej}^{\max} - \tau_{ej}^{\min}}{q_{95}(\alpha_{ej}, 0, 1) - q_5(\alpha_{ej}, 0, 1)}. \tag{6.54}$$

And finally, we choose the value of the parameter ξ which determines the location of the distribution:

$$\xi_{ej} = \tau_{ej}^{\max} - q_{95}(\alpha_{ej}, 0, 1)\omega_{ej}. \tag{6.55}$$

Next, we show how to approximate the expert-specific normalising constant, denoted by Z_e, for the product of skew-normal distributions in Equation 6.19:

$$Z_e = \int_{-\infty}^{+\infty} \left(\prod_{j=1}^{J_e} \frac{2}{\omega_{ej}} \phi\left(\frac{\tau - \xi_{ej}}{\omega_{ej}} \right) \Phi\left(\alpha_{ej} \frac{\tau - \xi_{ej}}{\omega_{ej}} \right) \right) d\tau, \tag{6.56}$$

where $\phi(x)$ denotes the standard normal density and $\Phi(x)$ denotes the standard normal cumulative distribution function. This integral can be approximated using the Gauss-Hermite the weights $\mathbf{w} = \{w_1, \ldots, w_K\}$ and the nodes $\mathbf{y} = \{y_1, \ldots, y_K\}$:

$$Z_e \approx \frac{2}{\sqrt{\pi}} \sum_{i=1}^{K} w_i \Phi\left(\sqrt{2}\alpha_{e1} y_i\right) \times \prod_{j=2}^{J_e} \frac{2}{\omega_{ej}} \phi\left(\frac{\sqrt{2}\omega_{e1} y_i + \xi_{e1} - \xi_{ej}}{\omega_{ej}} \right) \Phi\left(\alpha_{ej} \frac{\sqrt{2}\omega_{e1} y_i + \xi_{e1} - \xi_{ej}}{\omega_{ej}} \right). \tag{6.57}$$

This integral has to be computed only once for each expert, therefore we can use a very large number of nodes to obtain an accurate approximation. In the empirical example we used $K = 20,000$.

Chapter 7

Can IRT solve the missing data problem in test equating?

[1] **Abstract.** In this paper test equating is considered as a missing data problem. The unobserved responses of the reference population to the new test must be imputed to specify a new cutscore. The proportion of students from the reference population that would have failed the new exam and those having failed the reference exam are made approximately the same. We investigate whether item response theory (IRT) makes it possible to identify the distribution of these missing responses and the distribution of test scores from the observed data without parametric assumptions for the ability distribution. We show that while the score distribution is not fully identifiable, the uncertainty about the score distribution on the new test due to non-identifiability is very small. Moreover, ignoring the non-identifiability issue and assuming a normal distribution for ability may lead to bias in test equating, which we illustrate in simulated and empirical data examples.
Keywords: item response theory, incomplete design, marginal Rasch model, missing data, non-identifiability, test equating.

7.1 Introduction

One of the advantages of item response theory (IRT) over classical test theory is its ability to handle incomplete designs. Among the important applications in which data are missing by design is test equating, where results of different test forms must be made comparable by accounting for the two key facts. The first is that the reference and the new tests need not be of the same difficulty, and

[1]This chapter has been published as Bolsinova, M. & Maris, G. (2016). Can IRT solve the missing data problem in test equating? *Frontiers in Psychology, 6*: 1956. doi: 10.3389/fpsyg.2015.01956. Author contributions: B.M. and M.G. designed the research, B.M. performed the research, B.M. wrote the paper, M.G. provided feedback on the manuscript.

the second is that the reference and the new populations need not have the same ability distribution (Kolen & Brennan, 2004; von Davier, 2011).

Suppose, that the same students respond both to the reference and to the new test. Assume, for the sake of the argument, that both tests are scored with a number correct score. It is clear that, if both tests represent the same underlying construct, both scores are automatically equated. The need for equating scores derives from the fact that for every student we only observe the response to either the reference or the new test. That is, it derives from the fact that there is a missing data problem.

Equating procedures are methods to overcome the missing data problem. There are many different methods for score equating with some methods based on IRT and other on classical test theory. These methods are covered in detail by, for example, Kolen and Brennan (2004), von Davier (2011), von Davier, Holland and Thayer (2004), Holland and Dorans (2006), Livingston(2004). Most all equating procedures are such that all students with the same score on the reference test get the same equated score on the new test. This in contrast to both the complete data case we considered above, and more modern (multiple) imputation based techniques (Rubin, 1987).

The central question we consider in this paper is whether the distribution of the missing data (marginal or conditionally on the observed data) is in principle identifiable from the observed data. If the marginal distribution is not identifiable, neither is the conditional distribution needed to impute the missing data. Regardless of the preferred equating method, if the distribution of the missing data is not identifiable, the missing data problem can not be solved.

Suppose we take the most modest form of equating: translating the scores on the new test to a pass/fail decision (i.e. selecting a cut-score below which a student fails) consistently with the pass/fail criterion on the reference test, i.e. such the passing percentage in the reference population would be the same on the new test as it is on the reference test. To specify a new cutscore, it is sufficient to estimate the distribution of the scores of the persons from the reference population to the new test, denoted by $p(X_{+mis})$[2] . As we will show in the paper, this is not possible using an IRT model given the observed data only. Hence, solving more complicated problems of equating (obtaining a full correspondences between the scores on the two tests) is also not possible.

When IRT is used for test equating, the joint distribution of the observed data (responses of the reference population on the reference test, denoted by $p(\mathbf{X}_{obs})$) and the missing data (responses of the reference population on new test, denoted by $p(\mathbf{X}_{mis})$) is modelled by a marginal IRT model that consists of a conditional distribution of the data given a latent variable θ and a population distribution

[2]For simplicity, we considered a situation in which the new and the reference test do not have any common items. In the general case, the missing data are responses to the items that belong to the new test but not the reference test

$f(\theta)$. Two elements are required to estimate the distribution of missing responses $p(\mathbf{X}_{mis})$. First, the parameters of the items from the new test and from the reference test must be placed on the common scale. Second, the ability distribution of the reference population given the observed data $f(\theta \mid \mathbf{X}_{obs})$ must be estimated. In this paper, we have assumed that the tests are well connected through a linking design[3] and the IRT model is correctly specified and, therefore, the first element of equating is fully satisfied. We have focused on the second element, which is usually ignored in test equating practice. The problem is that the full distribution of ability $f(\theta)$ is not identifiable, as has been shown by Cressie and Holland (1983). Consequently, as we show in this paper, the distribution $p(X_{+mis})$ is also not identified from the observed data only. This issue is usually ignored in test equating practice, and instead a parametric distribution, usually a normal distribution, is assumed for $f(\theta)$. This assumption is not guaranteed to hold in practice, therefore it is important to consider to what extent the problem of inferring the distribution of missing responses can be solved without extra distributional assumptions.

We will discuss the problem of non-identifiability of $p(X_{+mis})$ using the marginal Rasch model (RM) for dichotomous data, which has only one parameter in the conditional model, as an example. The RM is chosen here for convenience; the identifiability issues are present at the level of the marginal model and are therefore not affected by the choice of a particular parametric conditional model.

In this study we investigate the extent to which the unavoidable uncertainty about the score distribution $p(X_{+mis})$ that comes from non-identifiability is problematic in practice. The main purpose of this study is not to introduce a new method for test equating, but to highlight a fundamental property of marginal IRT models. This property is that in IRT equating the score distribution $p(X_{+mis})$ can not be identified without making extra assumptions about the parametric shape of the ability distribution, and the practical consequences of ignoring this property.

7.2 Why IRT cannot solve the missing data problem

In this section we describe a simple model for test equating that tries (unsuccessfully) to predict missing responses from the observed data without additional distributional assumptions. The marginal RM is:

$$p(\mathbf{X}_{obs} = \mathbf{x}) = \int_{\mathbb{R}} \prod_i \frac{\exp(x_i(\theta - \delta_i))}{1 + \exp(\theta - \delta_i)} f(\theta)\, d\theta, \tag{7.1}$$

where $\mathbf{x}$ is a vector of dichotomous responses with $x_i = 1$ if item i is answered correctly and 0 otherwise; δ_i is the difficulty parameter of item i. There is assumed

[3]For a review of different linking designs see, for example, Angoff (1971), Wright and Stone (1979), Lord (1980), Petersen, Kolen and Hoover (1989), Kolen and Brennan (2004). Some of these linking designs are presented in Appendix D.

to be a population distribution $f(\theta)$; however, its parametric shape is not known.

Following Cressie and Holland (1988), the marginal RM in (7.1) can be re-written as

$$p(\mathbf{X}_{obs} = \mathbf{x}) = \prod_i (\exp(-\delta_i))^{x_i} \int_{\mathbb{R}} (\exp(\theta))^{x_+} \prod_i \frac{1}{1 + \exp(\theta - \delta_i)} f(\theta)\, d\theta, \quad (7.2)$$

where x_+ is the number of items answered correctly. It can be seen that

$$f(\theta \,|\, \mathbf{X}_{obs} = \mathbf{0}) \propto \prod_i \frac{1}{1 + \exp(\theta - \delta_i)} f(\theta), \quad (7.3)$$

which is the posterior distribution of ability given that the responses to all items are incorrect. Therefore,

$$p(\mathbf{X}_{obs} = \mathbf{x}) \propto \prod_i (\exp(-\delta_i))^{x_i} E((\exp(\Theta))^{x_+} \,|\, \mathbf{X}_{obs} = \mathbf{0}). \quad (7.4)$$

To make $p(\mathbf{X}_{obs} = \mathbf{x})$ a proper density, a normalising constant should be added. A convenient parameterisation of the marginal RM (Maris, Bechger, & San Martin, 2015) is:

$$p(\mathbf{X}_{obs} = \mathbf{x}) = \frac{\prod_i b_i^{x_i} \lambda_{x_+}}{\sum_{s=0}^{n} \gamma_s(\mathbf{b}) \lambda_s}, \quad (7.5)$$

where $\mathbf{b} = \{b_1, b_2, \ldots, b_n\}$ is a vector of item parameters that are transformations of difficulty parameters: $b_i = \exp(-\delta_i)$; $\boldsymbol{\lambda} = \{\lambda_0, \lambda_1, \ldots, \lambda_n\}$ is a vector of population parameters, and $\gamma_t(\mathbf{b})$ denotes a t-th order elementary symmetric polynomial (Verhelst, Glas, & van der Sluis, 1984). The denominator ensures that the distribution integrates to 1. The model in (7.5) is a marginal Rasch model if and only if $\boldsymbol{\lambda}$ is a sequence of moments of a distribution. This imposes a set of inequality constraints on the parameters (Shohat & Tamarkin, 1943):

$$\det \begin{bmatrix} \lambda_0 & \lambda_1 & \ldots & \lambda_m \\ \lambda_1 & \lambda_2 & \ldots & \lambda_{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_m & \lambda_{m+1} & \ldots & \lambda_{2m} \end{bmatrix} \geq 0, m = 0, 1, 2, \ldots \quad (7.6)$$

and

$$\det \begin{bmatrix} \lambda_1 & \lambda_2 & \ldots & \lambda_{m+1} \\ \lambda_2 & \lambda_3 & \ldots & \lambda_{m+2} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{m+1} & \lambda_{m+2} & \ldots & \lambda_{2m+1} \end{bmatrix} \geq 0, m = 0, 1, 2, \ldots \quad (7.7)$$

The extended Rasch model [ERM] (Cressie & Holland, 1983; Tjur, 1982; Maris et

al., 2015) does not have these restrictions.

We now apply the ERM to test equating. Let us consider the joint density of the response vectors $\mathbf{X}_{obs}$ and $\mathbf{X}_{mis}$:

$$p(\mathbf{X}_{obs} = \mathbf{x}, \mathbf{X}_{mis} = \mathbf{x}^*) = \frac{\prod_{i=1}^{n} b_i^{x_i} \prod_{j=1}^{m} d_j^{x_j^*} \eta_{x_+ + x_+^*}}{\sum_{t=0}^{n+m} \gamma_t(\mathbf{b}, \mathbf{d}) \eta_t}, \tag{7.8}$$

where $\mathbf{d} = \{d_1, \ldots, d_m\}$ are the parameters of the items in the new test (analogous to $\mathbf{b}$) and $\boldsymbol{\eta} = \{\eta_0, \eta_1, \ldots, \eta_{n+m}\}$ is a vector of $(n+m+1)$ population parameters corresponding to a combined test consisting of the items from both the reference and the new exams. It can be derived that the marginal distribution of the scores of the reference population on the new test is (see Appendix A for details):

$$Pr(X_{+mis} \leq T) = \sum_{t=0}^{T} p(X_{+mis} = t) = \sum_{t=0}^{T} \frac{\gamma_t(\mathbf{d}) \sum_{s=0}^{n} \gamma_s(\mathbf{b}) \eta_{s+t}}{\sum_{u=0}^{n+m} \gamma_t(\mathbf{b}, \mathbf{d}) \eta_u}. \tag{7.9}$$

The expression for this distribution contains parameters $\boldsymbol{\eta}$, whereas the density of the observed data contains parameters $\boldsymbol{\lambda}$. The parameters $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$ are related to each other as follows (see Appendix A for details):

$$\lambda_s = \sum_{t=0}^{m} \gamma_t(\mathbf{d}) \eta_{t+s}, \forall s \in [0, n]. \tag{7.10}$$

The parameters $\boldsymbol{\lambda}$ are identified from the data (up to a multiplicative constant), whereas parameters $\boldsymbol{\eta}$ are not; this is because in this system of $(n+1)$ equations (4) there are $(n+m+1)$ unknowns. Therefore, having observed only data $\mathbf{X}_{obs}$, we cannot make direct inferences about the distribution of X_{+mis}. Hence, IRT cannot solve the missing data problem.

7.3 What IRT allows us to infer about the distribution of missing responses

The conclusion at the end of the previous section does not mean that we do not know anything about the parameters $\boldsymbol{\eta}$ or the score distribution. The relations between $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$ impose restrictions on the values that $\boldsymbol{\eta}$ can take, and therefore, on the score distribution. Before considering what is and is not known about the score distribution $p(X_{+mis})$, we should discussed some additional constraints for parameters $\boldsymbol{\eta}$.

Along with the restriction given by the relations with the identified parameters (7.10), there are other restrictions that the parameters $\boldsymbol{\eta}$ must satisfy in

order to be parameters of the ERM. First, they must be positive to ensure that all probabilities in (7.9) are positive. To derive a second constraint, consider the probability of answering item i correctly given the rest score on the test:

$$\Pr(X_i = 1 \mid X^{(i)}_{+obs} + X_{+mis} = s) = \frac{\Pr(X_i = 1, X^{(i)}_{+obs} + X_{+mis} = s)}{\Pr(X^{(i)}_{+obs} + X_{+mis} = s)} = \\ = \frac{b_i \gamma_s(\mathbf{b}^{(i)}, \mathbf{d})\eta_{s+1}}{b_i \gamma_s(\mathbf{b}^{(i)}, \mathbf{d})\eta_{s+1} + \gamma_s(\mathbf{b}^{(i)})\eta_s} = \frac{b_i \frac{\eta_{s+1}}{\eta_s}}{1 + b_i \frac{\eta_{s+1}}{\eta_s}}, \quad (7.11)$$

where $\mathbf{b}^{(i)}$ denotes a vector of item parameters of all items in the reference test except item i, and $X^{(i)}_{+obs}$ is the sum score on these items; that is, the rest score. From the measurement perspective, this probability should increase when s increases (Junker, 1993; Junker & Sijtsma, 2000). This ensures that all item-rest correlations are positive, so that it makes sense to score the particular set of items together as one test. For this to be true, the ratios $\frac{\eta_{s+1}}{\eta_s}$ must form a monotonically increasing sequence. The inequality constraint

$$\frac{\eta_1}{\eta_0} \leq \frac{\eta_2}{\eta_1} \leq \frac{\eta_3}{\eta_2} \leq \cdots \leq \frac{\eta_{n+m}}{\eta_{n+m-1}} \quad (7.12)$$

can be specifies as a part of in the prior distribution of the population parameters (see Appendix E for details).

An alternative motivation for using the constraints in Equation 7.12 is that they follow from an important feature of the marginal RM, namely that

$$\frac{\eta_{s+2}}{\eta_s} - \left(\frac{\eta_{s+1}}{\eta_s}\right)^2 \quad (7.13)$$

is the (posterior) variance of $\exp(\theta)$ of a person with a score of s (Maris et al., 2015). The monotonicity constraints in Equation 7.12 follow from non-negativity of variance. Therefore, the constraints in (7.12) are necessary but not sufficient for the parameters to satisfy the moment constraints of the marginal RM. Hence, the model we are using for equating is an ERM with the monotonicity constraints. As will be shown in the next subsection this restriction enables to reduce the uncertainty about the score distribution on the new test.

7.3.1 A simple case: $m = 1$

In this subsection we derive the uncertainty about the marginal probability of answering a new item correctly, given the observed responses to n items. Let us consider the simplest case in which the number of items in the new test is equal to one ($m = 1$). Because we are ignoring the effect of sampling variability on the

uncertainty, we consider all identifiable parameters ($\mathbf{b}$ and $\boldsymbol{\lambda}$) known.

Let $\boldsymbol{\lambda} = \{\lambda_0, \lambda_1, \ldots, \lambda_n\}$ denote the set of identifiable population parameters; $\boldsymbol{\eta} = \{\eta_0, \eta_1, \eta_2, \ldots, \eta_{n+1}\}$ the set parameters for the combined test; and d the item parameter of the new item. The relations between $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$ form a system of linear equations:

$$\begin{pmatrix} \lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} = \begin{pmatrix} 1 & \mathrm{d} & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & \mathrm{d} \end{pmatrix} \begin{pmatrix} \eta_0 \\ \eta_1 \\ \vdots \\ \eta_n \\ \eta_{n+1} \end{pmatrix}. \tag{7.14}$$

This system of $n+1$ equations does not have a unique solution because the number of unknowns $(n+2)$ is larger than the number of equations. The general solution is:

$$\begin{pmatrix} \eta_0 \\ \eta_1 \\ \vdots \\ \eta_{n+1} \end{pmatrix} = \begin{pmatrix} k \\ \frac{\lambda_0}{d} - \frac{k}{d} \\ \vdots \\ \sum_{t=0}^{n} \frac{(-1)^{n-t}\lambda_t}{d^{n+1-t}} + (-1)^{n+1}\frac{k}{d^{n+1}} \end{pmatrix}, \tag{7.15}$$

where k is a parameter that captures all uncertainty about $\boldsymbol{\eta}$, such that the unique solution to the system of equations can be computed when k is known. This parameter is not completely free because $\boldsymbol{\eta}$ must satisfy the set of inequalities:

$$\begin{cases} \eta_s > 0, \forall s \in [0 : (n+1)], \\ \frac{\eta_{s+1}}{\eta_s} \geq \frac{\eta_s}{\eta_{s-1}}, \forall s \in [1 : n]. \end{cases} \tag{7.16}$$

We are interested in the probability of answering the new item correctly, which can be written as a function of k:

$$\Pr(X_{mis} = 1) = \pi_+(k) = \frac{d\sum_{t=0}^{n}\gamma_t(\mathbf{b})\eta_{t+1}}{d\sum_{t=0}^{n}\gamma_t(\mathbf{b})\eta_{t+1} + \sum_{t=0}^{n}\gamma_t(\mathbf{b})\eta_t}. \tag{7.17}$$

Using the solutions of the system of equations, one can derive (for details, see Appendix B):

$$\pi_+(k) = 1 - \frac{k\sum_{t=0}^{n}\frac{(-1)^{t-1}\gamma_t(\mathbf{b})}{d^t}}{\sum_{t=0}^{n}\gamma_t(\mathbf{b})\lambda_t} + \frac{\sum_{t=1}^{n}\sum_{s=0}^{t-1}\frac{(-1)^{t-s}\gamma_t(\mathbf{b})\lambda_t}{d^{t-s}}}{\sum_{t=0}^{n}\gamma_t(\mathbf{b})\lambda_t}. \tag{7.18}$$

This expression is linear in k. Therefore, the uncertainty about the probability of answering the new item correctly depends on the difference between the maximum and the minimum of k. The upper and the lower bounds for k can be derived from the inequalities for $\boldsymbol{\eta}$ in (7.16).

From the non-negativity of the parameters $\boldsymbol{\eta}$ (the first set of inequalities

Table 7.1: Item and population parameters used in the illustrative example

n	$\mathbf{b}$	$\boldsymbol{\lambda}$
3	$\{1.00, 0.58, 0.41\}$	$\{1.00, 0.80, 1.16, 3.25\}$
4	$\{8.90, 1.00, 0.58, 0.41\}$	$\{1.00, 0.52, 0.45, 0.68, 1.99\}$
5	$\{8.91, 1.12, 1.00, 0.58, 0.41\}$	$\{1.00, 0.42, 0.29, 0.32, 0.60, 2.01\}$
6	$\{8.86, 1.12, 1.00, 0.85, 0.58, 0.41\}$	$\{1.00, 0.36, 0.22, 0.19, 0.27, 0.63, 2.43\}$

in (7.16)), we have (see Appendix B for details):

$$\max(0, \max_{u=1}^{\lfloor \frac{n+1}{2} \rfloor} \sum_{t=0}^{2u-1} (-1)^t \lambda_t d^t) < k < \min_{u=0}^{\lfloor \frac{n}{2} \rfloor} \sum_{t=0}^{2u} (-1)^t \lambda_t d^t. \tag{7.19}$$

Moreover, the second set of inequalities in (7.16) leads to (see Appendix B):

$$\max_{u=0}^{\lfloor \frac{n-1}{2} \rfloor} \left(\frac{\lambda_{2u}^2 d^{2u}}{\lambda_{2u+1} d + \lambda_{2u}} + \sum_{t=0}^{2u-1} (-1)^t \lambda_t d^t \right) \leq k \leq \min_{u=1}^{\lfloor \frac{n}{2} \rfloor} \left(\sum_{t=0}^{2u-2} (-1)^t \lambda_t d^t - \frac{\lambda_{2u-1}^2 d^{2u-1}}{\lambda_{2u} d + \lambda_{2u-1}} \right). \tag{7.20}$$

Equations 7.19 and 7.20 together provide the lower and the upper bounds for k.

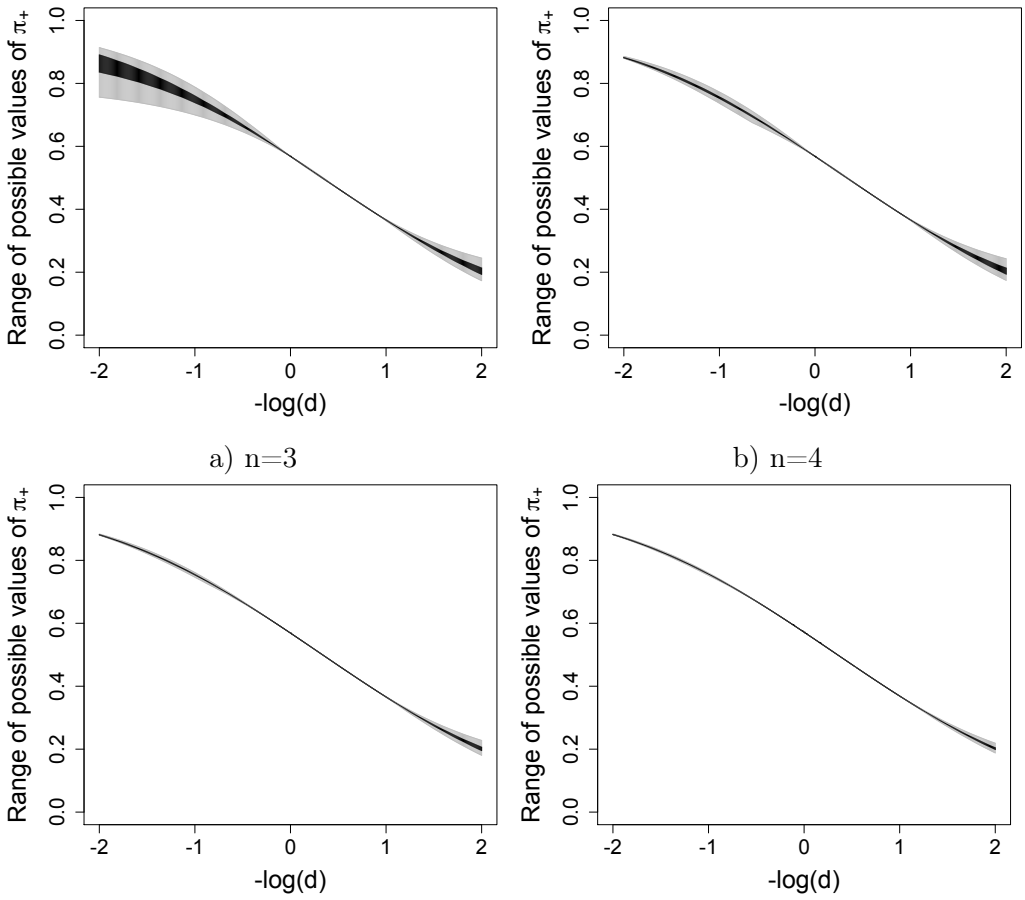
Next, we present a small example to show how the bounds on k change and what the uncertainty about the marginal probability of a correct response to the new item under the ERM is for different values of n. The item parameter d of this item varied from $\exp(-2)$ to $\exp(2)$, corresponding to the difficulty parameter varying from 2 to -2. We show how large the uncertainty is when only the non-negativity constraints are used, and when both the non-negativity and monotonicity constraints are used.

A data set with responses of persons sampled from a population with an ability distribution $\mathcal{N}(0, 1)$ to a test of six items with difficulties sampled from $\ln(b_i) \sim \mathcal{N}(0, 1)$ was simulated. First, only three items were taken into account, then four items, five items and, finally, all six items. We considered the identifiable parameters $\mathbf{b}$ and $\boldsymbol{\lambda}$ known in order to evaluate the uncertainty about π_+ coming only from the non-identifiability of $\boldsymbol{\eta}$. The identifiable parameters were fixed at their EAP-estimates obtained with a Gibbs sampler for the ERM (Maris et al., 2015), see Table 7.1.

The possible range of values for the free parameter k, and therefore for the probability of interest π_+ (given the fixed values of $\mathbf{b}$ and $\boldsymbol{\lambda}$) was evaluated for different values of the difficulty of the new item. Figure 7.1 shows the possible ranges of values for the probability of answering the new item correctly when only the constraints in (7.19) were used (in grey) and when the constraints in (7.19) and (7.20) were used (in black).

The uncertainty about π_+ decreases when n increases. For $n = 3$, the difference

between the maximum and the minimum of π_+ is for some d larger than .15 when only the constraints in (7.19) were used and larger than .05 when all the constraints were used; however, when $n = 6$, the maximum discrepancy is .03 and .006 when only non-negativity constraints and all the constraints were used, respectively. Moreover, the uncertainty about π_+ for the items with the difficulty parameter close to the items that have been answered is already very small if $n = 3$. In general, the uncertainty is larger for items with extreme difficulty.[4]



a) n=3 b) n=4

c) n=5 d) n=6

Figure 7.1: Uncertainty about the marginal probability of answering a new item correctly (grey - without monotonicity constraints, black - with monotonicity constraints) given the difficulty of the new item (on the x-axis)

We have used this small example to explicitly show that it is not possible to

[4]The graphs are not symmetric because the difficulties of the items in the reference test ($-\ln \mathbf{b}$) were also not symmetric around zero.

compute the marginal probability of answering the new item correctly. However, there uncertainty about this probability is not large.

It is difficult to extend the analytic solution described in this section to realistic settings, with n and m being usual test lengths, because of the accumulation of error while computing the bounds for k. Therefore, below we present a simulation-based approach to the problem. Appendix C presents a proof of the fact that a simulation based approach is justifiable.

7.3.2 Simulated examples

This subsection provides two simulated examples to illustrate the following:

1. the size of the uncertainty about the score distribution and which part of it is due to the non-identifiability of the parameters;

2. the practical consequences of ignoring the issue of non-identifiability of $f(\theta)$ when the true ability distribution is not normal.

In the first example, the data were simulated according to the non-equivalent group design with three linking groups. Each group consisted of 500 persons who gave responses to 15 items from the new test and 15 items from the reference test. The relevant equating designs are described in the Appendix D. The following parameters were used: $n = m = 60$, $N = M = 5,000$. Responses were simulated according to the simple RM, with person parameters sampled from $\mathcal{N}(0,1)$ for the reference population, $\mathcal{N}(0.5, 0.8^2)$ for the new population and $\mathcal{N}(-0.5, 2^2)$, $\mathcal{N}(-0.2, 2^2)$, $\mathcal{N}(-0.1, 2^2)$ for the three linking groups[5]. The item difficulties $(-\ln b_i)$ were sampled from a standard normal distribution.

First, the data augmented Gibbs sampler was used to estimate the total uncertainty about the score distribution. Second, to eliminate the uncertainty coming from the sampling variability, the new data were simulated with the same parameters but larger sample sizes ($N = M = 1,000,000$) and the algorithm was used with all the item parameters fixed at their true values. The posterior variance of the score distribution that remained was almost entirely due to the non-identifiability of the population parameters. Figure 7.2 presents the widths of the 95% credibility intervals of $\Pr(X_{+mis} \leq T), \forall T \in [0 : m]$ based on 50,000 draws from the posterior distribution after 10,000 iterations of burn-in. With a large N and fixed item parameters, the uncertainty about the score distribution becomes very small, not exceeding .002 on the probability scale.

In the second example, we compared the results of test equating using a marginal RM assuming a normal distribution of ability in the population with the results of test equating using the ERM without the normality assumption.

[5]These values could be seen as matching empirical practice in the sense that the persons in the linking groups perform worse than in the examination conditions and are more heterogeneous.

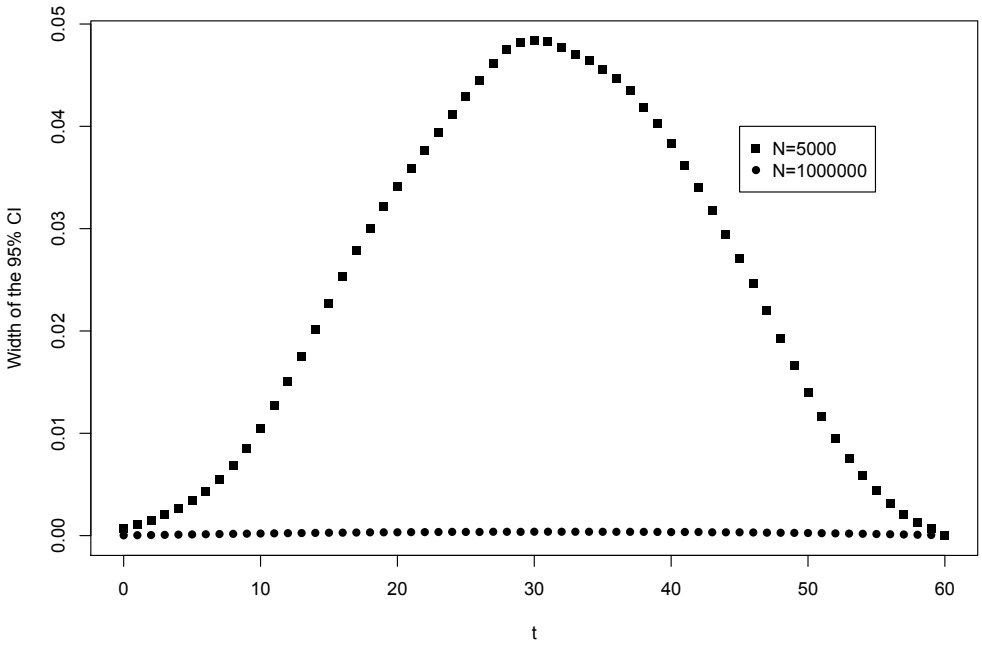


Figure 7.2: Uncertainty about the score distribution of the reference population on the new test

The data with different distributions of ability in the reference population were simulated. To show what happens if normality is violated, we used skew-normal distribution for ability (Azzalini, 2005). The parameters of the skew-normal distribution were chosen such that the mean was equal to 0, variance was equal to 1, and skewness was varied $\gamma = -0.25, -0.5, -0.75$. These distributions can be seen in Figure 7.3 (dotted lines) next to the standard normal distribution (solid line).

For each of the three degrees of skewness, we simulated the data of 5,000 persons from both the reference and the new populations taking the tests, which consisted of 40 items each, connected through three linking groups consisting of 500 persons responding to 20 items (10 from the reference test and 10 from the new test). For the new population and the three linking groups, person parameters were sampled from a normal distribution ($\mathcal{N}(0.5, 0.9^2)$, $\mathcal{N}(-0.5, 2^2)$, $\mathcal{N}(-0.2, 2^2)$, $\mathcal{N}(-0.1, 2^2)$, respectively). Item difficulties were sampled from $\mathcal{N}(0, 1)$. The data were simulated according to a RM.

The score distribution $\Pr(X_{+mis} \leq T)$ was estimated with marginal maximum likelihood (MML) assuming a normal distribution and with the Gibbs Sampler for the ERM. The ERM score distribution together with the 95% credibility intervals

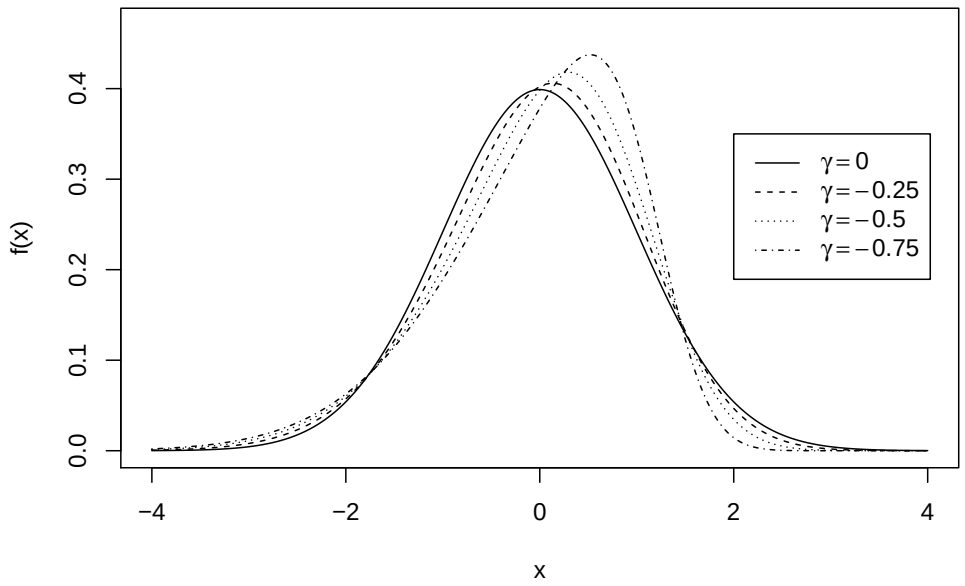
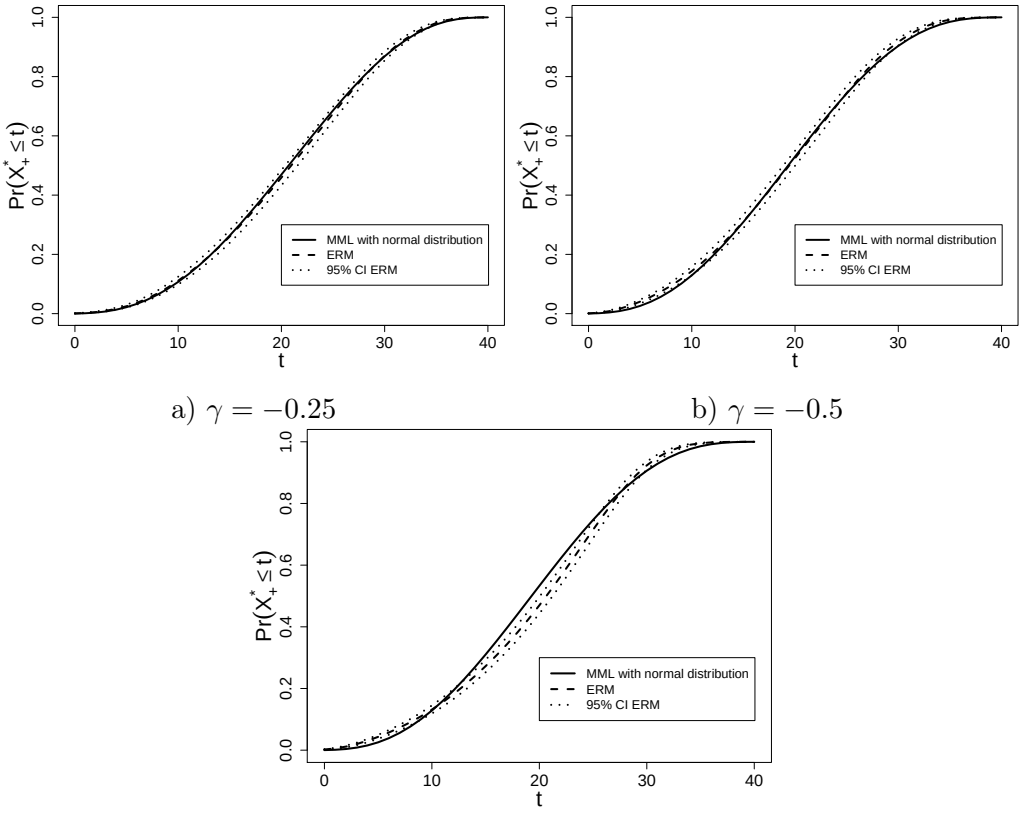


Figure 7.3: Specification of the skewed ability distributions

of $\Pr(X_{+mis} \leq T)$ based on 50000 draws from the posterior distribution (after 10000 iterations of burn-in) are presented in Figures 7.4, together with the MML-estimate of the score distribution. The more skewed the ability distribution is, the greater the difference between equating results for the MML and ERM approaches. When $\gamma = -0.25$, the MML-estimate does not fall outside of the 95% credibility interval obtained with the ERM. When $\gamma = -0.5$, the estimate based on the normality assumption is outside the credible interval for low and high scores, but within the interval for the middle range of the scores. Finally, when $\gamma = -0.75$, the MML-estimate is also outside the credible bound in the middle range of test scores. This is the range of scores within which the cut-score is usually placed, which means that different score distributions are likely to result in different cut-scores. This has consequences for the pass/fail decision for hundreds of students.

7.4 Empirical example

Using an empirical example we show the consequences of ignoring the problem of non-identifiability of $f(\theta)$ and assuming a normal distribution. We do this by comparing the estimated score distributions with and without the normality



a) $\gamma = -0.25$ b) $\gamma = -0.5$

c) $\gamma = -0.75$

Figure 7.4: Estimated score distributions with MML and ERM when the ability distribution in the reference population is skewed

assumption.

7.4.1 Method and data

We analysed data from the paper-and-pencil French language test for preparatory middle-level applied secondary education from examinations in 2011 and 2012. The sample sizes were 5518 for the reference exam and 5606 for the new exam. Both tests consisted of 41 items, but only dichotomous items were selected for analysis (35 and 34 in the reference and the new exams, respectively). The tests were linked through seven linking groups (with sample sizes ranging from 337 to 460) that responded to some items from either the reference test or the new test and some external anchor items (14 per group). The equating design is shown in Figure 7.5. There were 30 items from the reference test and 25 items from the new test answered by the linking groups. The items taken by the linking groups had been also answered by students in 2008.

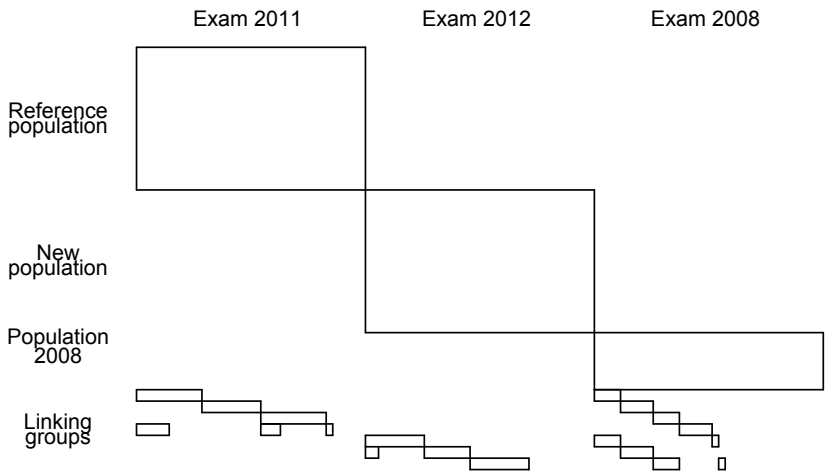


Figure 7.5: Equating design

First, the parameters of the ERM were estimated using the data augmented Gibbs sampler (see Appendix E). The algorithm was run for 60,000 iterations, of which the first 10,000 were discarded as a burn-in. The score distribution of the reference population on the new test was calculated at every iteration of the algorithm. Second, the marginal Rasch model with the normal distribution was fitted to the data and the MML-estimate of the score distribution was obtained.

7.4.2 Results

Figure 7.6 shows the posterior mean of the score distribution estimated with the ERM (together with the 95% credible interval) and the MML-estimate of the the score distribution. The estimated score distributions differ and the MML-estimate is outside of the credible interval at the lower and the higher scores. The posterior mean is also different from the MML-estimate in the middle range of scores, which could have consequences for establishing the new cut-score t_{new}. For example, if the desired proportion of persons from the reference population failing the new test was 55%, then the MML procedure would result in a cut-score of 17, whereas the ERM procedure would result in a cut-score of 18 as illustrated in Figure 7.6. The consequence of this would be that 476 students would have passed the test if a normal distribution were assumed, but would have failed if the ERM were used.

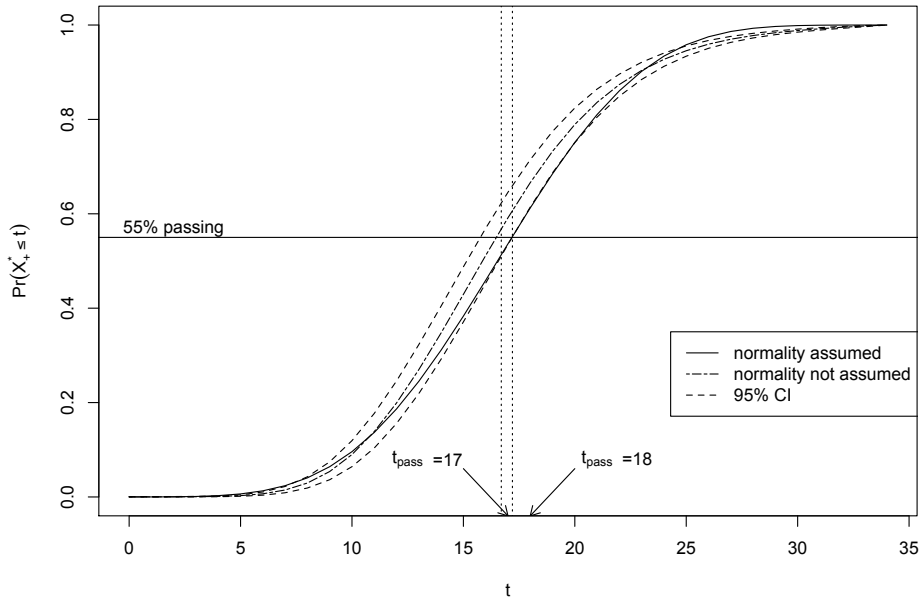


Figure 7.6: Score distribution of the reference population on the new test: posterior mean for the ERM (dashed line) and the MML-estimate based on the assumption of the normal distribution (solid line)

7.5 Discussion

Using a simple case, we have shown that, without the assumption of a parametric distribution, the score distribution on the new test is not identified. Knowing the

difficulty parameter of the new item is not enough to predict the proportion of correct responses to this item in the population, after observing the responses to a finite set of items. When the number of items observed increases, the uncertainty about the score distribution decreases. This uncertainty tends to zero with n going to infinity, but is always there. Hence, IRT cannot, strictly speaking, solve the missing data problem, since it does not allow us to impute the unobserved responses of the reference population on the new test. We have investigated the degree of uncertainty about the score distribution in realistic applications. With realistic test lengths, the uncertainty coming from non-identifiability of population parameters is small enough to be ignored for practical purposes. Therefore, test equating can be done effectively without the not-fully-testable assumption of a particular parametric shape of the ability distribution, despite the non-identifiability issue.

The theoretical importance of this paper is that it has shown what one can and cannot do with respect to test equating using IRT based only on the observed data without the assumption of a parametric shape of the distribution. Although we have used the marginal RM for illustration, the issue of non-identifiability that is discussed holds in more general marginal IRT models, since the problem of the ability distribution not being identified will not go away if more parameters are added to the conditional model.

7.6 Appendices

Appendix A

From (7.8), we can derive the joint distribution of the scores on the reference test and the new test:

$$p(X_{+obs}=x_+, X_{+mis}=x_+^*)=\frac{p(\mathbf{X}_{obs}=\mathbf{x}, \mathbf{X}_{mis}=\mathbf{x}^*)}{p(\mathbf{X}_{obs}=\mathbf{x}, \mathbf{X}_{mis}=\mathbf{x}^* \mid X_{+obs}=x_+, X_{+mis}=x_+^*)}=$$

$$p(\mathbf{X}_{obs}=\mathbf{x}, \mathbf{X}_{mis}=\mathbf{x}^*)\Big/ \frac{\prod_{i=1}^n b_i^{x_i}\prod_{j=1}^m d_j^{x_j^*}}{\gamma_{x_+}(\mathbf{b})\gamma_{x_+^*}(\mathbf{d})}=\frac{\gamma_{x_+}(\mathbf{b})\gamma_{x_+^*}(\mathbf{d})\eta_{x_++x_+^*}}{\sum_{t=0}^{n+m}\gamma_t(\mathbf{b},\mathbf{d})\eta_t}. \quad (7.21)$$

The marginal probability of obtaining a particular score on the new exam is then:

$$p(X_{+mis}=x_+^*)=\sum_{s=0}^{n} p(X_{+obs}=s, X_{+mis}=x_+^*)=\frac{\gamma_{x_+^*}(\mathbf{d})\sum_{s=0}^{n}\gamma_s(\mathbf{b})\eta_{s+x_+^*}}{\sum_{u=0}^{n+m}\gamma_u(\mathbf{b},\mathbf{d})\eta_u}. \quad (7.22)$$

To derive the relations between the parameters λ and η, let us consider the probability of observing a particular response vector $\mathbf{X}_{obs}$. On the one hand, it is

given in (7.5). On the other hand, it can be presented as follows:

$$p(\mathbf{X}_{obs}=\mathbf{x})=\sum_{t=0}^{m}p(\mathbf{X}_{obs}=\mathbf{x},X_{+mis}=t)=$$
$$=\sum_{t=0}^{m}\frac{\prod_{i=1}^{n}b_i^{x_i}\gamma_t(\mathbf{d})\eta_{x_++t}}{\sum_{u=0}^{n+m}\gamma_t(\mathbf{b},\mathbf{d})\eta_u}=\frac{\prod_{i=1}^{n}b_i^{x_i}\sum_{t=0}^{m}\gamma_t(\mathbf{d})\eta_{x_++t}}{\sum_{s=0}^{n}\gamma_s(\mathbf{b})\left(\sum_{t=0}^{m}\gamma_t(\mathbf{d})\eta_{s+t}\right)}. \quad (7.23)$$

Hence,

$$\lambda_s=\sum_{t=0}^{m}\gamma_t(\mathbf{d})\eta_{t+s},\forall s\in[0,n]. \quad (7.24)$$

Appendix B

The probability of answering the new item correctly is:

$$\Pr(X_{mis}=1)=\sum_{\mathbf{x}}\Pr(X_mis=1,\mathbf{X}_{obs}=\mathbf{x})=\frac{d\sum_{t=0}^{n}\gamma_t(\mathbf{b})\eta_{t+1}}{d\sum_{t=0}^{n}\gamma_t(\mathbf{b})\eta_{t+1}+\sum_{t=0}^{n}\gamma_t(\mathbf{b})\eta_t} \quad (7.25)$$

Using the general solution of the system of equations in (7.14), the two sums in this expression can be re-written as:

$$d\sum_{t=0}^{n}\gamma_t(\mathbf{b})\eta_{t+1}=d\sum_{t=0}^{n}\left(\gamma_t(\mathbf{b})\sum_{s=0}^{t}\frac{(-1)^{t-s}\lambda_s}{d^{t+1-s}}+(-1)^t\frac{k}{d^{t+1}}\right)=$$
$$=\sum_{t=0}^{n}\gamma_t(\mathbf{b})\sum_{s=0}^{t}\frac{(-1)^{t-s}\lambda_s}{d^{t-s}}+k\sum_{t=0}^{n}\gamma_t(\mathbf{b})\frac{(-1)^t}{d^t}=$$
$$=\sum_{t=0}^{n}\gamma_t(\mathbf{b})\sum_{s=0}^{t-1}\frac{(-1)^{t-s}\lambda_s}{d^{t-s}}+\sum_{t=0}^{n}\gamma_t(\mathbf{b})\lambda_t+k\sum_{t=0}^{n}\gamma_t(\mathbf{b})\frac{(-1)^t}{d^t} \quad (7.26)$$

and

$$\sum_{t=0}^{n}\gamma_t(\mathbf{b})\eta_t=\sum_{t=0}^{n}\left(\gamma_t(\mathbf{b})\sum_{s=0}^{t-1}\frac{(-1)^{t-1-s}\lambda_s}{d^{t-s}}+(-1)^{t-1}\frac{k}{d^t}\right)=$$
$$=-\sum_{t=0}^{n}\gamma_t(\mathbf{b})\sum_{s=0}^{t-1}\frac{(-1)^{t-s}\lambda_s}{d^{t-s}}-k\sum_{t=0}^{n}\gamma_t(\mathbf{b})\frac{(-1)^t}{d^t}. \quad (7.27)$$

Hence,

$$\Pr(X_{mis}=1)=\frac{\sum\limits_{t=0}^{n}\gamma_t(\mathbf{b})\sum\limits_{s=0}^{t-1}\frac{(-1)^{t-s}\lambda_s}{d^{t-s}}+\sum\limits_{t=0}^{n}\gamma_t(\mathbf{b})\lambda_t+k\sum\limits_{t=0}^{n}\gamma_t(\mathbf{b})\frac{(-1)^t}{d^t}}{\sum_{t=0}^{n}\gamma_t(\mathbf{b})\lambda_t}=$$
$$=1-\frac{k\sum_{t=0}^{n}\frac{(-1)^{t-1}\gamma_t(\mathbf{b})}{d^t}}{\sum_{t=0}^{n}\gamma_t(\mathbf{b})\lambda_t}+\frac{\sum_{t=1}^{n}\sum_{s=0}^{t-1}\frac{(-1)^{t-s}\gamma_t(\mathbf{b})\lambda_t}{d^{t-s}}}{\sum_{t=0}^{n}\gamma_t(\mathbf{b})\lambda_t}. \quad (7.28)$$

First, we will consider the constraints on k, following from the parameters $\boldsymbol{\eta}$ being positive:

$$\begin{cases} k=\eta_0>0, \\ \sum_{t=0}^{s-1}\frac{(-1)^{s-t+1}\lambda_t}{d^{s-t}}+(-1)^s\frac{k}{d^s}>0, \forall s\in[1:(n+1)]. \end{cases} \quad (7.29)$$

For even indices $s=2u, u=1,2,\ldots,\lfloor\frac{n+1}{2}\rfloor$, we have:

$$\frac{k}{d^{2u+1}}>\sum_{t=0}^{2u}\frac{(-1)^t\lambda_t}{d^{2u+1-t}}\Leftrightarrow k>\sum_{t=0}^{2u}(-1)^t\lambda_t d^t. \quad (7.30)$$

For odd indices $s=2u+1, u=0,1,\ldots,\lfloor\frac{n}{2}\rfloor$, we have:

$$\frac{k}{d^{2u}}<\sum_{t=0}^{2u-1}\frac{(-1)^t\lambda_t}{d^{2u-t}}\Leftrightarrow k<\sum_{t=0}^{2u-1}(-1)^t\lambda_t d^t. \quad (7.31)$$

Second, we consider the monotonicity constraints (7.12): $\eta_{s+1}\eta_{s-1}>\eta_s^2, \forall s\in[1:n]$. Using the the general solution of the system of equations in (7.10), we have:

$$\left(\sum_{t=0}^{s}\frac{(-1)^{s-t}\lambda_t}{d^{s+1-t}}+(-1)^{s+1}\frac{k}{d^{s+1}}\right)\left(\sum_{t=0}^{s-2}\frac{(-1)^{s-t-2}\lambda_t}{d^{s-1-t}}+(-1)^{s-1}\frac{k}{d^{s-1}}\right)>$$
$$>\left(\sum_{t=0}^{s-1}\frac{(-1)^{s-t-1}\lambda_t}{d^{s-t}}+(-1)^s\frac{k}{d^s}\right)^2. \quad (7.32)$$

If we multiply both sides by d^{2s} and denote $S=\sum\limits_{t=0}^{s-2}(-1)^{s-t}\lambda_t d^t$, then we get

$$\left(S+\lambda_s d^s-\lambda_{s-1}d^{s-1}-(-1)^s k\right)\left(S-(-1)^s k\right)>\left(-S+\lambda_{s-1}d^{s-1}+(-1)^s k\right)^2. \quad (7.33)$$

When multiplying the elements on the left side and taking a square on the right side, most of the element on the both sides are the same, hence they cancel out,

and the remaining inequality is:

$$(S-(-1)^s k)\lambda_s d^s > -(S-(-1)^s k)\lambda_{s-1}d^{s-1} + (\lambda_{s-1}d^{s-1})^2 \Leftrightarrow$$
$$S-(-1)^s k > \frac{\lambda_{s-1}^2 d^{s-1}}{\lambda_s d + \lambda_{s-1}}. \quad (7.34)$$

For even indices $s = 2u, u = 1, 2, \ldots, \lfloor \frac{n}{2} \rfloor$, we have:

$$k < \sum_{t=0}^{2u-2} (-1)^t \lambda_t d^t - \frac{\lambda_{2u-1}^2 d^{2u-1}}{\lambda_{2u} d + \lambda_{2u-1}}. \quad (7.35)$$

For odd indices $s = 2u+1, u = 0, 1, \ldots, \lfloor \frac{n-1}{2} \rfloor$, we have:

$$k > \frac{\lambda_{2u}^2 d^{2u}}{\lambda_{2u+1} d + \lambda_{2u}} + \sum_{t=0}^{2u-1} (-1)^t \lambda_t d^t. \quad (7.36)$$

Appendix C

For a simulation approach (such as a Gibbs Sampler) to be applicable, we have to show that the solutions of the system of equations (7.10) and inequalities (7.16) constitute a convex and bounded set, which ensures that the sampler can easily cover the full subspace of possible values of the non-identified parameters. All coefficients in the system of equations are positive, so are the parameters $\boldsymbol{\lambda}$, and therefore each of the parameters η_s is bounded:

$$0 < \eta_s < \min_{t=\max(0,s-m)}^{\min(s,n)} \frac{\lambda_t}{\gamma_{s-t}(\mathbf{b})}, \forall s \in [0 : (n+m)]. \quad (7.37)$$

For every $s \in [1 : (n+m-1)]$ the solutions of the following set of inequalities:

$$\begin{cases} \frac{\eta_{s+1}}{\eta_s} \geq \frac{\eta_s}{\eta_{s-1}} \\ \eta_{s-1} > 0 \\ \eta_s > 0 \\ \eta_{s+1} > 0 \end{cases} \quad (7.38)$$

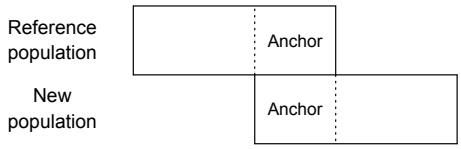
form a convex set. The interaction of convex sets from each s is itself a convex set. The intersection of the set formed by solutions of all inequalities and the set formed by the system of linear equations (which is always a convex set) is also a convex set. Therefore, all possible values of $\boldsymbol{\eta}$ constitute a convex set, and for each individual parameter there is only one range of possible values. Although the

parameters are not identified, it is still possible to sample from their joint posterior distribution. The data augmented Gibbs Sampler for test equating with the ERM which is an extension of the algorithm of Maris et al. (2015) was developed for this. The details of our algorithm can be found in the Appendix E.

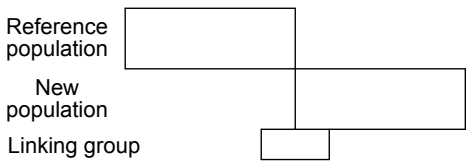
Appendix D

Non-equivalent group equating designs

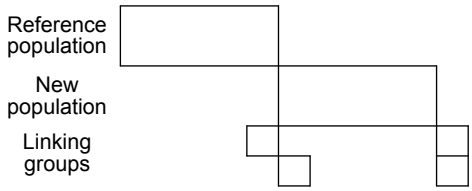
The most simple non-equivalent group design one is the anchor-item design, in which both the reference and the new tests include a common set of items. The second design is a post-equating design, in which the link between the two tests is established through the data collected in the so called linking groups, answering some items from the reference test together with some items from the new test. The third design is a variation of the post-equating design, in which persons from some of the linking groups answer the items from the reference test and some other items from the item bank, while persons from the other linking groups answer the items from the new test and the same items from the item bank. The items that do not belong to either the reference test or the new test might also be taken by students from some historic population in one of the previous years. The simplest forms of these designs are visualised in Figure 7.7.



a) Design 1



b) Design 2



c) Design 3

Figure 7.7: Three non-equivalent group equating designs

Let us by $\mathbf{Y}$ denote the $M \times m$ data matrix with responses of a sample of persons from the new population to the new test, by $\boldsymbol{\kappa}$ denote the $m+1$ identified population parameters of the new population, by $\{r\}$ the set of items in the reference test and by $\{c\}$ the set of items in the new test. If the design includes

linking groups, then $\mathbf{Z}$ is the data coming from the equating groups with $K^{(g)}$ and $k^{(g)}$ being the number of persons in the g-th linking group and the number of items answered by them; $\{e^{(g)}\}$ denotes the set of items answered by the g-th linking group and $\boldsymbol{\tau}^{(g)}$ are the population parameters of this linking group. Then the density of the observed data is:

$$f(\mathbf{X}_{obs}, \mathbf{Y}, \mathbf{Z}) = \frac{\prod_i b_i^{u_i} \prod_{s=0}^{n} \left(\sum_{t=0}^{|c/r|} \gamma_t(\mathbf{b}_{c/r}) \eta_{t+s} \right)^{N_s} \prod_{s=0}^{m} \kappa_s^{M_s}}{\left(\sum_{t=0}^{|r \cup c|} \gamma_t(\mathbf{b}_{r \cup c}) \eta_t \right)^{N} \left(\sum_{t=0}^{m} \gamma_t(\mathbf{b}_c) \kappa_t \right)^{M}} \prod_g \frac{\prod_{s=0}^{k^{(g)}} \left(\tau_s^{(g)} \right)^{K_s^{(g)}}}{\left(\sum_{t=0}^{k^{(g)}} \gamma_t \left(\mathbf{b}_{e^{(g)}} \right) \tau_t^{(g)} \right)^{K^{(g)}}}, \tag{7.39}$$

where u_i is the total number of correct responses to item i by all students which answered this item, N_s, M_s and $K_s^{(g)}$ are the number of persons from the reference population, new population and the g-th linking group, respectively, that gave exactly s correct responses to the items in the corresponding tests.

The score distribution of the reference population on the new test depends on the population parameters $\boldsymbol{\eta}$ and the item parameters $\mathbf{b}$:

$$p(X_{+mis} \leq T) = \sum_{t=0}^{T} \frac{\gamma_t(\mathbf{b}_{c/r}) (\sum_{s=0}^{n} \gamma_s(\mathbf{b}_{r/c}) \eta_{s+y})}{\sum_{u=0}^{|r \cup c|} \gamma_u(\mathbf{b}) \eta_u} \tag{7.40}$$

To make inferences about this distribution we obtain samples from the posterior distribution $p(\boldsymbol{\eta}, \mathbf{b} \mid \dots)$. This is done using a data augmented Gibbs sampler.

Appendix E

We describe here how the samples from the joint posterior distribution

$$p(\boldsymbol{\eta}, \mathbf{b} \mid \dots) \tag{7.41}$$

can be obtained using a Markov chain Monte Carlo algorithm. We describe it for the post-equating non-equivalent groups design (see Figure 7.7b) with G linking groups. The density of the data given this equating design is given in Equation 7.39 in Appendix D. The algorithm can be easily altered for the different kinds of non-equivalent group designs.

Data augmented Gibbs sampler.

For computational convenience, instead of parameters $\boldsymbol{\eta}$, we use a different parametrization with the ratios of the consecutive parameters $\frac{\eta_{s+1}}{\eta_s}$. To place the

parameters on the scale common in IRT we consider logarithms of these ratios:

$$p_s = \ln\left(\frac{\eta_{s+1}}{\eta_s}\right), \forall s \in [0 : n+m-1]. \tag{7.42}$$

We use a prior which in addition to the monotonicity constraint (7.12) has a lower and an upper bound for the parameters:

$$p(\mathbf{p}) \propto \prod_{s=0}^{n+m-1} \mathcal{I}_{[p_{s-1},p_{s+1}]}(p_s), \tag{7.43}$$

where $p_{-1} = -100$ and $p_{n+m} = 100$. This is a reasonable constraint, since it follows from the Dutch identity (Holland, 1990) that

$$p_s = \ln\left(\mathcal{E}(\exp(\Theta) \,|\, X_{+obs} + X_{+mis} = s)\right). \tag{7.44}$$

A priori, item and population parameters are independent. For item parameters we choose a uniform prior for difficulty parameters $-\ln(b_i)$, which is $p(b_i) \propto \frac{1}{b_i}$.

After the re-parametrization, the density of the observed data is:

$$\begin{aligned} &f(\mathbf{X}_{obs}, \mathbf{Y}, \mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(G)}) = \\ &= \prod_{g=1}^{G} \frac{\prod\limits_{i\in\{r\cap e_g\}} b_i^{x_{+i}+z_{+i}^{(g)}} \prod\limits_{i\in\{c\cap e_g\}} b_i^{y_{+i}+z_{+i}^{(g)}} \prod\limits_{s=1}^{k^{(g)}-1} \exp(r_s^{(g)})^{\sum_{j>s} K_j^{(g)}}}{(1+\sum_{t=1}^{k^{(g)}} \gamma_t(\mathbf{b}_{e_g}) \prod\limits_{j<t} \exp(r_j^{(g)}))^{K^{(g)}}} \times \\ &\frac{\prod\limits_{i\in\{r/e\}} b_i^{x_{+i}} \prod\limits_{i\in\{c/e\}} b_i^{y_{+i}} \prod\limits_{s=0}^{n} (1+\sum\limits_{t=1}^{m} \gamma_t(\mathbf{b}_c) \prod\limits_{j<t} \exp(p_j))^{N_s} \prod\limits_{s=0}^{m-1} \exp(q_s)^{\sum_{j>s} M_j}}{(1+\sum\limits_{t=1}^{n+m} \gamma_t(\mathbf{b}) \prod_{j<t} \exp(p_j))^{N} (1+\sum\limits_{t=1}^{m} \gamma_t(\mathbf{b}_c) \prod_{j<t} \exp(q_j))^{M}}, \end{aligned} \tag{7.45}$$

where $\mathbf{p}, \mathbf{q}, \mathbf{r}^{(1)}, \ldots, \mathbf{r}^{(G)}$ are the population parameters of the reference population, the new population and G linking groups respectively.

Although, we are interested only in the parameters $\mathbf{b}_c$ and $\mathbf{p}$, the other parameters $(\mathbf{b}_r, \mathbf{q}, \mathbf{r}^{(1)}, \ldots, \mathbf{r}^{(G)})$ are also sampled as nuisance parameters. Moreover, to make the full conditional posterior distribution of p_s tractable, at every iteration we will sample augmented data $\mathbf{x}^*$: responses of persons from the reference group to the items of the new test (Tanner & Wong, 1987; Zeger & Karim, 1991). This amounts to sampling from the joint posterior:

$$p(\mathbf{p}, \mathbf{q}, \mathbf{r}^{(1)}, \ldots, \mathbf{r}^{(G)}, \mathbf{b}, \mathbf{x}^* \,|\, \mathbf{X}_{obs}, \mathbf{Y}, \mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(G)}). \tag{7.46}$$

A Gibbs sampler is used, i.e., all parameters are subsequently sampled from their full conditional distributions given the new values of all other parameters (Geman & Geman, 1984; Casella & George, 1992). After starting from the initial values (1 for all item parameters, and the population parameters equally distanced from -3 to 3), the algorithm goes through the following steps:

Step 1. Sample the augmented data $\mathbf{x}^*$.

For every person $j \in [1:N]$, sample a vector of responses $\mathbf{x}_j^*$ from its full conditional posterior $p(\mathbf{x}_j^* \mid \dots)$, which is factored in the following way:

$$p(\mathbf{x}_j^* \mid \dots) = p(x_{j+}^* | x_{j+}, \mathbf{b}_c, \mathbf{p}) p(\mathbf{x}_j^* \mid x_{j+}^*, \mathbf{b}_c) = p(x_{j+}^* \mid x_{j+}, \mathbf{b}_c, \mathbf{p}) \times \\ p(x_{j,1}^* | x_{j+}^*, \mathbf{b}_c) p(x_{j,2}^* \mid x_{j,1}, x_{j+}^*, \mathbf{b}_c) \dots p(x_{j,m}^* | x_{j,1}^*, \dots, x_{j,m-1}^*, x_{j+}^*, \mathbf{b}_c), \quad (7.47)$$

where x_{j+} is the sumscore of person j. First, sample x_{j+}^* from the categorical distribution with probabilities

$$\Pr(x_{j+}^* = s \mid x_{j+}, \mathbf{p}, \mathbf{b}_c) = \frac{\gamma_s(\mathbf{b}_c) \prod_{u<(x_{j+}+s)} \exp(p_u)}{1 + \sum_{t=1}^{m} \gamma_t(\mathbf{b}_c) \prod_{u<(x_{p+}+t)} \exp(p_u)}. \quad (7.48)$$

And then for every item $i \in [1:m]$ sample $x_{j,i}^*$ from a Bernoulli distribution with probability:

$$\Pr(x_{j,i}^* = 1 \mid x_{j+}^*, \mathbf{x}_{j,s<i}^*, \mathbf{b}_c) = \frac{b_i \gamma_{x_{j+}^* - \sum_{s=0}^{i-1} x_{j,s}^* - 1}(b_{i+1}, \dots, b_m)}{\gamma_{x_{j+}^* - \sum_{s=0}^{i-1} x_{j,s}^*}(b_i, b_{i+1} \dots, b_m)} \quad (7.49)$$

Step 2. Sample from the full conditional posterior of the distribution of the item parameters.

For every $i \in \{r/e\}$, sample b_i from its full conditional posterior:

$$p(b_i \mid \dots) \propto \frac{b_i^{x_{+i}-1}}{(1+cb_i)^N}, \quad (7.50)$$

where $c = \frac{\sum_{t=1}^{n+m} \gamma_{s-1}(\mathbf{b}^{(i)}) \prod_{j<t} \exp(p_j)}{\sum_{t=0}^{n+m-1} \gamma_s(\mathbf{b}^{(i)}) \prod_{j<t} \exp(p_j)}$. This is an scaled beta-prime distribution, to sample from which first sample $y = \frac{cb_i}{1+cb_i}$ from $\mathcal{B}(x_{+i}, N - x_{+i})$, and then transform it: $b_i = \frac{1}{c} \frac{y}{1-y}$.

For every $g \in [1:G]$, for every $i \in \{r \cap e_g\}$, sample b_i from its full conditional posterior:

$$p(b_i \mid \dots) \propto \frac{b_i^{x_{+i}+z_{+i}^{(g)}-1}}{(1+c_1 b_i)^N (1+c_2 b_i)^{K^{(g)}}}, \quad (7.51)$$

where $c_1 = \frac{\sum_{t=1}^{n+m} \gamma_{s-1}(\mathbf{b}^{(i)}) \prod_{j<t} \exp(p_j)}{\sum_{t=0}^{n+m-1} \gamma_s(\mathbf{b}^{(i)}) \prod_{j<t} \exp(p_j)}$ and $c_2 = \frac{\sum_{t=1}^{k^g} \gamma_{s-1}(\mathbf{b}_{e_g}^{(i)}) \prod_{j<t} \exp(r_j^{(g)})}{\sum_{t=0}^{k^{(g)}-1} \gamma_s(\mathbf{b}_{e_g}^{(i)}) \prod_{j<t} \exp(r_j^{(g)})}$. Unlike

the full conditional of the item parameters of the items taken by persons from only one population, this distribution is not easy to sample from directly. It is more convenient to sample from the distribution of $\beta_i = -\ln(b_i)$ using a Metropolis-Hasting algorithm (Metropolis et al., 1953). We use $\mathcal{N}(-\ln(b_i), \tau^2 = 0.01)$ as a proposal density with b_i being the current value of the parameter.

For every $i \in \{c/e\}$, sample b_i from its full conditional posterior analogously to sampling b_i, $i \in \{r \cap e_g\}$, because these items are not only taken by the new population, but responses to these items by the reference population are imputed.

For every $g \in [1 : G]$, for every $i \in \{c \cap e_g\}$ sample b_i from its full conditional posterior:

$$p(b_i \mid \dots) \propto \frac{b_i^{y_{+i}+z_{+i}^{(g)}+x_{+i}^{*}-1}}{(1+c_1 b_i)^N (1+c_2 b_i)^{K^{(g)}} (1+c_3 b_i)^M}, \tag{7.52}$$

where $c_3 = \frac{\sum_{t=1}^{m} \gamma_{s-1}(\mathbf{b}_c^{(i)}) \prod_{j<t} \exp(q_j)}{\sum_{t=0}^{m-1} \gamma_s(\mathbf{b}_c^{(i)}) \prod_{j<t} \exp(q_j)}$. Use the same Metropolis-Hastings algorithm as for the items, taken by two populations. If the equating design specifies more than 3 populations taking some of the items, then the full conditional posteriors of those items can be extended accordingly.

Step 3. Sample the population parameters.

For every $s \in [0 : (n+m-1)]$, sample p_s from its full conditional posterior:

$$p(p_s \mid \dots) \propto \frac{\exp(p_s)^{\sum_{j>s} N_s^*}}{(1+c\exp(p_s))^N} \mathcal{I}_{[p_{s-1}, p_{s+1}]}(p_s), \tag{7.53}$$

where $c = \frac{\sum_{t=s+1}^{n} \gamma_t(\mathbf{b}) \prod_{j\neq s, j=0}^{t-1} \exp(p_j)}{1+\sum_{t=1}^{s} \gamma_t(\mathbf{b}) \prod_{j=0}^{t-1} \exp(p_j)}$. To sample from this distribution, we first sample $y = \frac{c\exp(p_s)}{1+c\exp(p_s)}$ from the truncated beta distribution

$$f(y) \propto y^{\sum_{j>s} N_s^* - 1} (1-y)^{N - \sum_{j>s} N_s^* - 1} \mathcal{I}_{[a_1, a_2]}(y), \tag{7.54}$$

where $a_1 = \frac{c\exp(p_{s-1})}{1+c\exp(p_{s-1})}$ and $a_2 = \frac{c\exp(p_{s+1})}{1+c\exp(p_{s+1})}$, using rejection sampling with $U(a_1, a_2)$ as a proposal distribution, and then transform it: $p_s = \ln(\frac{1}{c}\frac{y}{1-y})$.

For every $s \in [0 : (m-1)]$, sample q_s from its full conditional posterior analogously to sampling p_s. For every $g \in [1 : G]$, for every $s \in [0 : k^{(g)} - 1]$, sample $r_s^{(g)}$ form its full conditional posterior analogously to sampling p_s.

At every iteration of the Gibbs sampler, we compute the expected score distribution for the reference population on the new exam $Pr(X_{+mis} \leq T)$.

Chapter 8

Epilogue

The first part of the dissertation dealt with conditional independence and conditional dependence between response time and accuracy. Conditional independence - as formulated by most joint models for response times and accuracy applied in the context of educational measurement - means that given the latent variables of speed and ability the response accuracy and the response time of the same item are independent. Conditional dependence, on the other hand, means that the relationship between the response time and accuracy cannot be fully explained by the higher-level relationship between the persons' latent variables, or between the item characteristics related to time and accuracy. The first two chapters of Part I developed tools that help to answer the question of whether relatively simple response time and accuracy models assuming conditional independence adequately represent the complex relationships between time and accuracy in the data. The last chapter of Part I proposed to give up some of the model simplicity in order to explain the complex structure in the data and gain more insight into the substantively interesting response processes.

The Kolmogorov-Smirnov tests proposed in Chapter 2 proved their usability for detecting different types of violations of conditional independence. Even when an exponential family model for response accuracy is only an approximation (for example, if a Rasch model is used when the true model is the two-parameter-logistic model) the procedure performs reasonably well. Moreover, the semi-parametric nature of the procedure (i.e., neither the distribution of the response times nor the type of violation against which the assumption is tested need to be specified) makes it very general. However, due to the fact that separate tests are performed within each subgroup of persons with the same level of the sufficient statistic, the sample size needed to achieve high power is rather large ($N = 5000, 10000$).

Because the Kolmogorov-Smirnov tests put a strong demand on the required sample size, a second procedure for testing the assumption of conditional independence was developed in Chapter 3 which is more powerful but still relatively

flexible with the respect to the type of violations that can be detected. However, unlike the Kolmogorov-Smirnov tests this procedure does require one to specify the full model for the response time and accuracy, and requires one to consider specific consequences of conditional dependence, rather than testing conditional dependence against its general alternative. These restrictions that are imposed by the posterior predictive checks method do however greatly help in gaining more power to detect the violations of conditional independence for which it has been designed. The posterior predictive checks showed very good results in the simulation studies with adequate specificity and high sensitivity already with relatively small samples (compared to the sample sizes needed for the Kolmogorov-Smirnov tests). Furthermore, the issue of having to specify the full model and the concrete consequences of conditional dependence is partly compensated by the facts that the framework itself is rather general (i.e., different models can be used and different discrepancy measures can be used), and that the procedure is rather robust to misspecifications of the model. The three discrepancy measures that were used provided good results. However, further research might be needed to investigate whether differently formulated measures are more sensitive to residual dependencies, heteroscedasticity of response times, and interaction effect of response times and ability on response accuracy. Additionally, it is interesting to consider how other sources of conditional dependence which have not been addressed by the proposed measures can be tested within the framework of posterior predictive checks.

While a violation of conditional independence strictly speaking invalidates a model which assumes conditional independence, this model might still be useful (Box & Draper, 1987, p.424). An important direction of research that has not been the focus of this dissertation involves the robustness of the response time and accuracy models to violations of conditional independence. It is important to investigate to which extent the inferences based on, for example, the hierarchical model are influenced by the presence of the violations of conditional independence of different size and different type. In some situations a rather simple conditional independence model may still adequately capture the most important aspects of the data, even if not all of the complex data structure is represented (i.e, if conditional independence is rejected by a formal test). In other cases violations of conditional independence might be too severe or lead to an important misrepresentation of the response process, in which case alternate models need to be considered.

In Chapter 4 we proposed a way of dealing with conditional dependence when it is present in the empirical data. In our extension of the hierarchical model the effects of a response being relatively fast or slow on the item response function were explicitly modeled. This allows one to learn more about the relationship between response time and accuracy than can be captured in a single correlation between the speed and ability of the persons, and a single correlation between the difficulty

and time intensity of the items. Rather than considering conditional dependence as a source of misfit and a threat to the measurement properties of the model these dependencies can be considered as an opportunity to gain more insight into the response processes. In the proposed model the effects of residual response time are item dependent and in the presented empirical example part of the variation of the effects across items was explained by their difficulty. Alternatively, one could consider a model with the effect of residual response time being person-dependent or both item- and person-dependent. Moreover, the effect might depend on the difference between the person's ability and the item's difficulty.

The second part of the dissertation dealt with three different research questions, all answered using a Bayesian approach. In Chapter 5 the multi-scale Rasch model consisting of Rasch homogenous scales was proposed and a tool for unmixing Rasch scales was provided. The choice of the model was directly motivated by the idea of optimally balancing the simplicity of the model with the complexity of the data: It keeps an important measurement property of the simple Rasch model - sufficiency of the sumscore within each scale for the corresponding person parameter - but relaxes other assumptions of the Rasch model (equality of discriminations of all items in the test and unidimensionality) which often do not match the complexity of the educational data.

In Chapter 6 the use of informative priors to improve the quality of test linking was considered. The results suggested that this approach can improve the quality of linking: increase the precision of the linking results without introducing bias. The linking procedures that were used were based on the simple Rasch model. Although this strict model probably does not have a perfect fit to the data it is a very intuitive model with clear interpretations that are easy to communicate to the experts. Using this model also makes it possible to directly translate the expert judgements like "item i is more difficult than item j" to statements about the model parameters ($\delta_i > \delta_j$) which is crucial for prior elicitation. Hence, in this project the simplicity of the model was given relatively high priority over model fit, in order to be able to include expert knowledge and improve the quality of linking.

Within informative Bayesian analysis, there are different directions for further research, especially in the context of test linking and equating. In the proposed methodology the judgements of individual experts collected with the rulers method are combined in a mixture distribution. An alternative to the mathematical aggregation of the experts' opinions is the so called behavioural aggregation or group interaction methods. These procedures mean that the experts can discuss their individual judgements and come up with a consensus decision. For example, the Delphi technique and its variants (Pill, 1971) can be used. From a statistical perspective, the ways of taking the possible dependencies between the experts' judgements into account could be investigated. Furthermore, it is important to

investigate which item features experts use in judging their difficulty and which characteristics of experts are good predictors of high quality judgements.

In Chapter 7 we investigated if it is necessary to make an assumption about the parametric shape of the ability distribution to be able to make the results of two tests comparable. Using a simple example it was demonstrated that the distribution of the scores of the reference population on the new test is not identified without a parametric assumption for ability. However, it was also shown that the uncertainty about the score distribution stemming from its non-identifiability is very small and can be ignored for all practical purposes. In this chapter two approaches were compared: using a simpler model with a parametric distribution of ability (in the case of the normal distribution it has only two population parameters) or not including parametric assumptions and using a more complex model with a non-parametric ability distribution (the number of identified parameters being equal to the number of items plus one). On the one hand, a simple and more restrictive model does not have a problem of non-identifiability of the score distribution, but on the other hand, using it may lead to bias in estimating the cut-scores if the distribution of ability is not specified correctly, as was demonstrated both in the simulated and the empirical examples.

References

Abadie, A. (2002). Bootstrap tests for distributional treatment effect in instrumental variable models. *Journal of the American Statistical Association*, *97*(457), 284-292.

Adams, R., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *12*, 261-280.

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

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123-140.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42*, 69-81.

Angoff, W. (1971). Scales, norms and equivalent scores. In R. Thorndike (Ed.), Educational measurement. Washington, DC: American Council of Education.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, *32*, 159-188.

Béguin, A. A. (2000). *Robustness of equating high-stakes tests.* Unpublished doctoral dissertation, Enschede, The Netherlands.

Bejar, I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, *7*, 303-310.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (p. 395-479). Reading: Addison-Wesley.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113 (4)*, 700-765.

Bolsinova, M., de Boeck, P., & Tijmstra, J. (2015). *Modeling conditional dependence between response time and accuracy.* (Manuscript submitted for publication)

Bolsinova, M., & Maris, G. (2016). A test for conditional independence between response time and accuracy. *British Journal of Mathematical and Statistical Psychology*, *69*(1), 62-79.

Bolsinova, M., & Tijmstra, J. (2016). Posterior predictive checks for conditional independence between response time and accuracy. *Journal of Educational and Behavioral Statistics*, *41*, 123-145.

Box, G. E. P, & Draper, N. R. (1987). *Empirical model-building and response surfaces.* New York: Wiley.

Brooks, S., & Gelman, A. (1998). General methods for monitoring convergence

of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434-455.

Casella, G., & George, E. (1992). Explaining the Gibbs sampler. *The American Statistician*, *43*(3), 167-174.

Celeux, G., Hurn, M., & Robert, C. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, *95*, 957-970.

Chen, H., & De Boeck, P. (2014). *Are slow and fast ability test responses different?* (Manuscript submitted for publication)

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioural Statistics*, *22*(3), 265-289.

Cizek, G., & Bunch, M. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

College voor Toetsen en Examens: Staatsexamens NT2. (n.d.). Retrieved September 25, 2015, from http://www.staatsexamensnt2.nl.

Council of Europe. (2011). *Common European framework of reference for: Learning, teaching, assessment.* Council of Europe.

Coyle, T. (2003). A review of the worst performance rule: Evidence, theory, and alternative hypotheses. *Intelligence*, *31*(6), 567-587.

Cressie, N., & Holland, P. W. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika*, *48*, 129-141.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society*, *41*, 1-31.

Debelak, R., & Arendasy, M. (2012). An algorithm for testing unidimensionality and clustering items in Rasch measurement. *Educational and Psychological Measurement*, *72*, 375-387.

Diebolt, J., & Robert, C. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, *56*(363-375).

Fisher, G. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *36*, 359-374.

Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fisher & I. W. Molenaar (Eds.), *Rasch model: foundations, recent developments and applications.* (p. 15-38). New York: Springer-Verlag.

Fisher, R. (1925). *Statistical methods for research workers.* Edunburgh, UK: Oliver and Boyd.

Follmann, D. (1988). Consistent estimation in the Rasch model based on nonparametric margins. *Psychometrika*, *53*, 815-841.

Fox, J. P., & van der Linden, W. J. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical software*, *20*(7).

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Annales de l'Université de Lyon, Section A, Séries 3*, *14*, 53-77.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York, NY: Springer.

Gamerman, G., & Lopes, H. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*. Boca Raton, FL: Chapman & Hall/CRC.

Geisinger, K. (1991). Using standard setting data to establish cutoff scores. *Educational Measurement: Issues and Practice*, *10*(2), 17-22.

Gelfand, A., & Smith, A. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398-409.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical model. *Bayesian Analysis*, *1*(515-533).

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733-807.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457-472.

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

Ghosh, M., Ghosh, A., Chen, M., & Agresti, A. (2000). Noninformative priors for one-parameter item response models. *Journal of statistical planning and inference*, *88*, 99-115.

Gill, J. (2008). *Bayesian methods: A social and behavioural sciences approach*. New York, NY: Chapman & Hall/CRC.

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika*, *53*, 525-546.

Goldhammer, F., & Klein Entink, R. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence*, *39*, 108-119.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*(3), 608-626.

Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, *1*(3), 403-420.

Haberman, S. (2007). The interaction model. In M. von Davier & C. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: extensions and applications* (p. 201-216). Springer.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*, 144-149.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer Nijhof.

Hardouin, J.-B., & Mesbah, M. (2004). Clustering binary variables in subscales using an extended Rasch model and Akaike information criterion. *Communications in Statistics: Theory and Methods*, *33*, 1277-1294.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97-109.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods.* New York: Springer.

Holland, P. W. (1990). The Dutch identity: a new tool for the study of item response models. *Psychometrika*, *55*, 5-18.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4-th ed., pp. 189-220). Westport: Praeger.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65-70.

Huang, A., & Wand, M. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, *8*(2), 439-452.

Humphry, S. & Andrich, D. (2008). Understanding the unit in the Rasch model. *Journal of Applied Measurement*, *9(3)*, 249-264.

Humphry, S. (2011). The role of the unit in physics and psychometrics. *Measurement: Interdisciplinary Research and Perspective*, *9 (1)*, 1-24.

Humphry, S. (2012). Item set discrimination and the unit in the Rasch model. *Journal of Applied Measurement*, *13 (2)*, 165-24.

Ip, E.H. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika*, *66*(1), 109-132.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics*, *21*, 1359-1378.

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

Keizer-Mittelhaëuser, M. (2014). *Modeling the effect of differential motivation on linking educational tests.* Unpublished doctoral dissertation, Tilburg, The Netherlands.

Klein Entink, R., Fox, J. P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*(1), 21-48.

Klein Entink, R., Kuhn, J., Hornke, L., & Fox, J. P. (2009). Evaluating cognitive theory: a joint modeling approach using responses and response times. *Psychological methods*, *14*(1), 54-75.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: methods and practices.* New York, NY: Springer.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giorn. dell'Istituto Ital. degli Attuari*, *4*, 83-91.

Lee, Y., & Chen, H. (2011). A review of recent response-time analysis in educational testing. *Psychological Test and Assessment Modeling*, *53*, 359-379.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Statistica Sinica*, *6*, 733-807.

Levy, R., Xu, Y., Yel, N. , & Svetina, D. (2015). A standardised generalised dimensionality discrepancy measure and a standardised model-based covariance for dimensionality assessment for multidimensional models. *Journal of Educational Measurement*, *52*(2), 144-158.

Livingston, S. A. (2004). *Equating test scores (without IRT).* Princeton: Educational testing service.

Loeys, T., Rossel, Y., & Baten, K. (2011). A joint modeling approach for reaction time and accuracy in psycholinguistic experiments. *Psychometrika*, *76*(3), 487-503.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F.M., & Wingersky, M.S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, *8*, 452-461.

Loyd, B.H., & Hoover, H.D. (1980). Vertical equating using the Rasch Model. *Journal of Educational Measurement*, *17*, 179-193.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York: Oxford University Press.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20 (9)*, 1-20.

Marco, G.L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*, 139-160.

Maris, G., Bechger, T., & San Martin, E. (2015). A Gibbs sampler for the (Extended) marginal Rasch model. *Psychometrika*, *80*, 859-879.

Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: scoring rules based on response time and accuracy. *Psychometrika*, *77*, 615-633.

Marsman, M., Maris, G., Bechger, T., & Glas, C. A. (2015). *Markov chain Monte Carlo for large-scale Bayesian inference.* (Manuscript submitted for publication).

McLachlan, G., & Peel, D. (2000). *Finite mixture models.* New York, NY: John Wiley & Sons.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, *22*(3), 1142-1160.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087-1092.

Migrom, P. (1981). Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, *12*(2), 380-391.

Mislevy, R. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. *Applied Psychological Measurement*, *12*, 281-296.

Mislevy, R., Sheehan, K., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of educational measurement*, *30*, 55-78.

Mittelhaëuser, M., Béguin, A.A., & Sijtsma, K. (2015). Selecting a data collection design for linking in educational measurement: Taking differential motivation into account. In R.E. Milsap, L.A. van der Ark, D.M. Bolt, & W.-C. Wang (Eds.), New developments in quantitative psychology: Presentations from he 78th annual psychometric society meeting (pp. 181-193). New York, NY: Springer.

O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., & Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*. Wiley.

Ozaki, K., & Toyoda, H. (2006). A paired comparison IRT model using 3-value judgement: estimation of item difficulty parameters prior to administration of the test. *Behaviormetrika*, *33*(2), 131-147.

Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence*, *40*, 23-32.

Petersen, N.S., & Kolen, M.J., & Hoover, H.D. (1989). Scaling, norming and equating. In R.L. Linn (Ed.), Educational measurement. New York, NY: American Council of Education and Macmillan.

Petscher, Y., Mitchell, A., & Foorman, B. (2014). Improving the reliability of student scores from speeded assessments: an illustration of conditional item response theory using a computer-administered measure of vocabulary. *Reading and writing*, *28*(1), 31-56.

Pieters, J., & van der Ven, A. (1982). Precision, speed, and distraction in time-limit tests. *Applied Psychological Measurement*, *6*, 93-109.

Pike, R. (1973). Response latency models for signal detection. *Psychological Review*, *80*.

Pill, J. (1971). The Delphi method: substance, context, a critique an an annotated bibliography. *Socio-Economic planning science*, *5*, 57-71.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Ranger, J., & Ortner, T. (2012). The case of dependency of responses and response times: A modeling approach based on standard latent trait models. *Psychological Test and Assessment Modeling*, *54*(2), 128-148.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: The Danish Institute of Educational Research. (Expanded edition, 1980. Chicago, The University of Chicago Press)

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59-108.

Reckase, M. D. (1997). A linear logistic multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (p. 271-286). New York: Springer Verlag.

Reckase, M. D. (2008). *Multidimensional item response theory.* New York, NY: Springer.

Robert, C.P., & Casella, G. (2004). *Monte Carlo statistical methods.* New York, NY: Springer.

Roskam, E. E. (1987). Toward a psychometric theory of intelligence. In E. E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology* (p. 151-171). Amsterdam: North-Holland.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271-282.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, *12*, 1151-1172.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in survey.* New-York: Wiley.

San Martin, E., & Rolin, J. (2013). Identification of parametric Rasch-type models. *Journal of statistical planning and inference*, *143*, 116-130.

Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence*, *48*(37-50).

Schwarz, G. (1978). Estimating the dimension of the model. *Annals of Statistics*, *6*(2), 461-464.

Sekhon, J. (2011). Multivariate and propensity score matching software with automated balance optimization: The Matching package for r. *Journal of Statistical software*, *42*(7).

Shepard, L. (1980). Standard setting issues and methods. *Applied Psychological Measurement*, *4*(3), 447-467.

Sinharay, S., Johnson, M., & Stern, H. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*, 298-321.

Smith, B. (2007). boa: An R package for MCMC output convergence assessment and posterior inference. *Journal of Statistical software*, *21*, 1-37.

Spiegelhalter, D. J., Best, N. G., Carlin, B., & van der Linden, A. (2002). Bayesian

measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *64*, 583-639.

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201-210.

Stone, M. (1961). The opinion pool. *Annals of Mathematical Statistics*, *32*, 1339-1342.

Swaminathan, H., Hambleton, R., Sireci, S., Xing, D., & Rivazi, S. (2003). Small sample estimation in dichotomous item response models: effect of priors based on judgemental information on the accuracy of item parameter estimates. *Applied Psychological Measurement*, *27*, 27-51.

Tanner, M., & Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528-540.

Tatsuoka, K. (1987). Validation of cognitive sensitivity for item response curves. *Journal of Educational Measurement*, *24*, 233-245.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, *22*(4), 1701-1762.

Tjur, T. (1982). A connection between Rasch's item analysis and a multiplicative Poisson model. *Scandinavian Journal of Statistics*, *9*, 23-30.

Townsend, J., & Ashby, F. (1983). *Stochastic modeling of elementary psychological processes*. New York: Cambridge University Press.

Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika*, *70*(4), 629-650.

van Breukelen, G. J. P. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika*, *70*(2), 359-376.

van Breukelen, G., & Roskam, E. (1991). A Rasch model for the speed-accuracy trade of in time-limited tests. In J. Doignon & R. J. Falmagne (Eds.), *Mathematical psychology: Current developments* (p. 251-271). New York: Springer.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181-204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287-308. doi: 10.1007/s11336-006-1478-z

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*(1), 5-20.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*(3), 247-272.

van der Linden, W. J., & Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, *75*, 120-139.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365-384.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer.

Vanpaemel, W. (2011). Constructing informative model priors using hierarchical methods. *Journal of Mathematical Psychology*, *55*, 106-117.

Verhelst, N. D., Glas, C. A. W., & van der Sluis, A. (1984). Estimation problems in the Rasch model: The basic symmetric functions. *Computational Statistics Quarterly*, *1*(3), 245-262.

Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model: OPLM. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (p. 215-238). New York: Springer Verlag.

Verhelst, N.D., Verstralen, H.H.F.M., & Jansen, M. G. (1997). A logistic model for time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (p. 169-185). New York: Springer.

von Davier, A. A. (2011). *Statistical models for test equating, scaling, and linking.* New York, NY: Springer.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating.* New York: Springer.

Wang, T. (2006). *A model for the joint distribution of item response and response time using one-parameter Weibull distribution (CASMA Research Report 20).* Iowa City: IA: Center for Advanced Studies in Measurement and Assessment.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*, 323-339.

Wauters, K., Desmet, P., & van der Noordgate, W. (2012). Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers and Education*, *58*, 1183-1193.

Wingersky, M.S., & Lord, F.M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement*, *8*, 347-464.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago, IL: MESA Press University of Chicago.

Zeger, K., & Karim, M. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, *86*, 79-86.

Zeng, L., & Kolen, M. (1995). An alternative approach for IRT observed-score equating of number-correct scores. *Applied Psychological Measurement*, *19*, 231-240.

Nederlandse Samenvatting

Bij het onderwijskundig meten worden data verzameld voor zowel praktische en wetenschappelijke doeleinden, bijvoorbeeld voor het beoordelen van individuen of om de effecten van verschillende onderwijsmethodes te bestuderen. Hoewel deze data vaak een zeer complexe structuur hebben, proberen wij de belangrijkste aspecten ervan te beschrijven aan de hand van relatief simpele statistische modellen, die er bijvoorbeeld vanuit gaan dat er één onderliggende vaardigheid is die verklaart waarom sommige vragen goed en andere vragen fout gemaakt worden. Het gebruik van deze relatief simpele modellen is nodig om conclusies te trekken over de niet-waarneembare constructies die van belang zijn (bijvoorbeeld lees- of rekenenvaardigheid) op basis van de geobserveerde toetsdata. Een algemeen en veelgebruikte kader voor het modelleren van toetsdata is de item respons theorie [IRT].

IRT richt zich op de geobserveerde item responsen en gebruikt relatief simpele modellen om item responsen te voorspellen aan de hand van item- en persoonskenmerken en hun interacties. Er is een verscheidenheid aan parametrische IRT-modellen, die door middel van de zogenaamde item response functies de relatie tussen de geobserveerde respons en de latente variabele (meestal vaardigheid in het kader van educatieve meting denoemd) beschrijven. Hierdoor kan men het niveau van het vaardigheid schatten aan de hand van de responsdata. In dit proefschrift worden IRT modellen voor dichotome data (dat wil zeggen, elk antwoord is ofwel juist of onjuist) beschouwd. De lezer wordt verwezen naar Lord en Novick (1968), Lord (1980), Hambleton en Swaminathan (1985) en van der Linden en Hambleton (1997) voor een uitgebreid overzicht van de IRT.

Dit proefschrift presenteert verschillende bijdragen aan IRT in het onderwijskundige meten die op de een of andere manier zoeken naar een optimale balans tussen het gebruiken van simpele modellen en de complexiteit van de werkelijkheid die deze modellen proberen te beschrijven. Het proefschrift bestaat uit twee delen: Deel I presenteert bijdragen aan het simultaan modelleren van responstijd en responsaccuraatheid, en Deel II presenteert Bayesiaanse bijdragen aan de IRT.

Educatieve toetsen wordt in toenemende mate uitgevoerd in een geautomatiseerde vorm, in plaats van de traditionele pen-en-papier toetsen. Dit maakt het mogelijk om niet alleen de antwoorden te registreren, maar ook hoe lang het duurt

voordat deze antwoorden geleverd worden. Een groot deel van de IRT literatuur is bezig met het ontwikkelen van manieren voor het opnemen van data over responstijden in de meetmodellen. Als de reactietijden en de respons accuraatheid gezamenlijk gemodelleerd worden neemt men vaak aan dat deze gemodelleerd kunnen worden aan de hand van een beperkt aantal latente vaiabelen, meestal de persoonseigenschappen vaardigheid en snelheid. Bovendien wordt vaak aangenomen dat er sprake is van conditionele onafhankelijkheid tussen de responstijd en accuraatheid, gegeven snelheid en vaardigheid van de persoon. Dit houdt in dat, hoewel juiste antwoorden misschien gemiddeld langzamer of sneller dan onjuiste responsen, wanneer men rekening houdt met de latente variabelen er geen verschillen tussen de verdelingen van de responsietijden van juiste en onjuiste responsen meer zijn.

De aanname van conditionele onafhankelijkheid is belangrijk, zowel vanuit statistische als inhoudelijke oogpunt, maar het is een aanname die in de praktijk geschonden kan zijn, en het evaluaren van deze aanname is een belangrijke stap in het gezamenlijk modelleren responstijd en accuraatheid. De hoofdstukken 2 en 3 van dit proefschrift bieden twee methoden voor het testen van deze aanname. In hoofdstuk 2 stellen we voor om conditionele onafhankelijkheid met behulp van Kolmogorov-Smirnov tests (Kolmogorov, 1933) te testen. De gelijkheid van de reactietijd-verdelingen gegeven een juiste of onjuiste respons wordt getest in elke subgroep van personen met dezelfde waarde van de sufficiente statistiek voor vaardigheid. In simulatiestudies bleek dat de procedure bruikbaar is voor het detecteren van verschillende soorten schendingen van conditionele onafhankelijkheid. Vanwege het feit dat de steekproef bij deze methode onderverdeeld wordt in verschillende subgroepen geldt wel dat deze methode een relatief grote steekproefgrootte nodig heeft voor het bereiken van voldoende hoog onderscheidingsvermogen.

Omdat de Kolmogorov-Smirnov test hoge eisen stelt wat betreft de vereiste steekproefgrotte is in hoofdstuk 3 een tweede procedure voor het testen van de aanname van conditionele onafhankelijkheid ontwikkeld. Posterior predictieve checks voor conditionele onafhankelijkheid zijn voorgesteld die gericht zijn op het onderzoeken van verschillende mogelijke gevolgen van conditionele afhankelijkheid: residuele correlaties tussen responstijd en accuraatheid gegeven de vaardigheid en de snelheid, het verschil tussen de varianties van de responstijden van juiste en onjuiste responsen, en het verschil tussen de item-rest correlaties van langzame en snelle responsen. Deze posterior predictieve checks bleken in simulatiestudies al bij relatief kleine steekproeven tot voldoende specificiteit en een hoge gevoeligheid te leiden.

In de praktijk van onderwijskundig meten kunnen er resterende afhankelijkheden tussen responstijden en respons accuraatheid voorkomen die niet kunnen worden verklaard door de latente variabelen. In hoofdstuk 4 richten we ons op een applicatie waarin dergelijke conditionele afhankelijkheid voorkomt en stellen wij een uitbreiding van het hierarchische model voor de responstijd en accuraatheid (van der Linden, 2007) voor, die met de conditionele afhankelijkheden rekening

houdt. De effecten van de relatieve snelheid van de respons in vergelijking met wat voor deze persoon op dit item verwacht werd op zowel de moelijkheid als het onderscheidend vermogen van het item worden hierbij opgenomen in het model. In de empirische toepassing bleek de moeilijkheidsgraad in het algemeen hoger te liggen voor tragere responses dan voor snelle responsen, terwijl het onderscheidend vermogen hoger is voor relatief snelle responsen.

De hoofdstukken in het tweede deel van het proefschrift zijn met elkaar verbonden door het statistische kader dat ze gemeen hebben, namelijk Bayesian statistische methodologie. Twee belangrijke eigenschappen van de Bayesiaanse statistiek die het zeer nuttig maken in de context van de IRT zijn dat deze methodes het mogelijk maken om zeer complexe modellen te schatten via op simulatie gebaseerde technieken en dat dergelijke methodes het mogelijk maken om relevante achtergrondinformatie voorbij de geobserveerde data in de analyses mee te nemen.

Hoofdstuk 5 laat het nut van de Bayesiaanse aanpak voor het schatten van complexe multidimensionale modellen zien. In dit onderzoek bieden wij een oplossing voor het probleem van het kiezen van een scoringsregel voor onderwijskundige toetsen. We pleiten voor het gebruik van scoringsregels die eenvoudig en gemakkelijk te interpreteren zijn, maar die toch alle informatie over het vaardigheid van de persoon bevatten. De eenvoudigste scoringsregel die hier aan voldoet is het gebruik van de somscore (dat wil zeggen, het aantal juiste antwoorden), die volgt uit het Rasch model (Rasch, 1960). Echter, dit model is vaak te beperkt om goed op echte toetsdata te passen. Daarom stellen wij in hoofdstuk 5 een nieuwe uitbreiding van het Rasch model voor die ervan uitgaat dat de test bestaat uit een reeks van schalen waarvoor elk een Rasch model opgaat, maar waarbij de schaallidmaatschappen van de items a priori niet bekend zijn en moeten worden geschat. Zodra de schalen worden geïdentificeerd, kan de test worden gescoord aan de hand van een aantal somscores voor elk van de schalen.

In hoofdstuk 6 tonen we een tweede belangrijke voordeel van het Bayesiaanse statistische kader, namelijk de mogelijkheid om prior kennis mee te nemen in de analyse. Het inhoudelijke probleem waar we ons op richten is het vergelijkbaar maken van de resultaten van verschillende versies van een test met behulp van zogenaamde linking en equating procedures (voor een uitgebreid overzicht van linking en equating, zie Kolen en Brennan, 2004). We stellen twee methoden voor voor het eliciteren van prior kennis van vakexperts over het mogelijke verschil in moeilijkheidsgraad van de twee tests. Deze voorkennis wordt meegenomen bij het linken aan de hand van prior verdelingen, om daarmee het resultaat te verbeteren. De resultaten van onze twee empirische elicitatie studies voor de Entreetoets van groep 7 suggereren dat deze aanpak de kwaliteit van het linken kunnen verbeteren: de precisie van de resultatent verbeterd zonder dat er sprake is van vertekening.

In hoofdstuk 7 komt een derde aantrekkelijke eigenschap van het gebruik va Bayesiaanse schattingsmethodes aan bod: Het maakt het mogelijk om met verschillende bronnen van onzekerheid over de modelparameters rekening te houden.

Als sommige modelparameters niet volledig zijn geïdentificeerd is nog steeds mogelijk om trekkingen uit hun posteriori verdeling te halen, en hun variantie omvat zowel de onzekerheid ten gevolge van steekproefvariantie als de onzekerheid als gevolg van het feit dat het model niet volledig is geïdentificeerd. Met behulp van een simulatie benadering kan men proberen deze twee soorten onzekerheid te scheiden en daarmee het effect dat de niet-identificeerbaarheid heeft op de model inferenties onderzoeken. In dit hoofdstuk wordt dit onderzocht in de context van de verdeling van de ontbrekende data bij eenincomplete testopzet. We laten zien dat, terwijl de verdeling van de niet-waargenomen scores van de referentiepopulatie op de nieuwe test niet volledig geïdentificeerd is, de onzekerheid over de scoreverdeling zeer klein is en in de praktijk kan worden genegeerd. Verder laten we aan de hand van zowel voorbeelden bestaande uit gesimuleerde als echte toetsresultaten zien dat het negeren van het identificatieprobleem onder de aanname van een normale verdeling van de vaardigheid kan leiden tot een vertekening bij test equating.

Acknowledgements

First of all, I would like to thank my supervisors Gunter Maris and Herbert Hoijtink, I have learned a lot from both of you. I am very happy that I had both of you as my supervisors. Gunter, thank you for inspiring me with creative ideas and evil plans. Herbert, thank you for helping me keeping structure and not being carried away by wild ideas. And thank you both for giving me the freedom to work on various topics which I found interesting.

I worked on this thesis at the POK department at Cito and at the Department of Methodology and Statistics at Utrecht University and I am thankful to all my colleagues there for providing me with not one, but two great working environments. Dear Noémi, Jorine, Maarten, Timo, Mariëlle, Kees, Katarine, Shahab, Charlotte, Manon, Matthieu, Robert, Saskia, Rinke, Aniek, Maryam, Gerko, Marie-Anne, Nikky, and Silvia, thank you for all the fun and serious, psychometric and non-psychometric discussions that we have had.

Visiting Ohio State University in the last year of my PhD project was a great experience to me for which I am very grateful to Herbert, who helped to finance my trip to the States, and to Paul de Boeck, who was willing to receive me as a visiting scholar. Paul, our many discussions and exchange of ideas have been very fruitful and I have learned a lot from you. I would like to thank Paul, and also Min-Jeong, Hando, Leanne, and Robert for welcoming me in Columbus.

Chapter 6 of this dissertation required involvement of many people. I am thankful to Anton and Rianne van der Werff for the facilitating data collection, Jorine Vermeulen for helping with the expert meetings, Pascal van Kooten for writing the software, and Marije Fagginger Auer, Marian Hickendorff, Kees van Putten, Michiel Veldhuis, and Monica Wijers for finding the time in their busy research schedules to participate in the study.

Last but not least I would like to thank my parents for supporting my crazy idea of moving to the Netherlands, and Jesper for being there for me from Day 1 of my PhD project.

Curriculum Vitae

Maria Bolsinova was born in 1988 on October 15 in Moscow, Russia. From 2005 till 2010 she studied psychology at Moscow State University. After obtaining a cum laude degree in Work and Organisational Psychology she received Golden Excellence Scholarship from Leiden University to continue her education at the Methodology and Statistics Master programme at the Department of Psychology. She graduated cum laude for this master in 2011, after completing an internship at Cito, the Dutch National Institute for Educational Measurement.

In September 2011, Maria started working as a PhD candidate at the Department of Methodology and Statistics of Utrecht University and at the Psychometric Research Centre at Cito. Next to doing her dissertation research, she worked as a test expert on a variety of projects at Cito, was involved in statistical consultations to students and researchers, and in teaching undergraduate, graduate, and post-graduate courses in statistics at Utrecht University. In 2014, she went on a three-month research visit to Ohio State University with a research visit. Together with prof. Paul de Boeck she worked for three months on the psychometric modeling of response time and accuracy. Maria has presented her research at various national and international conferences and a number of her work has appeared in journals, such as the British Journal of Mathematical and Statistical Psychology and Frontiers in Psychology.

At present Maria is a post-doctoral researcher in Psychological Methods at the Department of Psychology at the University of Amsterdam.