The Annals of Applied Statistics
2016, Vol. 10, No. 2, 925–945
DOI: 10.1214/16-AOAS919

UNMIXING RASCH SCALES: HOW TO SCORE AN EDUCATIONAL TEST

BY MARIA BOLSINOVA[*,†], GUNTER MARIS[†,‡] AND HERBERT HOIJTINK[*]

Utrecht University[*], CITO Dutch Institute for Educational Measurement[†] and University of Amsterdam[‡]

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 to 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.

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 subquestions: Should the results be summarized 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 do we determine which items belong to which subscale? If different weights are used, how do we restrict the number of possible weights such that not every response pattern results in a unique weighted score? And how dow we 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.

Received May 2015; revised February 2016.

Key words and phrases. Educational testing, Markov chain Monte Carlo, mixture model, multidimensional IRT, one parameter logistic model, Rasch model, scoring rule.

The aim of this article is to choose a simple scoring rule for the state exam of Dutch as a foreign language [College voor Toetsen en Examens: Staatsexamen NT2, (n.d)]. 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 favorable because no information about the ability is lost by summarizing a vector of responses in one or more scores. One of the simplest IRT models—the Rasch model (RM) [Rasch (1980)]—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 a 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 summarizing 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 organized as follows. First, in Section 2 the state examination of Dutch as a 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 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 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, the estimation of the model is discussed. In Section 6, we evaluate the estimation procedure in a simulation study. After introducing the methodology and showing its usability in simulations, in Section 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.

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 computerized 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):

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

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 the 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 sumscores 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 and 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.

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 a 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, then a general model allowing for multidimensionality and different discriminations of items can be obtained [Reckase (2008)]:

$$\Pr(X_{pi} = 1 | \delta_i, \boldsymbol{\alpha}_i, \boldsymbol{\theta}_{p\cdot}) = \frac{\exp(\sum_{k=1}^{M} \alpha_{ik}\theta_{pk} - \delta_i)}{1 + \exp(\sum_{k=1}^{M} \alpha_{ik}\theta_{pk} - \delta_i)}, \tag{2}$$

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 (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 (2)] has only one nonzero 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 nonzero 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 summarized in a vector of subscores $\{\sum_i \alpha_{i1} X_{pi}, \sum_i \alpha_{i2} X_{pi}, \ldots, \sum_i \alpha_{iM} X_{pi}\}$. 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 (2), but with the constraints $\alpha_{ik} \in \{0, 1\}$ and $\sum_k \alpha_{ik} = 1$. Thus, $\boldsymbol{\alpha}_i = \{\alpha_{i1}, \ldots, \alpha_{iM}\}$ 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 and 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 reparameterize the difficulty parameter as $\delta_i^* = a_i \delta_i$. Then, within each set of items $\{i | a_i = \sigma_k\}$, a RM with person parameter $\boldsymbol{\theta}_{\cdot k}$ holds [Humphry (2011)], and the whole test is modeled 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:

$$\theta_{pk} = \frac{\sigma_k}{\sigma_l} \theta_{pl}. \tag{3}$$

These scales measure the same latent variable but represent different frames of reference and have different units of measurement [Humphry and 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 sum-scores, but also 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: $p(\boldsymbol{\alpha}|\mathbf{X})$. 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. For the proof see Section 1 of the supplementary article [Bolsinova, Maris and Hoijtink (2016)]. 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 and 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 prespecified, 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 determining 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 Section 3 of the supplementary article [Bolsinova, Maris and Hoijtink (2016)].

4. Model specification.

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 realization 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:

$$\Pr(\mathbf{X}_{p\cdot} = \mathbf{x}|\boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) = \int_{\mathbb{R}} \prod_{i=1}^{n} \frac{(\exp(\sum_{k=1}^{M} \alpha_{ik}\theta_k - \delta_i))^{x_i}}{1 + \exp(\sum_{k=1}^{M} \alpha_{ik}\theta_k - \delta_i)} p(\boldsymbol{\theta}|\boldsymbol{\Sigma})\, d\boldsymbol{\theta}, \tag{4}$$

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 (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 (4), the multi-scale RM assumes local independence, meaning that the item responses are independent given the vector of ability parameters.

In Section 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 (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 summarized in a subscore. For a model with four item clusters the covariance matrix has the form

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & \sigma_{1,4} \\ \sigma_{1,2} & \sigma_2^2 & \sigma_{2,3} & \sigma_{2,4} \\ \sigma_{1,3} & \sigma_{2,3} & \sigma_3^2 & \sigma_{3,4} \\ \sigma_{1,4} & \sigma_{2,4} & \sigma_{3,4} & \sigma_4^2 \end{bmatrix} \tag{5}$$

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:

$$\left\{ \sum_i \alpha_{i1} X_{pi}, \sum_i \alpha_{i2} X_{pi} \right\}. \tag{6}$$

Type 2. The test measures several abilities, but not each scale represents 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

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2 & \rho\sigma_1\sigma_3 & \rho\sigma_1\sigma_4 \\ \sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \rho\sigma_2\sigma_4 \\ \rho\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 & \sigma_3\sigma_4 \\ \rho\sigma_1\sigma_4 & \rho\sigma_2\sigma_4 & \sigma_3\sigma_4 & \sigma_4^2 \end{bmatrix}, \tag{7}$$

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 a bivariate normal distribution for the person parameters with a zero mean vector, unit variances and correlation ρ between the dimensions, 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 be 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 in the covariance matrix could have the

form

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2 & \sigma_1\sigma_3 & \sigma_1\sigma_4 \\ \sigma_1\sigma_2 & \sigma_2^2 & \sigma_2\sigma_3 & \sigma_2\sigma_4 \\ \sigma_1\sigma_3 & \sigma_2\sigma_3 & \sigma_3^2 & \sigma_3\sigma_4 \\ \sigma_1\sigma_4 & \sigma_2\sigma_4 & \sigma_3\sigma_4 & \sigma_4^2 \end{bmatrix}, \tag{8}$$

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 to estimate these weights.

The algorithm presented in this paper is exploratory, therefore, it need not be prespecified 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 prespecified beforehand, but can be decided upon based on the estimation results (see Section 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.

4.2. *Density of the data, prior and posterior distributions.* The density of the data is

$$f(\mathbf{X}|\boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) = \prod_{p=1}^{N} \Pr(\mathbf{X}_{p\cdot} = \mathbf{x}|\boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}), \tag{9}$$

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:

$$p(\boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} p(\delta_i) \prod_{i=1}^{n} p(\boldsymbol{\alpha}_i) p(\boldsymbol{\Sigma}). \tag{10}$$

We use noninformative 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 et al. (2000)]. For the item memberships a multinomial prior is used:

$$\Pr(\alpha_{ik} = 1, \alpha_{il} = 0, \forall l \neq k) = \frac{1}{M}, \qquad \forall k \in [1 : M], \forall i \in [1 : n], \tag{11}$$

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), page 110].

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

$$p(\boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma} | \mathbf{X}) \propto f(\mathbf{X} | \boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) p(\boldsymbol{\delta}) p(\boldsymbol{\alpha}) p(\boldsymbol{\Sigma}). \tag{12}$$

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

5. Estimation.

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 and Lopes (2006)] when the number of scales is prespecified. This algorithm makes it possible to obtain samples from the posterior distribution in equation (12). 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 initialization, 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 [Casella and George (1992), Geman and Geman (1984)]. Data augmentation is implemented [Tanner and Wong

(1987), Zeger and 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. In the first three steps individual person parameters, the covariance matrix and the item difficulties are sampled from their full conditional posteriors. Since during these steps the item memberships are fixed, sampling from these conditional posteriors does not differ from sampling from the posterior of a standard between-item multidimensional RM. The details about Steps 1–3 are presented in Section 2 of the supplementary article [Bolsinova, Maris and Hoijtink (2016)]. Step 4 is specific for the multi-scale RM: For every item $i \in [1:n]$, sample item scale membership $\boldsymbol{\alpha}_i$ from the full conditional posterior distribution:

$$
\begin{aligned}
p(\boldsymbol{\alpha}_i|\mathbf{X}, \boldsymbol{\delta}, \boldsymbol{\alpha}_{(i)}, \boldsymbol{\theta}, \boldsymbol{\Sigma}) &= p(\boldsymbol{\alpha}_i|\mathbf{X}_{\cdot i}, \delta_i, \boldsymbol{\theta}) \\
&\propto \prod_p \frac{\exp(X_{pi}(\sum_{k=1}^M \alpha_{ik}\theta_{pk} - \delta_i))}{1 + \exp(\sum_{k=1}^M \alpha_{ik}\theta_{pk} - \delta_i)},
\end{aligned} \tag{13}
$$

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

$$
p_{ik} = \Pr(\alpha_{ik} = 1, \alpha_{il} = 0, \forall l \neq k) = \frac{\prod_p \frac{\exp(X_{pi}(\theta_{pk} - \delta_i))}{1+\exp(\theta_{pk} - \delta_i)}}{\sum_{j=1}^M \prod_p \frac{\exp(X_{pi}(\theta_{pj} - \delta_i))}{1+\exp(\theta_{pj} - \delta_i)}}. \tag{14}
$$

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 and 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 and Robert (2000)]. It is more likely that the chain will stay in the neighborhood 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 the

following: (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

$$\bar{L}_c = \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n}\sum_{p=1}^{N}\ln\left(\frac{(\exp(\sum_{k=1}^{M}\alpha_{ik}^{gc}\theta_{pk}^{gc}-\delta_i^{gc}))^{X_{pi}}}{1+\exp(\sum_{k=1}^{M}\alpha_{ik}^{gc}\theta_{pk}^{gc}-\delta_i^{gc})}\right), \tag{15}$$

where G denotes the number of iterations after the burn-in and superscripts g and c denote the value of a parameter at the gth iteration in the cth 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 relabel 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 relabeled 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 relabeled chains can be combined. For each item i and each scale k compute the posterior probability of this item to belong to this scale:

$$\hat{\pi}_{ik} = \frac{\sum_{c\in\mathbf{C}}\sum_g \alpha_{ik}^{g,c}}{|\mathbf{C}|G}, \tag{16}$$

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 modeled 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.

5.2. *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 and 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 2 of the supplementary article [Bolsinova, Maris and Hoijtink (2016)] with $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$). 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 (i.e., the average values of the parameters across the iterations of the Gibbs Sampler after the burn-in), denoted by $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\Sigma}}$.

The modified AIC [Akaike (1974)] and the BIC [Schwarz (1978)] are computed as follows:[1]

$$\text{AIC} = -2\ln f(\mathbf{X}|\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Sigma}}) + 2\left(\frac{M(M+1)}{2} + n + (M-1)n\right) \tag{17}$$

and

$$\text{BIC} = -2\ln f(\mathbf{X}|\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Sigma}}) + \ln N\left(\frac{M(M+1)}{2} + n + (M-1)n\right). \tag{18}$$

The estimates $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\delta}}$ are used instead of the estimates based on the posterior in equation (12), since if throughout the iterations the items move frequently across the scales, the EAP estimates based on the draws from equation (12) 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 reassigned to a new scale or not. The evaluation of the log-likelihood in equations (17) and (18) involves integration over a multidimensional space, which is done here through numerical integration with a Gauss–Hermite quadrature. For details see Section 3 of the supplementary article [Bolsinova, Maris and Hoijtink (2016)].

[1]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.

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.

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.[2] The scales are correctly reconstructed if for every item the posterior 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

[2]In Section 4 of the supplementary article [Bolsinova, Maris and Hoijtink (2016)] 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 and Arendasy (2012)], which also aims at constructing a set of scales that each fit 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). The third study evaluates the autocorrelations in the Markov chain.

TABLE 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* $\hat{M} = M$

		Condition					
Method		**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

were fitted and was equal to $M \times 500, \forall M \in [2 : 6]$. The modified AIC and the modified BIC [see equations (17) and (18)] 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 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) 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 1.

7. Choosing a scoring rule for the NT2 exam.

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, and a testing set ($N = 898$) which was used for testing

TABLE 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	12,2494.2	41.5
3-scale RM	24/34/13/3	12,2452.7	0
4-scale RM	22/9/34/7/2	12,2608.7	156.0

whether the scales identified in the exploratory part are indeed Rasch scales, and testing hypotheses about the relations between the unmixed scales.

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 summarized in Table 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.

The estimated covariance matrix was

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 1.67\ [1.48, 1.88] & 1.13\ [1.01, 1.25] & 0.71\ [0.63, 0.80] \\ \textit{0.96 [0.95,0.97]} & 0.83\ [0.74, 0.93] & 0.49\ [0.44, 0.55] \\ \textit{0.92 [0.88,0.95]} & \textit{0.90 [0.86,0.94]} & 0.36\ [0.31, 0.42] \end{bmatrix}, \tag{19}$$

where the elements below the diagonal (italicized) are the correlation coefficients and the elements above the diagonal are the covariances, and 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.

TABLE 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

7.3. *Cross-validation of the unmixed scales.*

7.3.1. *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 scales 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 the R package `eRm` [Mair and Hatzinger (2007)].

We did not expect a perfect fit of the RM to the complete scales (see lines 1, 3 and 5 of Table 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 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.

7.3.2. *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 $\hat{\boldsymbol{\delta}}_{\text{type1}}$ and $\hat{\boldsymbol{\Sigma}}_{\text{type1}}$). As has been mentioned in

Section 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 Section 5 of the supplementary article [Bolsinova, Maris and Hoijtink (2016)]) and estimates of the item difficulties and the three discrimination parameters were obtained (denoted by $\hat{\boldsymbol{\delta}}_{\text{type3}}$ and $\hat{\boldsymbol{\sigma}}_{\text{type3}}$).

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

$$\ln\bigl(f(\mathbf{X}_{\text{test}}|\hat{\boldsymbol{\delta}}_{\text{type1}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Sigma}}_{\text{type1}})\bigr) = -34,776.93, \tag{20}$$

$$\ln\bigl(f(\mathbf{X}_{\text{test}}|\hat{\boldsymbol{\delta}}_{\text{type3}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\sigma}}_{\text{type3}})\bigr) = -34,767.83. \tag{21}$$

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.

7.3.3. *Does it make a difference?* Finally, we investigated whether using the chosen scoring rule $\sum_i (1.30\alpha_{i1} + 0.89\alpha_{i2} + 0.56\alpha_{i3})X_{pi}$ 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, denoted 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 cutoff value for the weighted sumscore leading to the same number of students passing the test is 48.21. Table 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 rules match each other, while for

TABLE 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

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 the noncompensatory scoring rule based on the two unweighted subscores.

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 scales 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 the 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 a 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 (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.

SUPPLEMENTARY MATERIAL

Supplement A: Supplement to "Unmixing Rasch scales: How to score an educational test." (DOI: 10.1214/16-AOAS919SUPP; .pdf). We provide the proof of identification of the multi-scale Rasch model in Section 1, details of the Gibbs Sampler for estimating the model in Section 2, details on approximating the likelihood of the model in Section 3, results of additional simulation studies in Section 4, and details on estimation of the model with fixed correlation parameters in Section 5.

REFERENCES

ADAMS, R., WILSON, M. and WANG, W. (1997). The multidimensional random coefficients multinomial logit model. *Appl. Psychol. Meas.* **12** 261–280.

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC-19** 716–723. MR0423716

ANDERSEN, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika* **38** 123–140. MR0311064

ANDERSEN, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika* **42** 69–81. MR0483255

BIRNBAUM, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores* (F. M. Lord and M. R. Novick, eds.) 395–479. Addison-Wesley, Reading, MA.

BOLSINOVA, M., MARIS, G. and HOIJTINK, H. (2016). Supplement to "Unmixing Rasch scales: How to score an educational test." DOI:10.1214/16-AOAS919SUPP.

CASELLA, G. and GEORGE, E. I. (1992). Explaining the Gibbs sampler. *Amer. Statist.* **46** 167–174. MR1183069

CELEUX, G., HURN, M. and ROBERT, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.* **95** 957–970. MR1804450

COLLEGE VOOR TOETSEN EN EXAMENS: STAATSEXAMENS NT2 (n.d.). Retrieved September 25, 2015. Available at http://www.staatsexamensnt2.nl.

COUNCIL OF EUROPE (2011). Common European Framework of Reference for Learning, Teaching, Assessment. Council of Europe.

DEBELAK, R. and ARENDASY, M. (2012). An algorithm for testing unidimensionality and clustering items in Rasch measurement. *Educ. Psychol. Meas.* **72** 375–387.

DIEBOLT, J. and ROBERT, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. Ser. B* **56** 363–375. MR1281940

FISCHER, G. H. (1995). Derivations of the Rasch model. In *Rasch Models* (*Vienna*, 1993) (G. H. Fisher and I. W. Molenaar, eds.) 15–38. Springer, New York. MR1367343

FRÜHWIRTH-SCHNATTER, S. (2006). *Finite Mixture and Markov Switching Models*. Springer, New York. MR2265601

GAMERMAN, D. and LOPES, H. F. (2006). *Markov Chain Monte Carlo*: *Stochastic Simulation for Bayesian Inference*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL. MR2260716

GEMAN, S. and 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. and AGRESTI, A. (2000). Noninformative priors for one-parameter item response models. *J. Statist. Plann. Inference* **88** 99–115.

HARDOUIN, J.-B. and MESBAH, M. (2004). Clustering binary variables in subscales using an extended Rasch model and Akaike information criterion. *Comm. Statist. Theory Methods* **33** 1277–1294. MR2069568

HOFF, P. D. (2009). *A First Course in Bayesian Statistical Methods*. *Springer Texts in Statistics*. Springer, New York. MR2648134

HUMPHRY, S. (2011). The role of the unit in physics and psychometrics. *Measurement*: *Interdisciplinary Research and Perspective* **9** 1–24.

HUMPHRY, S. (2012). Item set discrimination and the unit in the Rasch model. *J. Appl. Meas.* **13** 165–224.

HUMPHRY, S. and ANDRICH, D. (2008). Understanding the unit in the Rasch model. *J. Appl. Meas.* **9** 249–264.

LORD, F. M. and NOVICK, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA.

MAIR, P. and HATZINGER, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *J. Stat. Softw.* **20** 1–20.

MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. Wiley, New York. MR1789474

RASCH, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*, expanded edition. The Univ. Chicago Press, Chicago.

RECKASE, M. (2008). *Multidimensional Item Response Theory*. Springer, New York, NY.

ROST, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement* **14** 271–282.

SCHWARZ, G. (1978). Estimating the dimension of the model. *Ann. Statist.* **6** 461–464.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550. MR0898357

VERHELST, N. D. and GLAS, C. A. W. (1995). The one parameter logistic model: OPLM. In *Rasch Models*: *Foundations*, *Recent Developments and Applications* (G. H. Fischer and I. W. Molenaar, eds.) 215–238. Springer, New York.

ZEGER, K. and KARIM, M. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *J. Amer. Statist. Assoc.* **86** 79–86.

M. BOLSINOVA
G. MARIS
DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF AMSTERDAM
NIEUWE ACHTERGRACHT 129-B
AMSTERDAM, 1018 WC
THE NETHERLANDS
E-MAIL: m.a.bolsinova@uva.nl
g.k.j.maris@uva.nl

H. HOIJTINK
DEPARTMENT OF METHODOLOGY AND STATISTICS
UTRECHT UNIVERSITY
PADUALAAN 14
UTRECHT, 3584 CH
THE NETHERLANDS
E-MAIL: h.hoijtink@uu.nl