APPLIED MATHEMATICS SERIES

Rasch Models in Health

Edited by Karl Bang Christensen
Svend Kreiner, Mounir Mesbah

iSTE
WILEY

Table of Contents

PART 6 Analyzing and Reporting Rasch Models

Rasch Models in Health

Edited by
Karl Bang Christensen
Svend Kreiner
Mounir Mesbah

iSTE

WILEY

First published 2013 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

ISTE Ltd

27-37 St George's Road

London SW19 4EU

UK

www.iste.co.uk

John Wiley & Sons, Inc.

111 River Street

Hoboken, NJ 07030

USA

www.wiley.com

Library of Congress Control Number: 2012950096

British Library Cataloguing-in-Publication Data

A CIP record for this book is available from the British Library

ISBN: 978-1-84821-222-0

MIX
Paper from responsible sources
FSC® C013604
FSC
www.fsc.org

Preface

The family of statistical models known as Rasch models started with a simple model for responses to questions in educational tests presented together with a number of related models that the Danish mathematician Georg Rasch referred to as models for measurement. Since the beginning of the 1950s the use of Rasch models has grown and has spread from education to the measurement of health status. This book contains a comprehensive overview of the statistical theory of Rasch models.

Because of the seminal work of Georg Rasch [RAS 60] a large number of research papers discussing and using the model have been published. The views taken of the model are somewhat different. Some regard it as a measurement model and focus on the special features of measurement by items from Rasch models. Other publications see the Rasch model as a special case of the more general class of statistical models known as item response theory (IRT) models [VAN 97]. And, finally, some regard the Rasch model as a statistical model and focus on statistical inference using these models.

The statistical point of view is taken in this book, but it is important to stress that we see no real conflict between the different ways that the model is regarded. The Rasch model is one of the several measurement models defined by Rasch [RAS 60, RAS 61] and is, of course, also an IRT model. And even if measurement is the only concern, we need observed data and statistical estimates of person parameters to calculate the measures.

The statistical point of view is thus unavoidable. From this point of view, the sufficiency of the raw score is crucial and, following in the footsteps of Georg Rasch and his student Erling B. Andersen, we focus on methods depending on the conditional distribution of item responses given the raw score. The relationship between Rasch models and the family of multivariate models called graphical models [WHI 90, LAU 96] is also highlighted because this relationship enables analysis and modeling of properties like local dependence and non-differential item functioning in a very transparent way.

The book is structured as follows: Part I contains the probabilistic definition of Rasch models; Part II describes estimation of item and person parameters; Part III is about the assessment of the data-model fit of Rasch models; Part IV contains applications of Rasch models; Part V discusses how to develop health-related instruments for Rasch models; and Part VI describes how to perform Rasch analysis and document results.

The focus on the Rasch model as a statistical model with a latent variable means that little will be said about other IRT models, such as the two parameter logistic (2PL) model and the graded response model. This does not reflect a strong "religious" belief, that the Rasch model is the only interesting and useful IRT or measurement model, but only reflects our choice of a point of view for this book.

The book owes a lot to discussions at a series of workshops on Rasch models held in Stockholm (Sweden, 2001), Leeds (UK, 2002), Perth (Australia, 2003), Skagen (Denmark, 2005), Vannes (France, 2006), Bled (Slovenia, 2007), Perth (Australia, 2008 and 2012), Copenhagen (Denmark, 2010) and Dubrovnik (Croatia, 2011). Many of the authors have taken part and have helped create an atmosphere where topics relating to the Rasch model could be discussed in an open, friendly and productive manner.

The participants do not agree on everything and do not share all the points of views expressed. However, everyone agrees on the importance of Rasch's contributions to

measurement and statistics, and it is fair to say that this book would not exist if it had not been for these workshops.

Karl Bang CHRISTENSEN, Svend KREINER and Mounir MESBAH
Copenhagen, November 2012

Bibliography

[LAU 96] LAURITZEN S. *Graphical Models*, Clarendon Press, 1996.

[RAS 60] RASCH G., *Probabilistic Models for Some Intelligence and Attainment Tests*, Danish National Institute for Educational Research, Copenhagen, 1960.

[RAS 61] RASCH G., "On general laws and the meaning of measurement in psychology", *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, CA, pp. 321–334, 1961.

[VAN 97] VAN DER LINDEN W.J., HAMBLETON R.K., *Handbook of Modern Item Response Theory*, Springer-Verlag, New York, NY, 1997.

[WHI 90] WHITTAKER J., *Graphical Models in Applied Multivariate Statistics*, Wiley, Chichester, UK, 1990.

PART 1

Probabilistic Models

Introduction

This part introduces the models that are analyzed in the book. The Rasch model was originally formulated by Georg Rasch for dichotomous items [RAS 60]. This model is described in Chapter 1, where different parameterizations are also introduced. The sources of polytomous Rasch models are less clear. Georg Rasch formulated a quite general polytomous model where each item measures several latent variables [RAS 61]. However, this model has seen little use. Later, several authors [AND 77, AND 78, MAS 82] formulated models where items with more than two response categories measure a single underlying latent variable.

Bibliography

[AND 77] ANDERSEN E.B., "Sufficient statistics and latent trait models", *Psychometrika*, vol. 42, pp. 69–81, 1977.

[AND 78] ANDRICH D., "A rating formulation for ordered response categories", *Psychometrika*, vol. 43, pp. 561–573, 1978.

[MAS 82] MASTERS G.N., "A Rasch model for partial credit scoring", *Psychometrika*, vol. 47, pp. 149–174, 1982.

[RAS 60] RASCH G., *Probabilistic Models for Some Intelligence and Attainment Tests*, Danish National Institute for Educational Research, Copenhagen, 1960.

[RAS 61] RASCH G., “On general laws and the meaning of measurement in psychology”, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, IV*, University of California Press, Berkeley, CA, pp. 321–334, 1961.

Chapter 1

The Rasch Model for Dichotomous Items

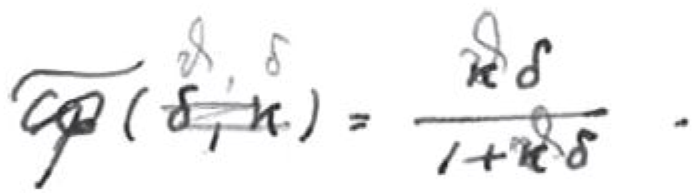
1.1. Introduction

The family of statistical models, which is known as Rasch models, was first introduced with a simple model for responses to dichotomous items (questions) in educational tests [RAS 60]. It was presented together with a number of related models that the Danish mathematician Georg Rasch called models for measurement. Since then, the family of Rasch models has grown to encompass a number of statistical models.

1.1.1. *Original formulation of the model*

All Rasch models share a number of fundamental properties, and we introduce this book with a brief recapitulation of the very first Rasch model: the Rasch model for dichotomous items. This model was developed during the 1950s when Georg Rasch got involved in educational research. The model describes responses to a number of items by a number of persons assuming that responses are stochastically independent, depending on unknown items and person parameters. In Rasch's original conception of the model (see Figure 1.1), the structure of the model was multiplicative. In this model, the probability of a positive response to an item depends on a *person parameter* ξ and an *item parameter* δ in such a way that the probability of a positive response to an item depends on the product of the person parameter and the item parameter.

Figure 1.1. *The Rasch model 1952/1953*



If we refer to the response of person v to item i as X_{vi} and code a positive response as 1 and a negative response as 0, the Rasch model asserts that

[1.1] $$Pr(X_{vi} = 1) = \frac{\xi_v \delta_i}{1 + \xi_v \delta_i}$$

where both parameters are non-negative real numbers. It follows from [1.1] that

[1.2] $$Pr(X_{vi} = 0) = 1 - Pr(X_{vi} = 1) = \frac{1}{1 + \xi_v \delta_i}$$

The interpretation of the parameters in this model is straightforward: the probability of a positive response increases as the parameters increase toward infinity. In educational testing, the person parameter represents the ability of the student and the item parameter represents the easiness of the item: the better the ability and the easier the item, the larger the probability of a correct response to the item. In health sciences, the person parameter could represent the level of depression whereas the item parameters could represent the risk of experiencing certain symptoms relating to depression.

EXAMPLE 1.1.– Consider the following dichotomous items intended to measure depression:

1) Did you have sleep disturbance every day for a period of two weeks or more?

2) Did you have a loss or decrease in activities every day for a period of two weeks or more?

3) Did you have low self-esteem every day for a period of two weeks or more?

4) Did you have decreased appetite every day for a period of two weeks or more?

Items like these appear in several questionnaires. According to the Rasch model, responses to these items depend on the level of depression measured by the ξ parameter and on four item parameters δ_1–δ_4. In a recent study, the item parameters were found to be 2.57, 1.57, 0.52 and 0.48, respectively [FRE 09, MES 09]. The interpretation of these numbers is that sleep disturbance is the most common and loss of appetite is the least common of the four symptoms. To better understand the role of the item parameters, we have to look at the relationships between the probabilities of positive responses to two questions. This is shown in Table 1.1, where it can be seen that the *ratio* between the two item parameters is the odds ratio (OR) comparing the odds of encountering the symptoms described by the items *irrespective* of the level of depression ξ of the persons. This interpretation should be familiar to persons with a working knowledge of epidemiological methods. According to the Rasch model, the level of depression does not modify the *relative* risk of the symptoms. In the theory of Rasch models, this is sometimes called *no item-trait interaction.*

Table 1.1. *Response probabilities for two items when the person parameter is* ξ_v

Item	parameter	$P(X=0)$	$P(X=1)$	Odds
1	δ_1	$\frac{1}{1+\xi_v\delta_1}$	$\frac{\xi_v\delta_1}{1+\xi_v\delta_1}$	$\xi_v\delta_1$
2	δ_2	$\frac{1}{1+\xi_v\delta_2}$	$\frac{\xi_v\delta_2}{1+\xi_v\delta_2}$	$\xi_v\delta_2$
				$\mathrm{OR} = \frac{\delta_2}{\delta_1}$

EXAMPLE 1.2.– Since the item parameters for the first two items are 2.57 and 1.57, we see that the odds ratio relating the risk of loss of or reduction of activities to the risk of sleep disturbances is equal to 1.57/2.57 = 0.613. Because of the symmetry in formula [1.1] the same argument applies to comparisons of persons. Table 1.2 considers the risk of encountering a specific symptom for each of two persons with different levels of depression. As for the items, we interpret the ratio between the person parameters as the odds ratio comparing the risk for person two to the risk for person one.

Table 1.2. *Response probabilities for an item with an item parameter equal to δ_i*

Person	Parameter	$P(X=0)$	$P(X=1)$	Odds
1	ξ_1	$\frac{1}{1+\xi_1\delta_i}$	$\frac{\xi_1\delta_i}{1+\xi_1\delta_i}$	$\xi_1\delta_i$
2	ξ_2	$\frac{1}{1+\xi_2\delta_i}$	$\frac{\xi_2\delta_i}{1+\xi_2\delta_i}$	$\xi_2\delta_i$
				$\mathrm{OR} = \frac{\xi_2}{\xi_1}$

To measure the level of depression, we have to estimate the parameter ξ based on observed item responses. However, this parameter is not identifiable in absolute terms because the probabilities [1.1] depend on the product of the person and item parameters. Multiplying all person parameters by a constant κ and dividing all item parameters by the same constant results in a reparameterized model

[1.3] $$P(X_{vi}=1) = \frac{(\xi_v\kappa)(\delta_i/\kappa)}{1+(\xi_v\kappa)(\delta_i/\kappa)}$$

[1.4] $$P(X_{vi} = 0) = \frac{1}{1 + (\xi_v \kappa)(\delta_i / \kappa)}$$

with exactly the same formal structure and the same probabilities as the original model, and where the odds ratio comparing response probabilities for the two persons is the same as in Table 1.2. To identify the parameters, we consequently have to impose restrictions on the parameters. The standard way of doing this is to fix the parameters such that the product of the item parameters is equal to one. The parameters of the depression items above were fixed in this way. Another way that may be more natural for an epidemiologist would be to select a reference item where the item parameter is equal to one. The item parameters for other items are then interpretable as ORs comparing the item to the reference item (see Table 1.1). Because of the symmetry in formula [1.1], similar arguments apply to the person parameters, that is requiring that the product of the person parameters be equal to one or fixing the value for a single (reference) person. All these parameterizations are valid and characterized by invariant ratios of both the person parameters and item parameters.

Multiplication of quantitative measurements with a constant corresponds to a change of unit of the measurement scale on which the values are measured. Because ratios of person parameters are the same for all choices of a measurement unit, the measurement scale on which ξ is measured is a ratio scale. This argument was very important for Georg Rasch who repeatedly stressed the similarity with measurement in physics, stating [RAS 60]

> If for any two objects we find a certain ratio of their accelerations produced by one instrument, then the same ratio will be found for any other instruments.

Measurement using Rasch models is relative rather than absolute. We can use estimates of ξ to compare the level of depression for two persons, but we cannot use a single ξ

measure to say that a person has a high or a low level of depression. Michell [MIC 97] claims that “scientific measurement is properly defined as the estimation of the ratio of some magnitude of a quantitative attribute to a unit of the same attribute” and also points out that measurement is relative rather than absolute depending on the choice of unit.

One further aspect of Rasch models is worth mentioning. Persons and items are completely symmetrical in the sense that there is no major difference between inference on item parameters and inference on person parameters using the simple model [1.1]. However, in the majority of applications, we will not exchange persons and items. The main purpose of constructing depression items like those discussed above is to measure a trait or the property of persons, whereas the risks associated with the four symptoms are of no special significance being only the means to the ends. Typically, covariates like age, gender and socioeconomic status are attached to people but not to items. Hence, conceptually there is a big difference between persons and items.

1.1.2. *Modern formulations of the model*

Over time, as the use of the model spread from educational testing to other research areas, the formal representation and the terminology associated with the model got changed. Today, the model is typically written as an additive logistic model, replacing ξ by $\theta = \log(\xi)$ and δ by $\beta = -\log(\delta)$. Furthermore, the unobservable (latent) nature of the person parameter is acknowledged by stating that Θ_V is a latent variable and θ_V is the unobserved realization of Θ_V and formulating the model in terms of the conditional probabilities

$$P(X_{vi}=1|\Theta_v=\theta_v)=\frac{\exp(\theta_v-\beta_i)}{1+\exp(\theta_v-\beta_i)} \quad [1.5]$$

and thus

$$P(X_{vi}=0|\Theta_v=\theta_v)=\frac{1}{1+\exp(\theta_v-\beta_i)} \quad [1.6]$$

In the above formulation, β_i is called an *item threshold parameter* or an *item location parameter.* The logit function logit(p) = log(p/(1 − p)) of the probability of a positive response is

$$\text{logit}\left(P(X_{vi}=1|\Theta_v=\theta_v)\right)=\theta_v-\beta_i \quad [1.7]$$

and therefore θ_v and β_i are often said to be on a logit scale. This terminology is not justifiable because the logit is a function of probabilities and we could argue that it is the *difference* between θ_v and β_i that is measured on a logit scale, similarly as probabilities are measured on a probability scale, but the name is popular and probably difficult to avoid. The two different representations of the model, [1.1] and [1.5], are mathematically equivalent. During statistical analysis of data by the Rasch model, it does not matter whether you use one or the other representation.

The scale on which θ is measured is often claimed to be an interval scale. This is not difficult to understand because changing the unit of the original ratio scale measure and then taking logarithms to get the value of θ after the change of the unit of ξ means changing the origin of the scale on which ξ is measured. When the unit on the multiplicative ξ scale is arbitrary, it follows that the origin on the θ scale is also arbitrary.

The symmetry of persons and items in the Rasch models and the fact that the probabilities in the Rasch models depend on the difference between person and item parameters show that items and persons are measured on the same scale. An item threshold can be interpreted as the person parameter value for which the probability of a positive response equals 0.5.

EXAMPLE 1.3.– The thresholds of the depression items are $\beta_1 = -\log(2.57) = -0.94$, $\beta_2 = -\log(1.57) = -0.45$, $\beta_3 = -\log(0.52) = 0.65$ and $\beta_4 = -\log(0.48) = 0.73$. Because the multiplicative parameters are restricted such that the product is equal to one, it follows that the sum of item thresholds is equal to zero (disregarding rounding error). Again, the risk of suffering from sleep disturbances is larger than the risk of loss of appetite, and the threshold of sleep disturbances is lower than the threshold of loss of appetite.

Finally, the assumption that the complete matrix consists of stochastically independent item responses has been replaced by the assumption that the set of item responses for a person is jointly *conditionally* independent given the variable Θ_v

$$P(\mathrm{X_v} = \mathrm{x} | \Theta_v = \theta_v) = \prod_{i=1}^{k} \frac{\exp(x_{vi}(\theta_v - \beta_i))}{1 + \exp(\theta_v - \beta_i)} \quad [1.8]$$

where $\mathrm{X_v} = (X_{v1}, \ldots, X_{vk})$ and $\mathrm{x} = (x_1, \ldots, x_k)$. Of course, responses from different persons are also considered to be independent.

The assumption of joint conditional independence means that any subset of item responses is jointly independent given Θ_v and therefore items are pairwise conditionally independent; but the reverse is not true, meaning that pairwise conditional independence does not imply joint conditional independence. We will return to this topic in section 1.9.

1.1.3. *Psychometric properties*

Viewed as a statistical model, the latent variable Θ in the model [1.5] can be characterized as a random effect explaining the covariation among items. In statistical models with random effects, we are rarely interested in the actual value of the random effect variables, and in this sense, the Rasch model is a different kind of model. The main purpose of the model is to estimate either the θ_V values or functions of the θ_V values.

On the basis of this, it is more useful to describe the Rasch model as a member of the class of statistical models known as item response theory (IRT) models [VAN 97]. Before we proceed to the discussion of the statistical features of the Rasch model, we summarize a number of requirements of IRT models that also apply to items from the Rasch model.

1.1.3.1. *Requirements of IRT modelsmodels*

Unidimensionality: The Rasch model [1.5] is a unidimensional latent trait model since Θ is a single scalar. Had Θ been a vector of variables, we would have said that the model is multidimensional.

Monotonicity: Because the probability [1.5] of a positive response to an item is a monotonously increasing function of θ, we say that the items satisfy the requirement of monotonicity.

Homogeneity: For any value of θ, the ordering of the item in terms of the probabilities is the same. Therefore, the set of items is called homogeneous. In the context of an educational test, this means that the easiest item is easiest for everybody.

Local independence: The assumption that item responses are conditionally independent given Θ is called by psychometricians the assumption of *local independence.*

Consistency: Psychometricians call a set of positively correlated items a consistent set of items. Because unidimensionality, monotonicity and local independence imply that all monotonously increasing functions of item responses – including the items in themselves – are positively correlated [HOL 86], it follows that items from Rasch models are consistent.

Absence of differential item functioning (DIF): Note, that the Rasch model only contains two types of variables: the latent variable and the items. When used, it is implicitly assumed that the model applies to all persons within a specific population (often called a specific frame of reference) and that partitioning into subpopulations does not change the model. If the frame of reference contains both men and women, it is assumed that the model [1.5] and the set of item parameters are the same for both men and women. This property is called the property of no DIF.

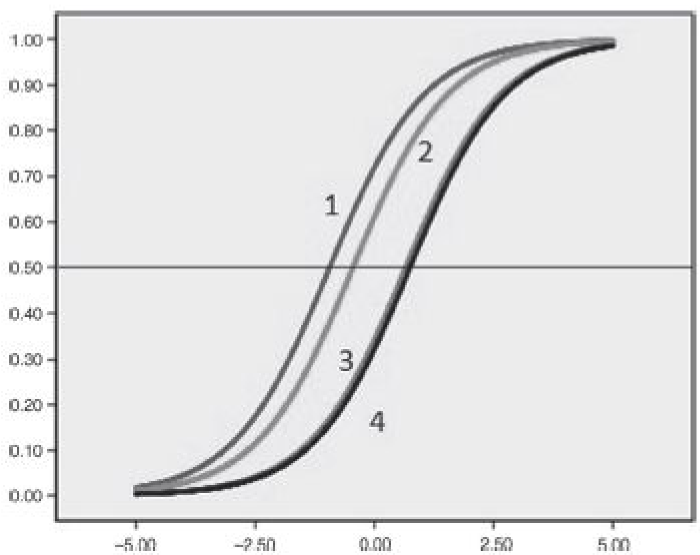
Criterion validity: The results concerning positive correlations among functions of items extend to relationships with other variables: if an exogenous variable is positively correlated with the latent variable; if items are unidimensional, monotonous and locally independent; and if there is no DIF, it follows that the exogenous variable must be positively correlated to all monotonous functions of the items, including the total score on all items. This result lies behind the psychometric notion of criterion validity.

Criterion-related construct validity: The ultimate requirement of measurement by items from IRT models is that the measurement is *construct* valid. Construct validity can be defined in several ways, for example by reference to an external *nomological* network of variables that theory insists are related to Θ [CRO 55], or by requirements of the way in which item responses depend on Θ. Rosenbaum collects all these points of views in a definition of criterion-related construct validity [ROS 89]. According to Rosenbaum, indirect measurement by a set of item responses is criterion-related construct valid if the requirements unidimensionality, monotonocity, local independence and absence of DIF are met by the items. Therefore, we claim that measurement by Rasch model items is construct valid.

1.2. Item characteristic curves

The functions $\theta \mapsto P(X_{Vi} = 1|\Theta_V = \theta)$ are called *item characteristic curves* (ICCs). Figure 1.2 shows the item characteristic curves of the four depression items under the Rasch models. In addition to being monotonous, those curves never cross. IRT models with this property are called *double monotonous* IRT models. In fact, the curves are not only double monotonous but also parallel.

Figure 1.2. *Item characteristic curves for four depression items under the Rasch model. Thresholds are −0.94 (1), −0.45 (2), 0.65 (3) and 0.74 (4)*



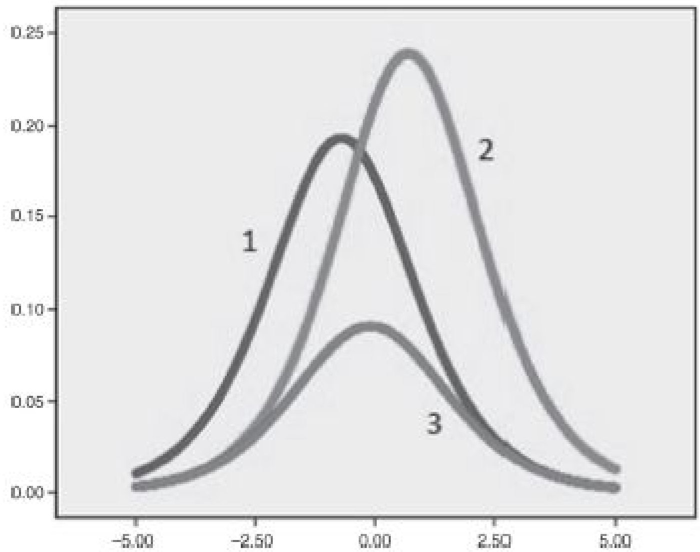
Because the items are double monotonous, the rank of the items with respect to the probabilities of positive responses to items is the same for all values of θ. At all levels of θ, the probability of a positive response to item two is smaller than the probability of a positive response to item one, but larger than the probability of a positive response to item three. Items from Rasch models are therefore homogeneous.

1.3. Guttman errors

Homogeneity is closely related to the notion of Guttman errors. Let X_{va} and X_{vb} be two item responses and assume that $\beta_a < \beta_b$. We say that a Guttman error occurs when a person has a positive response to the item with the largest threshold and a negative response to the other item, $X_{va} = 0$ and $X_{vb} = 1$. Analyses of Guttman errors play an important role in IRT models with homogeneous double monotonous items.

In Rasch models, the risk of Guttman errors depends on both item and person parameters. The closer the thresholds of the two items *a* and *b*, the larger the risk of a Guttman error. The larger the numerical value of the person parameter, the smaller the risk of a Guttman error. The risk of Guttman errors for pairs of depression items across the level of depression is shown in Figure 1.3.

Figure 1.3. *The risk of Guttman errors among responses to items 1 and 2 (1), items 3 and 4 (2) and items 1 and 4 (3)*



1.4. Test characteristic curve

From the score probabilities, it is easy to calculate the expected (mean) score of R for different values of ξ. These are called true scores and the function describing the true score as a function of $\theta = \log(\xi)$ is called the test characteristic curve (TCC). The TCC for the four depression items is shown in Figure 1.4. Note that the TCC is not linear.

1.5. Implicit assumptions

In statistical terms, the requirement of no DIF can best be described as the requirement that item responses (X_{v1}, ..., X_{vk}) are conditionally independent of all exogenous variables (Z_{v1}, ..., Z_{vk}) given Θ. The absence of DIF is a fundamental validity assumption in psychometrics. The Rasch model shares this assumption, but only as an implicit assumption, because the exogenous variables do not enter this model at all. Similarly, the Rasch model also shares the