

APPLIED MATHEMATICS SERIES



Rasch Models in Health

**Edited by Karl Bang Christensen
Svend Kreiner, Mounir Mesbah**

ISTE

 **WILEY**

Rasch Models in Health

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.

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms and licenses issued by the CLA. Enquiries concerning reproduction outside these terms should be sent to the publishers at the undermentioned address:

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

© ISTE Ltd 2013

The rights of Karl Bang Christensen, Svend Kreiner and Mounir Mesbah to be identified as the author of this work have been asserted by them in accordance with the Copyright, Designs and Patents Act 1988.

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



Printed and bound in Great Britain by CPI Group (UK) Ltd., Croydon, Surrey CR0 4YY

Table of Contents

Preface	xv
Karl Bang CHRISTENSEN, Svend KREINER and Mounir MESBAH	
PART 1. PROBABILISTIC MODELS	1
Chapter 1. The Rasch Model for Dichotomous Items	5
Svend KREINER	
1.1. Introduction	5
1.1.1. Original formulation of the model	5
1.1.2. Modern formulations of the model	9
1.1.3. Psychometric properties	10
1.1.3.1. Requirements of IRT models	11
1.2. Item characteristic curves	12
1.3. Guttman errors	12
1.4. Test characteristic curve	13
1.5. Implicit assumptions	13
1.6. Statistical properties	14
1.6.1. The distribution of the total score	15
1.6.2. Symmetrical polynomials	16
1.6.3. Partial credit model parameterization of the score distribution	17
1.6.4. Rasch models for subscores	17
1.7. Inference frames	18
1.8. Specific objectivity	20
1.9. Rasch models as graphical models	21
1.10. Summary	22
1.11. Bibliography	24

Chapter 2. Rasch Models for Ordered Polytomous Items	27
Mounir MESBAH and Svend KREINER	
2.1. Introduction	27
2.1.1. Example	27
2.1.2. Ordered categories	28
2.1.3. Properties of the polytomous Rasch model	31
2.1.4. Assumptions	33
2.2. Derivation from the dichotomous model	33
2.3. Distributions derived from Rasch models	37
2.3.1. The score distribution	39
2.3.2. Conditional distribution of item responses given the total score	40
2.4. Bibliography	41
PART 2. INFERENCE IN THE RASCH MODEL	43
Chapter 3. Estimation of Item Parameters	49
Karl Bang CHRISTENSEN	
3.1. Introduction	49
3.2. Estimation of item parameters	51
3.2.1. Estimation using the conditional likelihood function	52
3.2.2. Pairwise conditional estimation	54
3.2.3. Marginal likelihood function	56
3.2.4. Extended likelihood function	57
3.2.5. Reduced rank parameterization	58
3.2.6. Parameter estimation in more general Rasch models	59
3.3. Example	59
3.4. Bibliography	60
Chapter 4. Person Parameter Estimation and Measurement in Rasch Models	63
Svend KREINER and Karl Bang CHRISTENSEN	
4.1. Introduction and notation	63
4.2. Maximum likelihood estimation of person parameters	65
4.3. Item and test information functions	66
4.4. Weighted likelihood estimation of person parameters	67
4.5. Example	67
4.6. Measurement quality	70
4.6.1. Reliability in classical test theory	70
4.6.2. Reliability in Rasch models	71
4.6.3. Expected measurement precision	73
4.6.4. Targeting	74
4.7. Bibliography	76

PART 3. CHECKING THE RASCH MODEL	79
Chapter 5. Item Fit Statistics	83
Karl Bang CHRISTENSEN and Svend KREINER	
5.1. Introduction	83
5.2. Rasch model residuals	84
5.2.1. Notation	84
5.2.2. Individual response residuals: outfits and infits	86
5.2.3. Problem 1: the distribution of outfit and infit test statistics	87
5.2.4. Problem 2: calculating E_{vi}	88
5.2.5. Group residuals	90
5.2.6. Group residuals for analysis of homogeneity	91
5.3. Molenaar's U	93
5.4. Analysis of item-restscore association	94
5.5. Group residuals and analysis of DIF	96
5.6. Kelderman's conditional likelihood ratio test of no DIF	96
5.7. Test for conditional independence in three-way tables	98
5.8. Discussion and recommendations	100
5.8.1. Technical issues	100
5.8.2. What to do when items do not agree with the Rasch model	101
5.9. Bibliography	102
Chapter 6. Overall Tests of the Rasch Model	105
Svend KREINER and Karl Bang CHRISTENSEN	
6.1. Introduction	105
6.2. The conditional likelihood ratio test	105
6.3. Other overall tests of fit	109
6.4. Bibliography	109
Chapter 7. Local Dependence	111
Ida MARAIS	
7.1. Introduction	111
7.1.1. Reduced rank parameterization model for subtests	112
7.1.2. Reliability indices	112
7.2. Local dependence in Rasch models	113
7.2.1. Response dependence	113
7.3. Effects of response dependence on measurement	114
7.4. Diagnosing and detecting response dependence	118
7.4.1. Item fit	118
7.4.2. Item residual correlations	120

7.4.3. Subtests and reliability	122
7.4.4. Estimating the magnitude of response dependence	122
7.4.5. Illustration	122
7.5. Summary	127
7.6. Bibliography	128
Chapter 8. Two Tests of Local Independence	131
Svend KREINER and Karl Bang CHRISTENSEN	
8.1. Introduction	131
8.2. Kelderman's conditional likelihood ratio test of local independence	132
8.3. Simple conditional independence tests	133
8.4. Discussion and recommendations	135
8.5. Bibliography	136
Chapter 9. Dimensionality	137
Mike HORTON, Ida MARAIS and Karl Bang CHRISTENSEN	
9.1. Introduction	137
9.1.1. Background	138
9.1.2. Multidimensionality in health outcome scales	139
9.1.3. Consequences of multidimensionality	140
9.1.4. Motivating example: the HADS data	140
9.2. Multidimensional models	141
9.2.1. Marginal likelihood function	142
9.2.2. Conditional likelihood function	142
9.3. Diagnostics for detection of multidimensionality	142
9.3.1. Analysis of residuals	143
9.3.2. Observed and expected counts	143
9.3.3. Observed and expected correlations	145
9.3.4. The t-test approach	146
9.3.5. Using reliability estimates as diagnostics of multidimensionality	147
9.4. Tests of unidimensionality	149
9.4.1. Tests based on diagnostics	149
9.4.2. Likelihood tests	149
9.5. Estimating the magnitude of multidimensionality	152
9.6. Implementation	152
9.7. Summary	152
9.8. Bibliography	154

PART 4. APPLYING THE RASCH MODEL	159
Chapter 10. The Polytomous Rasch Model and the Equating of Two Instruments	163
David ANDRICH	
10.1. Introduction	163
10.2. The Polytomous Rasch Model	165
10.2.1. Conditional probabilities	165
10.2.2. Conditional estimates of the instrument parameters	167
10.2.3. An illustrative small example	169
10.3. Reparameterization of the thresholds	171
10.3.1. Thresholds reparameterized to two parameters for each instrument	171
10.3.2. Thresholds reparameterized with more than two parameters	175
10.3.3. A reparameterization with four parameters	175
10.3.3.1. A solution algorithm	176
10.3.3.2. Leunbach's precedent	176
10.4. Tests of fit	177
10.4.1. The conditional test of fit based on cell frequencies	177
10.4.1.1. Degrees of freedom for the conditional test of fit based on cell frequencies	178
10.4.2. The conditional test of fit based on class intervals	178
10.4.2.1. Degrees of freedom for the conditional test of fit based on class intervals	179
10.4.3. Graphical test of fit based on total scores	180
10.4.4. Graphical test of fit based on person estimates	180
10.5. Equating procedures	181
10.5.1. Equating using conditioning on total scores	181
10.5.2. Equating through person estimates	181
10.6. Example	182
10.6.1. Person threshold distribution	183
10.6.2. The test of fit between the data and the model	183
10.6.2.1. Conditional χ^2 test of fit based on cells of the data matrix and four moments estimated	183
10.6.2.2. Conditional χ^2 test of fit based on class intervals of the data matrix and four moments estimated	184
10.6.2.3. Conditional χ^2 test of fit based on cells of the data matrix and two moments estimated	185
10.6.2.4. Conditional χ^2 test of fit based on class intervals of the data matrix and two moments estimated	185
10.6.3. Further analysis with the parameterization with two moments for each instrument	186
10.6.3.1. Parameter estimates from two moments	186

10.6.3.2. Score characteristic curves	186
10.6.3.3. Observed and expected frequencies in class intervals	186
10.6.3.4. Graphical test of fit based on conditioning on total scores	186
10.6.3.5. Graphical test of fit based on person estimates	187
10.6.4. Equated scores based on the parameterization with two moments of the thresholds	188
10.6.4.1. Equated scores conditional on the total score	189
10.6.4.2. Equated scores given the person estimate	190
10.7. Discussion	193
10.8. Bibliography	195

Chapter 11. A Multidimensional Latent Class Rasch Model for the Assessment of the Health-Related Quality of Life 197
Silvia BACCI and Francesco BARTOLUCCI

11.1. Introduction	197
11.2. The data set	200
11.3. The multidimensional latent class Rasch model	202
11.3.1. Model assumptions	202
11.3.2. Maximum likelihood estimation and model selection	205
11.3.3. Software details	207
11.3.4. Concluding remarks about the model	208
11.4. Correlation between latent traits	209
11.5. Application results	212
11.6. Acknowledgments	215
11.7. Bibliography	216

Chapter 12. Analysis of Rater Agreement by Rasch and IRT Models 219
Jørgen Holm PETERSEN

12.1. Introduction	219
12.2. An IRT model for modeling inter-rater agreement	220
12.3. Umbilical artery Doppler velocimetry and perinatal mortality	221
12.4. Quantifying the rater agreement in the Rasch model	222
12.4.1. Fixed-effects approach	222
12.4.2. Random Effects approach and the median odds ratio	225
12.5. Doppler velocimetry and perinatal mortality	227
12.6. Quantifying the rater agreement in the IRT model	229
12.7. Discussion	231
12.8. Bibliography	232

Chapter 13. From Measurement to Analysis	235
Mounir MESBAH	
13.1. Introduction	235
13.2. Likelihood	237
13.2.1. Two-step model	238
13.2.2. Latent regression model	238
13.3. First step: measurement models	238
13.4. Statistical validation of measurement instrument	241
13.5. Construction of scores	245
13.6. Two-step method to analyze change between groups	246
13.6.1. Health-related quality of life and housing in europe	246
13.6.2. Use of surrogate in an clinical oncology trial	248
13.7. Latent regression to analyze change between groups	250
13.8. Conclusion	253
13.9. Bibliography	254
Chapter 14. Analysis with Repeatedly Measured Binary Item Response Data by <i>Ad Hoc</i> Rasch Scales	257
Volkert SIERSMA and Paolo EUSEBI	
14.1. Introduction	257
14.2. The generalized multilevel Rasch model	260
14.2.1. The multilevel form of the conventional Rasch model for binary items	260
14.2.2. Group comparison and repeated measurement	262
14.2.3. Differential item functioning and local dependence	263
14.3. The analysis of an <i>ad hoc</i> scale	264
14.4. Simulation study	268
14.5. Discussion	272
14.6. Bibliography	275
PART 5. CREATING, TRANSLATING AND IMPROVING RASCH SCALES	277
Chapter 15. Writing Health-Related Items for Rasch Models – Patient-Reported Outcome Scales for Health Sciences: From Medical Paternalism to Patient Autonomy	281
John BRODERSEN, Lynda C. DOWARD, Hanne THORSEN and Stephen P. MCKENNA	
15.1. Introduction	281
15.1.1. The emergence of the biopsychosocial model of illness	282
15.1.2. Changes in the consultation process in general medicine	283
15.2. The use of patient-reported outcome questionnaires	284
15.2.1. Defining PRO constructs	285

15.2.1.1. Measures of impairment, activity limitations and participation restrictions	285
15.2.1.2. Health status/health-related quality of life	287
15.2.1.3. Generic and specific questionnaires	288
15.2.2. Quality requirements for PRO questionnaires	290
15.2.2.1. Instrument development standards	290
15.2.2.2. Psychometric and scaling standards	291
15.3. Writing new health-related items for new PRO scales	294
15.3.1. Consideration of measurement issues	294
15.3.2. Questionnaire development	294
15.4. Selecting PROs for a clinical setting	297
15.5. Conclusions	297
15.6. Bibliography	298

Chapter 16. Adapting Patient-Reported Outcome Measures for Use in New Languages and Cultures 303

Stephen P. MCKENNA, Jeanette WILBURN, Hanne THORSEN and John BRODERSEN

16.1. Introduction	303
16.1.1. Background	303
16.1.2. Aim of the adaptation process	304
16.2. Suitability for adaptation	305
16.3. Translation process	305
16.3.1. Linguistic issues	305
16.3.2. Conceptual issues	306
16.3.3. Technical issues	306
16.4. Translation methodology	306
16.4.1. Forward–backward translation	307
16.4.1.1. Situation 1: The forward translation is good	307
16.4.1.2. Situation 2: The forward translation is good, but the back translation is poor	308
16.4.1.3. Situation 3: The forward translation is poor	308
16.5. Dual-panel translation	308
16.5.1. Bilingual panel	308
16.5.2. Lay panel	309
16.6. Assessment of psychometric and scaling properties	310
16.6.1. Cognitive debriefing interviews	310
16.6.1.1. Interview setting	311
16.6.1.2. Materials	311
16.6.1.3. Reporting on the interviews	311
16.6.2. Determining the psychometric properties of the new language version of the measure	312
16.6.3. Practice guidelines	313
16.7. Bibliography	315

Chapter 17. Improving Items That Do Not Fit the Rasch Model	317
Tine NIELSEN and Svend KREINER	
17.1. Introduction	317
17.2. The RM and the graphical log-linear RM	318
17.3. The scale improvement strategy	320
17.3.1. Choice of modification action	322
17.3.2. Result of applying the scale improvement strategy	325
17.4. Application of the strategy to the Physical Functioning Scale of the SF-36	326
17.4.1. Results of the GLLRM	326
17.4.2. Results of the subject matter analysis	327
17.4.3. Suggestions according to the strategy	328
17.5. Closing remark	331
17.6. Bibliography	331
PART 6. ANALYZING AND REPORTING RASCH MODELS	335
Chapter 18. Software for Rasch Analysis	337
Mounir MESBAH	
18.1. Introduction	337
18.2. Stand alone softwares packages	338
18.2.1. WINSTEPS	338
18.2.2. RUMM	338
18.2.3. CONQUEST	338
18.2.4. DIGRAM	339
18.3. Implementations in standard software	339
18.3.1. SAS macro for MML estimation	339
18.3.2. SAS macros based on CML estimation	340
18.3.3. eRm: an R Package	340
18.4. Fitting the Rasch model in SAS	340
18.4.1. Simulation of Rasch dichotomous items	340
18.4.2. MML estimation using PROC NLMIXED	341
18.4.3. MML estimation of using PROC GLIMMIX	342
18.4.4. JML estimation using PROC LOGISTIC	342
18.4.5. CML estimation using PROC GENMOD	343
18.4.6. JML estimation using PROC LOGISTIC	343
18.4.7. Results	344
18.5. Bibliography	344
Chapter 19. Reporting a Rasch Analysis	347
Thomas SALZBERGER	
19.1. Introduction	347
19.1.1. Objectives	347

19.1.2. Factors impacting a Rasch analysis report	348
19.1.3. The role of the substantive theory of the latent variable	349
19.1.4. The frame of reference	350
19.2. Suggested elements	350
19.2.1. Construct: definition and operationalization of the latent variable	351
19.2.2. Response format and scoring	351
19.2.3. Sample and sampling design	352
19.2.4. Data	353
19.2.5. Measurement model and technical aspects	353
19.2.6. Fit analysis	354
19.2.7. Response scale suitability	355
19.2.8. Item fit assessment	355
19.2.9. Person fit assessment	356
19.2.10. Information	357
19.2.11. Validated scale	357
19.2.12. Application and usefulness	358
19.2.13. Further issues	359
19.3. Bibliography	360
List of Authors	363
Index	365

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

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

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

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

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*.

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	$\frac{1}{1 + \xi_v \delta_2}$	$\frac{\xi_v \delta_2}{1 + \xi_v \delta_2}$	$\xi_v \delta_2$
				$OR = \frac{\delta_2}{\delta_1}$

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

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.

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	$\frac{1}{1 + \xi_2 \delta_i}$	$\frac{\xi_2 \delta_i}{1 + \xi_2 \delta_i}$	$\xi_2 \delta_i$
				$OR = \frac{\xi_2}{\xi_1}$

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

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

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

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

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 $\text{logit}(p) = \log(p/(1-p))$ of the probability of a positive response is

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

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

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

where $\mathbf{X}_v = (X_{v1}, \dots, X_{vk})$ and $\mathbf{x} = (x_1, \dots, x_k)$. 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 models

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, monotonicity, 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.

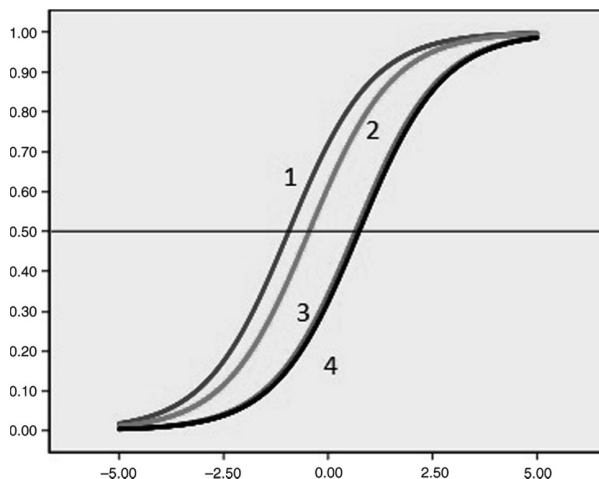


Figure 1.2. Item characteristic curves for four depression items under the Rasch model. Thresholds are -0.94 (1), -0.45 (2), 0.65 (3) and 0.74 (4)

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