21

Validity and Objectivity in Health-Related Scales: Analysis by Graphical Loglinear Rasch Models

Svend Kreiner[1] and Karl Bang Christensen[2]

[1] Department of Biostatistics, University of Copenhagen
[2] Institute of Occupational Health

21.1 Introduction

The total score of items fitting a Rasch model (RM) satisfies assumptions relating to validity and a number of technical requirements. For this reason, the RM is often used as a "gold standard" expressing ideal measurement requirements.

Most summated rating scales in health research that we have worked with have shown evidence of differential item functioning (DIF) and local dependence (LD), thus violating the assumptions of the RM, even though items appear to be face valid. In this situation, Rasch analysis is a destructive process: a large number of face valid items are rejected in order to obtain fit to the model. This can seem unacceptable when items are face valid. Data from a large health survey in Copenhagen County in 1995 is used for illustration focusing on responses to items measuring physical functioning in the SF-36 questionnaire (Ware Jr. et al., 1993).

This chapter views Rasch analysis as an examination of the items given the requirements of ideal measurements, yielding a summary of problems and an evaluation of their relevance. Graphical loglinear RMs (GLLRM) incorporating uniform DIF and uniform LD (Kreiner & Christensen, 2002, 2004) are used for this. This leads to reflection on measurement requirements: We suggest that DIF is more serious than LD, and that sufficiency and reliability are more important than specific objectivity. Items fitting a GLLRM provide measurement that is essentially valid and objective and the total score is sufficient.

Section 21.2 describes the PF subscale of the SF-36. Section 21.3 describes conditional-independence and chain-graph models and their global Markov properties. Section 21.4 introduces Criterion-related construct validity. Section 21.5 defines graphical RMs (Kreiner & Christensen, 2002) describing the latent-trait variable, the set of items, and the exogenous variables in the framework of graphical models (Lauritzen, 1996). Section 21.6 extends these

by allowing uniform DIF and uniform LD in the well-known manner of loglinear RMs (Kelderman, 1984, 1992, 1995). This yields graphical loglinear RMs (GLLRM). Section 21.7 presents the analysis of the SF-36 data and the measurement implications of the departures from the RM. Section 21.8 discusses essential objectivity and validity, and section 21.9 presents a summary and discussion.

Items are denoted by $Y=(Y_1,\ldots,Y_k)$, the total score by $S = \sum_i Y_i$, the latent variable by Θ, and exogenous variables by $X=(X_1,\ldots,X_m)$. We assume, without loss of generality, that all items have c+1 ordinal categories coded 0,1, ... , c. Exogenous variables may include response variables depending on Θ, criterion variables known to be monotonously related to Θ, covariates with a potential effect on Θ or simply variables that may be associated with Θ and/or items.

21.2 The Physical Functioning SubScale of the SF-36

The SF-36 (Ware Jr. et al., 1993) is a widely used questionnaire measuring aspects of general health status. It contains 36 items summarized into eight subscales. The physical functioning (PF) subscale summarizes responses to ten items under the common heading "Does your health now limit you in these activities? If so, how much?"

- Vigorous activities, e.g., running, heavy lifting, strenuous sport (PF1)
- Moderate activities (PF2)
- Lifting or carrying groceries (PF3)
- Climbing several fligths of stairs (PF4)
- Climbing one flight of stairs (PF5)
- Bending, kneeling, or stooping (PF6)
- Walking more than a mile (PF7)
- Walking several blocks (PF8)
- Walking one block (PF9)
- Bathing or dressing yourself (PF10)

Three ordinal response categories ("Not limited," "Limited a little," "Limited a lot") are used. The developers claim that "Studies to date have yielded content, concurrent, criterion, construct, and predictive evidence of validity" (Ware, J.E. (undated): SF-36® Health Survey Update. http://www.sf-36.org/tools/sf36.shtml). Scrutinizing the items will show that LD between PF4 and PF5 and between PF7, PF8, and PF9 must be expected if responses are rational and consistent. Whether the reported analyses of construct validity may have overlooked this is not the focus of the present chapter. Problems of this kind are not unusual in health scales, often while items are highly face valid.

21.3 Conditional-Independence and Chain-Graph Models

Conditional independence is the unifying concept of importance for item response models and chain-graph models. We write $X \perp Y|Z$ to indicate that two sets of variables, $X = (X_1,\ldots,X_a)$ and $Y = (Y_1,\ldots,Y_b)$, are conditionally independent given a third set, $Z = (Z_1,\ldots,Z_c)$, in the sense that $P(X|Y,Z) = P(X|Z)$. Chain-graph models are multidimensional block recursive statistical models defined by pairwise conditional independence of variables in the following way. Let $V = \bigcup_i V_i$ be a partitioning of the variables into ordered subsets, $V_1 \leftarrow \cdots \leftarrow V_r$ defining a block recursive statistical model $P(V) = \prod_i P(V_i|V_{i+1},\ldots,V_r)$. Assume that X and Y are variables belonging to block numbers a and b, respectively, where $a \leq b$. Set $Z_{rest}(X,Y) = \bigcup_{i=a}^{r} V_i \backslash \{X,Y\}$ such that $Z_{rest}(X,Y)$ contains all variables that are concurrent or prior to X according to the recursive structure of the model. A chain-graph model is defined by a set of assumptions concerning pairwise conditional independence, $\{X_i \perp Y_i | Z_{rest}(X_i,Y_i) : i = 1,\ldots,m\}$.

Graphical models are characterized by Markov independence graphs: networks where variables are represented by nodes. Nodes are disconnected if the variables are conditionally independent given all concurrent or prior variables Variables in the same recursive block are connected by undirected edges, whereas variables in different blocks are connected by arrows representing temporal and/or causal direction. The Markov graphs of graphical models are used both as visual diagrams illustrating the structure of the statistical model and as mathematical models—mathematical graphs—where mathematical graph theory may reveal properties of the statistical model that may be helpful both during the analysis of data and for interpretation of what the model conveys about the distribution of the variables. Examples of Markov graphs are shown in Figures 21.1 to 21.4 below. A comprehensive introduction to the theory of graphical models and the way the properties of the Markov graphs correspond to properties of the statistical model may be found in Lauritzen (1996).

21.3.1 Global Markov Properties of Chain-Graph Models

The global Markov properties of chain-graph models are of particular interest here. The global Markov properties tells us that conditional independence between two variables, X and Y, in a chain-graph model sometimes applies under conditioning with respect to subsets of $Z_{rest}(X,Y)$. To find such subsets, we have to examine the moral graph defined by replacing arrows by (undirected) edges and linking "parents" (see Figure 21.4).

The global Markov properties are linked to the concept of separation in undirected graphs. To subsets, A and B, of nodes in an undirected graph are separated by a subset of nodes, S, if every path from a node in A to a node

in B contains at least on node in S. The global Markov property of chain-graph models (Lauritzen, 1996, p. 55) implies that two set of variables, A and B, in a chain-graph model are conditionally independent given any subset of variables, S, that separates A and B in the moral graph.

21.4 Criterion-Related Construct Validity

Criterion-related construct validity requires unidimensionality, monotonicity, local independence, and the absence of DIF (Rosenbaum, 1989). The last assumption requires the relation between the latent trait and the items to be the same in any subpopulation and implies criterion validity, which thus is a necessary, but not sufficient condition for construct validity. These assumptions also define nonparametric item response models (Sijtsma & Molenaar, 2002).

The requirement of no DIF in this definition is somewhat vague. We assume that it refers to meaningful and relevant partitions of the persons defined by an exogenous variable, but notice that in most studies a limited number of such variables will be available. Absence of DIF can be stated as the requirement, $\mathrm{Y} \perp \mathrm{X} \mid \Theta$, of conditional independence and because local independence implies pairwise conditional independence criterion-related construct validity defines a chain-graph model.

21.5 Graphical Rasch Models

The RM for ordinal items (Andersen, 1977; Andrich, 1978; Masters, 1982)

$$P(Y_i = y|\Theta = \theta) = \exp(\alpha_{i0} + \theta y + \alpha_{iy}) \tag{21.1}$$

where $\alpha_{i0} = -\ln\left(\sum_{y=0}^{c} \exp(\theta y + \alpha_{iy})\right)$ satisfies the first three requirements of criterion related construct validity. The joint conditional distribution

$$P(Y_1 = y_1, \ldots, Y_k = y_k|\Theta = \theta) = \exp\left(\alpha_0 + \sum_{i=1}^{k} (\theta y_i + \alpha_{iy_i})\right) \tag{21.2}$$

is a loglinear model for a multivariate contingency table with main effects depending on the latent variable and no interaction parameters. Restrictions are needed for parameters to be identifiable. These are imposed by setting $\alpha_{i0} = 0$ for all items and $\sum_i \alpha_{ic} = 0$.

Different data generating processes may lead to this model. Reparameterization replacing item parameters with thresholds,$\tau_{ij} = \alpha_{i(j-1)} - \alpha_{ij}$, yielding a partial credit interpretation (Masters, 1982) where $P(Y_i = y|\Theta = \theta) =$

$\exp(\sum_{j=1}^{y}(\theta - \tau_{ij})\ /\ \Gamma_i$ can be useful, even though it may not be a valid description of the response behavior to the type of questions included in SF-36.

The total score, $S = \sum_i Y_i$, is sufficient for θ in the conditional distribution of items given $\Theta = \theta$ This implies Bayesian sufficiency, (Kolmogoroff, 1942; Arnold, 1988) and conditional independence of items and Θ given S.

The distribution of S is given by

$$P(S = s|\Theta = \theta) = \frac{\exp(\theta s + \varphi_s)}{\Phi} \tag{21.3}$$

where $\gamma_s = \exp(\varphi_s)$ are referred to as elementary symmetric functions (Andersen, 1973, Fischer, 1974; 1995). We refer to the φ-parameters in (21.3) as score parameters. The probabilities (21.3) can be expressed in terms of threshold parameters in the same way as (21.1).

The RM satisfies construct validity requirements and provides objective measurement by sufficient raw scores. DIF and criterion validity can not be addressed in formal terms within the framework of RMs, but in a larger framework including exogenous variables. One way to do this is to assume that the joint distribution of $(Y_1, \ldots, Y_k, \Theta, X_1, \ldots, X_m)$ is a graphical RM.

A graphical RM is a chain-graph model characterized by two Markov graphs (Figure 21.1): (1) an IRT graph expressing construct validity (items are conditionally independent of each other and of exogenous variables), and (2) A Rasch graph adding the score S separating items from Θ. Note that edges between items are added because items are not conditionally independent given the score. The only requirement of construct validity that is not an explicit part of the IRT graph is that the relationship between the latent variable and items must be monotonous. The IRT graph also describes relationships among exogenous variables.

It follows from the Markov properties of the IRT and Rasch graphs that the distribution (21.2) reappears as the conditional distribution of item responses given Θ and X,

$$P(Y_1 = y_1, \ldots, Y_k = y_k|\Theta\theta, X_1 = x_1, \ldots, X_m = x_m) = \tag{21.4}$$

$$\exp\left(\alpha_0 + \sum_{i=1}^{k}(\theta y_i + \alpha_{iy_i})\right) \tag{21.5}$$

Marginalizing over Θ in the Rasch graphs results in a *marginal Rasch graph* (not shown) defining a chain-graph model for the manifest variables (Whittaker, 1990, p. 395). The marginal Rasch graph contains edges or arrows between any pair of variables connected to Θ by an arrow originating from Θ.

The IRT and Rasch graphs (Figure 21.1) define the model and provide a visual display of the model structure. The moralized Rasch graph (Figure 21.2) provides information on conditional independencies among the manifest variables of the model. The moral Rasch graph is an undirected graph defined by the marginal Rasch graph where separation implies conditional

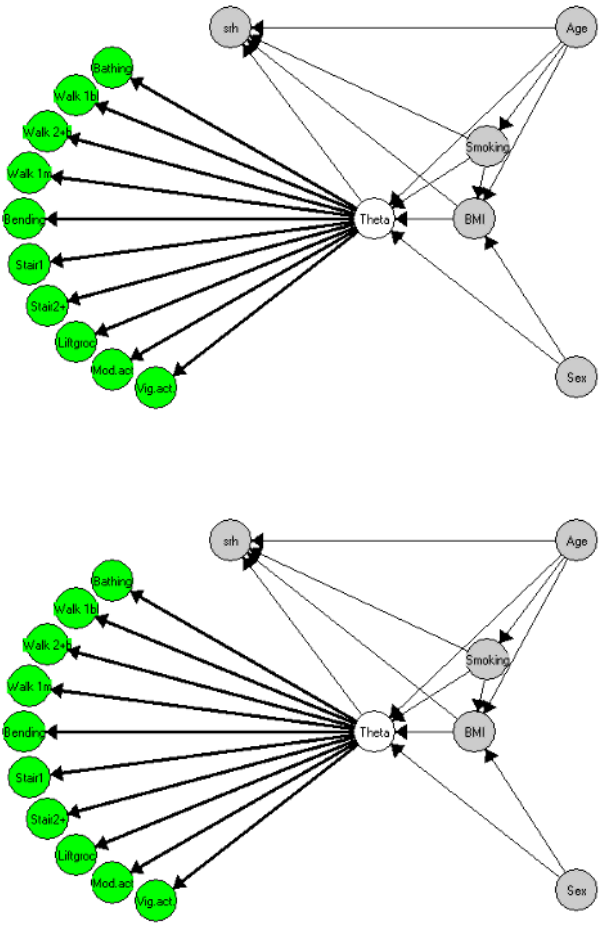

Fig. 21.1. The IRT and Rasch graphs defining the graphical RM for the ten PF items. The IRT describes relationships among exogenous variables: sex and age are marginally independent, smoking and sex are conditionally independent given age, SRH and sex are conditionally independent given Θ, BMI, smoking, and age.

independence due to the global Markov properties of chain-graph models. It follows from this that all pairs of items and exogenous variables are conditionally independent given S. This result lies behind the Mantel–Haenszel test for DIF (Holland & Thayer, 1988) and the global Markov properties of the Rasch graph shows that the result applies to all types of items and exogenous variables in graphical RMs.

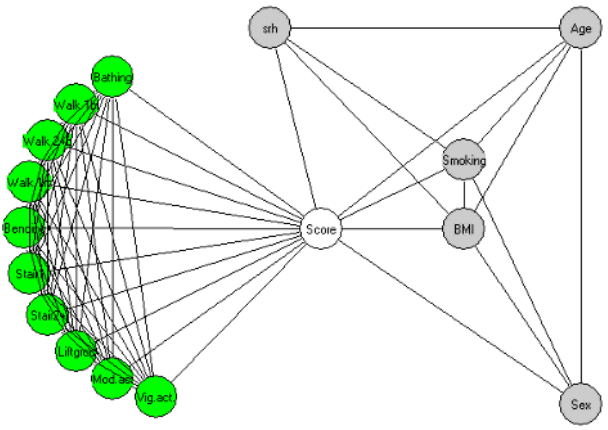

Fig. 21.2. The moral Rasch graph of the graphical model for the ten PF items

DIF is absent in graphical RMs in two ways: items and exogenous variables are conditionally independent given Θ and given S. This property appears to be unique to the RMs. We refer to Kreiner & Christensen (2002, 2004) for further discussions of properties of graphical RMs derived from the global Markov properties of Rasch graphs.

21.5.1 Inference in Graphical Rasch Models

Graphical RMs address two problems: (1) the quality of measurement (regarded as optimal if item responses fit the graphical RM) and (2) latent regression analysis

$$\theta = X_1\beta_1 + \cdots + X_m\beta_m + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

describing the association between the latent variable and covariates $X_1, \ldots, X_m$. Since the pioneering work of Andersen & Madsen (1977), models of this kind have been studied extensively (Zwinderman, 1991, 1997; Andersen, 1994; Hoijtink, 1995; Kamata, 2001; Maier, 2001 Christensen et al, 2004; De Boeck & Wilson, 2004; Adams & Wu, this volume).

In this chapter, item analysis is separated from latent regression and we are thus able to distinguish between lack of fit of the measurement model and misspecified latent structure (Zwinderman & van den Wollenberg, 1990; Christensen et al., 2004). This means conditional item analysis is used, because marginal inference relies on assumptions about the distribution of the latent variable.

Conditional inference in graphical RMs may be carried out in two ways. The first is a parametric approach fitting the conditional distribution of item responses given the total score, comparing item parameters in different subpopulations and calculating item-fit statistics. The presence of exogenous variables in the graphical RM defines explicit requirements of groups to be compared during the analysis. The second approach is nonparametric, testing the assumptions expressed by the moral Rasch graph. Mantel–Haenszel tests (Holland & Thayer, 1988) can be used for testing conditional independence for pairs of dichotomous items and dichotomous exogenous variables. Partial gamma coefficients (Agresti, 1984, p. 171) may be used when items or exogenous variables are ordinal. The tests of conditional independence will often be tests in large sparse tables and and Monte Carlo tests (Kreiner, 1987; von Davier, 1997) can be used to avoid the problem of inadequate approximation of p-values by conventional asymptotic methods.

The RM applies for any subset of items and therefore LD between an item, Y_i, and the other items can be tested as conditional independence given the rest score, $R_i = S - Y_i$ (Kreiner & Christensen, 2004). This test is one example of a less than conventional approach suggested by the graphical structure of these models.

A starting point for the latent structure analysis can be obtained by nonparametric analysis of manifest variables based on the moral graphs. For the SF-36, analysis of the effect of a covariate on physical disability may be performed as a test of conditional independence in a multi-way table containing these two variables together with the variables separating the two in the moral graph. If conditional independence is rejected the covariate should be included in the latent-regression model.

21.6 Graphical Loglinear Rasch Models

A graphical loglinear RM (GLLRM) adds interaction parameters to the conditional distribution of item responses (21.4): DIF parameters describing interaction between an item and an exogenous variable and LD parameters describe interactions between two items. It is convenient to distinguish between second-order DIF and LD parameters and general higher order interaction parameters. The only restriction imposed on the interaction parameters in GLLRMs is that they must not depend on the latent-trait variable. To simplify the discussion of validity and objectivity in GLLRMs, we first consider models with DIF parameters and present three ways to look at these models. Following this, we then consider models with LD parameters and finally the general family of GLLRM.

21.6.1 Uniform DIF

The model (21.6) adds interaction between items Y_a and X_b

$$P(Y_1 = y_1, \ldots, Y_k = y_k | \Theta = \theta, X_1 = x_1, \ldots, X_m = x_m) = \exp\left(\alpha_0 + \sum_{i=1}^{k} ((\theta y_i + \alpha_{iy_i}) + \delta_{ab}(y_a, x_b)\right) \quad (21.6)$$

For the parameters to be identifiable, we must impose additional restrictions on the δ-parameters in addition to those already imposed on the item main effect parameters. One convenient way to do so is to assume that δ_{ab}(0,x) = 0 and δ_{ab}(y,1) = 0 where we assume that categories of exogenous variables are integer coded from 1 to the number of categories of the variables. We regard the model defined by (21.6) as a model describing uniform DIF, where item parameter of Y_a in the subpopulation given by X_a = x is equal to α_a(y)+δ_{ab}(y,x). Alternatively, Y_b can be interpreted as a set of "virtual" items given only in a subpopulation (Tennant et al., 2004). Finally, of course, (21.6) is an example of a mixed RM. The mixture is manifest, but apart from that, the model satisfies all assumptions underlying the mixed RM.

21.6.2 Uniform LD

Adding interaction between two items, Y_a and Y_b, to (21.4) leads to a model with LD between the items:

$$P(Y_1 = y_1, \ldots, Y_k = y_k | \Theta = \theta, X_1 = x_1, \ldots, X_m = x_m) = \exp\left(\alpha_0 + \sum_{i=1}^{k} ((\theta y_i + \alpha_{iy_i}) + \lambda_{ab}(y_a, y_b)\right) \quad (21.7)$$

We once again assume that the interaction parameter do not depend on θ and set λ_{ab}(0,y) = λ_{ab}(y,0) = 0. If we remove Y_b from the score and treat it as an exogenous variable it follows from (21.7) that the conditional distribution of the items remaining in the rest score follows a loglinear RMs similar to (21.6) with uniform DIF of Y_a relative to Y_b. We have therefore coined the term uniform LD to cover the kind of local dependence implied by the interaction parameter in (21.7).

21.6.3 Graphical Loglinear Rasch Models

Expanding model (21.6) and (21.7) to models with several cases of uniform DIF and LD as well as higher order interactions terms is straightforward. A general GLLRM is defined by three types of loglinear generators. First, DIF generators, D = (D_1,...,D_r), where D_i= (A_i,Z_i) with $A_i \in$\{Y_1,...,Y_k\} and $Z_i \in$\{X_1,...,X_m\}. Second, LD generators, L = (L_1,...,L_s) consisting of pairs of items L_i = (U_i,V_i) where \{U_i,V_i\}$\subset$ \{Y_1,...,Y_k\}. Finally higher order interactions, G = (G_1,...,G_s), where each $G_i \subset$ \{Y_1,...,Y_k,X_1,...,X_m\} contains at least three variables one of which has to be an item. The GLLRM defined by these generators is given by

$$P(Y_1 = y_1, \ldots, Y_k = y_k | \Theta = \theta, X_1 = x_1, \ldots, X_m = x_m)$$
$$=$$
$$\exp\left(\alpha_0 + \sum_{i=1}^{k}(\theta y_i + \alpha_{iy_i}) + \sum_i \delta_i(a_i, z_i) + \sum_i \lambda_i(u_i, v_i) + \sum_i \mu_i(g_i)\right) \quad (21.8)$$
$$=$$
$$\exp\left(\alpha_0 + s\theta + \sum_{i=1}^{k}\alpha_{iy_i} + \sum_i \delta_i(a_i, z_i) + \sum_i \lambda_i(u_i, v_i) + \sum_i \mu_i(g_i)\right)$$

where s = Σ_iy$_i$ and (a$_i$,z$_i$), (u$_i$,v$_i$) and g$_i$ is the observed outcomes of the variables in the generators. It is usually assumed that the model is hierarchical. We refer to the δ and λ parameters as DIF and LD parameters, respectively, even though the interpretation in these terms is questionable when G is not empty. Note that while the main effects, θy+α_{iy}, are increasing functions of θ, not all marginal relationships between Θ and items are monotonously increasing when items may be negatively locally dependent. Items fitting a general GLLRM violate all but one of the assumptions of criterion related construct validity and conventional psychometric considerations would reject the scale as invalid. The measurement properties of items fitting a GLLRM models are discussed below based on the example.

21.6.4 Inference in GLLRMs

GLLRM's have moral Rasch graphs that may be used as a starting point for the same kind of tests as for the GRMs. The separation properties are a little more complicated in moral Rasch graphs from GLLRMs but graph theoretical algorithms exist that will take care of these problems.

Item, DIF, LD, and interaction parameters can be estimated by conditional maximum likelihood estimates evaluated by item-fit statistics comparing observed and expected item-characteristic curves and tested by conditional likelihood ratio tests (Kelderman, 1984, 1992, 1995). The Martin-Löf test of unidimenisonality (Martin-Löf, 1970; Glas & Verhelst, 1995; Verhelst, 2001; Christensen et al., 2002) generalize with few problems to GLLRMs (Kreiner & Christensen, 2004).

In a GLLRM the distribution of S given Θ is a power-series distribution similar to (21.3) with score parameters depending on item, DIF, LD, and interaction parameters. Estimation of person parameters and latent regression where estimates of score parameters are inserted can be done in the same way as in conventional models.

21.7 SF-36 Analysis

Data for this example originated in a Danish health survey including 2334 persons responding to the SF-36 items and the five exogenous variables (self-reported health, BMI, smoking status, sex, and age) included in the analysis.

All variables are potential sources of DIF and self-reported health is also used as a criterion variable. The primary purpose of the study is not validation of the measurement instrument, but rather to examine the effect of BMI on physical functioning. This is done using latent regression analysis (Christensen et al., 2004) controlling for the confounding effect of the other variables. Rather than a pure validity study the item analysis is meant to check that the result of the latent regression analysis is not confounded by systematic measurement errors.

The item "vigorous activities" (PF01) discriminates very poorly, $U = -5.94$, $p < 0.001$ (Molenaar, 1983) and is excluded. Presumably this has to do with problems responding when one does not participate in vigorous activities for other reasons than poor health (Fayers & Machin, 2000, p. 19). The complete analysis leading to the model will not be documented here, but evidence against the conventional RM and results supporting the adequacy of the GLLRM for the remaining nine items model is presented.

Conditional likelihood ratio tests (Andersen, 1973c), comparing item parameters in different groups, show evidence against the model (Table 21.1, columns marked RM). The reason for the discrepancy between model and data is not clear from overall test statistics.

Table 21.1. Conditional likelihood ratio tests of homogeneity of item parameters in subpopulations. Results presented for the RM and for the graphical loglinear RM.

	RM			GLLRM		
Variable Defining Subpopulations	CLR	df	P	CLR	df	P
Score groups (1-17, 18-19)	105.1	19	< 0.0005	70.0	67	0.379
SRH—five categories	261.2	76	< 0.0005	212.4	208	0.402
BMI—six categories	125.0	95	0.021	309.3	285	0.154
Smoking—three categories	50.5	38	0.085	168.6	134	0.023
Sex	72.5	19	< 0.0005	70.3	65	0.304
Age—six categories	179.0	95	< 0.0005	311.0	285	0.139

The risk of type I error is inherent in testing for LD of 36 item pairs and for DIF with 45 combinations of items and exogenous variables, and the level of significance is adjusted in order to control the false discovery rate at a 5% level (Benjamini & Hochberg, 1995). Moreover DIF or LD can lead to spurious evidence of DIF and/or LD for other items and/or exogenous variables and subsequent analyses are needed.

Partial gamma coefficients (Agresti, 1984, p. 171) showed strong evidence of LD for the item pairs (PF1, PF2), (PF2, PF3), (PF4, PF5), (PF7, PF8), and (PF8, PF9). Two-sided Monte Carlo estimates of exact conditional p-values were used. Table 21.2 shows evidence of DIF, only evidence from analysis taking several potential DIF sources into account are shown.

Table 21.2. Evidence of DIF disclosed by partial gamma coefficients. p-values are two-sided Monte Carlo estimates of exact conditional p-values. The false discovery rate has been controlled at 5% and only evidence from analysis taking several potential DIF sources into account support the evidence of DIF relative to variables written with bold letters.

Item	Exogenous Variable	Gamma	p
Vigourous activities (PF1)	Sex	0.23	0.012
Moderate activities (PF2)	BMI	0.31	0.000
Lifting groceries (PF3)	Sex	−0.45	0.000
Stairs—2+ flights (PF4)	Smoking	0.23	0.013
Stairs—1 flight (PF5)	Age	−0.24	0.008
Bending (PF6)	BMI	−0.17	0.012

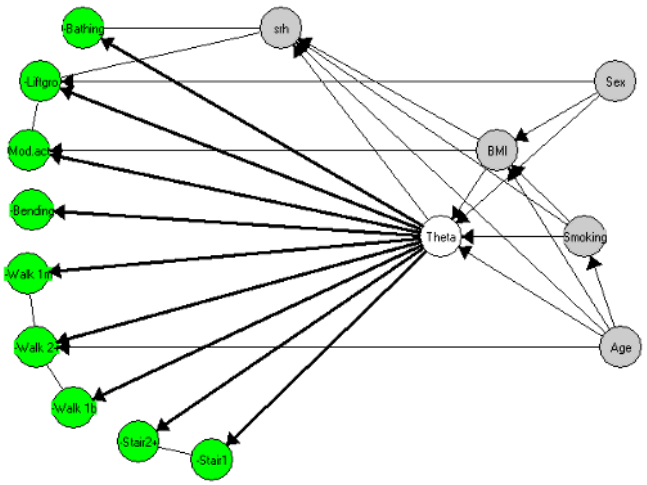

Fig. 21.3. The IRT graph of the final GLLRM for items PF2-PF10

All significant interactions were added to yielding a relatively simple GLLRM for the nine PF items. Figures 21.3 and 21.4 show the IRT graph and moral Rasch graph of this model. Significance levels are shown in Table 21.3. Conditional likelihood ratio tests comparing parameter estimates in subpopulations shows that this model fits the data better (Table 21.1, columns marked GLLRM). Observed and expected item-mean scores in each score group were compared and this also showed a good model fit.

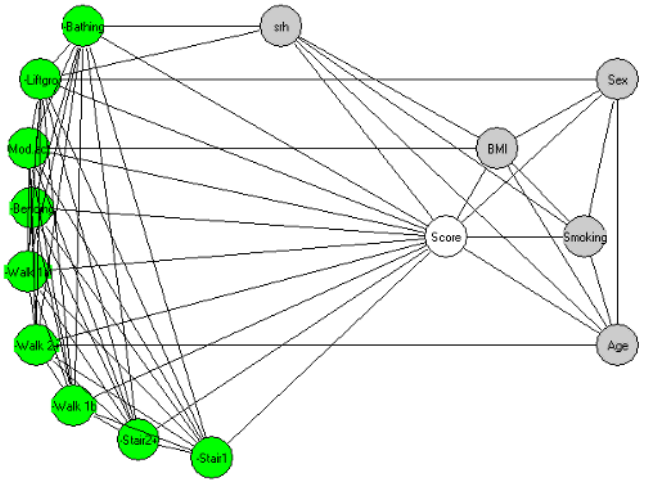

Fig. 21.4. The moral Rasch graph of the final GLLRM for items PF2-PF10

Table 21.3. Tests of vanishing DIF and LD parameters

Type of Interaction	Variables	CLR	df	p
Local dependence	PF2 & PF3	168.2	4	< 0.00005
	PF4 & PF5	116.5	4	< 0.00005
	PF7 & PF8	77.8	4	< 0.00005
	PF8 & PF9	203.5	4	< 0.00005
DIF	PF2 & BMI	34.9	10	0.00001
	PF3 & SRH	22.4	8	0.00430
	PF3 & Sex	20.1	2	< 0.00005
	PF8 & Age	27.6	10	0.00210
	PF10 & SRH	23.7	8	0.00260

21.7.1 Interpretation of Parameters

One item, "Bending and kneeling" (PF6), behaves like an ordinary RM item. Threshold parameters, for a partial-credit interpretation, are -1.83 and 0.11, implying a range of latent-trait values where each response is the most probable.

The items PF4 and PF5 concerning stair walking are locally dependent, but function in the same way relative to all other variables. Locally dependent items can be grouped together as a composite item defined as the sum of item scores. If there is no DIF composite items are distributed as items from an RM. Item parameters for the composite item $\mathrm{PF}_{4+5} = \mathrm{PF}_4 + \mathrm{PF}_5$ can be computed from the item and LD parameters, $\alpha_{4y_4} + \alpha_{5y_5} + \lambda_{45}(y_4, y_5)$. Reparametrization,

for a partial interpretation, thresholds show that thresholds are nicely ordered $(-1.75, -0.61, -0.09, 1.39)$.

DIF can be presented as loglinear-item and DIF parameters, but the effect is easier to interpret in terms of virtual items. The item "bathing" (PF10) is biased relative to self-reported health, but no evidence of LD was found. Partial credit thresholds of five "virtual" items in the subpopulations defined by self-reported health are shown in Table 21.4. Apart from some response categories not being used in the healthiest groups these appear to present a consistent picture with decreasing thresholds with failing health.

Table 21.4. Estimated thresholds of five virtual PF10-items in groups defined by self-reported health

	Thresholds	
SRH	1	2
Very good	0.93	+ inf.
Good	0.92	0.80
Fair	0.82	4.26
Bad	−0.21	1.92
Very bad	−0.27	2.48

For items with both DIF and LD, the situation is complicated and "virtual composite items" do not present an easy interpretation. As an example, consider $PF_{2+3} = PF_2+PF_3$ with DIF of PF2 (relative to BMI) and of PF3 (relative to SRH and sex): thresholds would have to be calculated for 60 virtual items to get a comprehensive description. The items relating to walking are a simpler example: the composite item, PF_{7+8+9}, is biased relative to age because DIF was disclosed for one of the three items. Disordered thresholds are common for "virtual composite items" and while the GLLRM appears to provide adequate description of the relationship between the variables of the model. An easy interpretation is not at hand.

21.7.2 The Effect of DIF on the Score

The score distribution (21.3) applies in GLLRM with the reservation that the score parameters depend on the exogenous variable (the sources of DIF)

$$P(S = s|\Theta = \theta, X = x) = \frac{\exp(\theta s + \varphi_s(x))}{\Phi(x)} \tag{21.9}$$

The score parameters are functions of item, DIF, and LD parameters and can be used to calculate estimates of θ or of the parameters of the distribution of Θ in the same way as for RMs.

Person parameters and their distribution can be compared on the latent-trait scale and this is preferable to the raw scores because the latent-trait scale

may be regarded as an interval scale. It can, however, be difficult to decide whether a difference on the latent-trait scale is relevant. DIF equation of true scores between groups can be useful: Let $T_0(\theta) = E(S \mid \Theta = \theta, X = \text{ref})$ be the true score of a person from the reference group and let $\hat{\theta}(s, x)$ be the estimate of θ for a person with S=s in the group defined by X=x. The DIF equated score of this person is equal to $T_0\left(\hat{\theta}(s, x)\right)$.

Figure 21.5 illustrates the effect of DIF w.r.t BMI of the item "Moderate activities" (PF2): persons with high BMI underestimate the degree of physical disability due to health. This is probably of minor consequence for those with BMI $= 22.5 - 25.0$, where the largest adjustment by DIF equation adds about .2 points to low scores, but of some consequence for those with BMI > 30 where DIF equation adds 0.44–0.65 point to scores between 2 and 13.

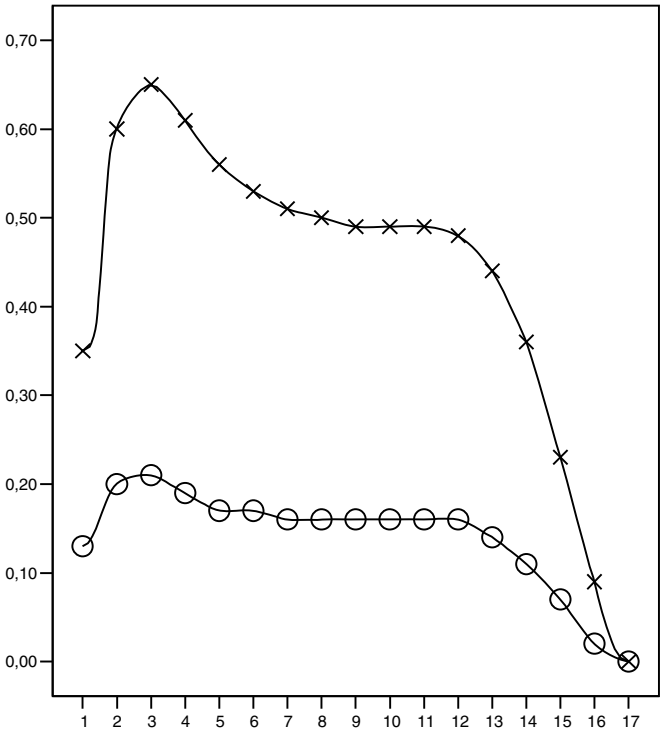

Fig. 21.5. DIF equated adjustment of scores for two groups of 18-to-29-year-old males with very good health (x = BMI = 22.5–25.0, o = BMI = 30+. The reference group consists of 18-to-29-year-old males with very good health and BMI ≤ 20.

21.7.3 Latent Regression

We now examine the effect of BMI on physical disability. Tests of conditional independence of the score and exogenous variables given the separators of the moral Rasch graph (Figure 21.4) yields a list of covariates that should be included. These show strong effects for all variables except smoking and can be seen as a stronger requirement of criterion validity (insisting that the association does not disappear when covariates related to both variables are taken into account—a requirement that is obviously met here). The estimated score parameters in (21.9) are used for latent regression (Christensen et al., 2004) using SAS (Christensen & Bjorner, 2003): A significant effect of BMI on physical functioning when controlling for sex, age, and self-reported health was found (LRT = 12.5, $df = 5, p = 0.03$).

Table 21.5. Difference between BMI groups controlled for sex, age, and self-reported health

BMI Group	Difference	95% CI
≤ 20	0.02	(−0.25, 0.30)
20.1–22.5	0.06	(−0.17, 0.29)
22.6–25.0	0	-
25.1–27.5	0.21	(−0.02, 0.44)
27.6–30.0	0.14	(−0,15, 0.43)
30.1 +	0.45	(0.18, 0.72)

Table 21.5 shows the estimated differences at the latent-trait scale between the six BMI groups. Physical disability appears to be at a minimum in the reference groups (BMI = 22.5–25.0) with a marked increase in physical disability when BMI is larger than 30. The evidence of increased disability in groups with BMI less than 22.5 is of course not significant.

21.8 Essential Validity and Objectivity

The previous section illustrates how LD and DIF may be dealt with if item responses fit a GLLRM. Latent-trait parameters can be estimated and compared as in the RMs. The question remains, however, whether measurement by these items can be regarded as valid and objective: All assumptions defining criterion-related construct validity except unidimensionality has been violated. We claim that validity and objectivity essentially has been preserved in GLLRMs. We return to the simple models given by (21.6) and (21.7) for the arguments supporting these claims and notice that the arguments carry over without problems to the general class of GLLRM.

The model defined by (21.7) includes one pair of uniformly locally dependent items, Y_a and Y_b. Replacing these two items by the composite item

$Y_{a+b} = Y_a + Y_b$ however results in a set of items satisfying all requirements of ideal scales except, perhaps, monotonicity of the composite item. Given the fact that the total scores are the same it is difficult to argument that Y_a and Y_b violates validity in any important way. The total score is sufficient for θ such that person and item parameters—among which we include the LD interaction parameters—may be separated during the analysis. The fundamental property of RMs supporting claims of objectivity therefore survives intact in (21.7), with one restriction compared to Rasch's definition of specific objectivity: we can not select items in a completely arbitrary way. One has to either include or exclude both items because an item subset including but one of the two dependent items does not fit an RM. This is, in our mind, a small price to pay during construction of a summary scale. Measurement may not be construct valid and objective according to conventional psychometric thinking, but it makes no sense to claim that measurement is invalid and biased or prone to systematic errors due to some arbitrary decision by the person constructing the test.

The model (21.6) with uniform DIF of Y_i relative to X_j, is a little more complicated. One may of course eliminate Y_i to obtain a smaller set of items satisfying requirements of ideal measurement. The set of items therefore is inherently valid and objective. When addressing problems relating to one of the groups defined by X_j, one would prefer to keep Y_i to increase reliability, because measurements are valid and objective in this specific population. From the point of view of the virtual Y_i we may also claim that test equating actually satisfies the requirements of specific objectivity because missing item responses is no hindrance to validity and objectivity. Regarding DIF parameters as item rather than incidental person parameters implies that conditioning with respect to the total score separates item parameters from the latent-trait parameters; the technicalities of objective analysis thus survives. Again a restriction applies: we are no longer free to make completely arbitrary choices during the design of the study. If we decide to include Y_i, we also have to include data on X_j, but apart from this measurements are essentially valid and essentially objective.

All arguments relating to models (21.6) and (21.7) apply without restriction to the general family of GLLRMs. The model may, of course, turn up to be so complicated that we prefer to reject the scale either because it is not practical to work with or because it is so far away from a conventional RM that we may be concerned that the substantive arguments behind the items do not hold water. If the GLLRM appears to fit the data, we should use these arguments and not arguments that measurements are invalid and systematically biased.

21.9 Discussion

This chapter discussed validity assumptions arguing from the point of view of a GLLRM fitting responses to the nine PF items. The SF-36 items violate conventional requirements of validity and objectivity due to unfortunate item-writing. Rather than rejecting the scale, we have taken a second look at requirements of valid and objective scales, partly because the items of SF-36 have a certain degree of face validity, but also because most scales we have worked with in health research suffer from similar problems. Our conclusion is that most requirements can be relaxed and that GLLRMs provide a sensible framework where all but a few properties of valid and objective measurements survive. Of the two types of departures from construct validity permitted by GLLRMs, the presence of uniform local dependency seems to be the least problematic. Regarding two locally dependent items as one composite item is a very small price to pay for the added reliability of the total score compared to the rest score without the items. Uniform DIF is a little more problematic. We may deal with uniform DIF, but it requires that all sources of DIF have to be included among the observations. The results of the analysis presented in this chapter implies that measurements of physical disability by the PF items will to some extent be confounded if sex, age, BMI, and SRH is not observed and taken into account. In addition, interpretation of the DIF of items relating to SRH is difficult. Is self-reported health worse because the person has problems bathing and/or carrying groceries home, or do these two tasks appear particular difficult because health as such is perceived as poor.

Quality of measurement is important and may be the only purpose of analysis. The widespread use of SF-36 is sufficient reason to examine validity, objectivity, and reliability. Often the measurement problem is subordinate to latent-structure analysis, as illustrated by analysis of the effect of BMI on physical disability. In this analysis, the evidence against the Rasch model is inescapable, even though four items fit an RM. The price to pay in terms of reliability is unacceptable because items appear to be face valid. The nine PF items provide essentially valid and objective measurements of physical disability. We base latent regression analysis on these nine items, taking DIF and LD into account and argue that this is better than using four items, even though these provide valid and objective measurements in the strict sense.