Item analysis in DIGRAM 5.01

Guided tours

Svend Kreiner and Tine Nielsen

February 2023

Table of contents

1 Introduction

1.1 SCD and DIGRAM

DIGRAM is part of a larger statistical package, SCD[1], containing facilities for analysis of discrete data. A general introduction to the program may be found in the Introduction to DIGRAM (Kreiner, 2003) and in notes on

Project management
Analysis of contingency tables by chain graph models
The graph module

The original version of DIGRAM (Kreiner, 1989) was a program dedicated to analysis of high-dimensional contingency tables by block recursive graphical models. While graphical modelling is still important for DIGRAM, the focus has to some degree shifted towards a larger range of problems where conditional independence plays important roles, but where graphical models are not regarded as full-fledged models, but rather as non-parametric skeletons on which specialized models may be build. In addition to graphical modelling DIGRAM now supports:

1) Analysis of collapsibility across categories in multidimensional contingency tables.
2) Analysis of inherent order and monotonous relationships among nominal or partially ordered variables.
3) MCA analysis of marginal and conditional homogeneity in multidimensional contingency tables.
4) Non-parametric log-linear modelling of ordinal categorical data.
5) Analysis of multidimensional Markov Chains.
6) Item analysis by Rasch models as well as graphical and log-linear Rasch models.

We assume that users of DIGRAM are familiar with graphical models and Rasch models. However, many users of Rasch model are only familiar with a version of the model for polytomous items referred to as a partial credit model. For this reason, Section 1.4 provides a brief introduction to the Rasch model for polytomous items, presenting the five equivalent versions of the model that DIGRAM applies. Since the versions are equivalent, it does not matter which version of the model

[1] SCD/DIGRAM is giftware. It comes without a charge and you are free to distribute copies of the program to anyone to whom it may be useful. You can download a copy of the program from Biostat.ku.dk/DIGRAM where updates of the programs and the user guides will be available as they appear.

that a program for item analysis by Rasch models uses. The results of the analysis will be the same. DIGRAM applies all five versions since they are useful for different purposes and because they provide different insights in the way that the polytomous Rasch items function. Section 1.4 elaborate on these issues.

DIGRAM supports item analysis by graphical log-linear Rasch models (GLLRMs). Sections 1.5 and 1.6 provides a short illustration of analysis by GLLRMs and a list of publications that describe and/or use GLLRMs. Readers that are familiar with the models may skip these section and proceed directly to Section 2.1 with the first guided tour.

1.2 Item analysis

The purpose of item analysis in DIGRAM is to

1) To examine whether a summated scale counting responses to a set of items provides a valid, objective and useful measure of a latent trait by an item analysis of the fit of item responses to a conventional Rasch model or a graphical and log-linear Rasch model.

2) To identify items and persons that do not fit the proposed model if the fit to the Rasch model is unsuccessful, and/or to find a graphical log-linear Rasch model (GLLRM) where uniform differential item functioning (DIF) and uniform local dependence (LD) is permitted.

3) To calculate estimates (measures) of the value of the person parameters and to assess measurement error, reliability and targeting of measurements.

Item parameters are always estimated and presented during the item analysis, but estimates of item parameters are often subordinate to the other purposes during tests-of-fit of the model and in connection with estimation of person parameters. However, they may also be of interest in themselves in connection with special applications, for instance during development and revision of of tests or summary scales, and in connection with criterion referenced classification of scores.

For this reason, it is important that estimates of item parameters provide information that can be used to assess and compare the qualities of items. It is to help with this that, that DIGRAM provides estimates of all the item parameters associated with the five different versions of the Rasch model (RM) for polytomous items.

1.3 Four examples

We include four DIGRAM projects with the guide. The data for these examples can be found in DIGRAM projects that are distributed together with the program.

1.3.1 The DHP project

The data in the DHP project originated in a study of The Diabetes Health Profile (DHP). The DHP is a multidimensional patient self-completion diabetes-specific inventory designed to identify psychosocial dysfunction among adult insulin dependent and insulin requiring patients. Factor analyses have suggested that responses to DHP items depend on three latent variables representing Psychological distress, Barriers to Activity and Disinhibited eating. Chwalow et.al (2007) describe a randomized study of the quality of life of type 2 diabetic patients. We use data from this study to illustrate item analysis of the Disinhibited eating (DE) subscale summarizing responses to the following five questions with four ordinal response categories that were coded in such a way that 0 represents no dysfunction and 3 represents a high degree of dysfunction:

A: DHP32 Do you wish there were not so many things to eat?

Responses: a) “Not at all”, b) “A little”, c) “A lot”, d) “Very much”

B: DHP34 How likely are you to eat something extra when you feel bored or fed up?

Responses: a) “Not at all likely”, b) “Not very likely”, c) “Quite likely”, d) “Very likely”

C: DHP36 When you start eating, how easy do you find it to stop?

Responses: a) “Very easy”, b) “Quite easy”, c) “Not very easy”, d) “Not at all easy”

D: DHP38 Do you have problems keeping to you diet because you eat to cheer yourself up?

Responses: a) “Never”, b) “Sometimes”, c) “Usually”, d) “Always”

E: DHP39 Do you have problems keeping to your diet because you find it hard saying no to food you like?

Responses: a) “Never”, b) “Sometimes”, c) “Usually”, d) “Always”

In addition to the items, the DHP project also includes information on sex and age.

1.3.2 The CONF08 project

The data for this project originated in the Danish component of the European values studies in 2008. We use five polytomous items measuring confidence in the following public institutions.

D) The police
E) The parliament
G) The social security system
K) The health system
L) The courts.

Response categories were “High degree of confidence” (0), “Some degree of confidence” (1), “Little degree of confidence” (2) and “No confidence” (3). Note that responses were coded in such a way that a high item score indicate a high degree of distrust. We refer to the instrument as the CONF scale even though a high CONF score is an indication of lack of confidence of public institutions.

1.3.3 The PF3 project

The second project originated in a Danish Health survey. We will here be concerned with the validity of the SF-36 subscale measuring physical functioning. The scale summarizes responses to the following ten items:

Does your health now limit you in these activities? If so, how much?

A) PF1: Vigorous activities
B) PF2: Moderate activities
C) PF3: Lifting or carrying groceries
D) PF4: Climbing several flights of stairs
E) PF5: Climbing one flight of stairs
F) PF6: Bending, kneeling, or stooping
G) PF7: Walking more than a mile
H) PF8: Walking several blocks
I) PF9: Walking one block
J) PF10: Bathing or dressing yourself

The responses to these questions were coded so that a low score indicates physical impairment.

0 : Limited a lot 1: Limited a little 2: Not limited

Gender and Age are also included in this project.

1.3.4 The ADLtired project

The majority of the features implemented in DIGRAM apply for both polytomous and dichotomous items, but DIGRAM also supports a number of methods for item analysis by Rasch's model for dichotomous items. To illustrate these methods we use data on from a study of the construct validity of a so-called PADL (Physical Activities of Daily Living) measure of functional ability of healthy elderly (Avlund et.al., 1993).

In this study, data was collected from 734 70-year old in the County of Copenhagen, Denmark. The PADL scale consisted of 16 items covering the three domains shown in Table 1.1.

Responses for the example used throughout these notes were coded as 0 = "Cannot do it at all, or cannot do it without getting tired" and 1 = "can do it without getting tired"

Table 1.1 PADL items.

Mobility function	Lower limb function	Upper limb function
A: Are you able to walk indoors?	G: Are you able to wash the lower part of the body?	L: Are you able to wash the upper part of the body?
B: Are you able to walk out of doors in nice weather?	H: Are you able to cut your toenails?	M: Are you able to cut your fingernails?
C: Are you able to walk out of doors in nice weather?	I: Are you able to go to the toilet yourself?	N: Are you able to comb your hair?
D: Are you able to manage stairs?	J: Are You able to dress the lower part of the body?	O: Are you able to wash your hair?
E: Are you able to get outdoors?	K:Are you able to take shoes/stockings on/off?	P: Are You able to dress the upper part of the body?
F: Are you able to get up from a chair or bed?		

Social class, sex and pension age are included in this project.

1.4 Parameterizing Rasch models for polytomous items

We assume that users of DIGRAM are familiar with Rasch models for polytomous items and are comfortable with the version of the model known as the partial credit model (PCM). However, the PCM is only one of five equivalent versions of the Rasch model for polytomous items that DIGRAM use. This section defines these versions and explain why we use them.

It is important to underscore that the five versions are equivalent. Tests of fit and estimates of person parameters are the same for all versions. For practical reasons, DIGRAM calculates conditional maximum likelihood (CML) estimates of the parameters defined by the multiplicative PSD version of the model where item are power series distributed and transform these estimates to CML estimates of the parameters defined by the other versions. Had it been more practical to start by estimating the parameters of another version, the results would have been the same.

1.4.1 Rasch's model for polytomous items

It is informative to look at Rasch's original models for polytomous items. Let X_{vi} be the response to a polytomous item with m response categories where m>2. The model for polytomous items presented for the first time by Rasch (1961) and Rasch (1962) was a multivariate model depending on a vector of person parameters

$$PR(X_{vi} = x) = \frac{\exp(a_x\theta_{vx} + b_x\sigma_{ix} + c_x\theta_{vx}\sigma_{ix} + \omega_x)}{\sum_{h=0}^{m}\exp(a_h\theta_{vh} + b_h\sigma_{ih} + c_h\theta_{vh}\sigma_{ih} + \omega_h)} \tag{1.1}$$

In (1.1), a_x, b_x and c_x are known scoring functions and θ_{vx}, σ_{ix} and ω_x are unknown parameters. Obviously, model (1.1) violated the principle of parsimony. Rasch (1962, p. 103) asked "whether it is possible to reduce the number, m, of parameters per person and per item" and suggested that the "simplest case would seem to be that each person as well as each item is fully characterized by a one-dimensional parameters θ_v and σ_i". He also discarded the multiplicative $c_x\theta_{vx}\sigma_{ix}$ term and assumed that the scoring functions should be the same for the person and item parameters. The result was a special case of (1.1) with unidimensional person and item parameters and a vector of category parameters $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_m)$

$$PR(X_{vi} = x) = \frac{\exp\left(a_x\left(\theta_v + \sigma_i\right) + \omega_x\right)}{\sum_{h=0}^{m}\exp\left(a_h\left(\theta_v + \sigma_i\right) + \omega_h\right)} \tag{1.2}$$

Except for the scoring function a_x, the model defined by (1.2) is equivalent to the so-called rating scale model by Andrich (1978). Rasch never documented examples of applications of this model, but Andersen (1973) provided one example. Rasch did not discuss whether it was reasonable to

assume that the vector of category parameters had to be the same for all items, and Andersen (1977) extended (1.2) to a model where category effects depend on the item

$$PR(X_{vi}=x) = \frac{\exp\left(a_x\left(\theta_v+\sigma_i\right)+\omega_{ix}\right)}{\sum_{h=0}^{m}\exp\left(a_h\left(\theta_v+\sigma_i\right)+\omega_{ih}\right)} \tag{1.3}$$

Finally, setting $\sigma_{ix} = a_x\sigma_i+\omega_{ix}$ leads to the model that we today – except for the scoring functions – refer to as the Rasch model for polytomous items.

$$PR(X_{vi}=x) = \frac{\exp\left(a_x\theta_v+\sigma_{ix}\right)}{\sum_{h=0}^{m}\exp\left(a_h\theta_v+\sigma_{ih}\right)} \tag{1.4}$$

1.4.2 The Rasch model for polytomous items

To Rasch, the advantage of the distributions defined by (1.1) - (1.4) is that they are exponential family distributions with sufficient statistics.

Let $\mathbf{X_v} = (X_{v1},\ldots,X_{vk})$ be the vector of item responses and let $A(x) = a_x$ be the score function. In formulas (1.2) – (1.4), the person score $R_v = \Sigma_i A(X_{vi})$ is sufficient for the person parameter θ_v. Andersen (1977) showed that data reduction is as efficient as possible if the scores are equidistant and preferred scores for m+1 ordinal categories to be decreasing from m to zero. Today, the preferred scores refer to the ranks of the categories starting with zero and ending with m. It is the (1.4) model defined by such scores that we today refer to as the Rasch model for polytomous items.

1.4.3 The exponential family model (EFM)

We refer to statistical models in which the distribution of variables are members of exponential families as exponential family models (EFM) from which it follows that the Rasch model for polytomous items defined by (1.4) is as an EFM. The model that we refer to as the EFM version of the Rasch model for polytomous items modifies (1.4) in three ways.

First to let the polytomous items have different number of ordinal categories with item sores from zero to m_i. In this way, the rasch model for polytomous items include dichotomous items with m_i=1.

Second by assuming that the person parameter of (1.4) is the outcome of a latent variable. In other words, the Rasch model for polytomous items describe the *conditional* distributions of responses to items given the outcome of an unobservable variable. Since this outcome appear as an unknown parameter on the Rasch model it follows that we can provide indirect measures of the unobservable outcome by statistical estimates of the person parameter.

The exponential family version (EFM)

Let $I_z(x)$ be the indicator function with $I_z(x)=1$ if x=z and zero otherwise. The exponential family version of the Rasch model for polytomous items is defined by

$$PR(X_{vi}=x \mid \theta_v) = \frac{\exp\left(x\theta_v+\sigma_{ix}\right)}{\sum_{z=0}^{m_i}\exp\left(z\theta_v+\sigma_{iz}\right)} = \frac{\exp\left(x\theta_v+\sum_{z=0}^{m_i} I_z(x)\sigma_{iz}\right)}{\sum_{z=0}^{m_i}\exp\left(z\theta_v+\sigma_{iz}\right)} \qquad (1.5)$$

The model is overparameterized. In DIGRAM, we assume that $\sigma_{i0} = 0$ for all items and that $\sum_i \sigma_{im_i} = 0$.

The exponential version represent the statistical view on the Rasch model. Statistical models defined by exponential family distributions have sufficient statistics.

In exponential family models, you can read the sufficient statistics and the ways to calculate maximum likelihood (ML) estimates directly off formulas like (1.5). In the Rasch model for polytomous items, the total person scores are sufficient for the person parameters and the item *margins* counting the number of times that the response categories of an item has been observed are sufficient for the item parameters. Finally, the ML estimates are the parameters where the expected sufficient statistics are equal to the observed statistics.

1.4.4 The item and category effects (ICE) version

The EFM version (1.5) of the Rasch model has one serious problem. There is no simple way to interpret of the meaning of the item parameters because we cannot separate the effect of the content of the item from the effect of the categories.

Polytomous items are defined by an item-issue that respondents have to address and a set of ordinal categories. The problem is that several sets of response categories will be available when we construct items.

Consider for instance, the DHP and CONF08 projects described in Sections 1.3.1 and 1.3.2 with information on five polytomous items.

The five items of CONF08 use the following response categories for items addressing degree of confidence in public institutions:

"High degree", "Some degree", Little degree", "No confidence"

The five DHP items use four different sets of response categories for questions addressing issues related to disinhibited reading

"Not at all", "A little", "A lot" and "Very much"
"Not at all likely", "Not very likely", "Quite likely" and "Very Likely"
"Very easy", "Quite easy", "Not very easy" and "Not at all easy"
"Never", Sometimes", "Usually" and "Always"

Given the different sets of response categories (and the many other options that must have been available) the unavoidable question is how the different sets of categories influence the way the items function and whether the effect of a given set of categories depend on the item issue. In which ways is item functioning different for the four sets of DHP categories and does "Some degree of confidence" mean the same for the police and the health care system?

Unfortunately, the parameters of the EFM version is of no help because they do not separate the effect of the item issue from the effect of the categories. To compare the different sets of categories versions we have to reparametrize the model in a way that attempt to do that; and one way to do that is to return to Rasch's original parameterization in (1.3) with a parameter describing the item effect and a set of parameters describing category effects.

We refer to this version as an item and category effect (ICE) version. The fundamental assumption is that the item issue defines the item effect in a way that does not depend on the choice of categories. That the general level of confidence in the police is the same whether we use the four categories of the CONF08 project, or another meaningful set of response categories. And that a

respondent will select a category that accommodates his or her their experiences of confidence. If this is correct, the choice of a category must depend on the degree to which the different categories accommodate different experiences of problems relating to the issue.

The ICE model defined in Formula (1.6) provides one way to separate item effects and category effects. The person effect is the same as the person parameter in the EFM model, the item effect is the EFM item parameter of the highest item score (m_i) divided by m_i and the category effect is the EFM parameter of the category minus the item effect multiplied by the item score.

The item and category effects (ICE) version

Parameter	Canonical parameters	ICE parameters
Person effect	$\boldsymbol{\theta_v}$	$\boldsymbol{\theta_v}$
Item effect	$\boldsymbol{\sigma_{ix}}$	$\sigma_i = \sigma_{im_i}/m_i$
Category effect		$\boldsymbol{\omega_{ix} = \sigma_{ix} - x\sigma_i}$

$$PR(X_{vi}=x \mid \theta_v) = \frac{\exp(x\theta_v+\sigma_{ix})}{\sum_{z=0}^{m_i}\exp(z\theta_v+\sigma_{iz})} = \frac{\exp(x(\theta_v+\sigma_i)+\omega_{ix})}{\sum_{z=0}^{m_i}\exp(z\theta_v+\sigma_{iz})} \quad (1.6)$$

It follows from the restriction imposed on the EFM version that $\omega_{i0} = \omega_{im_i} = 0$ and $\sum_i \sigma_i = 0$.

The ICE version of the Rasch model for polytomous items is an exponential distribution. It suggests that it makes sense to include the total item score as a sufficient statistic for the item effect together with the item margins that are sufficient for the category effects. However, the item score is a function of the item margins and the set of sufficient statistics is not minimal. For this reason, there are no *practical* reasons to include the item score among the sufficient statistics during the analysis. The easiest way to estimate the item effect is to estimate the parameters of (1.5) and then reparametrize as in (1.6) to obtain estimates of item and category effects.

The estimates of the item and category effects of the CONF08 and DHP items can be found in Table 1.4 and Figure 2.2.6. The differences of the category effects are striking. The category effects of the extreme categories of the CONF items (“High degree of confidence” and “No confidence”)

are very weak compared to the effects of some or little degree of confidence. The DHP items are characterized of extreme category effects that are closer effects of the second and third category.

1.4.5 The multiplicative power series model (PSD)

The next versions of the Rasch model rewrite (1.5) and (1.6) as multiplicative models where items have power series distributions (PSDs) with unknown score parameters. We refer to the multiplicative version of the EFM as the PSD version and the multiplicative version of ICE as the MICE version.

The advantages of using the PSD version of the RMP instead of exponential family versions is that parameters have values on ratio scales and that formulas are more transparent.

A second advantage is that we can use results on power series distributions. The most important is that the sum of two PSDs depending on the same person parameter belong to the same family of PSDs. Since we can rewrite item distributions as PSD distributions, it follows that the sum of two or more items also has PSD distributions. In other words, that scores and subscores over polytomous Rasch items can be regarded as as polytomous rasch items.

Formulas (1.7) and (1.8) define the PSD and MICE version. They add an extra benefit: the PSD version permits some of the score parameters to be equal to zero. Assume for instance, that responses to a question with five response categories for substantive reasons are scored 1, 2, 4, 6 and 7. To redefine this as a polytomous Rasch item we subtract 1 from the proposed score. In this way, the range of item scores change to {0,1,3,5,6}. This would be a problem for the EFM and ICE versions, but using the PSD version (1.7) instead (1.5) we may redefine the items as a polytomous item with seven categories with score parameters $\delta_{i2} = \delta_{i4} = 0$.

In this way, the PSD and MICE versions represent extensions of the Rasch model for polytomous items. Since the implementation of Rasch models in DIGRAM use the multiplicative versions for estimation of parameters and tests of fit, item analysis by DIGRAM include the extended PSD and MICE versions of the Rasch model for polytomous items.

A final convenient consequence of using the PSD and MICE versions is that we may assume that all items have the same number of response categories. For instance, if the set of items include

dichotomous and trinary items we just have to treat the third score parameter of the dichotomous items as a structural zero.

The multiplicative PSD and MICE versions

Parameter	Canonical parameters	ICE parameters
Person effect	$\boldsymbol{\theta_v}$	$\xi_v = \exp(\theta_v)$
Score parameter	$\boldsymbol{\sigma_{ix}}$	$\delta_{ix} = \exp(\sigma_{ix})$

The PSD version rewrites (1.5) as

$$\Pr(X_i = x|\xi_v) = \xi_v^x \delta_{ix} \Big/ \sum_{j=0}^{m} \xi_v^j \delta_{ij} \qquad (1.7)$$

Parameter	Canonical	ICE parameters	MICE parameters
Person effect	$\boldsymbol{\theta_v}$	$\boldsymbol{\theta_v}$	$\xi_v = \exp(\theta_v)$
Item effect	$\boldsymbol{\sigma_{ix}}$	$\sigma_i = \sigma_{im_i}/m_i$	$\psi_i = \exp(\sigma_i)$
Category effect		$\boldsymbol{\omega_{ix} = \sigma_{ix} - x\sigma_i}$	$\alpha_{ix} = \exp(\omega_{ix})$

The MICE version is a multiplicative reparameterization of the ICE (1.6)

$$\Pr(X_i = x|\xi) = \frac{(\xi\psi_i)^x \alpha_{ix}}{\sum_{j=0}^{m} (\xi\psi_i)^x \alpha_{ix}} \qquad (1.8)$$

1.4.6 The partial credit version (PCM)

Masters (1980 and 1982), Wright & Masters (1982) and Andrich (1978) reparametrized the Rasch model for polytomous in a way where the probabilities of responses depend on differences between person parameters and item parameters and referred to it as a partial credit *model* (PCM). However, since the PCM is a trivial reparameterization of the EFM version we prefer to refer to it as the PCM version of the Rasch model for polytomous items.

The PCM version

Parameter	Canonical parameters	PCM parameters
Person effect	$\boldsymbol{\theta_v}$	$\boldsymbol{\theta_v}$
PCM threshold	$\boldsymbol{\sigma_{ix}}$	$\boldsymbol{\beta_{ix} = \sigma_{i,x-1} - \sigma_{i,x}}$

$$\Pr(X_{vi}=x|\theta_v) = \frac{\exp\left(x\theta_v - \sum_{j=1}^{x}\beta_{ij}\right)}{G_{vi}} = \frac{\exp\left(\sum_{j=1}^{x}(\theta_v - \beta_{ij})\right)}{G_{vi}} \qquad (1.9)$$

During analysis by the PCM version, we define the location β_i of the item as the average of the PCM thresholds, $\beta_i = \sum_{x=1}^{m_i}\beta_{ix} \Big/ m_i$.

Calculating a statistic as the average of a number of parameters does not necessarily define a statistic with a meaningful interpretation. However, it is not difficult to show that the location defined by the PCM is equal to minus the item effect defined by the ICE, $\beta_i = -\sigma_{i..}$ The location of a polytomous items is a meaningful measure of item effect that is comparable to the item effect defined by ICE in the same way that the difficulty of a dichotomous item relates to the easiness of the item.

1.4.7 Interpretation of item parameters

To understand how a polytomous item function we have to study the conditional distribution of item responses given different values of the person parameter and since there are no simple functions describing these distributions we have to calculate these differences ourselves. DIGRAM provide several ways to obtain these distributions, but in order to understand how the item parameters shape the distributions it is useful to recognize the relationships between the item parameters and a subset of the conditional item distributions.

At the origin of the θ scale where the multiplicative person parameter ξ is equal to 1, the distribution of an item is a simple function (1.10) of the score parameters of the PSD version.

At the location of the item, the distribution of the item is a simple function (1.11) of the multiplicative category effects and the probability of an item score equal to zero is the same as the probability of the maximum item score as show in Formula (1.12).

Finally, (1.12) shows that the PCM thresholds define the locations where adjacent categories have the same probability.

Conditional distributions of responses defined by item parameters

The score parameters of the PSD version (1.7) define the item distribution anchored at the origin of the θ scale

$$\Pr(X_i = x \mid \theta_v = 0) = \delta_{ix} \Big/ \sum_{z=0}^{m_i} \delta_{ix} \qquad (1.10)$$

The multiplicative category effects of the MICE (1.8) define the item distribution at the location of the item

$$\Pr(X_i = x \mid \theta_v = \beta_i) = \alpha_{ix} \Big/ \sum_{z=0}^{m_i} \alpha_{ix} \qquad (1.11)$$

Since the effect of the extreme categories are equal to 1 it follows from (1.11) that the probabilities of extreme categories are the same at the location of the item

$$\Pr(X_i = 0 \mid \theta_v = \beta_i) = \Pr(X_i = 0 \mid \theta_v = \beta_i) \qquad (1.12)$$

The PCM thresholds define location where the probabilities of adjacent categories are the same in the item distributions that are anchored at he PCM thresholds,

$$\Pr(X_i = x\text{-}1 \mid \theta_v = \beta_{ix}) = \Pr(X_i = x \mid \theta_v = \beta_{ix}) \text{ for } x = 1,\ldots,m_i \qquad (1.13)$$

To appreciate the category effects of ICE and MICE it is useful to compare how they characterize response categories and the way that the PCM version characterize the categories.

Let x be one of the categories of a polytomous item. The category effects of the ICE and MICE provide concrete information on the capacities of this category that we may compare to the

capacities of other categories of the same and other items. Compared to this, the PCM needs two statements to characterize category x

$$\Pr(X_i = x\text{-}1 \mid \theta_v = \beta_{ix}) = \Pr(X_i = x \mid \theta_v = \beta_{ix})$$
$$\&$$
$$\Pr(X_i = x \mid \theta_v = \beta_{i,x+1}) = \Pr(X_i = x+1 \mid \theta_v = \beta_{i,x+1})$$

Since (1.13) does not provide information on the size of the probabilities, the characterization of the category is opaque. For this reason, the PCM version is rarely helpful if you want to compare a category of a polytomous item with other categories. Therefore, you have to calculate the category effects of the ICE or MICE yourself if the PCM thresholds are the only item parameters available to you.

Fortunately, this is easy.

In terms of the PCM and MICE versions, the probabilities at the item location defined by (1.11) are

$$\Pr(X_i = 0 \mid \theta_v = \beta_i) = \frac{1}{1+\sum_{z=1}^{m_i} \exp(z\beta_i - \sum_{j=1}^{z} \beta_{ij})} = \frac{1}{\sum_{z=0}^{m_i} \alpha_{iz}}$$

and

$$\Pr(X_i = x \mid \theta_v = \beta_i) = \frac{\exp(x\beta_i - \sum_{j=1}^{x} \beta_{ij})}{1+\sum_{z=1}^{m_i} \exp(z\beta_i - \sum_{j=1}^{z} \beta_{ij})} = \frac{\alpha_{ix}}{\sum_{z=0}^{m_i} \alpha_{iz}}$$

because the category effect of extreme categories are equal to 1 in MICE.

From this, it follows the category effects defined by MICE and ICE are equal to

$$\alpha_{ix} = \exp(x\beta_i - \sum_{j=1}^{x} \beta_{ij}) \text{ and } \omega_{ix} = x\beta_i - \sum_{j=1}^{x} \beta_{ij} \qquad (1.14)$$

which should not be too difficult to calculate using a calculator app on a phone.

1.5 Graphical Rasch models and graphical log-linear Rasch models

Conventional Rasch models are IRT models that describe the conditional distribution of a set of locally independent items without DIF given the outcome θ of a latent variable, the value of which appears as a person *parameter* in the conditional distribution of items given the latent trait variable.

The models for item analysis by DIGRAM extend and generalize the conventional Rasch model in two ways. First, as a graphical Rasch model and second, as a graphical log-linear Rasch models.

1.5.1 The Rasch model as a graphical model

Chain graph models are multivariate recursive models defined by assumptions of pairwise conditional independence given all current and prior variables. The models are characterized by so-called Markov graphs where variables are represented by nodes and where a missing edge between two nodes mean that the two variables are conditionally independent. Since items are locally independent and because the sufficiency of the total score for the person parameter implies that items are conditionally independent of the latent variable given the total score we may regard the Rasch model as a chain graph model defined by two different Markov graphs.

We will use data in the CONF08 project with five polytomous items measuring confidence in public institutions in Denmark in 2008.

D – The police
E – The Parliament
G – The social security system
K – The health care system
L – The courts

Three assumptions characterize the Rasch model of these items.

1) The model is recursive because we assume that the relationship between the items and the latent variable is causal. The latent trait is the cause and responses to items are effects.

2) Items are locally independent.

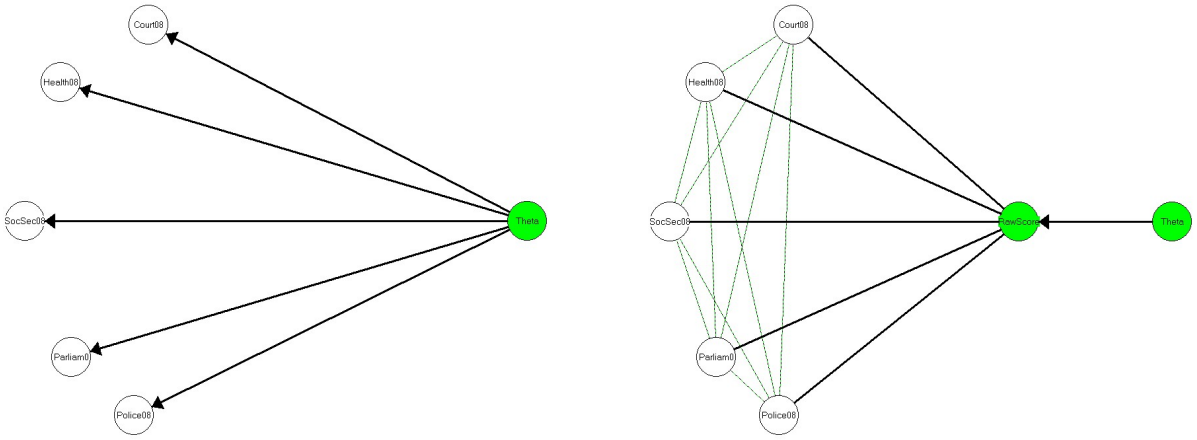
3) The person score is sufficient for the person parameter representing the outcome on the latent trait variable.

From these assumptions, follow two properties of Rasch models from which it follows that the Rasch model is a graphical model defined by two Markov graphs.

a) The first is that pairs of items are conditionally independent given the latent trait variable and the other items. E.g. that D ⊥ E | θ, G, K, L.

b) The second is that the vector of items is conditional independent of θ given the person score, (D,E,G,K,L) ⊥ θ | R = D+E+G+K+L because R is sufficient for θ.

It follows from a) that the Rasch model belongs to the family of graphical models. We refer to the Markov graph in Figure 1a as an IRT graph, because all conventional IRT models with locally independent items are graphical models defined by this graph. In addition to this, it follows from b) that the model including the person score is a graphical model defined by the Rasch graph in Figure 1b where the score separates items from θ. Since the Rasch model is the only IRT model with a sufficient person score it follows that the Rasch graph does not apply for other IRT models. Notice howevert, that items are not conditionally independent because conditioning with the raw score induce negative conditional association among items.

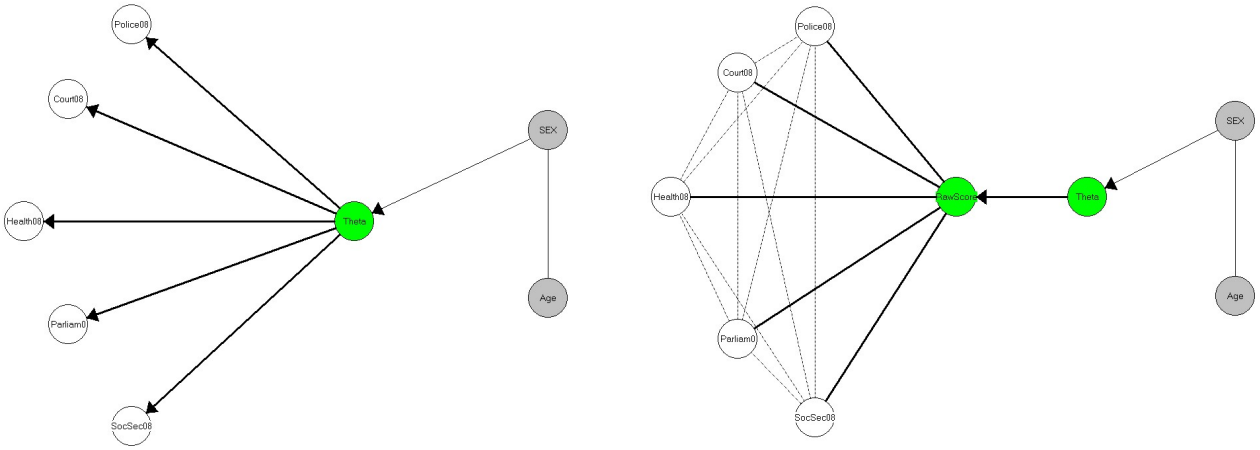


(a) The IRT graph (b) The Rasch graph

Figure 1.1 The two Markov graphs defining a Rasch model

1.5.2 Graphical Rasch models

During item analysis with Rasch model, it is customary to test for DIF relative to exogenous variables. The assumption of no DIF require that items are conditionally independent og exogenous variables given θ, but the analysis of DIF is informal relative to the Rasch model, because the Rasch model is a model without exogenous variables. Graphical Rasch models (GRMs) provide a riposte to this dilemma. Graphical Rasch models insert the Rasch model in a multivariate chain graph model together with all the covariates that may create DIF and other covariates that may be associated with the latent variable. To illustrate this here, Figure 1.2 defines a graphical Rasch model where sex and age are included as exogenous variables.



(a) The IRT graph (b) The Rasch graph

Figure 1.2. IRT and Rasch graphs of a graphical Rasch model

The GRM defined by Figure 1.2 the missing edges and pairs induce three sets of pairwise conditional independence

i. Items are locally independent
ii. There is no DIF among CONF items relative to Sex and Age because items are conditionally independent of Sex and Age given θ
iii. Items are conditionally independent of sex and age given the raw score
iv. Theta is conditionally independent of Age given Sex

i), ii) and iv) applies for all IRT models embedded in a graphical model with Sex and Age, but iii) requires a sufficient score and therefore a Rasch model.

Graphical Rasch models are similar to structural equation models (SEMs) for confirmatory factor analysis. Graphs also characterize SEMs, but the SEM graphs are not Markov graphs. A missing link between two variables of a SEM does not imply that the variables are conditionally independent.

1.5.3 Item analysis by Graphical Rasch models

Embedding the Rasch model in a graphical model including all the exogenous variables that may be sources of DIF is useful for two reasons. It suggests a simple test of no DIF in Rasch models that we may apply before we estimate the item parameters and it explains why spurious evidence of DIF is expected to turn up and how to distinguish between genuine and spurious DIF.

To test that there is no DIF of confidence in the Police relative to Sex we just have to count the three-way table with Police and Sex stratified by the Score and test that Police and Sex are conditional independent given the score. Had Police been a dichotomous item we could use the Mantel-Haenszel test[2]. Since items are polytomous we need tests defined for ordinal categories, but this is a minor technical issue. The point is that we only need to look at three-way tables to test for DIF in Rasch models.

Assume that there is DIF of confidence in the police relative to sex. The Markov graphs suggests that spurious evidence may turn op for two reasons.

Except for the tests of no DIF of the CONF items and the analysis of the dependence of the latent trait variable on Sex and Age there is little difference between conventional Rasch analysis and analysis by graphical Rasch models. Item parameters are estimated in the same way, the test of fit of the model include the same tests of item-fit and local independence, person parameters are estimated in the same way. Conventional Rasch analysis may use the same tests for no DIF, but DIF analysis is informal because the conventional Rasch model does not recognize exogenous variables.

The following tables present the CML estimates of the item parameters of the CONF items.

[2] The MH test of no DIF of dichotomous items is often proposed and used for general IRT models. The point is, that the MH-test is a Rasch test. No DIF insists that items are conditionally independent of exogenous variables given θ. In Rasch model, this implies that items are conditionally independent of exogenous variables given the score. Since this is not true for general IRT models without a sufficient raw score it follows that the MH test may provide spurious evidence of DIF, because the null-hypothesis of the MF test may be false even though there is no DIF.

Table 1.1 shows the distribution of item responses and in the manifest order of the items defined by the average item scores. The confidence is highest for the police and lowest for the social security system.

Table 1.1 Marginal distribution of confidence in public institutions.

	Degree of confidence				
Item	High (0)	Some (1)	Little (2)	No (3)	Average
G – Social security	0.082	0.601	0.295	0.021	1.26
E - Parliament	0.101	0.612	0.267	0.019	1.20
K – Health care	0.113	0.608	0.257	0.021	1.19
L – Courts	0.230	0.647	0.118	0.005	0.90
D – Police	0.265	0.649	0.083	0.003	0.82

The frequencies of Table 1.1 are the sufficient statistics for the item parameter. We use proportional fitting of these frequencies to obtain the CML estimates of the multiplicative parameters of the PSD version of the Rasch model for polytomous items. Table 1.2 puts these estimates on view.

Table 1.2 CML estimates of multiplicative item parameters (δ_{ix})

	Degree of confidence (Item scores)			
Item	High (0)	Some (1)	Little (2)	No (3)
G – Social secur.	1.000	70.883	98.992	7.035
E - Parliament	1.000	53.365	62.910	4.221
K - Health care	1.000	45.504	51.286	3.969
L – Courts	1.000	15.986	5.635	0.153
D - Police	1.000	12.680	2.889	0.056

DIGRAM calculated the parameters of Table 1.2 by iterative proportional fitting and used them to calculate the thresholds and the item location defined by the PCM (Table 1.3) and the item and category effects of the ICE and MICE (Table 1.4).

Table 1.3 CML estimates of PCM thresholds (β_{ix})

Item	1: High vs some	2: Some vs Little	3: Little vs None	Location
G – Social security	-4.261	-0.334	2.644	-0.650
E - Parliament	-3.977	-0.164	2.702	-0.480
K - Health care	-3.818	-0.120	2.559	-0.459
L – Courts	-2.772	1.043	3.609	0.627
D - Police	-2.540	1.479	3.951	0.963

Table 1.4 CML estimates of item and category effects. (ω_{ix}) and (α_{ix})

		Degree of confidence (Item scores)				
Version	Item	High (0)	Some (1)	Little (2)	No (3)	Item effect
	G – Social secur.	0.000	3.603	3.222	0.000	0.650
	E - Parliament	0.000	3.593	3.321	0.000	0.480
ICE	K - Health care	0.000	3.413	3.100	0.000	0.459
	L – Courts	0.000	3.462	3.067	0.000	-0.627
	D - Police	0.000	3.538	3.016	0.000	-0.963
	G – Social secur.	1	36.993	26.962	1	1.916
	E - Parliament	1	33.027	24.087	1	1.616
MICE	K - Health care	1	28.741	20.459	1	1.583
	L – Courts	1	29.914	19.730	1	0.534
	D - Police	1	33.222	19.835	1	0.382

The category effects draw the same picture for all items. The capacity is a little larger for "Some confidence" than for "Little confidence" and much larger for than the capacities of the extreme categories. The only way to interpret this is to assume that some and little confidence cover a wide range of different experiences with public institutions.

Attempting to interpret the PCM thresholds as statements on categories is less easy. The thresholds define locations where adjacent categories have the same probabilities. The probability of high confidence is equal to the probability of no confidence at the threshold between these categories. However, since the thresholds say nothing about the sizes of the probabilities, we find it difficult to compare the functioning of the items. Do they function in the same way or are some items better than other items? And why?

If we want to compare categories, the problem is the same. The category "Some confidence in the health care system" is characterized by two thresholds: -3.818 separating High and Some degree of confidence and -0.120 separating some and little degree of confidence whereas "Little confidence is characterized by -0.120 and 2.559. Since these threshold refer to all four categories is not obvious what they say about the "some confidence" and "little confidence" categories.

The best way to interpret the different parameters of the Rasch model for polytomous items is to examine the effects of the parameters on the distributions of the items. To do this, Formulas (1.10) – (1.13) may be useful.

It follows from formula (1.10) that the conditional probabilities of responses to items given $\theta = 0$ are proportional to the multiplicative parameters in Table 1.2. Table 1.5 presents these probabilities. They show that the items almost function as trinary items at the origin of the θ scale with close to no chance of a response in one of the extreme categories and with large probabilities of some or little confidence. Table (1.5) includes the mean and the item information at $\theta = 0$ defined by the variance of the item scores.

The large probabilities of some or little confidence are expected, but it is noteworthy that the probability of little confidence is larger than the probability of some confidence for three items even though the category effect of some confidence is stronger than the effect of little confidence.

Table 1.5 Conditional distribution of confidence in public distributions given $\theta = 0$.

	Confidence in public institutions					
Item	High (0)	Some (1)	Little (2)	No (3)	Mean	VAR/Inf
G – Social sec.	0.006	0.398	0.556	0.040	1.63	0.32
E - Parliament	0.008	0.439	0.518	0.035	1.58	0.33
K - Health care	0.010	0.447	0.504	0.039	1.57	0.34
L – Courts	0.044	0.702	0.247	0.007	1.22	0.27
D - Police	0.060	0.763	0.174	0.003	1.12	0.23

At the item locations, (1.11) claims that the probabilities of item responses are proportional to the multiplicative category effects of the MICE and (1.12) that the conditional probabilities of extreme item scores are the same. Table 1.6 confirms these claims.

Table 1.6 Conditional distribution of confidence in public distributions at the item locations.

	Location	Confidence in public institutions					
Item	θ	High (0)	Some (1)	Little (2)	No (3)	Mean	VAR/Inf
G	-0.650	0.015	0.561	0.409	0.015	1.42	0.30
E	-0.480	0.017	0.559	0.407	0.017	1.42	0.31
K	-0.459	0.020	0.561	0.400	0.029	1.42	0.32
L	0.627	0.019	0.579	0.382	0.019	1.40	0.32
D	0.963	0.018	0.603	0.360	0.018	1.38	0.31

At the item locations, the CONF items function as dichotomous items with little risk of an extreme category. The dichotomous distributions are close to uniform for items G, E and K and the item information a little better than a true dichotomous item could provide.

Finally, Formula (1.13) shows that the conditional probabilities of adjacent categories are the same at the locations defined by the PCM threshold that separating the categories. Tables 1.7 and 1.8 present these distributions for items G and D.

Table 1.7 The distribution of confidence in social security (G) at the pcm thresholds.

	Confidence in public institutions					
θ	High (0)	Some (1)	Little (2)	No (3)	Mean	VAR/Inf
-4.261	0.495	0.495	0.010	0.000	0.51	0.27
-0.334	0.010	0.483	0.483	0.025	1.52	0.32
2.644	0.000	0.025	0.488	0.488	2.46	0.30

Table 1.8 The distribution of confidence in the police (D) at the pcm thresholds.

	Confidence in public institutions					
θ	High (0)	Some (1)	Little (2)	No (3)	Mean	VAR/Inf
-2.540	0.496	0.496	0.009	0.000	0.51	0.27
1.479	0.009	0.476	0.476	0.040	1.55	0.35
3.951	0.000	0.040	0.480	0.480	2.46	0.30

The similarities between the probabilities of Table 1.7 and Table 1.8 are remarkable. Together with Tables 1.5 and 1.6, the conclusion is that except for the location of the items defined by the item effects, the five CONF items functions in the same ways and as close to dichotomous items in a wide range on the θ scale. To add to and modify this picture, we have to calculate the conditional distributions for other values of θ. The longer tour through the Rasch model will show you how to do that in DIGRAM.

1.5.4 Graphical log-linear Rasch models

A Log-linear Rasch model (Kelderman, 1984) doES not require that items are locally independent and without DIF. Despite this, it a member of the Family of Rasch models for measurement because the total score is sufficient for the person parameter θ. Log-linear Rasch models permit *uniform* DIF and *uniform* local dependence if the associations between items and exogenous variables do not depend on θ. To accommodate such situations, log-linear Rasch models include log-linear interaction parameters that do not depend on the person parameter representing DIF and LD.

Graphical log-linear Rasch models (GLLRMs) are log-linear Rasch models embedded in chain graph models in the same way that Rasch models insert conventional Rasch models in graphical models. Therefore, GLLRMs are generalizations of GRMs that may be helpful when item analyses have disclosed DIF and local dependence among items.

DIGRAM provides facilities for analysis by both GRMs and GLLRMs. Some of the methods capitalize on techniques associated with inference in chain graph models, but conventional inference in Rasch models is also supported and extended to inference in log-linear Rasch models.

In GLLRMs, the total score R is sufficient for θ in exactly the same way as in conventional Rasch models. Inference in GLLRMs can therefore be conditional in exactly the same way as in conventional Rasch models and all estimates and fit statistics that apply for the Rasch models also work for the GLLRMs. It is beyond the scope of this introduction to discuss the details of analysis by GLLRMs. For that purpose, we refer to a number of papers by Kreiner & Christensen (2002, 2004, 2006, 2007, 2011).

The tests of fit rejected the fit of the CONF items to a graphical Rasch model and disclosed evidence of both local dependence and DIF. Contrary to this, the graphical log-linear Rasch model defined by the IRT graph in Figure 1.3 survived all attempts to find evidence against the model.

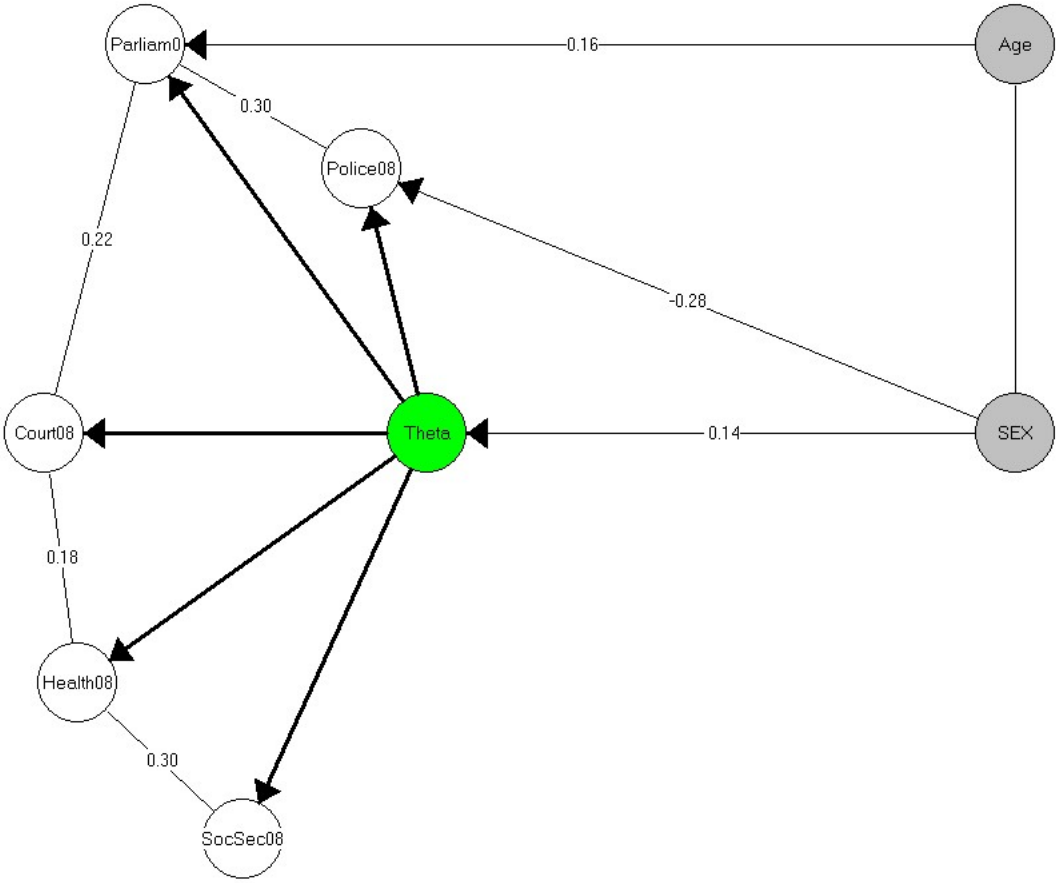


Figure 1.3. IRT graph of a graphical log-linear Rasch model.

The GLLRM defined by Figure 1.1 adds three sets of interaction parameters (one for each pair of locally dependent items and one for DHP36 and Sex) to the usual Rasch model structure in the following way

$$\Pr(D = d, E = e, G = g, K = k, L = l \mid \theta, SEX = y, Age = z) = \frac{\exp(r\theta + \sigma_d^D + \sigma_e^E + \sigma_g^G + \sigma_k^K + \sigma_l^L + \lambda_{gk}^{(GK)} + \lambda_{kl}^{(KL)} + \lambda_{el}^{(EL)} + \lambda_{de}^{(DE)} + \delta_{dy}^{D,SEX} + \delta_{ez}^{E,SEX})}{K} \quad (1.15)$$

where the σ parameters are the item parameters, λ parameters represent local dependence and the δ parameters are DIF parameters.

Models like (1.15) have many interesting features. The first is that the composite sum of items that are connected in the IRT graph have the same distribution as conventional polytomous Rasch items.

If it is convenient, we can define PCM thresholds and item and category effects of component scores in GLLRMs.

In addition to this, the CONF example illustrate that GLLRMs without locally independent items may fit data. That measurement can do without locally independence, also in cases, where there are no locally independent items. The total CONF score over all items has the same distribution as a score over independent Rasch items and estimation of person parameters proceeds in exactly the same way as in the Rasch model.

In this way, GLLRMs may provide solutions to measurement problems caused by the disagreement of the data and the pure Rasch model. There is, however, no guarantee that the solution works, so for this reason we have to check the GLLRM just as carefully as we checked the Rasch model. The sufficiency of the total score under the GLLRM means that we can do this in exactly the same way as for the Rasch model. The tours through the graphical log-linear Rasch models will illustrate how to do it on another data set. The CONF example is included with DIGRAM and we suggest that you check if the model defined by Figure 1.3 fits, on your own after the tours and compare the estimates of the person parameters defined by a conventional Rasch model to estimates by the GLLRM.

1.6 Publications on graphical (log-linear) Rasch models

The primary purpose of the guided tours is to show you how to analyse item response data in DIGRAM using graphical Rasch models and graphical log-linear Rasch models. We will add a few notes in the detours on technical details that are not covered elsewhere, but many technical details relating to these models are not discussed here because they have been documented in papers on these models and in the book by Christensen et.al. (2013) that is included in the following list of publications that includes a number of papers where GRMs and GLLRMs have been used.

1.6.1 Technicalities

Kreiner S (1987) Analysis of multidimensional contingency tables by exact conditional tests: Techniques and Strategies. *Scandinavian Journal of Statistics* 14, 97 - 112.

Kreiner S (1993/2006) Validation of Index Scales for Analysis of Survey data: The Symptom Index. In Bartholomew, DJ (ed) *Measurement* VOL III: 297-328

Kreiner S, Christensen KB. (2002) Graphical Rasch Models. In Mesbah et.al. (2002): *Statistical Methods for Quality of Life Studies. Design, Measurement and Analysis*: 169-184.

Kreiner S, Christensen, KB. (2004) Analysis of local dependency and multidimensionality in graphical log-linear Rasch models. *Communications in Statistics*, 33: 1239-1276

Kreiner S, Hansen M, Hansen CR (2006) On local homogeneity and stochastically ordered Mixed Rasch models. *Journal of Applied Psychological measurement*, 30: 271-297

Christensen KB, Kreiner S (2007) A Monte Carlo approach to unidimensionality testing in polytomous Rasch models. *Journal of Applied Psychological Measurement*, 31: 20-30

Kreiner S, Christensen KB (2007) Validity and Objectivity in health-related Scales: Analysis by Graphical Log-linear Rasch models. In von Davier & Carstensen (2007). *Multivariate and Mixture Distribution Rasch Models*: 329-346. Springer.

Kreiner S (2007) Validity and objectivity. Reflections on the role and nature of Rasch Models. *Nordic Psychology*, 59: 268-298

Kreiner S (2007) Determination of Diagnostic Cut-Points Using Stochastically Ordered Mixed Rasch Models. In von Davier & Carstensen (2007). *Multivariate and Mixture Distribution Rasch Models*; 131-146. Springer.

Christensen K.B. & Kreiner S. (2010) Monte Carlo tests of the Rasch model based on scalability coefficients. *British Journal of mathematical and Statistical Psychology*, 63, 101-111.

Kreiner, S & Christensen KA (2011) Item Screening in Graphical Log-linear Rasch models. *Psychometrika*,76, 228-256

Kreiner S, Christensen KB (2011) Exact evaluation of Bias in Rasch model residuals. *Advances in Mathematics Research*, 12, 19-40

Kreiner S (2011) Item-rest-score association. *Applied Psychological Measurement*, 35, 557-561

Christensen KB, Kreiner S, Mesbah M (eds.) (2013) *Rasch Models in Health.* London: ISTE Wiley

Nielsen, T & Kreiner S. Improving items that do not fit the Rasch models: exemplified with the physical functioning scale of SF36. *Pub.Inst.Stat.Univ.Paris*, 57, fasc. 1-2, 85-90

Kreiner S. (2013) A note on α-curves for dichotomous items. *Pub.Inst.Stat.Univ.Paris*, 57, fasc. 1-2, 91-108

Eusebi P, Kreiner S. (2015) Differential item functioning analysis by applying multiple comparison procedures. *Journal of Applied Measurement*. 16, 13-23.

Mûller M & Kreiner S. (2015) Item Fit Statistics in Common Software for Rasch Analysis. *Department of Biostatistics, Univ. of Copenhagen*. Research report 15/06.19 pages.

1.6.2 Applications

Kreiner S, Simonsen E, Mogensen J (1990) Validation of a Personality Inventory Scale: The MCMI P-Scale (Paranoia) *Journal of Personality Disorders*, 4: 303-311

Schultz-Larsen K, Kreiner S, Lomholt RK (2007) Mini-Mental Status Examination: A short form of MMSE was as accurate as the original MMSE in predicting dementia *Journal of Clinical Epidemiology* 60: 260-267

Schultz-Larsen K, Lomholt RK, Kreiner S (2007) Mini-Mental Status Examination: Mixed Rasch model item analysis derived two different cognitive dimensions of the MMSE *Journal of Clinical Epidemiology* 60: 268-279

Nielsen, T., Kreiner, S. & Styles, I. (2007). Mental self-government: Development of the additional democratic learning styles using Rasch measurement models. *Journal of Applied Measurement,* 2(8), pp. 124-148.

Nielsen, T. & Kreiner, S. (2011). Reducing the item number to obtain same-length self-assessment scales: A systematic approach using result of graphical loglinear Rasch modeling. *Journal of Applied Measurement. 12*(4).

Soee A-B L, Skov L, Kreiner S, Tornoe B, Thomsen LL (2013) Pain Sensitivity and Pericranial Tenderness in Children with Tension-Type headache: a Controlled Study. Jo*urnal of Pain Research*, 6, 425-434

Soee, A-B L, Thomsen, LL, Kreiner S, Tornoe, B & Skov, L. (2013) Altered pain perception in children with chronic tension-type headache: Is this a sign of central sensitization? *Cephalalgia*, **33**, 454-462

Ringsted TK, Wildgaard K, Kreiner S. (2013) Pain-related impairment of daily activities after

Thoraic surgery. *The Clinical Journal of Pain*, **29,** 791-799

Nielsen, T. & Kreiner S. (2013). Improving items that do not fit the Rasch model: exemplified with the physical functioning scale of the SF-36. *Annales de L'I.S.U.P. Publications de L'Institut de Statistique de L'Université de Paris*, *Numero Special, 57*(1-2), 91-108.

Kreiner S & Christensen, K B (2014) Analysis of Model Fit and Robustness. A New Look at the PISA Scaling model Underlying Ranking of Countries According to Reading Literacy. *Psychometrika*, 79, 210-231

Lauritsen M-B G, Kreiner S, Söderström M, Dørup J, Lous J. (2015) A speech reception in noise test for preschool children (the Galker-test): Validity, reliability and acceptance. *International Journal of Pediatric Otorhinolaryngolo*, 79, 1694-1701.

Thewes, B., Zachariae, R., Christensen, S., Nielsen, T., & Butow, P. (2015). The Concerns About Recurrence Questionnaire: Validation of a brief measure of Fear of Cancer Recurrence Amongst Danish and Australian Breast Cancer Survivors. *Journal of Cancer Survivorship*, *9*(1), 68-79. doi: 10.1007/s11764-014-0383-1

Lauritsen, M-B G, Söderström M, Kreiner S, Dørup J, Lous J. (2016) The Galker test of speech reception in noise; assocoations with background variables, middle ear status, hearing, and language in Danish preschool children. *International Journal of Pediatric Otorhinolaryngolo*, 80, 53-60.

Nielsen, T & Kreiner S. (2017) Course Evaluation for the purpose of development: What can learning styles contribute? *Studies in Educational Evaluation*, 54, 58-79

Sjö NM & Kreiner S. SEAM™. (2017) *Social-Emotional Assessment/Evaluation Measure. Dansk udgave*. Hogrefe Psykologisk Forlag A/S

Karstoft, K-I.; Nielsen, A.B.S. & Nielsen, T. (2017). Assessment of depression in veterans across missions: A validity study using Rasch measurement models. *European Journal of Psychotraumatology, 8*(1), 10 pages, http://dx.doi.org/10.1080/20008198.2017.1326798

.

Nielsen, T.; Makransky, G.; Vang, M. L. & Dammeyer, J (2017): How specific is specific self-efficacy? A construct validity study using Rasch measurement models. *Studies in Educational Evaluation, 57,* 87-97.

Lindkvist EB, Kristensen LJ, Sildorf SM, Kreiner S, Svensson J, Mose AH, Birkebæk N, Thastum M. (2018) A Danish version of self-efficacy in diabetes management: A valid and reliable questionnaire affected by age and sex. *Pediatricdiabetes*, 3, 544-542.

Rayce SB, Kreiner S, Damsgaard MT, Nielsen T, Holstein BE. (2018) Measurement of alienation among adolescents: construct validity of three scales of powerlessness, meaninglessness and social isolation. J*ournal of Patient-Reported Outcomes*, 2, Online

Andersen TJ & Kreiner S. (2018) Examining higher executive Functions in the Formation and Expression of Opinions: Preliminary Data on the RETEPP™ Test. C*linical Neuropschological Consultation*. 41 sider. Online

Nielsen T & Kreiner S. (2018) Measuring statistical anxiety and attitudes towards statistics: The development of a comprehensive Danish Instrument (HFS-R). *Educational Assessment and Evaluation*. Online.

Sjoe NM, Kiil A, Bleses D, Dybdal L, Kreiner S, Jensen P. (2018) Assessing strengths and difficulties in social development: a comparison of the Social Emotional Assessment Measure (SEAM) with two established developmental psychopathological questionnaires. E*uropean Journal of Developmental Psychology*.Online

Poulsen I, Kreiner S, Engberg AW. (2018) Validation of the Early Functional Abilities Scale: An Assessment of Four Dimensions in Early Recovery After Traumatic Brain Injury. *Journal of Rehabilitation Medicine*, 50, 165-172.

Hovaldt, H. B.; Nielsen, T. & Dammeyer, J. (2018). Are symptoms of Depression and sensory impairments linked? Validation of the Major Depression Inventory among elderly with acquired dual sensory loss. *Innovation in Aging, 2*(1), 1-11. DOI: 10.1093/geroni/igy010

Karstoft, K-I*.; Nielsen, T*. & Nielsen, A.B.S. (2018). Perceived danger during deployment: a Rasch validation of an instrument assessing the exposure to and witnessing of danger in a war zone. *European Journal of Psychotraumatology, 9 (1).* DOI: 10.1080/20008198.2018.1487224

Nielsen, T. (2018) The intrinsic and extrinsic motivation subscales of the Motivated Strategies for Learning Questionnaire: A Rasch-based construct validity study, Cogent Education, 5,(1), 1-19. DOI: 10.1080/2331186X.2018.1504485

Nielsen, T.; Dammeyer, J.; Vang, M. L. & Makransky, G. (2018). Gender Fairness in Self-Efficacy? A Rasch-Based Validity Study of the General Academic Self-Efficacy Scale (GASE), Scandinavian Journal of Educational Research, *62*(5), 664-681, DOI: 10.1080/00313831.2017.1306796

Pontoppidan, M.*; Nielsen, T*. & Kristensen, I. H. (2018). Psychometric properties of the Danish Parental Stress Scale: Rasch analysis in a sample of mothers with infants. *PLOS ONE 13*(11): e0205662. https://doi.org/10.1371/journal.pone.0205662 .

Amnitzbøll J, Skovgaard AM, Holstein BE, Andersen A, Kreiner S, Nielsen T. (2019) Construct validity of a service-setting based neasure to identify mental health problems in infancy. *PLOS|ONE*. Online.

Pedersen MAM, Kristensen LJ, Sildorf SM, Kreiner S, Svensson J, Mose AH, Thastum M, Birkebaek N. (2019) Assessment of family functioning in families with a child diagnosed with type 1 diabetes: Validation and Clinical relevance of the general functioning subscale of the Mcmaster family assessment device. *PediatriDiabetes*, 6, 785-793

Sjoe NM, Bleses D, Dybdal L, Nielsen H, Sehested KK, Kirkeby H, Kreiner S, Jensen P. (2019) Measurement Properties of the SEAM Questionnaire Using Rasch Analysis on Data from a Representative Danish Sample of 0- to 6-years-Olds. *Journal of Psychoeducational Assessment*, 37, 320-337.

Adroher ND, Kreiner S, Young C, Tennant A. (2019) Test equating sleep scales: applying the Leunbach's model. *BMC Medical research Methodology*.Online.

Bundsgaard J & Kreiner S. (2019) *Undersøgelse af De nationale Tests måleegenskaber*. Århus Universitet, DpU.

Sjoe NM, Bleses D, Dybdal L, Tideman E, Kirkeby H, Sehested KK, Kreiner S, Jensen P. (2019) Short Danish Version of the Tools for early Assessment in math (TEAM) for 3-6-Year-Olds. *Early Education and Development*, 2. Online.

Karstoft, K-I.; Nielsen, T. & Nielsen, A.B.S. (2019). Measuring social support among soldiers with the Experienced Post-deployment Social Support Scale (EPSSS): a Rasch-based construct validity study. *Behavioral Medicine.* DOI: 10.1080/08964289.2019.1676192

Nielsen, T. & Dammeyer, J. (2019). Measuring higher education students' perceived stress: an IRT-based construct validity study of the PSS-10. *Journal of Studies in Educational Evaluation, 63*, 17-25. https://doi.org/10.1016/j.stueduc.2019.06.007

Nielsen, T.; Friederichsen, I. S. & Hartkopf, B. T. (2019). Measuring academic learning and exam self-efficacy at admission to university and its relation to first-year attrition: an IRT-based multi-program validity study. *Frontline learning Research, 7*(3), 91-118.

Svensson J, Sildorf SM, Bøjstrup J, Kreiner S, Skrivarhaug T, Hanberger L, Petersson C, Åkesson K, Frøisland DH, Chaplin J. (2020) The DISABKIDS generic and diabetes-specific modules are valid but not directly comparable between Denmark, Sweden, and Norway. *PediatricDiabetes*, Vol 21, 900-908.

Sjoe, NM, Kiil, A, Bleses D, Dybdal L, Kreiner S Jesen P. (2019) Assessing strengths and difficulties in social development: a comparison of the Social Emotional Assessment Measure (SEAM) with two established developmental psychopathological questionnaires. *European Journal of Development Psychology,* Vol 17, 103-122

Nielsen T, Kreiner S, Teasdale, TW. (2020) Assessment of cognitive ability for the Danish army: Is a single score sufficient? *Scandinavian Journal of Psychology*. Vol 61, 161-167

Nair, R.; Dutt, A. & Nielsen, T. (2020). Construct validity of revised version of Ability in Behaviour Assessment and Interventions for Teachers (ABAIT). *Journal of Autism and Developmental Disorders, 59,* 1081–1087, DOI: 10.1007/s10803-019-04286-5

Nielsen, T. (2020). The Specific Academic Learning Self-efficacy and the Specific Academic Exam Self-Efficacy scales: construct and criterion validity revisited using Rasch models. *Cogent Education, 7*(1), 1-15. Article 1840009. DOI: 10.1080/2331186X.2020.1840009

Nielsen, T.; Pontoppidan, M. & Rayce, S. B. (2020). The Parental Stress Scale revisited: Rasch-based construct validity for Danish parents of children 2-18 years old with and without behavioral problems. *BMC Health and Quality of Life Outcomes,* 18, article 281 (2020).

Ribeiro Santiago, P.H., Nielsen, T., Smithers, L.G. Roberts, R. & Jamieson, L. (2020). Measuring stress in Australia: validation of the perceived stress scale (PSS-14) in a national sample. *Health Qual Life Outcomes 18,* 100. https://doi.org/10.1186/s12955-020-01343-x

Santiago, P. H. R.*; Nielsen, T.*; Roberts, R.; Smithers, L. & Jamieson, L. (2020). Sense of Control: can it be measured culturally unbiased across Australian aboriginal and non-aboriginal populations? Shared first authorship. *PLoS ONE*

Upegui-Arango. L. D.; Forkmann, T.; Nielsen, T.; Hallensleben, N.; Glaesmer, H.; Spangenberg, L.; Teismann, T.; Juckel, G. & Boecker, M. (2020) Psychometric evaluation of the Interpersonal Needs Questionnaire (INQ) using item analysis according to the Rasch model. *PLoS ONE*

Vindbjerg, E.; Carlsson, J.; Mortensen, E. L.; Makransky, G. & Nielsen. T. (2020). A Rasch-based validity study of the Harvard Trauma Questionnaire. *Journal of Affective Disorders*

Nielsen, T. & Santiago, P. H. R. (2020). Chapter 14: Using graphical loglinear Rasch models to investigate the construct validity of the Perceived Stress Scale. In Myint Khine (Ed.) *Rasch Measurement: Applications in Quantitative Educational Research.* Singapore, Springer Nature, pp. 261-281. ISBN: 978-981-15-1799-0

Nielsen, T. (2021b)**.** Psychometric evaluation of the Danish language version of the Field Practice Experiences Questionnaire for teacher students (FPE-DK) using item analysis according to the Rasch model. *PlOS ONE,* 16(10):e0258459. doi: 10.1371/journal.pone.0258459.

Nielsen, T. (2021a). Pre-Academic Learning Self-Efficacy revisited: Validation in the Danish Academy Profession Degree context and differences across degree programs. *Scandinavian Journal of Educational Research.* DOI: 10.1080/00313831.2021.1910559

Nielsen, T.; Friderichsen, I. S. & Rayce, S. B. (2021). Classification of Loneliness using the T-ILS: Is it as simple as it seems? *Scandinavian Journal Psychology, 62*(1):104-115. doi: 10.1111/sjop.12697

Nielsen, T. & Kreiner, S. (2021). Statistical Anxiety and Attitudes Towards Statistics: criterion-related construct validity of the HFS-R questionnaire revisited using Rasch models. *Cogent Education,* 8:1, DOI: 10.1080/2331186X.2021.1947941

Nielsen, T.**;** Martínez-García, I. & Alestor-García, E. (2021*)*. Critical Thinking of Psychology Students: A Within- and Cross-Cultural Study using Rasch models. *Scandinavian Journal of Psychology, 62*(3):426-435. DOI: 10.1111/sjop.12714

Vindbjerg, E.; Mortensen, E. L.; Makransky, G.; Nielsen. T. & Carlsson, J. (2021). A Rasch-based validity study of the HSCL-25. *Journal of Affective Disorders Reports,* ttps://doi.org/10.1016/j.jadr.2021.100096.

Nielsen, T. (2021). Measuring differences in Current Academic Learning Self-Efficacy (CAL-SE) and Current Academic Exam Self-Efficacy (CAE-SE) for social science students: a Rasch-based multi degree-program validity study. In Edith Braun, Rachelle Esterhazy & Robert Kordts-Freudinger (Eds.) *Research on Teaching and Learningin Higher Education*, Waxmann, Berlin, pp. 57-81.

Nielsen, T.; Martinez-Garcia, I. & Alastor, E. (2022). Exploring first semester changes in domain-specific critical thinking. *Frontiers in Education: Higher Education, Research topic "Generic Skills in Higher Education".* doi: 10.3389/feduc.2022.884635

Nielsen, T. (2022). Predicting Student Teacher's Academic Learning Self-Efficacy at the Second Semester from their Pre-Academic Learning Self-Efficacy. In J.J. Carmona (Ed.) *The Importance of Self-Efficacy and Self-compassion.* Nova Science Publishers, pp. 1-32.

Nielsen, T.; Martínez-García, I. & Alestor-García, E. (2022). Chapter 5: Psychometric properties of the Spanish translation of the Specific Academic Learning Self-Efficacy and the Specific Academic Exam Self-Efficacy scales in a higher education context. In Myint Swe Khine & Tine Nielsen (Eds.) *Academic Self-efficacy: Nature, Measurement, and Research.* Springer Nature.

1.7 Item analysis commands

Tables 1.7 shows commands providing information on the current version of DIGRAM and Table 1.8 shows the commands for item analysis that are illustrated in the guided tours. The following five types of commands are the most important:

1) ITEMS, FLIP, CUT, and EXO defines the set-up of the analysis
2) SHOW I and S provides information on distribution of items and scores
3) DIF, SCREEN I and S, CHECK I end D, and MDIF are used for analyses of the manifest variables of the models by tests of the global Markov properties of the models.
4) GRM, RASCH and PERSONFIT are used for parametric analyses of the models.
5) STABLE are used to create multivariate contingency table describing the joint distribution over score groups together with other variables.

You can use DIGRAM's graph module to display the IRT graphs and to redefine the models by adding or deleting edges and arrows between items and exogenous variables.

Table 1.7 Information on DIGRAM

Commands	Parameters	Purpose
Information on DIGRAM		
SHOW	**N**	Shows additions to the program since 2003
SHOW	**L**	Shows the current limitations of DIGRAM
SHOW	**E**	Provides information on the environment
SHOW	**P**	Provides a list of the DIGRAM projects that you have worked with
Information on commands		
HELP		Lists all available commands
command	**?**	Provides information on a specific command

Table 1.8 Commands for item analysis

Commands	Parameters	Purpose	Shown in section
Select and define variables			
ITEMS	variables	Selects items and defines scores and score groups	2.1.1
FLIP		Changes the orientation of items	2.2.1
IMISS		Provide info on missing item responses	2.1.1
CUT	Cut points	Redefines score groups	2.2.2
EXO	variables	Selects exogenous variables	2.1.2
Information on variables			
SHOW	**I**	Provides information on items	2.4.1
SHOW	**S**	Provides information on scores	2.4.1
Analysis of global Markov properties of Rasch models			
DIF	variables	Performs analyses of DIF	2.2.3
LDE	variables	Tests of local independence	2.2.3
SCREEN	**I**	Define a model by item screening	2.4.2
SCREEN	**J**	Performs item screening by do not define a model	2.4.2
SCREEN	**E**	Analyze the effect of exogenous variables on the score	2.4.2
CHECK	**I**	Check global Markov properties of the model	2.4.5
CHECK	**D**	Check the global Markov properties relating to DIF	2.4.5
COMP		Provide info on item components of the current GLLRM	2.4.7
Parametric analyses			
GRM	Generators	Analysis by graphical log-linear Rasch models	2.1.3
WML		Calculates the WML estimates of person parameters	
SAVE	**R**	Saves a command file with the definition of the current GLLRM so that you can easily return to this model if you want to continue the analyses	2.3.5
PERSONFIT		Analysis of response patterns and exact person fit test	2.4.6
IPR		Creates scale anchored item distributions	
SPR			2.2.7
TPR			
Score tables			
STABULATE	variables	Creates a contingency table containing the score groups and other variables.	

1.8 Abbreviations and acronyms

		Mentioned the first time on page
CML	Conditional maximum likelihood estimates	9
DE	Disinhibited eating	6
DHP	Diabetes Health Profile	6
DIF	Differential item functioning	5
EFM	Exponential family model	10
GLLRM	Graphical log-linear Rasch model	5
GRM	Graphical Rasch model	
ICE	The item and category version of the RM	12
IRT	Item response theory	19
LD	Local dependence	5
MICE	The multiplicative item and category version of the RM	14
ML	Maximum likelihood estimates	61
PF	Physical functioning	7
PSD	Power series distribution	9
RM	Rasch model	5
RMSE	Root mean squared error	61
WML	Weighted maximum likelihood estimate of person parameters	61
WPG	Weighted mean of the partial gamma coefficient testing the hypothesis of local independence during item screening	127

2 Guided tours

This chapter describes four tours through DIGRAM where you will get a chance to see what DIGRAM has to offer for item analysis by Rasch models.

We start with a very short tour, where you will learn how to select items and exogenous variable, how to estimate item parameters and person parameters, and how to perform a rudimentary check of whether the Rasch model provides a reasonable description of the distribution of item responses.

The next tour is somewhat longer. You will learn how to manipulate items and score groups, how to test that the model is able to explain observed correlations among item, how to test for unidimensionality, and how to assess how well the items actually target the study population. During this tour, we will also look at a number of graphical descriptions of the model: item and test characteristic curves, item maps, and IRT and Rasch graphs encapsulating the assumptions on which the Rasch model is built.

The third tour is also relatively short. During this tour, you will learn how to define log-linear Rasch models with uniform DIF and/or uniform local dependence and you will see that, there is no difference between inference in Rasch models and inference in graphical log-linear Rasch models. You estimate parameters and test the fit of models in exactly the same way that you did it for the standard Rasch models.

The fourth tour is a long tour through graphical log-linear Rasch modeling. It is during this tour that the differences between inference by Rasch models and inference by graphical log-linear Rasch models become apparent.

2.1 Rasch models. The very short tour

During the first tour where we use the Diabetes Health profile (DHP) project to illustrate the basics of item analysis by Rasch models. During this tour, you will learn how to

1) select items and exogenous covariates,
2) estimate the item parameters of the Rasch model,
3) test the model,
4) estimate the person parameters.

We need three DIGRAM commands on this tour: ITEMS, EXOGENOUS, and GRM. The GRM command invokes a dialog where you have to select among a number of options guiding your analysis and where output will be displayed. The dialog is described in Section 2.1.3 and shown in Figure 2.1.5. We suggest that you pay particular attention to this dialog since most of the item analysis will happen here.

2.1.1 Selecting items

Use the ITEMS command to select items. "**ITEMS ABCDE**" selects the DHP items, recodes item responses so that item are scored from zero to the number of categories of the items minus one, calculates the total score as the sum of item scores, and defines two score groups in such a way that the number of respondents with non-extreme scores is as close to being the same as possible in the two groups. Figure 2.1.1 shows DIGRAM's main form after selection of items. Two buttons ("IRT graph" and "Graphical Rasch models") have been enabled, a list of items is shown in the panel below the main form, and the current model recognized by DIGRAM is written in red below the "Graphical Rasch model" button. The IRT and Rasch graphs will be discussed in Section 2.2.4 during the longer tour through the Rasch models.

Limitations:

The Rasch models accepts items with different numbers of response categories, but DIGRAM expects all items to have the same number of categories. DIGRAM therefore proposes to redefine the project variables creating dummy categories that are never used, because DIGRAM knows how to handle such variables during the Rasch analysis.

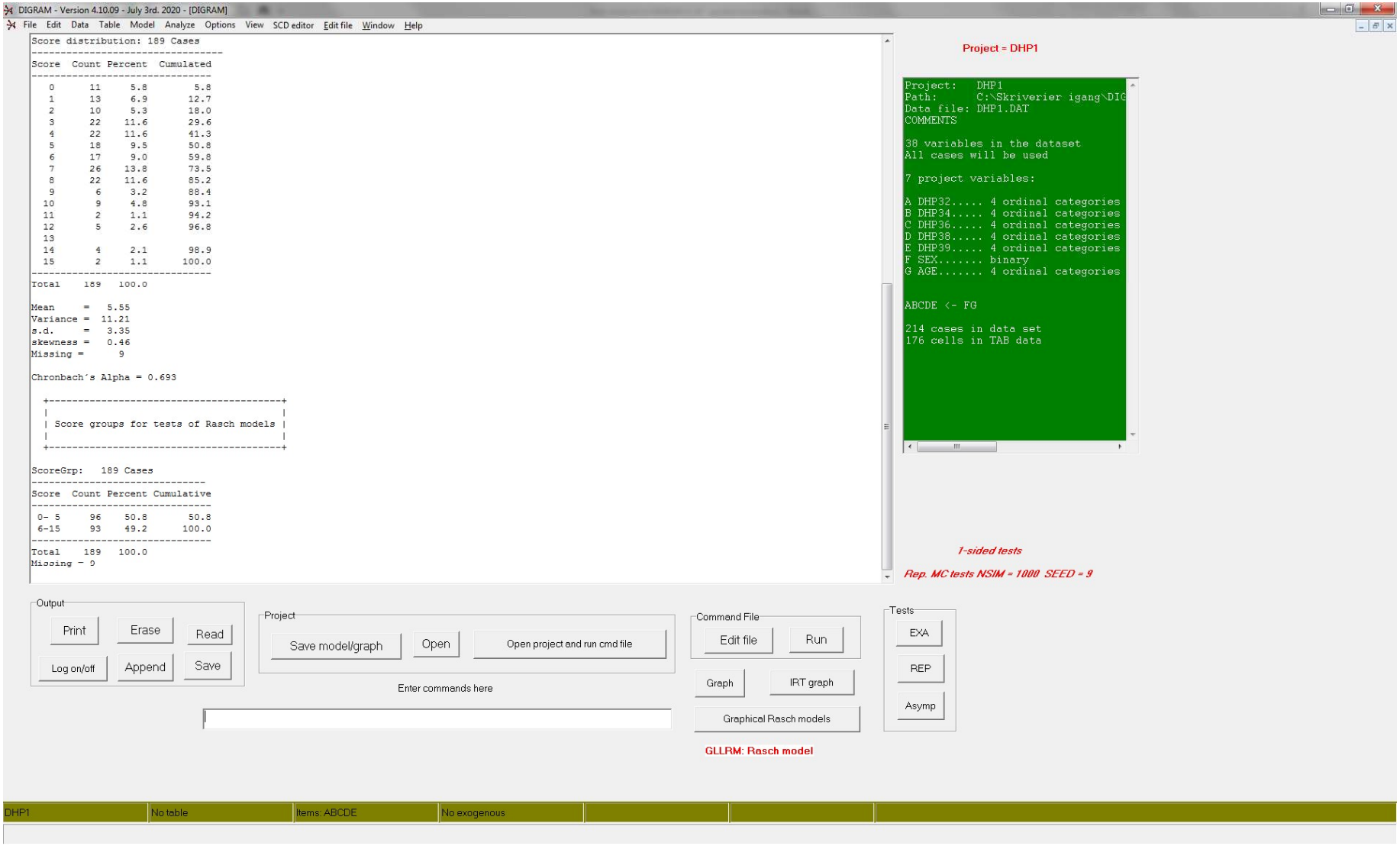


Figure 2.1.1 DIGRAM's main form after selection of items. The definition of the current GLLRM is written with red letters below the "Graphical Rasch model" button. When you select items, DIGRAM, assumes that the current model is a Rasch model.

Figures 2.1.2 and 2.1.3 show the output produced by item selection. Figure 2.1.2 provides information on items while Figure 2.1.2 gives information on the distribution of scores.

```
  +--------------------------------------+
  |                                      |
  | Variables selected for item analysis |
  |                                      |
  +--------------------------------------+

5 items: ABCDE
---------
A:    DHP32 - 4 ordinal categories.
B:    DHP34 - 4 ordinal categories.
C:    DHP36 - 4 ordinal categories.
D:    DHP38 - 4 ordinal categories.
E:    DHP39 - 4 ordinal categories.

Exogeneous variables have not been defined


  +-------------------------------------------+
  |                                           |
  | Average item scores and score distribution |
  |                                           |
  +-------------------------------------------+

                                  complete cases
  items      n       mean       mean    item range
-------------------------------------------------------
A:   DHP32  195     0.856     0.852      0 - 3
B:   DHP34  197     1.513     1.487      0 - 3
C:   DHP36  194     1.247     1.259      0 - 3
D:   DHP38  197     0.731     0.746      0 - 3
E:   DHP39  196     1.199     1.206      0 - 3

Obtainable score range:  0 - 15
```

Figure 2.1.2 Information on items

Information on average item scores is provided for all the persons that responded to an item and for persons with complete responses on all items. Use this information to check that item responses seem to be missing at random.

Use an **IMISS** command for information on missing items. Seven persons is missing response on one item and two on two items. Additional information on items is available if you invoke the "**SHOW I**" command. This will be described during the first detour in Chapter 3. We suggest that you wait with this detour until later.

Figure 2.1.3 shows the distribution of the score over all items.

```
Score distribution: 189 Cases
--------------------------------
Score  Count Percent  Cumulated
------------------------------
   0      11     5.8        5.8
   1      13     6.9       12.7
   2      10     5.3       18.0
   3      22    11.6       29.6
   4      22    11.6       41.3
   5      18     9.5       50.8
   6      17     9.0       59.8
   7      26    13.8       73.5
   8      22    11.6       85.2
   9       6     3.2       88.4
  10       9     4.8       93.1
  11       2     1.1       94.2
  12       5     2.6       96.8
  13
  14       4     2.1       98.9
  15       2     1.1      100.0
------------------------------
Total    189   100.0

Mean     =   5.55
Variance =  11.21
s.d.     =   3.35
skewness =   0.46
Missing =      9

Chronbach´s Alpha = 0.693

  +----------------------------------------+
  |                                        |
  | Score groups for tests of Rasch models |
  |                                        |
  +----------------------------------------+

ScoreGrp:   189 Cases
-----------------------------
Score  Count Percent Cumulative
------------------------------
 0- 5     96    50.8       50.8
 6-15     93    49.2      100.0
------------------------------
Total    189   100.0
```

Figure 2.1.3 Information on the score and score groups

The information on the score includes Cronbach's α. The score is missing if responses are missing for one or more items. 13 persons have extreme scores and 9 persons have missing responses on at least one item. Persons with extreme score can be of interest in themselves, but they do not provide information that can be used to estimate item parameters and test the fit of items to the Rasch

model. The data used for these purposes therefore consists of item responses from 176 persons. The skewness of the score distribution may influence the iterative procedure that DIGRAM uses to estimate item parameters. If it is positive as in Figure 2.1.3 where the distribution is right skewed, you should have no problems. If it is negative, you should consider changing the orientation of the items to save time and to avoid situations where the iterative procedure cannot find the estimates. (See Section 2.2.1 on how to do this).

In this example, a cut point equal to 5 defines two score groups where 85 (= 96-11) persons have scores between 1 and 5 and 91 (=93-2) persons have scores between 6 and 14. These score groups are used for tests of item homogeneity during the item analysis and in tables where you can examine the association between the score and other variables. During the next and somewhat longer tour, we will show you how to use the CUT command to redefine the score groups.

The second detour in Chapter 3 will describe the "SHOW S" command providing additional information on the score.

2.1.2 Selecting exogenous variables

Exogenous variables are covariates that we include in the model to test for DIF and for association with the latent variable. The DHP project has two exogenous variables, F = Age defined by four ordinal categories, 18-49, 50-59, 60-69, and 70-100 and G =Sex. To select these variables we invoke the "EXOGENOUS FG" command. Figures 2.1.4.a and 2.1.4.b show the results.

Figure 2.1.4.a provide a list of the exogenous variables and information on persons where some of the information on exogenous variables is missing. DIGRAM reports the number of cases that are lost for this reason and compares the mean scores for respondent with and without information on the exogenous variables. Missing outcomes on exogenous variables reduce the number of cases that are covered by the graphical Rasch model. For this reason, DIGRAM also shows the distribution of the score for all persons with complete responses to items and exogenous variables, but this table is not included in Figure 2.1.4a.

Next, DIGRAM, shows the recursive structure of the variables included in the model with items, the total score labelled '#', the latent variable labelled '¤', and the exogenous variables. The model assumes that items are located in the ultimate recursive block of the model. Exogenous variables,

appearing after items in the recursive structure defined by the DIGRAM project, are therefore included in the same recursive block as the items to enable analyses of DIF relative to these variables. However, evidence DIF relative to such variables should probably be interpreted as evidence of differential item functioning because the recursive project structure describe the association between items and these variables as asymmetric relationships pointing from items to the exogenous variables

```
  +---------------------------------+
  |                                 |
  | Overview of exogenous variables |
  |                                 |
  +---------------------------------+

2 Exogeneous variables:
-----------------------
F:      AGE - 4 ordinal categories.  Recursive level: 2
G:      SEX - 2 ordinal categories.  Recursive level: 2

   189 cases with complete item responses
   188 cases with complete item and exo responses

Frequency of missing values among cases with complete item responses

                    mean score  mean score
Variable    count   if missing   if known     t    p
-------------------------------------------------------
F:      AGE     1       14.0        5.5     35.40 0.000
G:      SEX     1       14.0        5.5     35.40 0.000

  +-----------------------------------------------------------+
  |                                                           |
  | Recursive structure among items and exogenous variables   |
  |                                                           |
  +-----------------------------------------------------------+

               ABCDE# <- ¤ <- FG
```

Figure 2.1.4a Exogenous variables

When exogenous variables have been selected, DIGRAM screens the relationships between the total score and the exogenous variables in order to identify the exogenous variables with a direct effect on the latent variable[3].

[3] Item screening including screening of the associations between the score and the exogenous variables are described in more details in Section 2.4.1 of these notes.

Figure 2.1.4.b shows the results for the DHP project. The first two lines with statistical tests show the test of marginal association between the score (labelled #) and sex (F) and age (G). Both tests provide significant evidence of marginal association. The next lines tests conditional independence of the score and one exogenous variable given the other. In both cases, the tests reject conditional independence. For this reason, DIGRAM concludes that both sex and age have direct effects on the DHP score.

```
  +-----------------------------------------------------------+
  |                                                           |
  | Analysis of the effects of exogenous variables on the score |
  |                                                           |
  +-----------------------------------------------------------+
2 variables with a marginal effect on the score: F G

                         p-values                 p-values (2-sided)
Hypothesis       X²  df asymp exact               Gamma asymp exact               nsim
-------------------------------------------------------------------------------------
#&F            48.9  42 0.215 0.214 (0.183-0.249) -0.26 0.000 0.000 (0.000-0.007) 1000 --
#&G            16.1  14 0.305 0.325 (0.288-0.364)  0.26 0.004 0.003 (0.001-0.012) 1000 ++
#&F|G          90.5  81 0.220 0.224 (0.192-0.260) -0.26 0.000 0.001 (0.000-0.008) 1000 --
#&G|F          47.9  42 0.245 0.220 (0.188-0.256)  0.21 0.035 0.045 (0.031-0.065) 1000 +
-------------------------------------------------------------------------------------
Benjamini Hochberg rejects if p <  0.025 for FDR = 0.05
                          and p <  0.004 for FDR = 0.01
Significance of
X²        xx : FDR = 0.01    x : FDR = 0.05
Gamma  ++/-- : FDR = 0.01  +/- : FDR = 0.05
```

Figure 2.1.4b Screening of the association between the score and the exogenous variables

If you need to select additional or other exogenous variables, you have to use the EXOGENOUS command again. When you do this, the current set of exogenous variables will be disposed and replaced with the new set. If you for some reason just want to get rid of the current set of exogenous variables, you must use a "DISPOSE E" command. In both cases, DIGRAM reinitializes the graphical Rasch model and the results obtained during the item analysis with the previous set of exogenous variables have to be recalculated.

2.1.3 Item analysis

Invoke the GRM command without parameters or click on the "Graphical Rasch model" button to initiate the item analysis. When you do this, the GRM dialog shown in Figure 2.1.5 turns up following which, DIGRAM estimates the item parameters and shows the results on the large output field at the right side of the GRM dialog.

Before proceeding, you should familiarize yourself with the GRM dialog form.

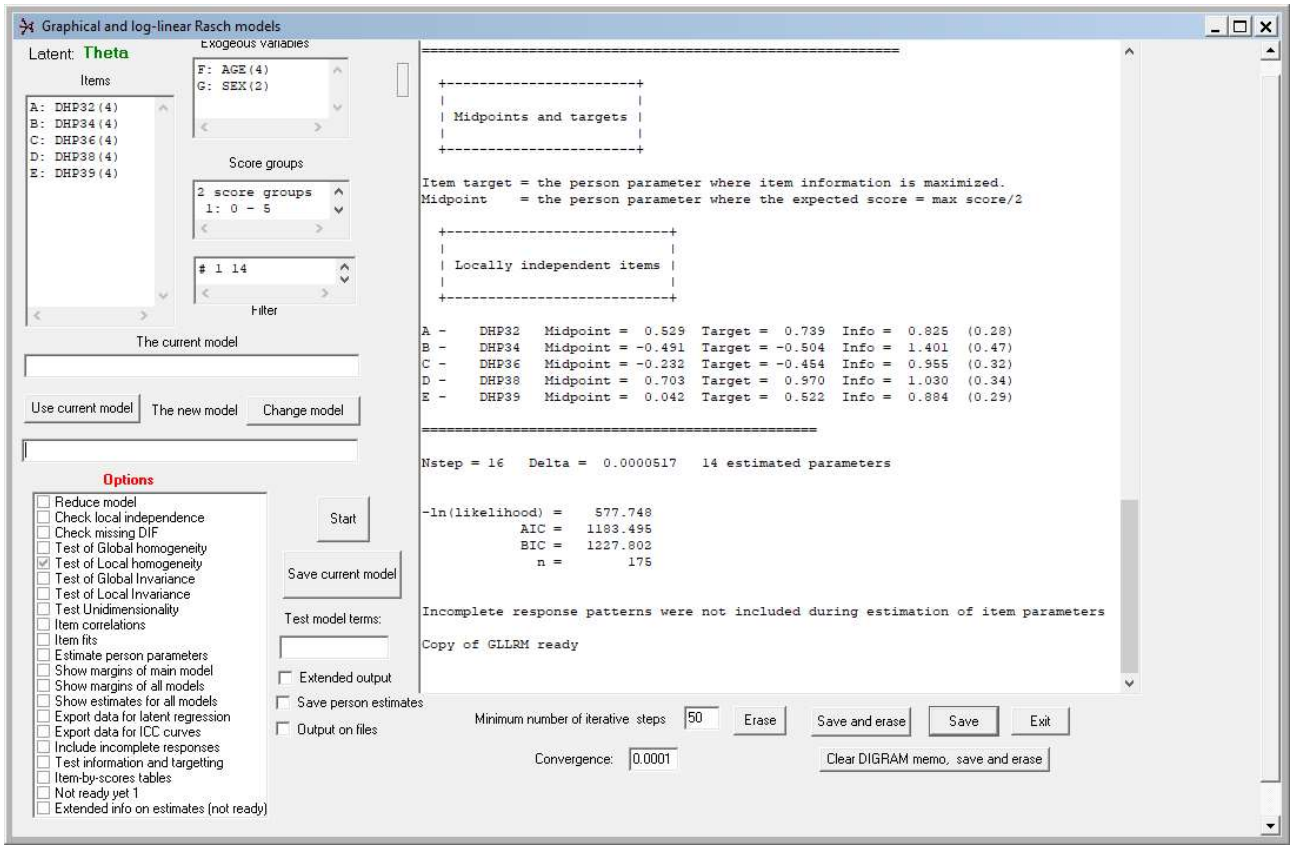


Figure 2.1.5 The GRM dialog. Item parameters are estimated when the GRM dialog is invoked.

At the upper left corner, the GRM dialog provides information on items and exogenous variables.

You cannot edit the information on items and exogenous variables from within the GRM dialog, except for the filter that defines the cases to be used during the item analysis. The default filter tells DIGRAM to consider persons with non-extreme scores (persons with scores larger than zero and less than 15 which is the maximum score on the 5 items). You can delete, change and add other filters as you wish[4] if you want to restrict the analysis to a specific subset of persons.

Below the information on items and exogenous variables, you find two fields with information on two models. You can use these fields to define graphical log-linear Rasch models with DIF and local dependence. Since this tour is about the pure Rasch models with locally independent items

[4] The format of a filter has to be "variable min max. Since # is the label for the score,. "# 1 14" means that the score should be in the [1,14] range.

without DIF, we do not need these fields for now, but return to them later. Below the models, the GRM dialog offers a number of options that you may select to guide your analysis.

The GRM dialog has a number of buttons and additional options. Use the Start button when you have selected decided on the options that you need. If no options are selected DIGRAM will estimate the item parameters of the model (unless it already has done so, in which case it will tell you that the model is the same as before). Output generated during the analysis will appear at the output field of the GRM dialog form. The rest of the buttons and options will be described below Figure 2.1.5.

Table 2.1.1 summarize the buttons and the analysis options on the GRM dialog. The first buttons relating to GLLRMs and will not be used during analysis by conventional Rasch models.

Table 2.1.1 GRM buttons

Button	Function
Use current model	Copies the current model to the new model
Change model	Copies the new model to the current model
Start	Estimate the model and initiate analysis
Save current model	Save the current model on a command file
Erase	Erase the GRM output memo
Save and Erase	Save the GRM output on the DIGRAM memo and erase the GRM output
Save	Save the GRM output on the DIGRAM memo
Exit	Exit the GRM dialog
Clear DIGRAM memo, save and erase	Clear the output on the DIGRAM memo and save and erase the GRM output

Table 2.1.2 summarize the GRM dialog's analysis options and refer where you can find information on them in these notes. In addition to using buttons and selecting analysis options, you can also ask for extended output and output on files. Table 2.1.2 include information on whether these options are available, but these tours provide no or little information on what you may obtain if you ask for it.

Table 2.1.2 Analysis options

Option	Described in Section	Extended output	Output on files
Reduce model			
Check local independence	2.1.3.5	yes	yes
Check missing DIF	2.1.3.4	yes	yes
Test of global homogeneity	2.1.3.2	no	yes
Test of local homogeneity	-		
Test of global invariance	2.1.3.2	yes	yes
Test of local invariance	-		
Test unidimensionality	2.2.5.1	yes	no
Item correlations	-		
Item fits	2.1.3.3	yes	yes
Estimate person parameters	2.1.3.6	yes	yes
Show margins of main model	-		
Show margins of all models	-		
Show estimates for all models	-		
Export data for latent regression	-		
Export data for ICC curves	-		
Include incomplete responses	-		
Test information and targeting	-		
Items-by-scores tables	-		
Not ready yet 1	-		
Extended info on estimates	2.1.3.1	no	no

2.1.3.1 Estimating item parameters

DIGRAM uses iterative proportional fitting to estimate the item parameters. The default options are to use at least 50 steps before it gives up and to stop when the largest difference between the observed and expected sufficient marginal is less than 0.0001. In cases with pure Rasch models, iteration always use less the 50 steps to reach the level of precision. For instance, Figure 2.1.5 shows that DIGRAM used 16 steps to estimate the item parameters of the DHP items and that the largest difference between the observed and expected item margins was equal to 0.0000517. However, feel free to change the limits if you are busy or if you want to increase the precision of the estimates.

DIGRAM calculates conditional maximum likelihood estimates of item parameters since these estimates are known to be consistent and do not require any assumptions on the distribution of the

person parameters. If item parameters have not been estimated, you have to click on “Start”. You may also choose some of the options if you want to do more than this, but you are not required to do so if you only want to have a look at the item parameters. Figures 2.1.6 - 2.1.9 show the results.

```
  +-----------------------------------------------+
  |                                               |
  | Analysis by graphical loglinear Rasch models  |
  |                                               |
  +-----------------------------------------------+

The model is estimated under the following assumptions:

   1 <= raw score <= 14
   1 <= F <= 2
   1 <= G <= 4

The assumed model:

All items are assumed to be locally independent
No DIF

175 valid cases included

Nvalid      = 175
Nincomplete = 9
Nresponses  = 4
Nuseful     = 8
Nuseless    = 1


Observed score range: 0 - 15

  +-----------------------------------------------------+
  |                                                     |
  | Estimated item parameters: complete item responses  |
  |                                                     |
  +-----------------------------------------------------+

       item          0         1         2         3
-----------------------------------------------------
A:    DHP32        1.000     2.048     0.862     0.272
B:    DHP34        1.000     1.073     2.103     4.171
C:    DHP36        1.000     1.365     3.377     1.308
D:    DHP38        1.000     1.903     0.286     0.230
E:    DHP39        1.000     9.285     2.479     2.932
```

Figure 2.1.6 The multiplicative items parameters of the PSD[5] version of the Rasch model

Section 1.4 describe five ways to parameterize the Rasch model, but the default analysis only present four of those ways. If you want to see the item parameters defined by the EFM[6] version of

[5] The PSD version of the Rasch model parameterizes the item distributions as power series distributions.
[6] The EFM version parameterizes item distributipns as exponential family distributions

the model, you have to select the "**Extended info on estimates**" option that also provide the so-called Thurstonian thresholds.

DIGRAM use proportional fitting to estimate the multiplicative PSD parameters and transform these estimates to the parameters defined by the PCM, ICE and MICE version.

Figure 2.1.7 shows the PCM thresholds. Disordered thresholds are of concern to some users of Rasch models, and DIGRAM include a ">"between thresholds if they are disordered[7]. Four out of five DE items have disordered thresholds and DHP34 have completely disordered thresholds[8].

```
  +----------------+
  |                |
  | PCM thresholds |
  |                |
  +----------------+

The location of the item is the average of the PCM thresholds
Locations may be meaningless if thresholds are disordered or
+/- infinity

  +---------------------------+
  |                           |
  | Locally independent items |
  |                           |
  +---------------------------+

  item             1        2        3    Location
-------------------------------------------------
A:    DHP32    -0.717    0.865    1.153     0.434
B:    DHP34    -0.071 > -0.673 > -0.685    -0.476
C:    DHP36    -0.311 > -0.906    0.949    -0.089
D:    DHP38    -0.643    1.896 >  0.218     0.490
E:    DHP39    -2.228    1.321 > -0.168    -0.359
```

Figure 2.1.7 The partial credit thresholds and information on item locations

The locations of the polytomous items are related to the item effects defined by the ICE version of the model shown in Figure 2.18

[7] Disordered thresholds have implications for the estimates of the person parameter, but is not evidence against the model and therefore not evidence against the validity and objectivity of measurement.
[8] DIGRAM offers no facilities for collapsing of categories during the item analysis. If you want to analyse items with collapsed response categories you must define a new DIGRAM project with the collapsed versions of the items.

```
  +---------------------------+
  |                           |
  | Item and category effects |
  |                           |
  +---------------------------+

  item              0        1        2        3   Item effect
-------------------------------------------------------------
A:    DHP32       0.000    1.151    0.719    0.000    -0.434
B:    DHP34       0.000   -0.405   -0.209   -0.000     0.476
C:    DHP36       0.000    0.222    1.038   -0.000     0.089
D:    DHP38       0.000    1.134   -0.272   -0.000    -0.490
E:    DHP39       0.000    1.870    0.190    0.000     0.359

              ---- MICE effects ----

A:    DHP32       1.000    3.160    2.053    1.000     0.648
B:    DHP34       1.000    0.667    0.812    1.000     1.610
C:    DHP36       1.000    1.249    2.824    1.000     1.094
D:    DHP38       1.000    3.107    0.762    1.000     0.613
E:    DHP39       1.000    6.487    1.210    1.000     1.431
```

Figure 2.1.8 Estimates of item and category (ICE) effects

In addition to the item location, DIGRAM offers three summary item statistics - the item midpoint, the item target, and the item information at target. Figure 2.1.9 shows these statistics.

The item midpoint is the value of the person parameter where the expected item score is equal to half the maximum item score, the item target is the person parameter where item information is maximized and target info is equal to the item information at the target.

The number in the parentheses following the item info is equal to target info divided by the maximum item score. If this number is close to 0.25, the target info corresponds to the information provided by three dichotomous items targeted at the item target of the polytomous item. In this example, some item provide much more information than targeted dichotomous items can offer.

Locations, midpoints and targets are often close to each other and always the same for dichotomous[9] items, but Figure 2.19 show that there are examples with considerable differences. The location of DHP36 is -0,089, the midpoint is -0.232 and the target is 0.454. This has

[9] Locations and midpoints are always the same for trinary items, but trinary items may have two targets lying far away from the location.

implications for the ranking of items according to these statistics. The rank order according to the location is B, E, C, A, E, while the rank order according to the midpoints is B, C, E, A, D.

The target of DHP36 is equal to -0.45 where the item information is equal to 0.96 and the target of DHP32 is equal to 0.74. In this case, the rank order of items from low to high target values is the same as for the midpoints while the rank order from low to high location is B, E, C, A, D.

```
  +-----------------------+
  |                       |
  | Midpoints and targets |
  |                       |
  +-----------------------+

Item target = the person parameter where item information is maximized.
Midpoint    = the person parameter where the expected score = max score/2.

  +---------------------------+
  |                           |
  | Locally independent items |
  |                           |
  +---------------------------+

A -    DHP32   Midpoint =  0.529  Target =  0.739  Info =  0.825  (0.28)
B -    DHP34   Midpoint = -0.491  Target = -0.504  Info =  1.401  (0.47)
C -    DHP36   Midpoint = -0.232  Target = -0.454  Info =  0.955  (0.32)
D -    DHP38   Midpoint =  0.703  Target =  0.970  Info =  1.030  (0.34)
E -    DHP39   Midpoint =  0.042  Target =  0.522  Info =  0.884  (0.29)
```

Figure 2.1.9 Item midpoints and targets

In addition to the estimates of the item parameters, DIGRAM includes information on how long it took to calculate the estimates, the logarithm of the likelihood function and the AIC and BIC information criteria. These results can be seen in Figure 2.1.5.

2.1.3.2 Overall tests of fit

DIGRAM uses Andersen's (1973) conditional likelihood ratio (CLR) test for overall tests of homogeneity and invariance. The test of homogeneity compares item parameter test in the two score groups defined during selection of items[10], whereas the tests of no DIF compares item parameters in groups defined by the exogenous variables.

[10] You can redefine the score groups if you are not satisfied with DIGRAM's initial proposal. Section 2.2.2 describes how to do that.

To invoke these tests, you must select the "**Test of global homogeneity**" and "**Test of global invariance**[11]" options. The results are collected in a small table presented at the end of the output produced during the calculation of the tests. Figure 2.1.10 shows the results.

```
Summary of global test results. Delta will
be reported if estimation did not converge.

             CLR   df   p         delta
-------------------------------------
scoregroups   27.2  14 0.018
F:      AGE   42.4  42 0.454
G:      SEX   23.7  14 0.050

Critical levels adjusted by the Benjamini-
Hochberg procedure:

 FDR = 0.05          reject if p<=  0.0500
 FDR = 0.01          reject if p<=  0.0033
 FDR = 0.001         reject if p<=  0.0003
```

Figure 2.1.10 Overall test of homogeneity and invariance of the Rasch model

In most cases, there will be no reason to look at anything but the final table containing the CLR tests where the critical limits of the tests has been adjusted for multiple testing by the Benjamini-Hochberg procedure. Situations may occur, however, where the rest of the output produced during the calculation of the tests could be of interest. For instance, the weak evidence against the hypothesis of homogeneity ($p = 0.018$) [12] motivates a closer look at what happened during the calculation of the test comparing estimates for persons with low and high scores. Figure 2.1.11 compares observed and expected average item scores together with standardized residuals in the two score groups. The results suggest that the reason for the weakly significant CLR test could have something to do with item DHP38 where the observed items scores are lower than expected among persons with a score between 1 and 5.

Taken by itself, the weak evidence against homogeneity and the significant residual for item DHP38 does not support the conclusion that DHP38 does not fit the Rasch model. Before this conclusion is drawn, we have to look at the item fit statistics described in Section 2.1.3.3.

[11] Measureement is invariant if the set of item function in the same way in different subpopulations.
[12] To us, weak evidence, e.g if $0.01 < p < 0.05$, is not enough to reject the model unless evidence of more specific problems turn up.

```
****  Score = 1 - 5  ****

Observed and expected item mean scores

                        mean
    item          n   obs   exp    res
-------------------------------------
A -    DHP32     85 0.612 0.502   1.74
B -    DHP34     85 0.659 0.714  -0.61
C -    DHP36     85 0.812 0.722   1.07
D -    DHP38     85 0.294 0.429  -2.36 -
E -    DHP39     85 0.882 0.892  -0.15

****  Score = 6 - 14  ****

Observed and expected item mean scores

                        mean
    item          n   obs   exp    res
-------------------------------------
A -    DHP32     90 1.122 1.226  -1.29
B -    DHP34     90 2.400 2.347   0.59
C -    DHP36     90 1.778 1.862  -1.00
D -    DHP38     90 1.189 1.062   1.63
E -    DHP39     90 1.600 1.591   0.11

Test of homogeneity of 2 score groups. 14 parameters

         CLR =   27.16  df =  14  p = 0.0184
```

Figure 2.1.11 Analysis of homogeneity of item responses

A final comment. The conditional likelihood ratio test is the fundamental overall fit statistic for the Rasch model. It addresses the same fit issue as other overall fit statistics like, for instance, the test of no item-trait interaction in RUMM, but it is based on solid statistical footing with well-known asymptotic properties as sample sizes increase towards infinity. Section 2.1.3.5 presents an overall fit statistic based on a completely different approach where asymptotics is a serious issue.

2.1.3.3 Item fit statistics

Select "**Item fits**" if you want to check whether the separate items fit the Rasch model.

DIGRAM calculates three item fit statistics. Outfits and Infits and item-rest-score gamma (IRS γ) coefficients measuring the correlation between items and rest scores without the items by Goodman and Kruskal's gamma coefficient.

Outfits and Infits are well known and much used item fit statistics going back to the early days of the theory of Rasch models. DIGRAM calculates *conditional* Outfits and Infits comparing observed item responses to the expected responses under the conditional distribution of responses given the total score. The advantage of using conditional Outfits and Infits is that they avoid bias and provide more realistic assessment of significance than conventional Outfits and Infits (see Kreiner & Christensen, 2011b for details). Both fit statistics have expected values equal to one under the Rasch model. Fit statistics above one indicate weaker item discrimination than expected under the Rasch model, whereas fit statistics below one suggest that the item discrimination is too strong for the items in a Rasch model.

The IRS γ compares the observed correlation between the score of a separate item and the total score on all other items to the expected score under the Rasch model. To make sure that the estimates and test statistics are consistent and unbiased, the significance of this coefficient is also assessed under the conditional distribution of item responses given the total score on all items (see Kreiner (2011) for details).

The issue of item discrimination is important during item analysis by general IRT models. Item discrimination is a question of the steepness of the item characteristic curves because the steepness determines how well an item is able to distinguish between persons. The degree of item discrimination is known to be the same for all dichotomous Rasch items, but not for polytomous Rasch items where ICC curves are rather flat when there are large differences between PCM thresholds and steeper if there is little difference or even disorder among thresholds. The expected IRS γ in Figure 2.1.2 take this into account.

Figure 2.1.12 provides evidence against the Rasch model. The observed correlation between DHP38 and the rest score without DHP38 is much stronger than expected by the Rasch model. The significant Outfit of DHP38 agree and significant Infits and Outfits suggest that the item discrimination of DHP32 and DHP34 are weaker than expected by the Rasch model. However, DIGRAM dismiss the significance of these statistics by adjusting for multiple testing.

```
  +--------------------------------+
  |                                |
  | Conditional outfits and infits |
  |                                |
  +--------------------------------+
              Outfit                          Infit
Item          observed     sd       p       observed     sd       p
-------------------------------------------------------------------------
A -    DHP32    1.237    0.107  0.02750       1.246    0.108  0.02249
B -    DHP34    1.042    0.146  0.77173       0.910    0.107  0.39952
C -    DHP36    1.260    0.110  0.01828       1.190    0.094  0.04391
D -    DHP38    0.754    0.125  0.04840       0.815    0.128  0.14664    high
E -    DHP39    0.937    0.128  0.62388       0.923    0.119  0.51693
-------------------------------------------------------------------------

  +-----------------------------+
  |                             |
  | Item rest-score association |
  |                             |
  +-----------------------------+

                  Item-rest-score gamma
Item           observed expected    sd      p
-------------------------------------------------
A -    DHP32    0.305    0.426    0.068  0.07385
B -    DHP34    0.500    0.465    0.060  0.56575
C -    DHP36    0.345    0.445    0.062  0.10452
D -    DHP38    0.686    0.432    0.072  0.00039**  high
E -    DHP39    0.522    0.470    0.070  0.45460
-------------------------------------------------

Critical levels adjusted by the Benjamini-Hochberg procedure:
* < 5 % FDR, ** < 1 % FDR, *** = FDR < 0.1 % FDR
```

Figure 2.1.12 Item fit statistics. The assessment of significance is adjusted by the Benjamini-Hochberg procedure controlling the false discovery rate at 5 % (*), 1 % (**) and 0.1 % (***)

Evidence of discrepant item discrimination is evidence against the fit to the Rasch model, but the interpretation of the departures from the Rasch model depends on whether item discrimination is weaker or stronger than expected by the Rasch model. Weaker discrimination is expected for corrupt items that do not relate exclusively to the latent variable. In such cases, items are often eliminated. Evidence of too strong item discrimination does not support such interpretations. To some, it suggests that the Rasch model should be abandoned in favour of another type of IRT model. While this may be one way to address the problem, we in general avoid taking this step until further investigation has confirmed that the evidence is not caused by other types of violations of the assumptions of Rasch models (e.g. by local dependency, and/or DIF) that because such problems require very different kinds of solutions.

2.1.3.4 Tests of no DIF

DIGRAM uses Kelderman's (1984) test of no DIF to test that there is no DIF relative to the two exogenous variables. To obtain these estimates you must select "**Check missing DIF**" from the list of options. Figure 2.1.13 shows the results. DIGRAM's default is to print the significant evidence of DIF. To obtain a complete list of the tests of DIF you have to check the "Extended output" int the GRM dialog.

```
Check assumptions of no DIF

A & F:   lr =    9.71  df =   9  p = 0.3745
B & F:   lr =    5.68  df =   9  p = 0.7710
C & F:   lr =   13.08  df =   9  p = 0.1592
D & F:   lr =    3.39  df =   9  p = 0.9470
E & F:   lr =    9.91  df =   9  p = 0.3577
A & G:   lr =    1.47  df =   3  p = 0.6890
B & G:   lr =    3.19  df =   3  p = 0.3637
C & G:   lr =   12.72  df =   3  p = 0.0053
D & G:   lr =    5.47  df =   3  p = 0.1406
E & G:   lr =    4.93  df =   3  p = 0.1773

Benjamini & Hochberg rejects at 0.00500
```

Figure 2.1.13 Tests of no DIF. The evidence of DIF is adjusted for multiple testing by the Benjamini-Hochberg procedure controlling the false discovery rate at 5 %.

Figure 2.1.13 discloses evidence of DIF for item C (DHP36) relative to G (Sex). However, adjustment for multiple testing suggests that the evidence should be discarded, confirming the results of the over-all test of invariance of DHP in Figure 2.1.10.

2.1.3.5 Tests of local independence

DIGRAM uses Kelderman's (1984) test of local independence to test that the assumption of local independence of the Rasch model is not violated. To obtain these estimates you must select the "**Check local independence**" options. DIGRAM will only show the significant test results unless you have checked the "Extended output" box.

Figure 2.1.14 shows the results. The tests of local independence disclose evidence of local dependence for two pairs of items: B & D (DHP34 & DHP38) and D & E (DHP38 & DHP39). The evidence is so strong that there is no doubt that the fit of item responses to the Rasch model has to be rejected. During the tours through graphical log-linear Rasch models, we will show you how to

deal with that. For now, we only need to point out that the results support the weak evidence against homogeneity in Figures 2.1.10 and 2.1.11 and that both cases of local dependence involves the item (DHP38) that the analysis of the IRS γ in Figure 2.1.13 had too strong item discrimination. This result illustrates that evidence of strong item discrimination often turns up together with evidence of local dependence.

```
Check assumptions of local independence

A & B:   lr =    6.22  df =   9  p = 0.7182
A & C:   lr =   15.87  df =   9  p = 0.0696
A & D:   lr =   17.57  df =   9  p = 0.0405
A & E:   lr =   14.69  df =   9  p = 0.0999
B & C:   lr =   19.82  df =   9  p = 0.0190
B & D:   lr =   41.76  df =   9  p = 0.0000
B & E:   lr =    5.61  df =   9  p = 0.7780
C & D:   lr =    4.39  df =   9  p = 0.8839
C & E:   lr =    6.10  df =   9  p = 0.7295
D & E:   lr =   38.09  df =   9  p = 0.0000

Benjamini & Hochberg rejects at 0.01000

Suggested additions to the model:

LD:           BD DE
```

Figure 2.1.14 Tests of local independence. The evidence of local dependence is adjusted by the Benjamini-Hochberg procedure controlling the false discovery rate at 5 %.

2.1.3.6 Estimating person parameters

Even though the tests of fit rejected the Rasch model, we proceed as if nothing was wrong to show you how to estimate the person parameters and how to assess the reliability and targeting of measurement by the five items.

Select "**Estimate person parameters**" to obtain estimates of person parameters together with assessment of the bias and errors of the estimates.

Person parameter estimates are monotonic functions of the total score over all items. The exact[13] distribution of the total score depends on the person parameter and on a set of so-called score parameters that are functions of the item parameters. To calculate the person estimates, DIGRAM 1) re-estimate the item parameters to make sure that they are estimated without error, 2) uses these parameters to calculate the score parameters, and 3) calculates likelihood based estimates of the person parameter for each value of the total score.

It follows from the monotonic relationship between the total score and the person parameter estimates that the exact distribution of the person parameter estimates is known. For this reason, assessment of the properties of the estimates does not depend on the assumption that the distribution of the person parameter estimate can be approximated by the normal distribution[14].

DIGRAM calculates three different types of person estimates: Höglund's exact estimates (Höglund, 1974) Maximum likelihood (ML) estimates, and weighted maximum likelihood estimates (WML).

Figures 2.1.15 – 2.1.17 show these estimates. The exact estimates are interval estimates defined by the range of person parameter values where the observed score is the most probable outcome. In addition to being of some interest in themselves, these estimates are also of interest because the thresholds between the intervals correspond to the thresholds of the distribution of the total score if this is re-parameterized as a partial credit distribution. Exact person parameter estimates therefore only exist, if the thresholds of the total score are ordered, which is the case in this example even though the majority of items had disordered thresholds.

Höglund shows that the ML estimate corresponding to a given score is always included in the interval defined by the exact estimate for the same score. Concerning the ML estimate, the only complication is that the ML estimate for extreme scores are infinite. In order to calculate the moments of the distribution of the ML estimate for given values of the person parameter, we have to assign finite estimates to extreme scores. DIGRAM calculates such estimates, assuming that the

[13] One of the important features that set the Rasch model apart from other IRT model is that the exact distribution of the score and the estimates of parameters are known and can be estimated. IRT models have to assume that estimates of person parameters have asymptotic normal distributions. Since this require the number of items to increase towards infinity, it is easy to show and illustrate that this is unrealistic. Rasch models can avoid this assumption, but DIGRAM appears to be the only Rasch models that have implemented the exact distributions of person parameters.

[14] It is one important difference between estimates of person parameters in Rasch and IRT models. In the Rasch model we have estimates of the exact distribution of the estimates of person parameters. In IORT models, they have to rely on approximations by the asymptotic distribution under the assumption that the number of items increase towards infinity.

expected score is equal to 0.25 and 14.75 respectively. These values lie within the intervals defined by the exact estimates for extreme scores, -4.061 ∈]-∞,2.752] and 3.573 ∈[2.122,+ ∞[.

```
  +---------------------------------+
  |                                 |
  | Estimates of person parameters. |
  |                                 |
  +---------------------------------+

189 cases.   Mean score =  5.55  s.d. of score =  3.35

            ML          Höglunds exact       WML
Score   estimate      interval estimate   estimate
-------------------------------------------------
    0      -inf.       -inf.  -  -2.752     -3.512
    1     -2.493      -2.752  -  -1.672     -2.064
    2     -1.672      -1.672  -  -1.189     -1.437
    3     -1.203      -1.189  -  -0.912     -1.066
    4     -0.865      -0.912  -  -0.663     -0.787
    5     -0.585      -0.663  -  -0.437     -0.545
    6     -0.329      -0.437  -  -0.229     -0.316
    7     -0.082      -0.229  -   0.024     -0.084
    8      0.167       0.024  -   0.299      0.156
    9      0.420       0.299  -   0.531      0.404
   10      0.683       0.531  -   0.797      0.655
   11      0.963       0.797  -   1.051      0.911
   12      1.279       1.051  -   1.335      1.185
   13      1.675       1.335  -   1.696      1.503
   14      2.304       1.696  -   2.122      1.933
   15      +inf.       2.122  -   +inf.      2.803
-------------------------------------------------

The lower limits of Höglund´s exact estimate corresponds to PCM
thresholds

Pseudo ML estimates for extreme scores:
   Score =   0    theta =      -4.061
   Score =  15    theta =       3.573
```

Figure 2.1.15 Maximum likelihood (ML) and weighted maximum likelihood (WML) estimates and exact interval estimates of person parameters.

Figures 2.1.16 and 2.1.17 summarizes the properties of the ML and WML estimates. For each value of the person parameter estimate, the tables provide information on the expected (true) score, the bias of the estimate, the root mean squared error (RMSE), and the standard error of the estimate of the expected score. To us, the most important information in Figure 2.1.16 and 2.1.17 is the bias and the RMSE of the estimates of the person parameters because these two statistics tell us how

well the person parameter estimate performs. In this case, we see that there is considerable bias outside a narrow range of person parameter values. The weighted ML estimate reduces both the bias and the RMSE for a relatively wide range of person parameter values.

Score	Theta estimate	True score	Bias	RMSE	Score SEM
0	-4.061	0.25	0.376	0.794	0.48
1	-2.493	1.00	-0.274	1.024	0.92
2	-1.672	2.00	-0.277	0.962	1.31
3	-1.203	3.00	-0.199	0.829	1.61
4	-0.865	4.00	-0.127	0.715	1.82
5	-0.585	5.00	-0.071	0.630	1.95
6	-0.329	6.00	-0.032	0.572	2.00
7	-0.082	7.00	-0.007	0.540	2.01
8	0.167	8.00	0.010	0.533	2.00
9	0.420	9.00	0.026	0.552	1.97
10	0.683	10.00	0.052	0.599	1.93
11	0.963	11.00	0.096	0.681	1.85
12	1.279	12.00	0.166	0.793	1.70
13	1.675	13.00	0.257	0.907	1.46
14	2.304	14.00	0.286	0.912	1.06
15	3.573	14.75	-0.294	0.644	0.51

Figure 2.1.16 Properties of the maximum likelihood estimates of person parameters.

Score	Theta estimate	True score	Bias	RMSE	Score SEM
0	-3.512	0.41	0.549	0.933	0.61
1	-2.064	1.43	0.052	0.839	1.10
2	-1.437	2.45	-0.007	0.721	1.46
3	-1.066	3.38	-0.013	0.633	1.70
4	-0.787	4.27	-0.008	0.573	1.86
5	-0.545	5.15	-0.001	0.533	1.96
6	-0.316	6.05	0.003	0.509	2.00
7	-0.084	6.99	0.003	0.496	2.01
8	0.156	7.96	0.001	0.495	2.00
9	0.404	8.94	-0.002	0.503	1.97
10	0.655	9.89	-0.003	0.522	1.93
11	0.911	10.82	-0.002	0.550	1.86
12	1.185	11.72	-0.002	0.585	1.75
13	1.503	12.60	-0.011	0.621	1.57
14	1.933	13.49	-0.069	0.637	1.29
15	2.803	14.43	-0.427	0.680	0.80

Figure 2.1.17 Properties of weighted maximum likelihood estimates of person parameters.

Finally, DIGRAM prints information summarizing other properties of the estimates together with an estimate of the distribution of the person parameter and the reliability of the score over all items. This is shown in Figure 2.1.18.

```
------------------------------------------------------------

Range of persons with bias < 0.01: [-0.893 - 1.488]
        True score range:              [3.9 - 12.6]

Range of persons with bias < 0.05: [-2.051 - 1.838]
        True score range:              [1.4 - 13.3]

------------------------------------------------------------

Test midpoint   =       0.042
Test target     =      -0.132   Test information =   4.051

------------------------------------------------------------

  +---------------------------------------+
  |                                       |
  | Monte Carlo estimates of Reliability  |
  |                                       |
  +---------------------------------------+

Normal trait distribution with mean = -0.52 s.d.(trait) =  0.85


Expected True Score =   5.59   Variance =     7.84
Observed mean Score =   5.59   Variance =    11.12

                                  Ratio =     0.71

               Test-retest correlation     =  0.71
               Test-true score correlation =  0.83

------------------------------------------------------------

Frequencies of persons with bias less than 0.01:  0.660
Frequencies of persons with bias less than 0.05:  0.962

------------------------------------------------------------
```

Figure 2.1.18 Midpoint, target and reliability of measurement by DHP items.

The information in Figure 2.1.18 includes:

1) The range of the person parameters where measurement by the WML estimate is unbiased. If we regard measurement as unbiased if bias less than 0.01 we conclude that measurement I unbiased for scores between 4 and 12. Close to unbiased measurement with bias less than 0.05 is provided by scores from 2 to 13.
2) The test ***midpoint*** defined by the person value with an expected (true) score equal to half the maximum score. The midpoint of the DHP score is 0.042.
3) The test ***target*** equal to the person value where test information is maximized. The target of the DHP score lies below the midpoint at -0.13. target info is 4.05.
4) Estimates of the mean and standard deviation of the person parameter distribution under the assumption that the distribution is normal. The mean is -0.52 and the standard deviation is 0.85.
5) Estimates of the exact reliability calculated under the assumption that the person parameter has a normal distribution with the estimated mean and variance. Reliability is equal to 0.72 or 0.71 (depending on the definition of reliability).
6) If the distribution of the person parameter is as estimated it follows that measurement will be unbiased for 66 % of the population and close to unbiased for 96 %.

Recall that Cronbach's α was equal to 0.69. This is in accordance with the theory of Cronbach's α that is supposed to provide a lower bound of the true reliability. The example illustrates what we find in many (almost all) cases, namely that Cronbach's α is very close to the true reliability.

The results concerning the bias and errors of person parameter estimates and the results concerning targeting and reliability represent to different viewpoints that we assume when we discuss the qualities of the measurement provided by the items. Bias and standard errors of measurement looks at the measurement instrument from the point of view of single persons whereas targeting and reliability attempt to assess measurement quality from a population point of view. We pursue these points of view during the next (somewhat longer) guided tour through DIGRAM.

2.1.3.7 Extended output during estimation of person parameters

DIGRAM offers a wide range of extended output during estimation of person parameters. Figure 2.1.19 shows what is available in the current version. We will not illustrate how they work here, but some of it may turn up during the next tours.

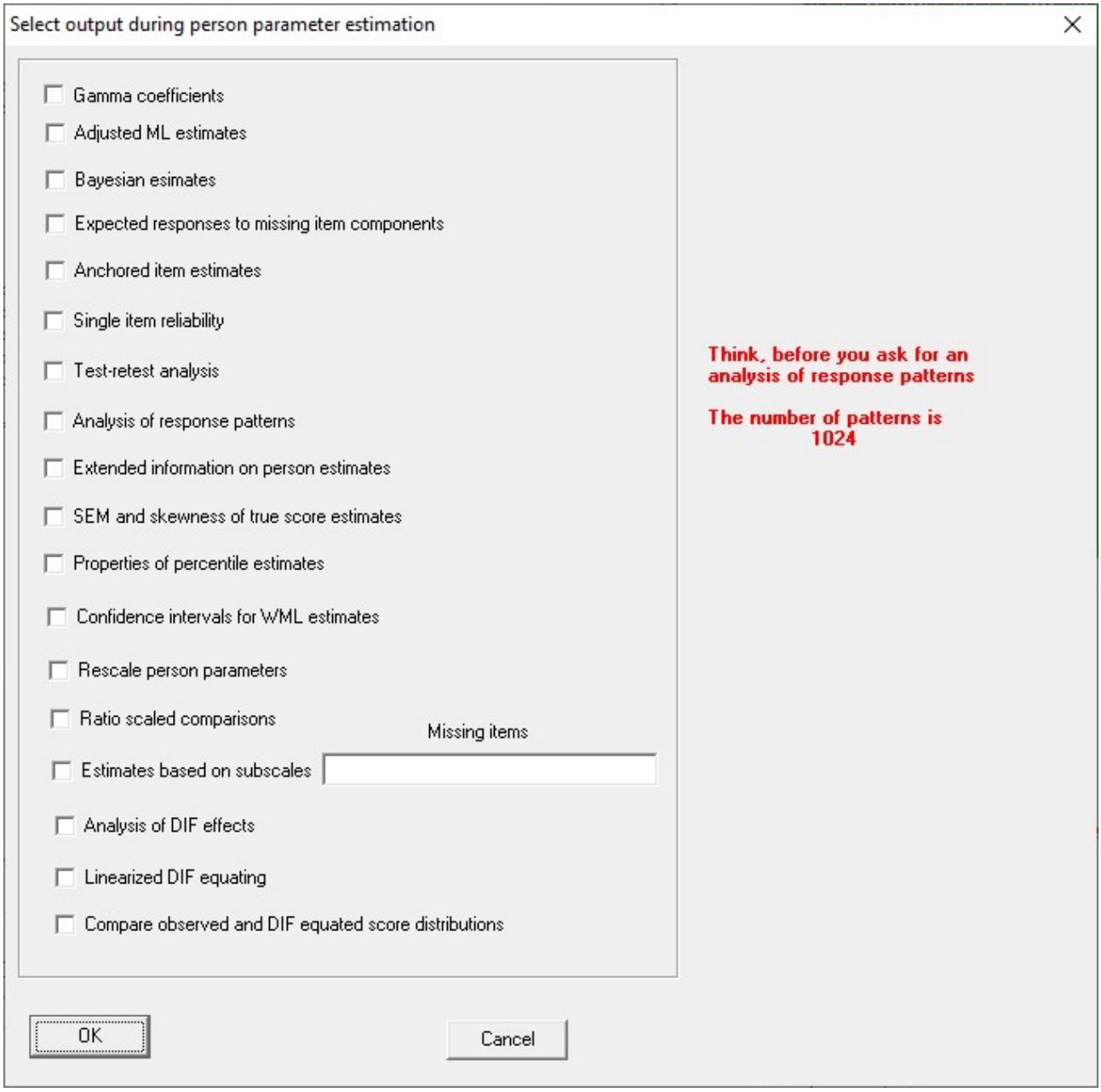


Figure 2.1.19 Extended output during estimation of person parameters

2.2 Rasch models. A longer tour.

During this tour, we trace the same path with the same DIGRAM project as we did in the previous tour, but this time we will take the time to show you facilities that sometimes are useful. On this tour, you will hear about

1) how to change the orientation of items,
2) how to redefine score groups,
3) how to test for DIF and local dependence,
4) how to create IRT and Rasch graphs,
5) how to assess targeting,
6) how to test for unidimensionality,
7) how to create plots with ICC, probability and information curves,
8) how to generate scale anchored item distributions

2.2.1 Changing the orientation of items

The DHP items are scored so that a low score means that the respondent has a high degree of control over his eating habits whereas a high score indicates low control. If you think that the interpretation of the scores is easier if a high score indicates high degree of control, you should change the orientation of the items. You can do this in two ways.

The first is to select items as before by followed by a "**FLIP**" command with or without reference to specific items. The second is by adding a "-" to the items when you invoke the ITEMS command.

"**ITEMS ABCDE**" followed by "**FLIP**" is, the same as "**ITEMS -A-B-C-D-E**" because "**FLIP**" without items flips the complete set of items.

"**ITEMS ABCDE**" followed by "**FLIP A**" is, the same as "**ITEMS -ABCDE**" because "**FLIP A**" changes the orientation of A, but keep the orientations of the other items.

Note also, that the possibility of flipping the orientation during item selection can also be used if you have some items phrased in a positive way and other items with a negative connotation. You therefore do not have to concern yourself about the orientation of items when you create your DIGRAM project.

Figures 2.2.1 and 2.2.2 show the results if you invoke the FLIP command. Take a little time to compare the results with Figures 2.1.2 and 2.1.3. Cronbach's Alpha and the variance of the score remains the same after you flipped the items.

```
  +--------------------------------------+
  |                                      |
  | Variables selected for item analysis |
  |                                      |
  +--------------------------------------+

5 items: ABCDE
---------
A:    DHP32 - 4 ordinal categories. * Flipped *
B:    DHP34 - 4 ordinal categories. * Flipped *
C:    DHP36 - 4 ordinal categories. * Flipped *
D:    DHP38 - 4 ordinal categories. * Flipped *
E:    DHP39 - 4 ordinal categories. * Flipped *

Exogeneous variables have not been defined

  +-------------------------------------------+
  |                                           |
  | Average item scores and score distribution |
  |                                           |
  +-------------------------------------------+

                                  complete cases
  items       n        mean       mean     item range
-------------------------------------------------------
-
A:   DHP32  195     2.144     2.148      0 - 3
B:   DHP34  197     1.487     1.513      0 - 3
C:   DHP36  194     1.753     1.741      0 - 3
D:   DHP38  197     2.269     2.254      0 - 3
E:   DHP39  196     1.801     1.794      0 - 3

Obtainable score range:  0 - 15
```

Figure 2.2.1 Information on flipped items

The flipped item scores are equal to 3 minus the original item scores. The same is therefore true for the average items scores. From this, it follows that the correlation among items are the same as before flipping. It is for this reason that Cronbach's Alpha was unchanged after flipping. Alpha is a function of the correlations among items.

Since the flipped score is equal to 15 minus the original score and Var(15-X) = Var(X) it follows that the variance of the flipped score is equal to the variance of the original score.

```
Score distribution: 189 Cases
--------------------------------
Score  Count Percent  Cumulated
------------------------------
   0       2     1.1        1.1
   1       4     2.1        3.2
   2
   3       5     2.6        5.8
   4       2     1.1        6.9
   5       9     4.8       11.6
   6       6     3.2       14.8
   7      22    11.6       26.5
   8      26    13.8       40.2
   9      17     9.0       49.2
  10      18     9.5       58.7
  11      22    11.6       70.4
  12      22    11.6       82.0
  13      10     5.3       87.3
  14      13     6.9       94.2
  15      11     5.8      100.0
------------------------------
Total    189   100.0

Mean     =   9.45
Variance =  11.21
s.d.     =   3.35
skewness =  -0.46
Missing =      9

Chronbach´s Alpha = 0.693

  +----------------------------------------+
  |                                        |
  | Score groups for tests of Rasch models |
  |                                        |
  +----------------------------------------+

ScoreGrp:   189 Cases
------------------------------
Score  Count Percent Cumulative
------------------------------
 0- 9     93    49.2       49.2
10-15     96    50.8      100.0
------------------------------
Total    189   100.0
```

Figure 2.2.2 Information on the flipped score and score groups

During this tour, you should also select F and G as exogenous variable ("**EXO FG**"). Note that DIGRAM will not flip the exogenous variables for you. If you want to change the orientation or order of the categories of exogenous variables, you have to use the "**RECODE**" command described in th notes on Project management.

2.2.2 Changing the score groups

DIGRAM's default is to define two score groups with approximately the same number of persons for tests of homogeneity of item responses across score groups and for tables that show the effect of exogenous variables on the score. If you are dissatisfied with these groups you can redefine them using the CUT command followed by list of parameters that define the score groups in the way you want them.

To define m score groups, CUT requires a minimum score s_0, m-1 cut points, $s_1,\ldots,s_{m-1}$, and a maximum score s_m[15], but CUT may also be used with fewer parameters as shown in Table 2.1.

Table 2.1 Definition of score groups

COMMAND	SCORE GROUPS
CUT	$0,1,2,..,s_{max}-1,s_{max}$
CUT s	$[0,s],[s+1,s_{max}]$
CUT s_0 s_1	$s_0,s_0+1,..,s_1-1,s_1$
CUT s_0 s_1 … s_m	$[s_0,s_1]$, $[s_1+1,s_2]$,.., $[s_{m-1}+1,s_m]$
CUT /k	Defines k score groups with a uniform distribution of scores
CUT #n	Defines score groups with at least n persons in each

Score groups are always defined up to and including the cut points.

"**CUT**" without parameters define 16 score groups, one for each separate score.

"**CUT 5**" defines the two score groups 0-9 and 10-15 shown in Figure 2.2.2.

Figure 2.2.3 show the results of four applications of the CUT commands.

In Figure 2.2.3a, "**CUT 0 7 10 15**" defines three score groups, 0-7, 8-10, and 11-15.

In Figure 2.2.3b, "**CUT /3**" defines three score groups where the distribution of non-extreme scores are as close to uniform as possible.

[15] The minimum and maximum scores will in most cases be equal to the extreme scores, but you may use s_0 and s_m to restrict some of the analyses to a subset of persons.

In Figure 2.2.3c, "**CUT #25**" defines score groups with at least 25 persons with non-extreme scores in each group.

Finally, Figure 2.2.3d illustrates that you can define score groups where low scores and high scores are excluded. "**CUT 4 7 10 12**" defines score groups 4-7, 8-10 and 11-12.

```
ScoreGrp:   189 Cases
------------------------------
Score  Count Percent Cumulative
------------------------------
 0- 7     50    26.5       26.5
 8-10     61    32.3       58.7
11-15     78    41.3      100.0
------------------------------
Total    189   100.0
Missing = 9
```

(a) Score groups after "CUT 0 7 10 15"

```
ScoreGrp:   189 Cases
------------------------------
Score  Count Percent Cumulative
------------------------------
 0- 8     76    40.2       40.2
 9-11     57    30.2       70.4
12-15     56    29.6      100.0
------------------------------
Total    189   100.0
Missing = 9
```

(b) Score groups after "CUT / 3"

```
ScoreGrp:   189 Cases
------------------------------
Score  Count Percent Cumulative
------------------------------
 0- 7     50    26.5       26.5
 8- 9     43    22.8       49.2
10-11     40    21.2       70.4
12-15     56    29.6      100.0
------------------------------
Total    189   100.0
Missing = 9
```

(c) Score groups after "CUT # 25"

```
Incomplete score distribution
minscore = 4  maxscore = 12
Score  Count Percent Cumulative
------------------------------
 4- 7     39    27.1       27.1
 8-10     61    42.4       69.4
11-12     44    30.6      100.0
------------------------------
Total    144   100.0
Missing = 54
```

(d) Score groups after "CUT 4 7 10 12"

Figure 2.2.3 Redefined Score groups after "CUT # 25"

Recall, that the score groups play important roles during the conditional likelihood ratio test of homogeneity comparing item parameters estimated in different score groups. During the rest of this tour we will used the flipped DHP items and the score groups defined by the "CUT /3" command.

2.2.3 Item analysis with flipped items and three score groups

This section present a little of the initial analysis of the flipped items (Figures 2.2.4 – 2.2.6). Flipping items change the signs and the order of PCM thresholds and the signs of item locations and

item effects. Figure 2.2.5 presents the category in the order defined by flipping, but category effects are the same as before flipping.

The orientation of the item has no effect on the CLR tests of invariance and would have had no effect if we had used the default score groups defined by DIGRAM. However, increasing the number of score groups provided stronger evidence against homogeneity.

item		1		2		3	Location
A:	DHP32	-1.153		-0.865		0.717	-0.434
B:	DHP34	0.685	>	0.673	>	0.071	0.476
C:	DHP36	-0.949		0.906	>	0.311	0.089
D:	DHP38	-0.218	>	-1.896		0.643	-0.490
E:	DHP39	0.168	>	-1.321		2.228	0.359

Figure 2.2.4 PCM thresholds and locations of flipped items.

item		0	1	2	3	Item effect
A:	DHP32	0.000	0.719	1.151	-0.000	0.434
B:	DHP34	0.000	-0.209	-0.405	0.000	-0.476
C:	DHP36	0.000	1.038	0.222	0.000	-0.089
D:	DHP38	0.000	-0.272	1.134	-0.000	0.490
E:	DHP39	0.000	0.190	1.870	-0.000	-0.359
		---- MICE effects ----				
A:	DHP32	1.000	2.053	3.160	1.000	1.543
B:	DHP34	1.000	0.812	0.667	1.000	0.621
C:	DHP36	1.000	2.824	1.249	1.000	0.914
D:	DHP38	1.000	0.762	3.107	1.000	1.633
E:	DHP39	1.000	1.210	6.487	1.000	0.699

Figure 2.2.5 Item and category effects

		CLR	df	p	delta
scoregroups		54.4	28	0.002	
F:	AGE	42.4	42	0.454	
G:	SEX	23.7	14	0.050	

Figure 2.2.6 CLR tests of homogeneity and invariance

The results of the tests of fit remain the same after flipping and the estimates of person parameters are the same except that the sign of the estimate has changed and that the estimates are presented in the opposite order, but it interpretation of parameters and interpretations of test results during analyses of DIF may be more challenging. We suggest that you check it yourself before we proceed.

2.2.4 Testing for DIF and local dependence

This section describes tests of DIF and local dependence by analysis of conditional independence in three-way tables that do not need estimates of item parameters. That is, in simple ways that could have been included as part of the initial descriptive analysis of data that we undertake before we try to fit the Rasch models.

2.2.4.1 DIF

Assume that Y_i is an item that X_j is an exogenous variable and that S is the total score over all items. If it is true that item responses fit a Rasch model then it follows that Y_i and X_j are conditionally independent given S. The hypothesis of conditional independence, $Y_i \perp X_j \mid S$, is a hypothesis relating to a simple three-way table where the association between the item and the exogenous variable is stratified by the total score over all items. Such hypotheses are easy to test with statistical programs with facilities for analysis of multidimensional contingency tables in general and very easy to test with DIGRAM, because DIGRAM is tailor made for such hypotheses.

To simplify these analyses, we have implemented a DIF command that you can use if you want to perform such DIF analyses for all items relative to some or all exogenous variables or even relative to variables that you have not designated as exogenous variables for your Rasch model[16]. Tables with responses to item, exogenous variables together with the total score (not the score groups) are counted and test statistics calculated, but the tables are not printed. If you want to see the tables, we suggest that you use the STABS[17] command to create tables where you can test conditional independence in the same way that you do with ordinary tables in DIGRAM.

[16] "**DIF**" without parameters tests for DIF relative to all exogenous variables, whereas "**DIF**" followed by a list of variables tests that there is no DIF relative to these variables.
[17] "STABS v1 v2" creates a three-way table with the v1, v2 and the score groups. "HYP v1 v2" defines the hypothesis that v1 and v2 are conditional independent given the score groups and "TEST" test the hypothesis. User guide to DIGRAM from 2003 provide mor details on analysis of contingency tables.

DIGRAM calculates χ^2 tests and partial γ coefficients and assess significance by repeated Monte Carlo tests. P-values for the γ coefficients are two-sided. The results are summarized in two different ways: first, for the separate items (Figure 2.2.4) and second, for the exogenous variables (Figure 2.2.5).

```
Analysis of DIF for A: DHP32
Scale : # - RawScore

Exogenous   X²   df asymp exact  gamma asymp exact nsim
-----------------------------------------------------
F:     AGE  86.0 63 0.029 0.061   0.05 0.683 0.693 1000
G:     SEX  26.6 26 0.429 0.667   0.07 0.595 0.524   21

Analysis of DIF for B: DHP34
Scale : # - RawScore

Exogenous   X²   df asymp exact  gamma asymp exact nsim
-----------------------------------------------------
F:     AGE  58.6 56 0.379 0.606  -0.12 0.299 0.242   33
G:     SEX  18.1 21 0.643 0.745   0.21 0.130 0.160   94

Analysis of DIF for C: DHP36
Scale : # - RawScore

Exogenous   X²   df asymp exact  gamma asymp exact nsim
-----------------------------------------------------
F:     AGE 101.4 66 0.003 0.008   0.14 0.233 0.213 1000 **
G:     SEX  28.0 26 0.360 0.530  -0.32 0.012 0.014 1000 -

Analysis of DIF for D: DHP38
Scale : # - RawScore

Exogenous   X²   df asymp exact  gamma asymp exact nsim
-----------------------------------------------------
F:     AGE  40.2 43 0.591 0.740  -0.22 0.113 0.136  154
G:     SEX  24.4 18 0.143 0.208   0.34 0.039 0.056 1000

Analysis of DIF for E: DHP39
Scale : # - RawScore

Exogenous   X²   df asymp exact  gamma asymp exact nsim
-----------------------------------------------------
F:     AGE  55.5 51 0.309 0.437   0.21 0.120 0.132  174
G:     SEX  27.1 21 0.168 0.188  -0.03 0.826 0.859   64
```

Figure 2.2.4 Overview of tests for DIF for different items. Significant χ^2 statistics are flagged with one or more *'s whereas significant γ coefficients are flagged with –'s or +'s depending on the sign of the γ coefficient.

Figure 2.2.4 disclose evidence of DIF two different ways for item C (DHP36): relative to Gender by the γ coefficient and relative to Age by the χ^2 statistic. The summary for the exogenous variables in Figure 2.2.5 tells the same story.

```
  +-----------------------------------------------+
  |                                               |
  | Test results for separate exogenous variables |
  |                                               |
  +-----------------------------------------------+

Analysis of DIF relative to F: AGE
Scale : # - RawScore

     Item   X²   df asymp exact  gamma asymp exact nsim
-------------------------------------------------------
A:   DHP32  86.0 63 0.029 0.061   0.05 0.683 0.693 1000
B:   DHP34  58.6 56 0.379 0.606  -0.12 0.299 0.242   33
C:   DHP36 101.4 66 0.003 0.008   0.14 0.233 0.213 1000  **
D:   DHP38  40.2 43 0.591 0.740  -0.22 0.113 0.136  154
E:   DHP39  55.5 51 0.309 0.437   0.21 0.120 0.132  174


Analysis of DIF relative to G: SEX
Scale : # - RawScore

     Item   X²   df asymp exact  gamma asymp exact nsim
-------------------------------------------------------
A:   DHP32  26.6 26 0.429 0.667   0.07 0.595 0.524   21
B:   DHP34  18.1 21 0.643 0.745   0.21 0.130 0.160   94
C:   DHP36  28.0 26 0.360 0.530  -0.32 0.012 0.014 1000      -
D:   DHP38  24.4 18 0.143 0.208   0.34 0.039 0.056 1000
E:   DHP39  27.1 21 0.168 0.188  -0.03 0.826 0.859   64
```

Figure 2.2.5 Overview of tests for DIF relative to the exogenous variables

2.2.4.2 Local dependence

The test of DIF described above is a simple generalization of the well-known Mantel-Haenszel test of DIF for dichotomous items and binary exogenous variables. The test of local dependence invoked by the **LDE** command is probably less well-known but it is based on the same Markov properties of graphical models as the DIF test.

Let $(X_1,\ldots,X_K)$ be items and $R=\sum_i X_i$ be the score. If items are conditionally independent and fits a Rasch model it follows that X_i is conditionally independent of X_j given $R-X_i$. DIGRAM uses this

property for a test of local independence testing conditional independence of X_i and X_j in the three-way table with X_i, X_j and R-X_i**.**

To invoke a test you of local independence you must use the "**LDE itemset1 + <itemset2>**" command causing DIGRAM to test conditional independence of items in itemset1 and itemset2 given the score over items in itemset2. If you have not included itemset2, DIGRAM assumes that it consists of all items except for those in itemset1.

The tests of item fit Figure 2.1.12 found evidence against the fit of DHP38 (D). The observed IRS γ is larger than the expected under the Rasch model. Since this would happens if DHP is positively dependent on other items, we invoke the "**LDE D**" to test the local dependence of D and the other items. Figure 2.2.6 shows the results.

```
Analysis of DIF for A: DHP32
Scale : # - ABCE

Exogenous   X²   df asymp exact  gamma asymp exact nsim
-----------------------------------------------------
D:   DHP38  35.1 43 0.799 0.952  -0.10 0.484 0.714   21


Analysis of DIF for B: DHP34
Scale : # - ABCE

Exogenous   X²   df asymp exact  gamma asymp exact nsim
-----------------------------------------------------
D:   DHP38  54.7 37 0.030 0.070   0.33 0.023 0.019 1000 +


Analysis of DIF for C: DHP36
Scale : # - ABCE

Exogenous   X²   df asymp exact  gamma asymp exact nsim
-----------------------------------------------------
D:   DHP38  48.0 42 0.243 0.353  -0.53 0.000 0.000 1000 ---


Analysis of DIF for E: DHP39
Scale : # - ABCE

Exogenous   X²   df asymp exact  gamma asymp exact nsim
-----------------------------------------------------
D:   DHP38  58.6 39 0.022 0.048   0.26 0.095 0.091 1000 *
```

Figure 2.2.6 Tests of local dependence of item D (DHP38)

The results in Figure 2.2.6 agree to some degree with the results of Kelderman's CLR test of local dependence (Figure 2.1.14). Kelderman found strong evidence of local dependence between B & D

and between D & E. Figure 2.2.6 agrees, but significance is weak. On the other hand, the CLR test of local independence accept the hypothesis for B & C where the test of conditional independence between C & D is highly significant.

A major difference between the tests of conditional independence in Figure 2.2.6 and the CLR tests in Figure 2.1.4 is that the tests of conditional independence differentiate between positive and negative local dependence whereas the CLR tests are tests for nominal variables and therefore cannot describe the association in these terms.

The differences between the two sets of results leave questions unanswered. We return to these issues in the GLLRM tours.

2.2.5 IRT and Rasch graphs

The tests for DIF and local dependence are derived from the Markov properties of the IRT and Rasch graphs of Graphical Rasch models. Since they become more and more important as we move along, it is convenient to take a first look at these graphs here.

Look at Figure 2.1.1 showing DIGRAM's main form after you have selected the items. In the lower right corner above the "Graphical Rasch model" button, you will find the IRT button. When you click this button, DIGRAM takes you to the graph module where the IRT graph is displayed (Figure 2.2.7)[18]. Next, click the "Rasch" button to see the Rasch graph shown in Figure 2.2.8.

For now, the main purpose of the IRT and Rasch graphs is to remind you of the assumptions of the Rasch model and to make it clear that we are assessing the Rasch model within a multivariate frame of reference defined as a graphical model. We return to these graphs during the tours through the graphical and-log-linear Rasch models where they play a much more important role and where you may define and modify models by adding or deleting connections between pairs of items and or connections between items and exogenous variables[19].

[18] It may happen that DIGRAM has problems generating these graphs. If this happens, you have to click on the "Graph" button and then click on the "IRT" button in the graph dialog.
[19] The notes describing DIGRAM's graph module provide information on how to work with the graphs.

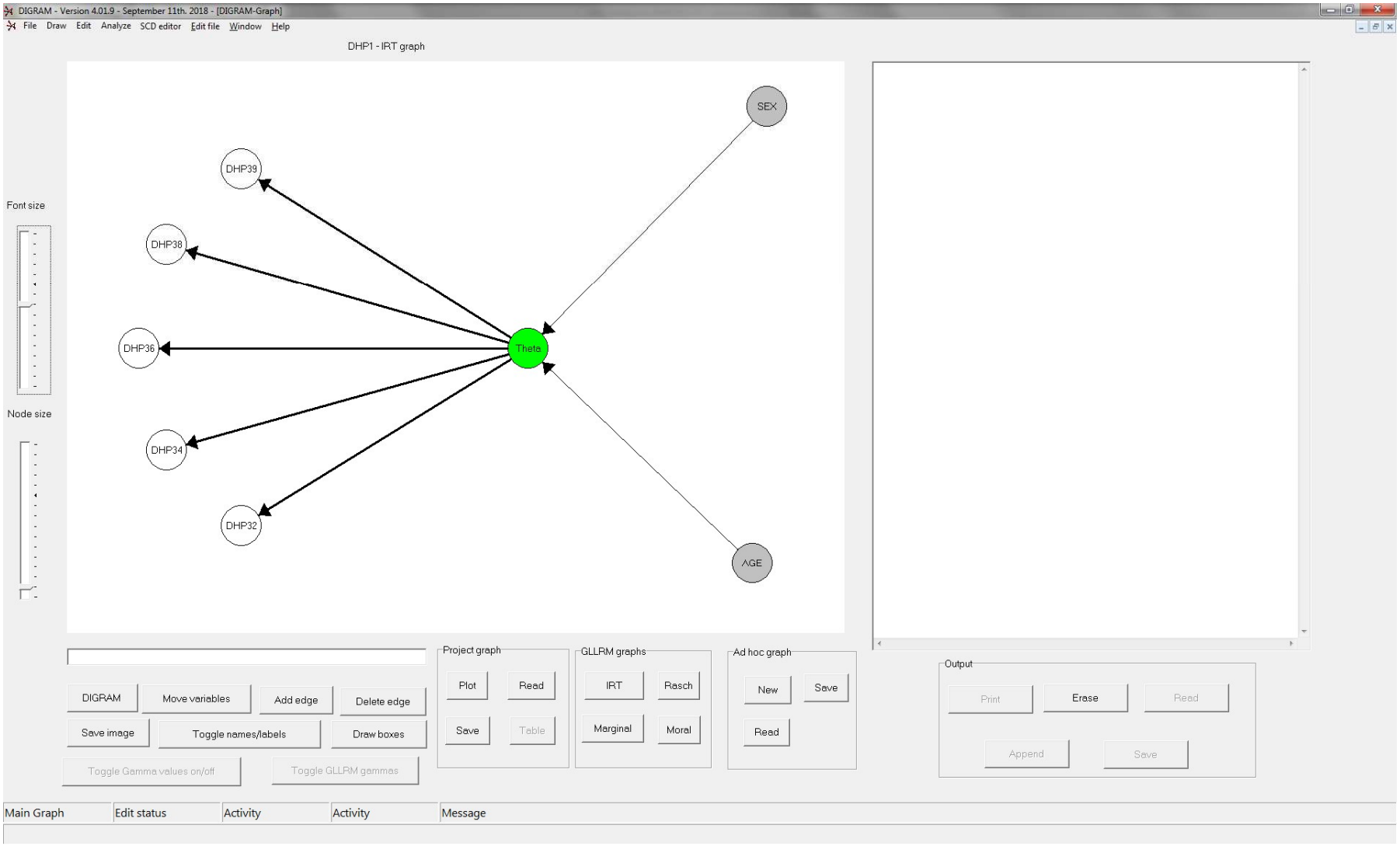


Figure 2.2.7. The IRT graph of the graphical Rasch model for the five DE items and with Age and Sex as exogenous variables. The graph has been edited and therefore do not look exactly like this when you try this for the first time[20].

[20] Consult the notes on the Graph model to see how to do this.

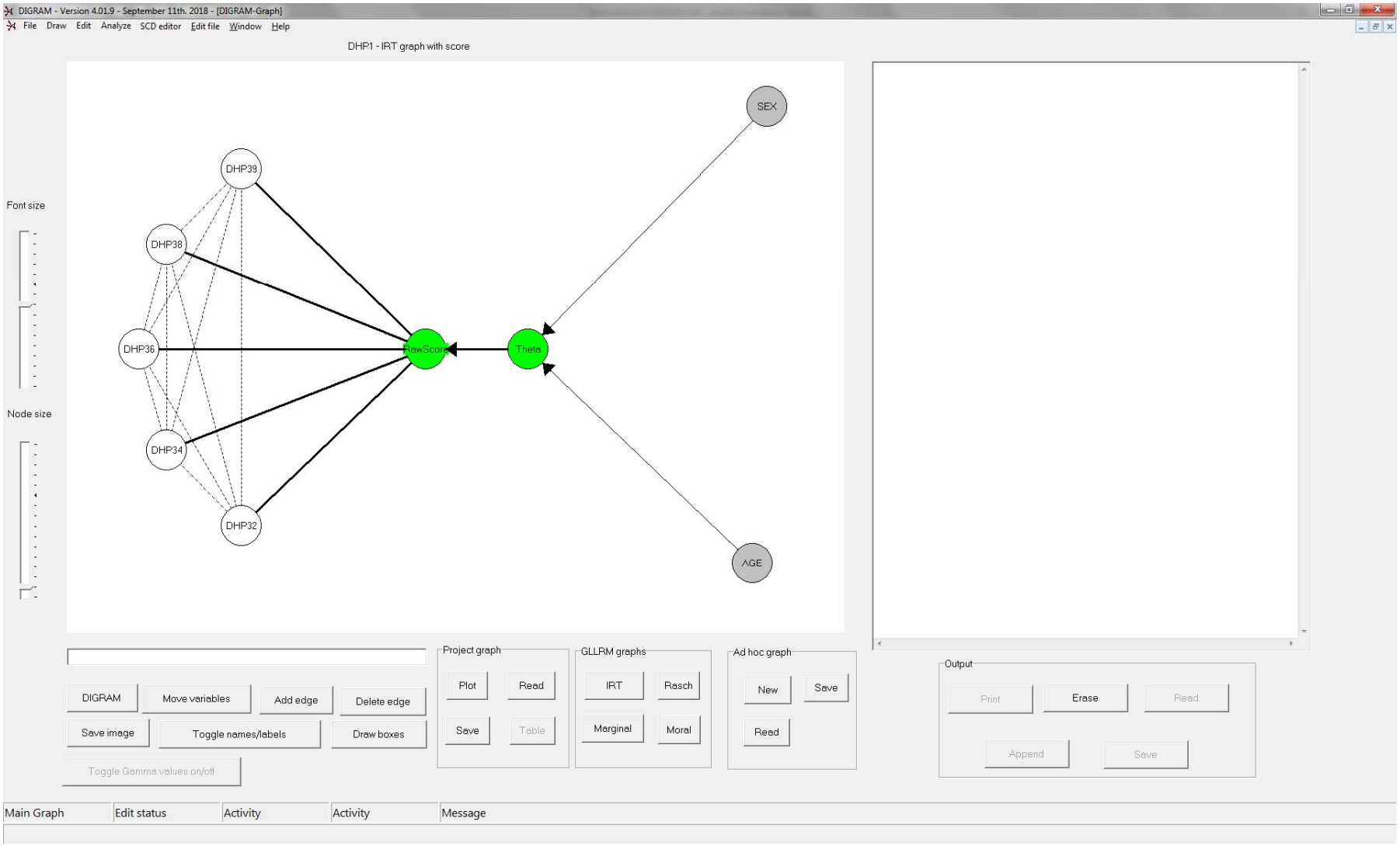


Figure 2.2.8. The Rasch graph of the graphical Rasch model for the five DE items with Age and Sex as exogenous variables. The graph has been edited

2.2.6 Tests of unidimensionality

If the set of items consists of two subsets depending on different latent variables, we would expect positive local dependence within subsets and negative local dependence between subsets. Figure 2.6 disclosed significant evidence of both positive between B, D and E and negative dependence between C and D. The natural next step in the analysis would be to test whether BDE and AC depend on different latent variables

To invoke a test of unidimensionality you have to select the "**Test unidimensionality**" option following which DIGRAM asks you to define the dimensions of the multidimensional alternative to unidimensionality. This is done in the little dialog box (Figure 2.2.9) where you enter the subsets[21] of items separated by a '+'.

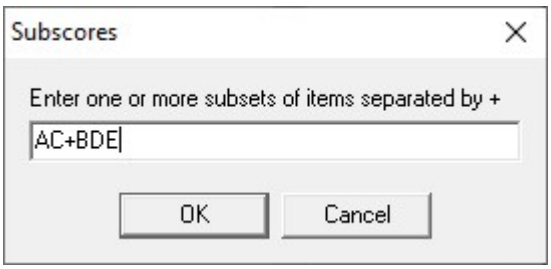


Figure 2.2.9 Definitions of subsets of items for tests of unidimensionality.

You only have to include a single subset of items if you just want to test whether the items in the subset measure the same construct as the rest of the items. For instance, "BDE" is the same as "BCE+AD" resulting in the same analysis as the one initiated in Figure 2.2.10 except that Subscore 1 = B+D+E and Subscore 2 = A+C.

To test the hypothesis of unidimensionality, DIGRAM compares the observed and expected joint distributions of the A+C and B+D+E subscores and reject the hypothesis of unidimensionality if the observed correlation of A+C and B+D+E is significantly lower than the expected correlation under the Rasch model. Figures 2.2.10 and 2.2.11 show the observed and expected distributions and Figures 2.2.12 and 2.2.14 present the results of the test[22].

[21] You can define more than two subsets of items.
[22] You will only see that tables in Figures 2.2.10 & 2.2.11 you select "extended output" during tests for unidimensionality. It is included here to show you what happens during the analysis.

Observed counts

Subscore 1: AC
Subscore 2: BDE

Subscore 2	Subscore 1: 0	1	2	3	4	5	6
0	2	2		2			
1	1		1		2		
2		1	1	4	1	2	2
3	1		2	3	4	1	
4	1	1	2	13	11	4	
5			3	9	5	5	4
6			2	5	6	5	4
7			3	6	10	5	3
8	1			2	11	5	9
9		1	1	2	2	4	11

Figure 2.2.10 Observed joint distribution of subscores A+C and B+D+E.

Expected counts (expected counts < 0.01 are not printed)

Subscore 1: AC
Subscore 2: BDE

Subscore 2	Subscore 1: 0	1	2	3	4	5	6
0	2.0	2.1		0.8	0.1	0.1	
1	0.9		1.3	0.3	0.4	0.1	0.1
2		2.3	0.9	3.1	0.9	1.2	0.4
3	0.7	0.5	3.1	2.0	3.9	2.2	0.5
4	0.2	2.0	2.3	10.0	8.4	3.3	1.6
5	0.3	0.7	5.7	10.7	6.4	5.1	3.5
6	0.1	1.1	3.6	4.8	5.8	6.6	4.8
7	0.1	0.7	1.7	4.5	7.8	9.5	4.1
8	0.0	0.2	1.0	3.8	6.9	5.0	10.2
9		0.0	0.2	0.8	0.8	2.8	11.0

Figure 2.2.11 Expected joint distribution of subscores A+C and B+D+E.

In addition to expecting the observed correlation between the subscores to be weaker than expected if the assumption of unidimensionality is false, we expect a larger than expected frequencies of persons with relatively high score on one subscale and relatively low score on the other. DIGRAM therefore include information on the observed and expected frequencies of persons with scores outside 95 and 50 percent confidence regions.

The results can be seen in Figure 2.2.12. Observed frequencies outside the confidence regions are larger than expected, but the differences are not significant. The observed correlation is weaker than the expected, which is also what we would find if unidimensionality was false, but the difference is far from significant. Taken at face value, this suggests that the hypothesis of unidimensionality should be accepted.

```
  +-------------------------------------------------------------+
  |                                                             |
  | Overview of significant differences between the two subscales |
  |                                                             |
  +-------------------------------------------------------------+

Count outside confidence regions

AC too low:

level observed expected residual    p
0.050        8     5.54    1.070 0.1424
0.250       42    39.51    0.460 0.3228

AC too high:

level observed expected residual    p
0.050        6     4.41    0.778 0.2182
0.250       38    36.48    0.295 0.3841

  +-------------------------------------------+
  |                                           |
  | Observed and expected subscore correlation |
  |                                           |
  +-------------------------------------------+

Expected Gamma =   0.464  s.e. =  0.0572
Observed Gamma =   0.407    p  =  0.3166  (Two-sided)
```

Figure 2.2.12 Analysis of unidimensionality

The p-value of the test that there is no difference between the correlations is based on the asymptotic distribution of Goodman and Kruskall's γ. However, since the table with expected frequencies in Figure 2.2.11 is large and sparse with many cells with close to no expected observations, there is a considerable risk that the asymptotic distribution of the γ does not approximate the exact distribution. For this reason, DIGRAM proposes (Figure 2.2.13) that you should calculate a Monte Carlo estimate of the exact p-value by parametric bootstrapping. Figure 2.2.14 shows the result. The importance of using Monte Carlo estimates of p-values is obvious. In this case, the difference between the observed and expected correlation between the subscores is weakly significant.

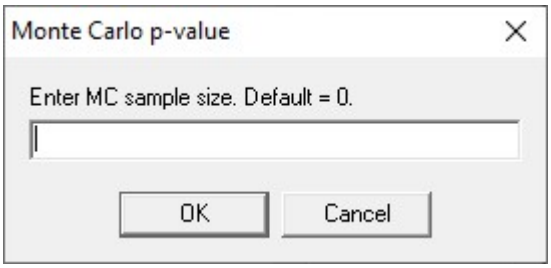


Figure 2.2.13 Monte Carlo estimate of p-value comparing observed and expected correlations between subscores.

```
   +-----------------------------------------------------+
   |                                                     |
   | Monte Carlo estimate of p-values. Sample size = 400 |
   |                                                     |
   +-----------------------------------------------------+

Gamma:                      p(below) =  0.033
```

Figure 2.2.14 Monte Carlo estimate of the one-sided p-value[23] of the test of unidimensionality

2.2.6.1 Assessment of practical unidimensionality

Disclosing evidence against unidimensionality is important for statistical purposes to make sure that we have a model that fits data and for substantive and conceptual reasons if we want to understand what goes on. However, it may be less than essential in connection with measurement if we can show that the score in practice function as if it was unidimensional.

We say that a set of items is *practically unidimensional* if subsets of items rank the persons in the same way as the complete set of items. To assess this, we compare the ranking of all pairs of persons by the total score and by the subscores and say that there is discordance if the total score and a subscore disagree. Even if the model fits data, we do expect a certain degree of discrepancy for random reasons. DIGRAM therefore estimate the expected frequencies of discordance under the unidimensional model calculate the ratio between the observed and expected discordance and test that it is significantly higher than one.

[23] We are using a one-sided p-value because the observed correlation will be smaller than the expected if the latent DHP structure is two-dimensional

In Figure 2.2.15, the ratio between the observed and expected discordance of the A+C subscore is equal to 1.31. This is significant confirming that A+C does not measure the same latent variable as the B+D+E subscore, and since more than the observed frequency of discordance is more than 30 % than the expected frequency[24] we conclude that DHP is not *practically* unidimensional. It is unreasonable.

```
  +-----------------------------------------+
  |                                         |
  | Assessment of practical unidimensionality |
  |                                         |
  +-----------------------------------------+

 Observed    Tied Concordance Discordance
Subscore1 =  0.26    0.857       0.143
Subscore2 =  0.18    0.945       0.055
Joint     =  0.39    0.768       0.232

 Expected
Subscore1 =  0.24    0.891       0.109
Subscore2 =  0.19    0.939       0.061
Joint     =  0.37    0.800       0.200

Ratios between observed and expected discordance

Subscore1 =  1.313   p = 0.003
Subscore2 =  0.906   p = 0.768
Joint     =  1.163   p = 0.033
```

Figure 2.2.15 Assessment of practical unidimensionality

2.2.7 Describing items

When item parameters have been estimated and if the fit of the Rasch model have been confirmed we often need to describe how the items function. There are both graphical and numerical ways to do that.

[24] The 30 % limit is not meant as a rule of thumb for assessment of practical unidimensionality. It is up to the user testing for unidimensionality to decide where the limit between practical and unreasonable unidimensional should be.

2.2.7.1 Item and test characteristic curves, information curves and probability curves

Three types of curves illustrate features of Rasch items:

1) *Item and test characteristic curves* (ICC and TCC) plot the expected item and test scores against values of the person parameter.
2) *Item information curves* (IIC) plot the test and item information against the values of the person parameter.
3) Category characteristic curves (CCC) plot the probabilities of item scores against the values of person parameter.

DIGRAM will not produce these curves for you, but will print information on a text file that you may enter into your standard statistical program, where the curves can be drawn.

Select the "**Export data for ICC curves**" in the GRM dialog box if you want DIGRAM to do this. Figure 2.2.16 shows the list of text files created by DIGRAM. In addition to the text files with the data for the plots, DIGRAM also creates SPSS syntax files with code that define the variables, read the data, and create the basic curves that DIGRAM assumes that you want to see.

```
SPSS users can use ICC.sps, INF.sps and PROB.sps to access the ICC, information and probability
curves. SAS users can use ICC.sas to access the ICC curves


Data for ICC curves under the CURRENT(!) model will be written on ICC.txt
Data for probability curves under the CURRENT(!) model will be written on ItemProbs.txt
Data for information curves under the CURRENT(!) model will be written on ItemInf.txt
```

Figure 2.2.16 Information on files with data for ICC, probability and information curves.

ICC curves

The file with data for the ICC curves contains the following variables.

Theta = the person parameter value.
Score = the expected total score.
A-E = the expected item scores or the average obtained item scores corresponding to person parameter estimates.
Type = 0 : expected item score 1: average observed item score.
Npersons = number of persons with the person parameter estimate.

Figures 2.2.17 and 2.2.18 show the content of ICC.txt with data. The format of the data on ICC_ABCDE.txt is free with variable names in the first row.

```
Theta score A B C D E type npersons
-5,000 0,057 0,022 0,003 0,017 0,009 0,006 0 0
-4,990 0,058 0,022 0,003 0,017 0,009 0,006 0 0
                    .........

-0,040 7,509 1,875 0,826 1,258 2,049 1,502 0 0
-0,030 7,549 1,882 0,836 1,266 2,055 1,509 0 0
                    .........

4,990 14,897 2,986 2,993 2,990 2,987 2,940 0 0
5,000 14,898 2,986 2,993 2,991 2,987 2,941 0 0
-2,304 1 0,750 0,000 0,000 0,250 0,000 1 4
-1,279 3 1,200 0,200 0,600 1,000 0,000 1 5
                    .........

1,672 13 2,500 2,700 2,600 2,800 2,400 1 10
2,493 14 2,923 2,769 2,769 3,000 2,538 1 13
```

Figure 2.2.17 The contents of the ICC.txt file with data for ICC curves.

Figure 2.2.18 shows the ICC curve for item D. If you are dissatisfied with SPSS's default plots, you have to edit the ICC plot yourself. We have done so in the ICC curve for item D (the item that the item fit statistics suggested had stronger item discrimination than expected by the Rasch model) where we have hidden the legends for item D. The marks refer to the average observed item scores for persons with a given total score.

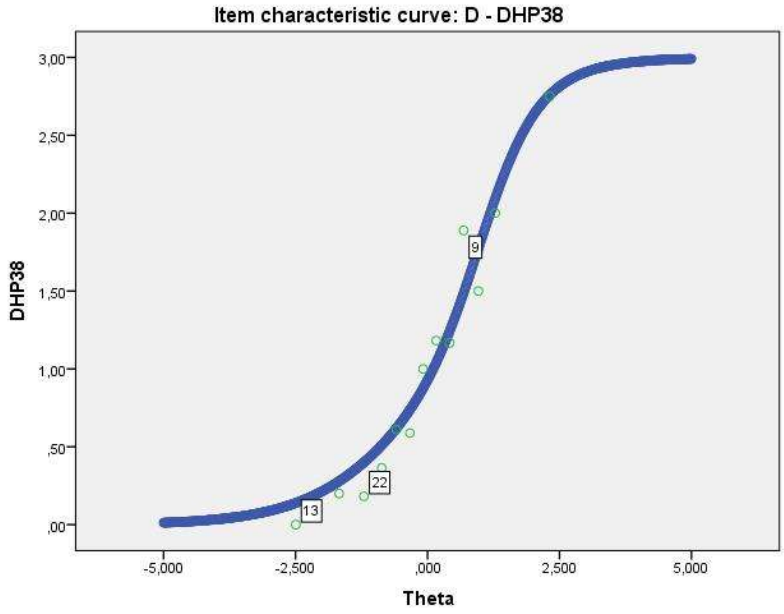


Figure 2.2.18 ICC curve for items D. The green points show the average observed item scores for ML estimates of person parameters corresponding to the total score on all items

Figure 2.2.19 shows the ICC curves of all items next to each other. It is easy to see that ICC curves for polytomous items do not satisfy the requirement of invariant item ordering. (IIO)

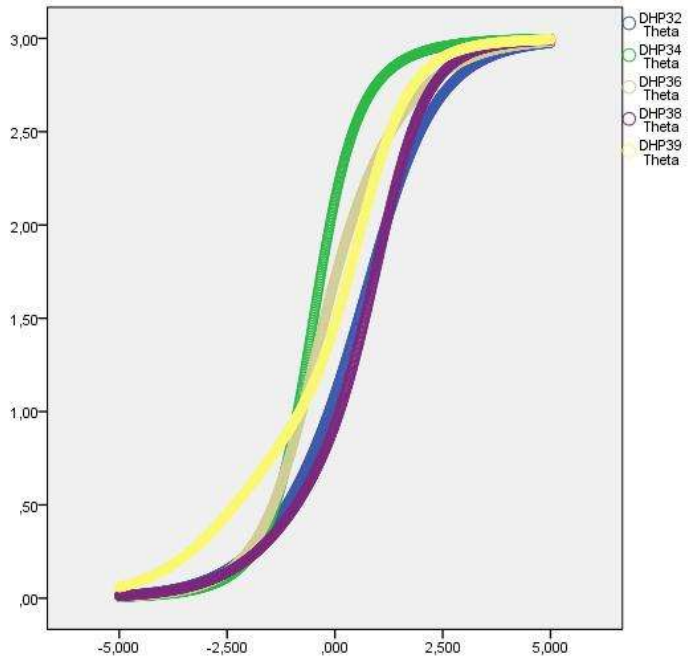


Figure 2.2.19. Overlay plot with ICC curves for all items.

Finally, Figure 2.2.20 shows the relationship between the person parameters and the expected scores (the true scores) on all items. Notice the close to linear relationship between the total score and the person parameters estimates for total scores equal between 3 and 12.

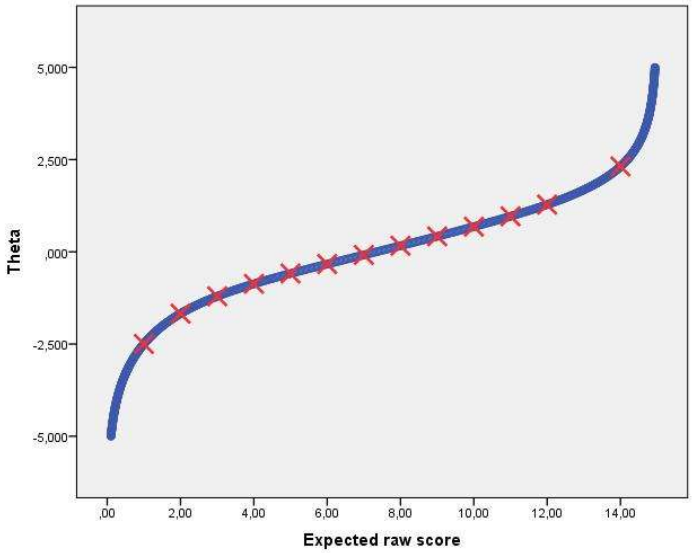


Figure 2.2.20 Person parameter values plotted against the expected (true) total score on all items. The red X's shows the maximum likelihood estimates of person parameters corresponding to observed scores on all items.

Test and item information

Figures 2.2.21 and 2.2.22 show the test information and the standard error of measurement (SEM). In many cases, the SEM curve is almost horizontal in a wide range of θ values even though the test info curve has a noticeable mode. This is not the case for the DHP scale.

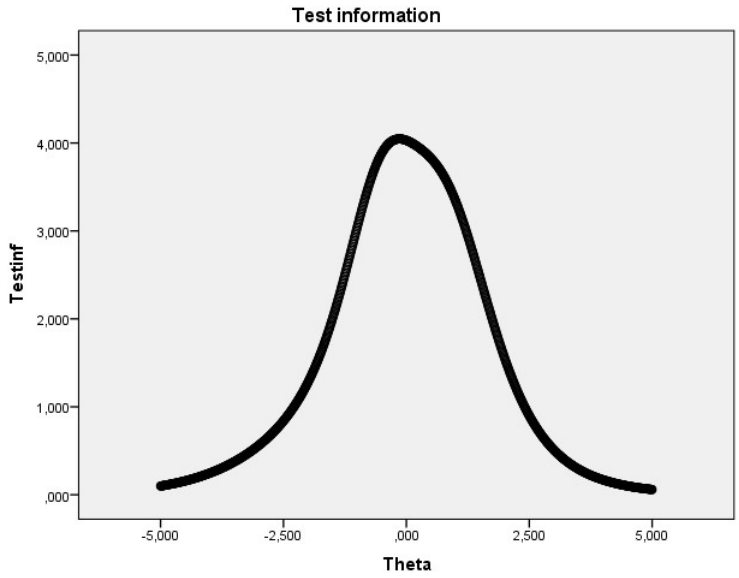


Figure 2.2.21. Test information under the current Rasch model

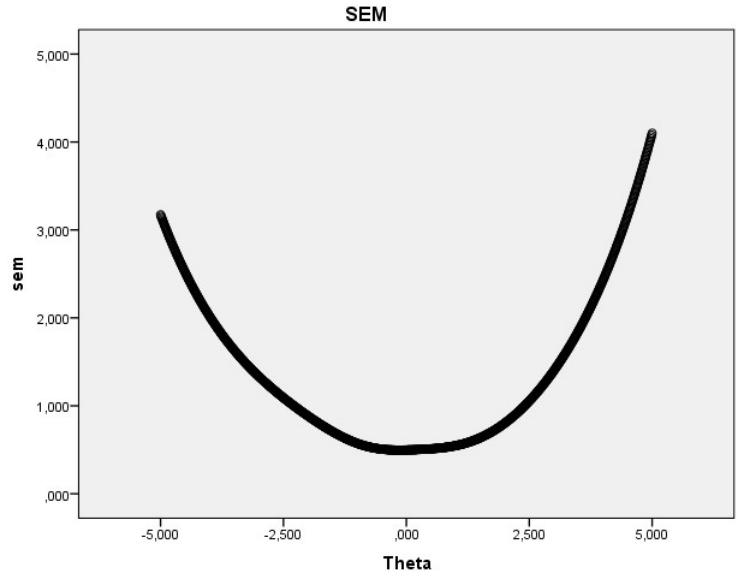


Figure 2.2.22. SEM under the current Rasch model

Figure 2.2.23 shows the item information curves. All the information curves are unimodal defining unique targets for the items.

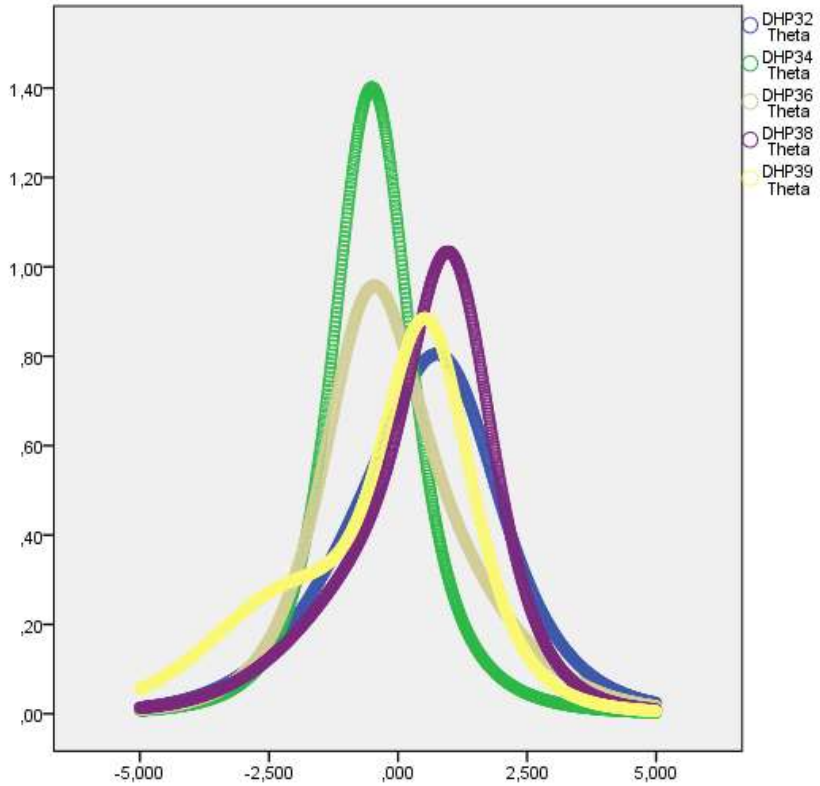


Figure 2.2.23. Item information curves for all items. The theta values where information is maximized define the targets of the item.

Category characteristic curves

The CCC curves plots the probabilities of item scores against θ. Figure

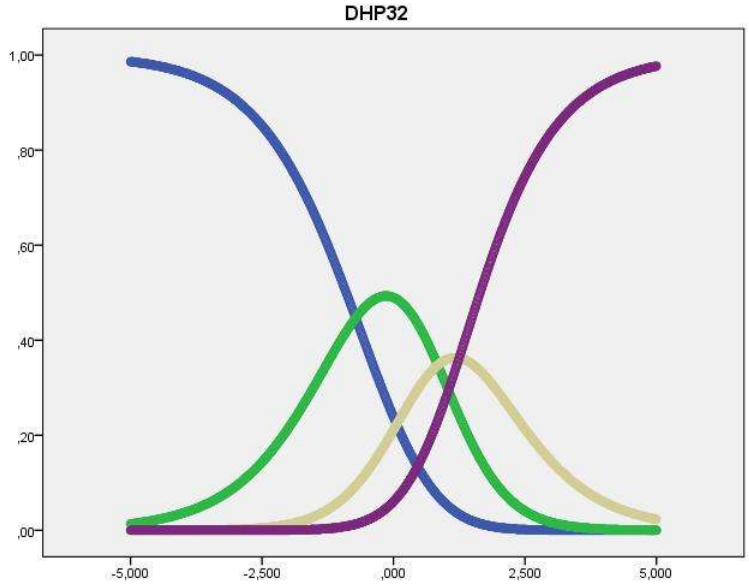


Figure 2.2.24. CCC curves for item A – DHP32

2.2.7.2 Scale anchored item distributions

A scale-anchored item distribution is the set of conditional probabilities of responses to items given (or anchored at) a specific value of the person parameter.

The CCC curve of Figure 2.2.24 describe the way the scale-anchored item distributions change as θ increase from -5 to +5, but the description is imprecise and less than transparent because the it is not easy to read this from CCC curves. For that we need tables with distributions anchored at specific θ values of interest.

DIGRAM has three commands that you may use to create such tables, **IPR**, **SPR** and **TPR**.

Item probabilities - IPR

Invoke the **IPR** command for tables relating to separate items with probabilities anchored at the WML estimates of the person parameters.

You do not have to ad parameters to this command, but DIGRAM need to know which items and WML estimates you are interested in so you have to respond to the two questions shown in Figure 2.2.25.

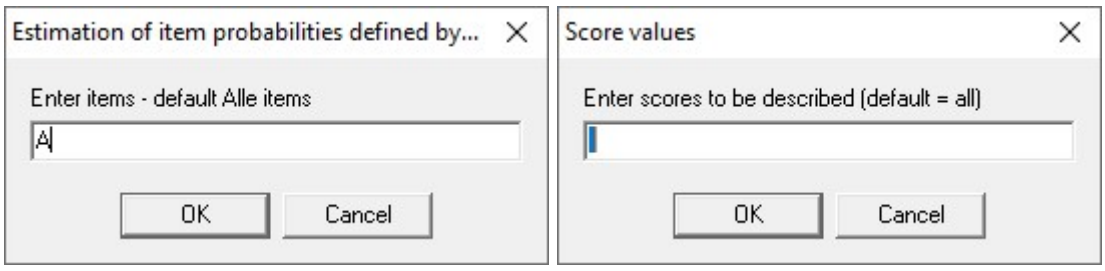


Figure 2.2.25 Information required by the IPR command

The result is shown in Figure 2.2.26. The IPR command produce on table for each item where you can ses how the response probabilities change as the person parameter (the WML) increase from lower to higher values. We suggest that you compare the results in Figure 2.2.26 with the CCC curve in Figure 2.2.24 and with similar tables for the other items that you have to create yourself.

```
  +-------+
  |       |
  | DHP32 |
  |       |
  +-------+

Score   WML     0       1       2       3 Expected  VAR/info
-------------------------------------------------------------
  1   -1.93  0.6145  0.2818  0.0968  0.0068   0.50    0.485
  2   -1.50  0.4722  0.3328  0.1758  0.0191   0.74    0.658
  3   -1.19  0.3601  0.3488  0.2533  0.0378   0.97    0.764
  4   -0.91  0.2681  0.3416  0.3262  0.0640   1.19    0.816
  5   -0.65  0.1926  0.3170  0.3912  0.0992   1.40    0.823
  6   -0.40  0.1322  0.2797  0.4435  0.1446   1.60    0.793
  7   -0.16  0.0868  0.2354  0.4781  0.1997   1.79    0.739
  8    0.08  0.0553  0.1906  0.4925  0.2616   1.96    0.672
  9    0.32  0.0345  0.1499  0.4884  0.3272   2.11    0.603
 10    0.55  0.0210  0.1146  0.4693  0.3952   2.24    0.537
 11    0.79  0.0120  0.0837  0.4364  0.4680   2.36    0.470
 12    1.07  0.0061  0.0562  0.3878  0.5498   2.48    0.399
 13    1.44  0.0024  0.0316  0.3161  0.6499   2.61    0.315
 14    2.06  0.0004  0.0109  0.2040  0.7846   2.77    0.200
```

Figure 2.2.26 Item distributions of DHP32 anchored at WML estimates

Score probabilities - SPR

Invoke the **SPR** command for tables relating to separate items with probabilities anchored at the WML estimates of the person parameters where output is organized in tables summarizing response probabilities for all items defined by a person parameter defined as the WML estimate for a single score. You have to select items and scores as in figure 2.2.25.

Tables with scale-anchored distributions are useful if you intend to generate criterion related categories that refer to probabilities of low and high scores. The default is to treat zero and maximum item score as low and high scores, but DIGRAM will let you select other limits.as shown in Figure 2.2.27.

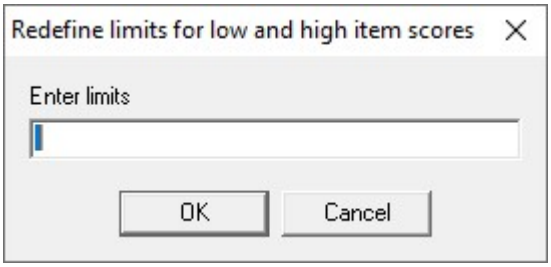


Figure 2.2.27. Definition of limits defining low and high scores.

We selected all items and all scores, but Figure 2.2.28 shows the tables for scores equal to 3, 8 and 13. An item is considered difficult at the given value of θ if the probability of a low score is larger than 0.75 and very difficult if the probability is larger than 0.90. Easy items are defined by similar probabilities for high scores. DHP34 is "difficult" at θ = -1.28 and easy at θ = 1.67.

```
  +-------------------------------------+
  |                                     |
  | Score = 3  Person parameter = -1.28 |
  |                                     |
  +-------------------------------------+

Difficult items:      DHP34

    Item    0     1     2     3       Mean Var/Inf
-------------------------------------------------
   DHP32 0.393 0.347 0.229 0.031     0.90  0.74
   DHP34 0.858 0.120 0.017 0.004     0.17  0.20   * difficult
   DHP36 0.551 0.396 0.045 0.009     0.51  0.39
   DHP38 0.480 0.166 0.308 0.045     0.92  0.96
   DHP39 0.672 0.158 0.165 0.005     0.50  0.61

Total info =  2.903

  +-------------------------------------+
  |                                     |
  | Score = 8  Person parameter =  0.08 |
  |                                     |
  +-------------------------------------+

Neither easy nor difficult items at this person level

    Item    0     1     2     3       Mean Var/Inf
-------------------------------------------------
   DHP32 0.056 0.191 0.492 0.261     1.96  0.67
   DHP34 0.464 0.254 0.140 0.142     0.96  1.17
   DHP36 0.166 0.466 0.205 0.163     1.36  0.89
   DHP38 0.057 0.076 0.552 0.315     2.13  0.60
   DHP39 0.164 0.151 0.613 0.072     1.59  0.71

Total info =  4.049

  +-------------------------------------+
  |                                     |
  | Score = 13  Person parameter =  1.67 |
  |                                     |
  +-------------------------------------+

Easy items:           DHP34

    Item    0     1     2     3       Mean Var/Inf
-------------------------------------------------
   DHP32 0.001 0.021 0.272 0.706     2.68  0.27
   DHP34 0.021 0.057 0.155 0.767     2.67  0.46   * easy
   DHP36 0.006 0.086 0.185 0.722     2.62  0.44
   DHP38 0.001 0.007 0.261 0.730     2.72  0.22
   DHP39 0.007 0.031 0.612 0.351     2.31  0.31

Total info =  1.713
```

Figure 2.2.28 Item distribution anchored at estimates of person parameters for scores equal to 4, 8 and 13.

The table generated by the SPR command is used to compare the response distributions for items at specific values of the person parameter. In addition to this, DIGRAM summarize the results counting the numbers of easy and difficult items at the estimates of the person parameters for each score. Figure 2.2.29 present the summary. There are no difficult items below θ = -0.68, neither and difficult nor easy items for -0.68 ≦ θ ≦ 1.20, and some easy if θ is larger than or equal to 1.67.

Summary

Score	Theta	Very easy	Easy	----	Difficult	Very difficult
1	-2.30	0	0	1	3	1
2	-1.67	0	0	3	1	1
3	-1.28	0	0	4	1	0
4	-0.96	0	0	4	1	0
5	-0.68	0	0	5	0	0
6	-0.42	0	0	5	0	0
7	-0.17	0	0	5	0	0
8	0.08	0	0	5	0	0
9	0.33	0	0	5	0	0
10	0.58	0	0	5	0	0
11	0.86	0	0	5	0	0
12	1.20	0	0	5	0	0
13	1.67	0	1	4	0	0
14	2.49	1	3	1	0	0

Figure 2.2.29 Summary of item distributions anchored at estimates of person parameters.

TPR – distributions anchored at θ values

The TPR command is similar to the SPR command, but it generates scale-anchored distributions at values of θ that you have to include as parameters when you invoke it. For instance, "**TPT -1.153 -0.865 -0.434 0.717**" generates distributions that are anchored at the PCM thresholds and the location of DHP32.

You also have to select the items for this command. We have selected DHP32 and got the results shown in Figure 2.2.30. The tables illustrate that probabilities of adjacent categories are the same if

the distribution is anchored at the threshold that separate them and the probabilities of extreme score are the same if the distribution is anchored at the location of the item.

```
  +--------------------------+
  |                          |
  | Person parameter = -1.15 |
  |                          |
  +--------------------------+

    Item    0     1     2     3       Mean Var/Inf
-----------------------------------------------------
   DHP32 0.349 0.349 0.262 0.040      0.99  0.77

  +--------------------------+
  |                          |
  | Person parameter = -0.87 |
  |                          |
  +--------------------------+

    Item    0     1     2     3       Mean Var/Inf
-----------------------------------------------------
   DHP32 0.254 0.338 0.338 0.070      1.22  0.82

  +--------------------------+
  |                          |
  | Person parameter = -0.43 |
  |                          |
  +--------------------------+

    Item    0     1     2     3       Mean Var/Inf
-----------------------------------------------------
   DHP32 0.139 0.285 0.438 0.139      1.58  0.80

  +--------------------------+
  |                          |
  | Person parameter =  0.72 |
  |                          |
  +--------------------------+

    Item    0     1     2     3       Mean Var/Inf
-----------------------------------------------------
   DHP32 0.014 0.092 0.447 0.447      2.33  0.49
```

Figure 2.2.30 Distributions of DHP32 anchored at the PCM thresholds (-1.15, -0.87, 0.72) and the item location (-0.43)

2.2.8 Assessment of the measurement quality

The fit of items to a Rasch model supports claims of the validity and objectivity of the measurement provided by the person score and estimates of person parameters. However, measurement quality require much more than this. Measurement has to be precise and unbiased and since the standard error and bias of measurement depend on the person parameter, we have to know whether the instrument provide precise and unbiased measurement in populations of interest.

To assess this issue, you have to select the "**Test information and targeting**" option. This section describe the results.

2.2.8.1 Targeting

We refer to the question of the degree to which measurement is precise and unbiased in a specific population as a question of targeting. We say that the instrument target a population if the average test information in the population is close to the maximum test info that the instrument provide and DIGRAM calculates a number of targeting indices to assess the degree of targeting for the study population.

If the analysis provided evidence of an effect of exogenous variables on the latent variable, we have to assess the degree of targeting in subpopulations defined by these variables. Since the DHP score depend on both sex and age, DIGRAM assess and compare targeting in all subgroups defined by these variables. Output from these analyses is verbose. For this reason, we only present the details for 60-70 years old women after which we present summary results for all the groups.

2.2.8.2 Test information and targeting of 60-70 years old women

If we want to assess the targeting of a measurement instrument in a specific population, we need to know the distribution of θ. During the analysis of targeting, DIGRAM assumes that the distribution is normal and estimates the mean and the variance. Since inference in DIGRAM is conditional without assumptions of the distribution, it is important to stress that DIGRAM only use the normal distribution for illustrative purposes. The results will show how the measurement instrument function *if* the population is normal. To make the results plausible, DIGRAM calculates a χ^2 test

that compare the observed distribution of the score to the expected distribution, but you should not take the result for more than that unless you have specific hypotheses about the distribution of θ.

According to Figure 2.2.31, the mean of the distribution of θ is equal to 0.512 among 60-69 years old women. The standard deviation is 0.823 and the test accepts the normal distribution (p = 0.55).

Figure 2.2.32 provides information on the test location[25], the test midpoint and the test target where the target info is equal to 4.05. The target lies a little above the location, but below the population mean. The degree to which this implies that DHP is out of target remains to be seen.

Table 2.2.33 presents estimates of sensitivity and specificity relative to cut points on the θ scale. Consider for instance a situation where you want to classify the persons as lying below or above the 10 % percentile of the normal population with mean = 0.512 and sd = 0.823. The 10 % percentile is -0542. The sensitivity is the probability that a person from this distribution has a WML estimate less than or equal -0.542 while the specificity is the probability that a person with θ > -0.542 has a WEML estimate above -0.542. Specificity is adequate, but we leave it to decide whether we should be satisfied with the sensitivities.

Figure 2.2.34 summarize the average precision and bias of the ML and WML estimates and define target indices comparing the average test information and SEM in the normal distribution with the values at test target. Target indices above 0.80 indicate relative good targeting compared to many examples of health related scales known to us.

Finally, Figure 2.2.35 presents measurer of reliability. To us, the probabilities of correct or no person separation are more meaningful than the other options, but you are free to disagree and decide whether reliability is adequate.

Figure 2.2.35 also provide estimate of the bias of the difference between two random persons from the normal distribution. Had measurement functioned as proper interval scaled measurement there should be no bias. The results show that WML estimation provide measurement that function as interval scaled measurement degree to a higher degree than ML estimates.

[25] Recall that we have fixed the parameters in such a way that the test location has to be at the origin of the θ scale

```
Group: SEX = Female   AGE = 60-69

  +-----------------------------------------------+
  |                                               |
  | Test of normal person parameter distribution  |
  |                                               |
  +-----------------------------------------------+

Observed and fitted score group distributions

         0 - 8  9 - 11 12 - 15
   --------------------------
   Obs    19.0    10.0    13.0
   Fit    16.0    12.8    13.2
   --------------------------

   Chi**2=   1.19  df = 2  p =  0.5506


  42 persons.   Mean score =   9.43    sd =   3.30  Mean Theta =  0.512    sd =  0.823
```

Figure 2.2.31 The distribution of 60-69 years old women

```
Location        =  -0.000  Test information  =  4.04  SEM =   0.498  True score =   7.7 SEM(TS) =  2.01
Test midpoint   =  -0.042  Test information  =  4.03  SEM =   0.498  True score =   7.5 SEM(TS) =  2.01
Test target     =   0.132  Max test info     =  4.05  SEM =   0.497  True score =   8.2 SEM(TS) =  2.01
```

Figure 2.2.32 Location, midpoints and target of DHP scores

```
*** Sensitivity and specificity ***

                                                    False     False
                    Theta  Sensitivity Specificity positive Negative
< 5% percentile    -0.841      0.727        0.948    0.575    0.015
< 10% percentile   -0.542      0.732        0.931    0.461    0.031
> 90% percentile    1.566      0.521        0.944    0.490    0.053
> 95% percentile    1.865      0.633        0.926    0.691    0.020
```

Figure 2.2.33 Analysis of sensitivity and specificity

```
**** ML estimates  ****
Mean bias =      0.107  sd =   0.093
Mean RMSE =      0.690  sd =   0.151
Mean SEM  =      0.678  sd =   0.137

 **** WML estimates ****
Mean bias =      0.010  sd =   0.026
Mean RMSE =      0.585  sd =   0.101
Mean SEM  =      0.584  sd =   0.100

Mean test information        =   3.296  sd =   0.865   Target index =   0.814
Mean SEM defined by test inf =   0.575  sd =   0.123   Target index =   0.864
```

Figure 2.2.34 Summary of bias and precision

```
var(true score)/var(score)      =  0.697
Person separation index(PSI)    =  0.676
test-retest correlation         =  0.693
test-true score correlation     =  0.845

Probability of correct person separation =  0.772
Probability of no person separation      =  0.086

Bias of interval estimates
ML estimate   : 0.1310
WML estimate  : 0.0155
```

Figure 2.2.35 Reliability and bias of interval estimation

Figures 2.2.36 and 2.2.37 summarize the results for six groups defined by sex and age. Targeting is generally good, but reliability could be better, suggesting that five polytomous items are not enough for quality measurement.

```
  +--------------------------------+
  |                                |
  | Targeting and test information |
  |                                |
  +--------------------------------+
```

SEX	AGE	Target	n	theta Mean	theta sd	test info Mean	test info max	target index	RMSE(WML) Mean	RMSE(WML) min	target index	PSI
Male	18-49	0.13	14	0.32	0.83	3.415	4.051	0.843	0.568	0.497	0.874	0.681
Female	18-49	0.13	17	-0.30	0.92	3.392	4.051	0.837	0.555	0.497	0.895	0.769
Male	50-59	0.13	33	0.59	0.59	3.432	4.051	0.847	0.572	0.497	0.869	0.509
Female	50-59	0.13	23	0.33	0.69	3.553	4.051	0.877	0.555	0.497	0.896	0.592
Male	60-69	0.13	47	0.80	0.83	3.025	4.051	0.747	0.618	0.497	0.804	0.669
Female	60-69	0.13	42	0.51	0.82	3.296	4.051	0.814	0.585	0.497	0.850	0.676

```
  +--------------------------------+
  |                                |
  | targeting and test info summary |
  |                                |
  +--------------------------------+
```

	Target	Theta	Info	max	n
Over all	0.130	0.486	3.301	4.051	176

SEX	Target	Theta	Info	max	n
Male	0.130	0.655	3.226	4.051	94
Female	0.130	0.292	3.388	4.051	82

AGE	Target	Theta	Info	max	n
18-49	0.130	-0.020	3.402	4.051	31
50-59	0.130	0.483	3.482	4.051	56
60-69	0.130	0.663	3.153	4.051	89

Figure 2.2.36 Summary of targeting for subpopulations defined by sex and age.

True score estimation and reliability

SEX	AGE	Target	n	Score Mean	Score sd	separation reliability	sep. prob	no sep. prob	Target SEM(TS)	Mean SEM(TS)
Male	18-49	8.21	14	8.71	3.43	0.709	0.776	0.081	2.01	1.83
Female	18-49	8.21	17	6.65	3.62	0.746	0.797	0.075	2.01	1.83
Male	50-59	8.21	33	9.85	2.74	0.548	0.719	0.097	2.01	1.84
Female	50-59	8.21	23	8.87	3.08	0.638	0.748	0.091	2.01	1.87
Male	60-69	8.21	47	10.34	3.10	0.679	0.748	0.098	2.01	1.71
Female	60-69	8.21	42	9.43	3.30	0.693	0.772	0.086	2.01	1.80

Weighted means: Reliability = 0.66 Person separation = 0.76

Targeting summary

SEX	AGE	n	Target Theta	Target SEM	Target TS	Target SEM	Population average Theta	Population average SEM	Population average TS	Population average SEM	reliability
Male	18-49	14	0.13	0.50	8.21	2.01	0.32	0.57	8.71	1.83	0.71
Female	18-49	17	0.13	0.50	8.21	2.01	-0.30	0.55	6.65	1.83	0.75
Male	50-59	33	0.13	0.50	8.21	2.01	0.59	0.57	9.85	1.84	0.55
Female	50-59	23	0.13	0.50	8.21	2.01	0.33	0.55	8.87	1.87	0.64
Male	60-69	47	0.13	0.50	8.21	2.01	0.80	0.62	10.34	1.71	0.68
Female	60-69	42	0.13	0.50	8.21	2.01	0.51	0.58	9.43	1.80	0.69

Figure 2.2.37 True score (TS) estimation and summary

2.2.8.2 Item Maps

Another way to compare the locations of the items and the persons is to use so-called item maps[26]. DIGRAM will not plot such maps but creates text files with information that you can use to create them with a statistical program of your own choice.

The data with information on item maps is called Imap.txt. SPSS user may use the Imap.sps created by DIGRAM together with the data file. Figure 2.2.38 show the beginning of Imap.txt. F and G are the exogenous variables on which the distribution of θ depends. Theta is the person parameter. Type indicates the ML estimates (0), the WML estimates (1), the estimated population distribution (2), the PCM thresholds (3), the item targets, the item info (5) and the SEM defined by the item info (6). and the score thresholds (4), . The weight indicates the size of the bars of the item maps[27].

```
F G theta type weight
1 1 -1.89 2 5
1 1 -1.80 2 2
1 1 -1.72 2 2
1 1 -1.63 2 3
1 1 -1.54 2 4
1 1 -1.46 2 4
1 1 -1.37 2 5
1 1 -1.29 2 7
1 1 -1.20 2 8
1 1 -1.12 2 10
```

Figure 2.2.38 The start of the IMAP.txt file.

The item maps plot the estimated distribution of the person parameter within the 99 % confidence range, the distribution of the item thresholds and the distribution of the thresholds of the distribution of the score.

Figure 2.2.39 show the item map of women aged 60-69 years. The map includes the WML, the estimated normal distribution, the PCM thresholds and the item targets.

Figure 2.2.40 show the info map plotting the distribution of the persons above the test info and SEM curves. It is easy to see that DHP is a little bit out of target for this population.

[26] Item maps are sometimes referred to as Wright maps, because Benjamin Wright was the first to suggest them.
[27] The formatting of the maps are described as part of the targeting output for the different groups defined by F and G.

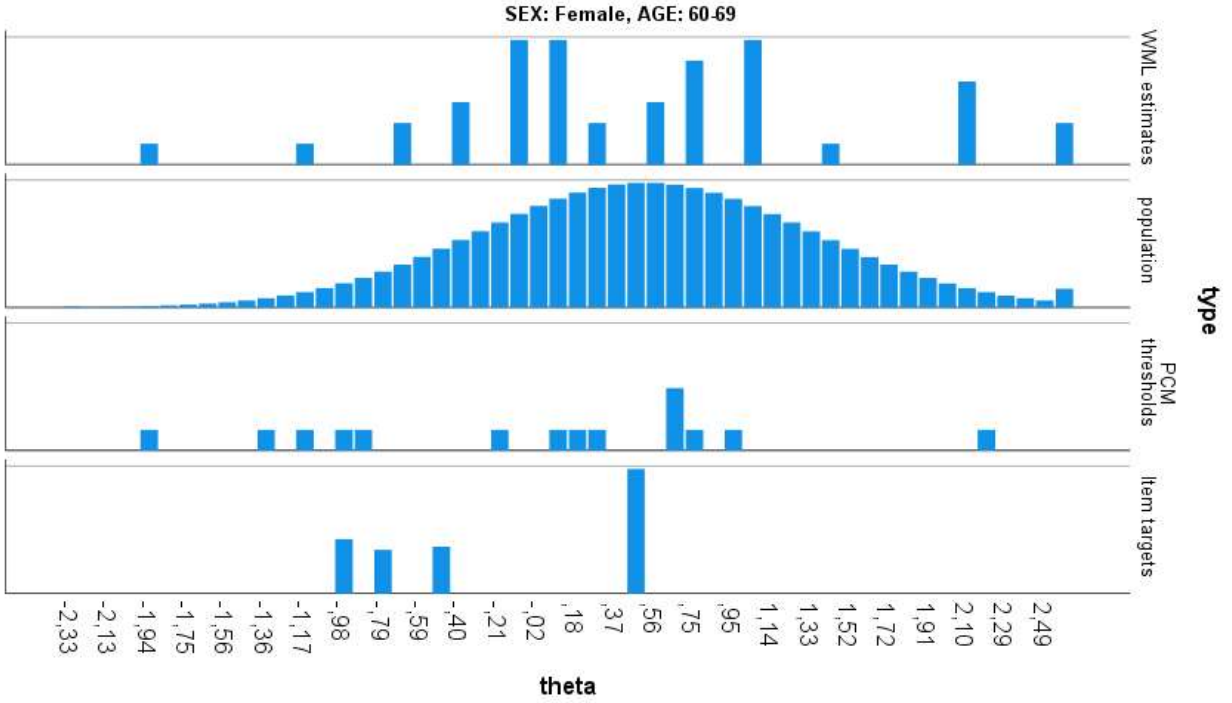


Figure 2.2.39 Item map for women, age 60-69.

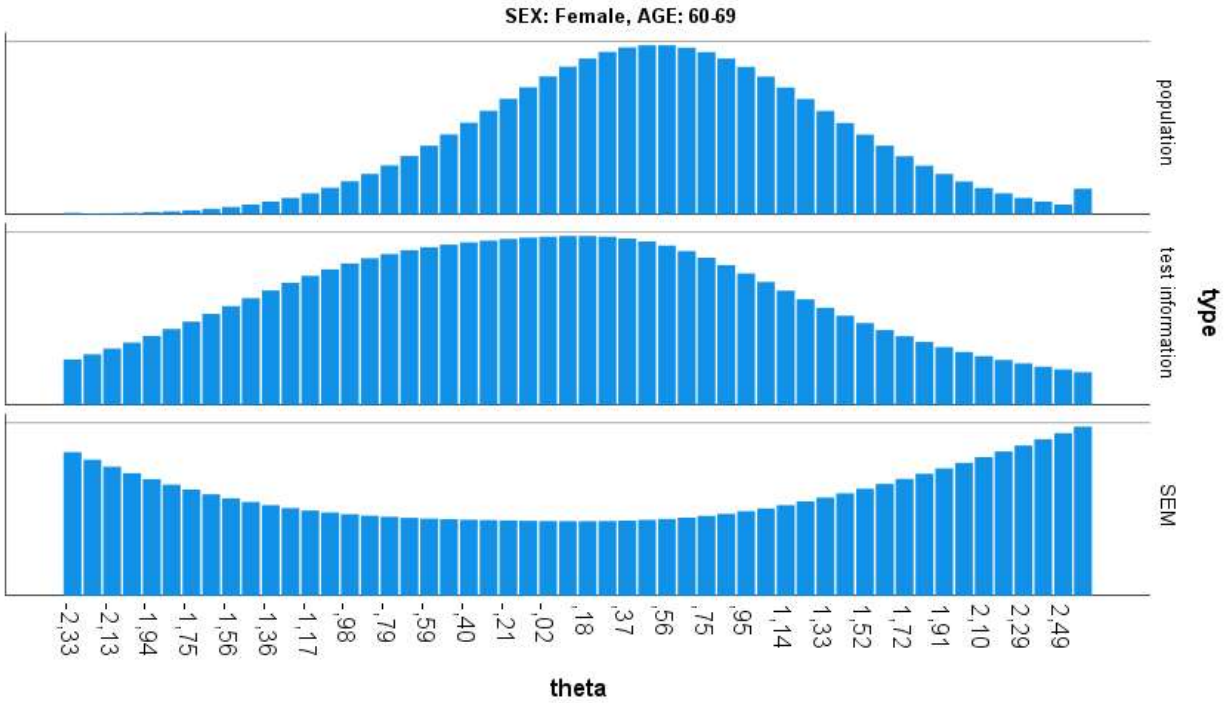


Figure 2.2.40 Info map for women men, age 60-69.

2.3 Graphical log-linear Rasch models. The Short tour.

On this tour, we return to the original version of the DHP scale where the total score measure the level of disinhibited eating. The initial analysis provided evidence suggesting 1) DIF of item C (DHP36) relative to G (Sex), 2) local dependence (LD) between items B (DHP34) and D (DHP38), 3) local dependence between items D (DHP38) and E (DHP39), and 4) that the item discrimination of item D was stronger than expected by the Rasch model. During this tour, we will define and test a graphical log-linear Rasch model that take this evidence into account and test whether this model fits data.

2.3.1 Definition of graphical log-linear Rasch models

The evidence against the fit of DHP items to the Rasch model is so comprehensive that the scale won't survive attempts to purify it by elimination of items. A much better option is to attempt to fit a GLLRM where uniform DIF and uniform local dependence[28] is accepted, because it follows from the sufficiency of the total score in GLLRMs that they possess the same fundamental properties as the Rasch models.

GLLRMs are defined by generating sets (subsets of items and exogenous variables) defining log-linear interaction among variables). The current version of DIGRAM only permits two-way interactions. The generating sets for a model defined by the evidence of DIF and local dependence is therefore equal to (BD, DE, CG).

There are three ways to tell DIGRAM that this is the model that you want to define.

i. You can invoke the "GRM BD DE CG" command. DIGRAM will understand that this is GLLRM that you will use as the current model.

ii. You can invoke the GRM command without parameters or click on the GRM button. DIGRAM will define the Rasch model as the current model, following which you have to define the GLLRM when the GRM dialog is enabled.

[28] DIF is uniform if the *strength* of the association between the item and the source of DIF does not depend on the person parameter. In the same way, we say that local dependence is uniform if the strength of the association between two variables is the same for all values of the person parameter.

iii. You can use and "ITA BD DE CG" command. DIGRAM will take you to the GRM dialog, define and estimate the parameters of the GLLRM and return to DIGRAM's main dialog

iv. Finally you can define click on the IRT graph and add edges between the variables to the IRT graph as shown in Figure 2.3.1 and then invoke the GRM command *without* parameters or click on the GRM button.

Figure 2.3.1.shows the IRT graph of the model. We assume that you have used the ITA command and the wanted to see the model. Notice, that the "Toggle GLLRM gammas" button has been enabled. We will return to that later. For now, you only need to know that the parameters of the item parameters have been estimated.

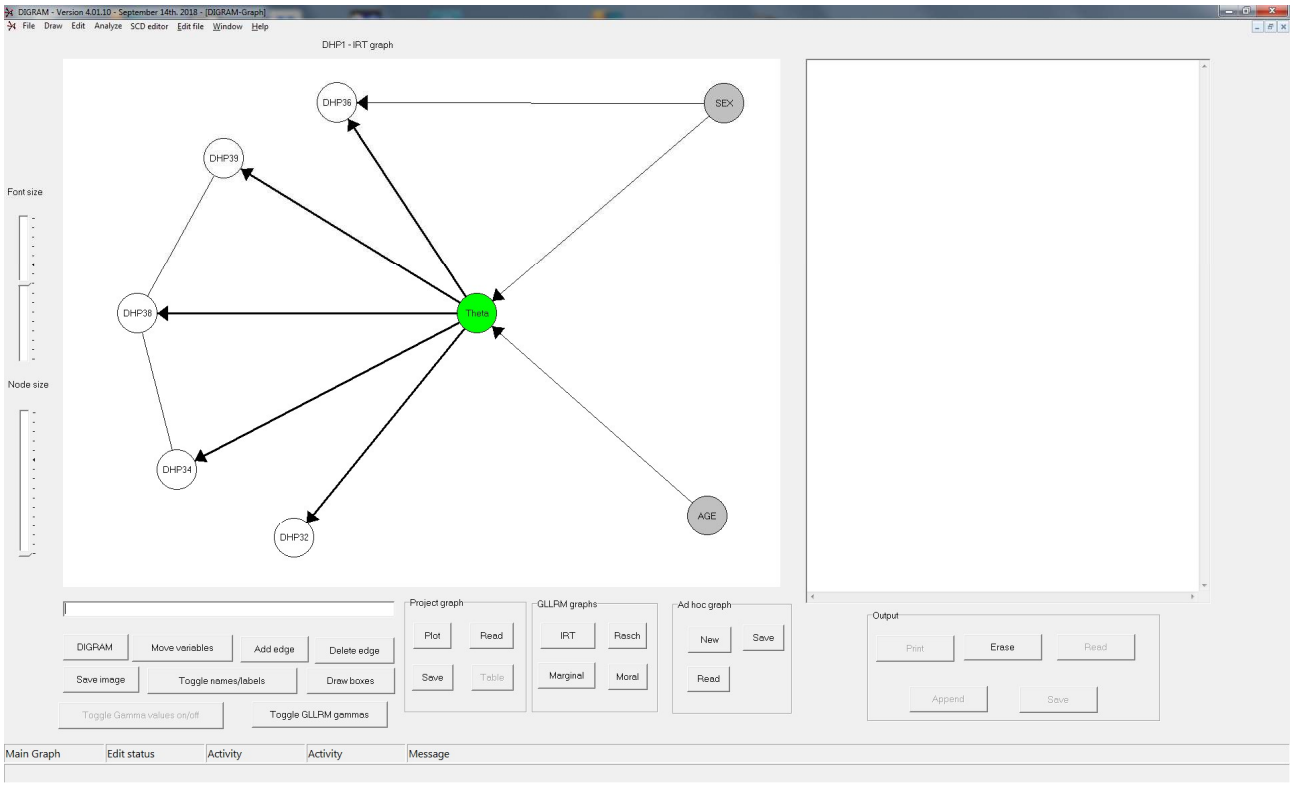


Figure 2.3.1 IRT graph of a GLLRM with DIF and local dependence. The graph has been edited to make it easier to read.

Before we go into the details of item analyses by GLLRMs it is useful to take a closer look at the GRM dialog (Figure 2.3.2) which appears when you click on the GRM button. The dialog box contains two model fields: the "Current model" field with model terms written in red and the "New model field" with model terms written in black.

The *current model* is the model defined by the IRT graph or by the parameters added to the GRM command. DIGRAM assumes that this is your preferred model. You cannot edit this field.

The new model is the model that DIGRAM will use for the analysis from the GRM dialogue. To begin with, DIGRAM copies the current model to the new model assuming that you want to examine this model, but if you want to fit a different model, you have to define this model in the "new model" field by adding and/or deleting model terms.

You cannot edit the "current model" field, but you may replace the current model with the new model by pushing the "Change model" button. If you, on the other hand, want to discard the new model and return to the current model you can press the "Use current model" button instead.

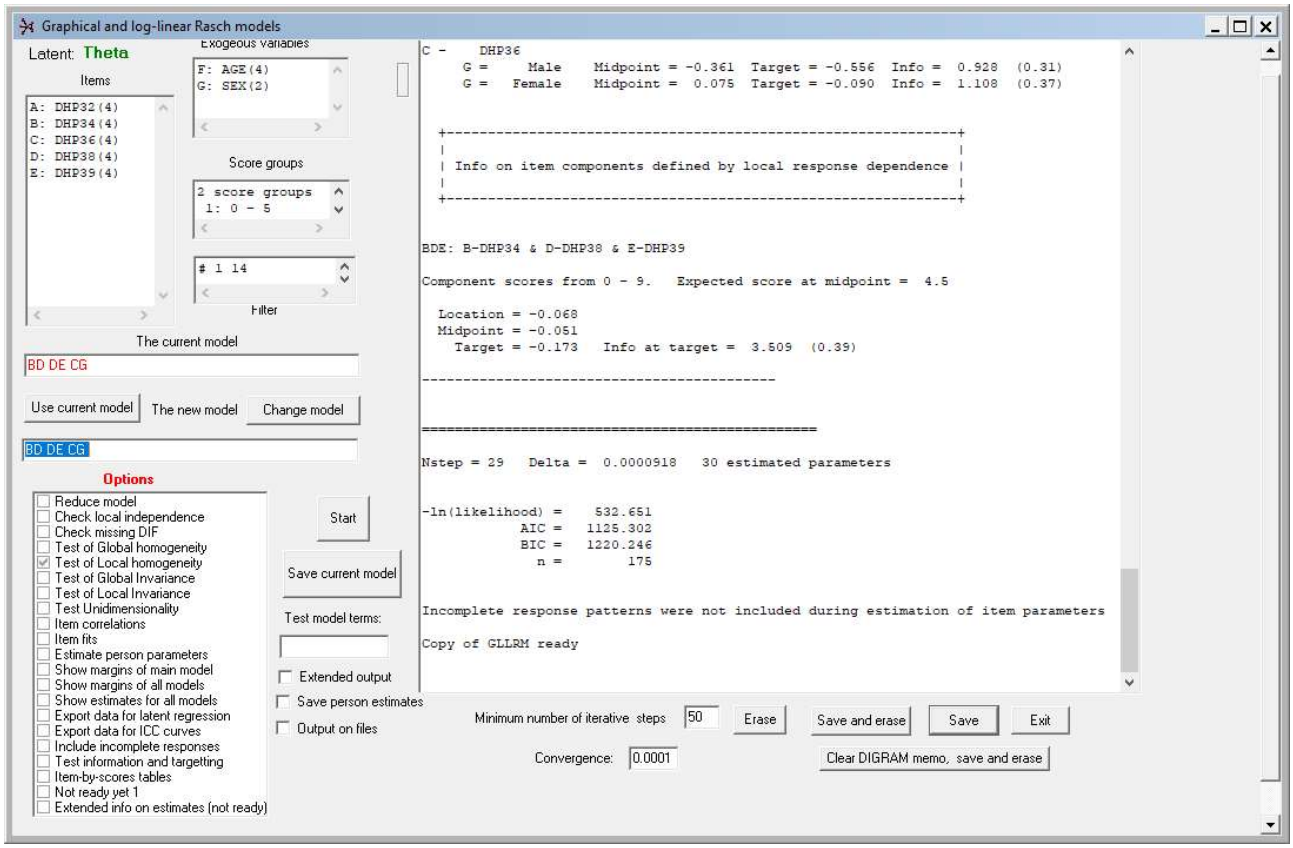


Figure 2.3.2 The GRM dialog form following either editing of the IRT graph or a "GRM BD DE CG" command. Item parameters have been estimated.

2.3.2 Item analysis by GLLRMs

Since we have just defined the current model, there is no reason to define another model. We proceed directly to item analysis by GLLRMs.

Until you want to define a new model, the analysis proceeds in exactly the same way for the GLLRM as for the ordinary Rasch model except that estimates of item parameters include estimates of interaction parameters relating to local dependence and DIF and parameters defining the distribution of so-called component scores over locally dependent items.

You have to

a) Estimate the item parameters by pushing the START button,

b) select "**Test global homogeneity**" for CML tests of homogeneity and invariance,

c) select "**Item fits**" for the same item fit statistics as for the Rasch model,

d) select "**Reduce model**" for confirmatory tests of the model's claims of local dependence and DIF.

e) select "**Check local independence**" and "**Check missing DIF**" to make sure that there is no evidence of DIF and LD above and beyond the DIF and LD in the model,

f) select "**Export data for ICC curves**" if you want to se the ICC, IIC and CCC curves,

g) select "**Estimate person parameters**" for estimates of person parameters,

h) select "**Test information and targeting**" to assess the appropriateness of the items for the current study population.

Steps a) to e) have to be repeated until you have an adequate model before it makes sense to take steps f) to h).

2.3.3 Estimates of item parameters

DIGRAM estimates the item parameters when you invoke the GRM dialog, but you have to click start to estimate the parameters again, if you revise the new model.

Figure 2.3.3 shows the estimates of the multiplicative item parameters.

The multiplicative parameters include main effect parameters corresponding to the multiplicative parameters of the Rasch model and multiplicative interaction parameters. The main effects

parameters are fixed so that the product of the parameters for the maximum item scores is equal to one.

The log-linear interaction parameters are fixed in such a way that the interaction parameters corresponding to reference categories for both variables in an interaction terms are equal to 1. DIGRAM prefers to use the first category as the reference category, but selects another reference category to avoid situations where combinations of a reference category on one variable and a value of another variable have not been observed.

```
       item           0         1         2         3
-------------------------------------------------------
A:    DHP32         1.000     1.750     0.741     0.257
B:    DHP34         1.000     0.724     0.662     0.827
C:    DHP36         1.000     1.863     4.460     2.103
D:    DHP38         1.000     4.031     1.662     1.707
E:    DHP39         1.000     3.792     0.918     1.312

LD: DHP34(B)  & DHP38(D)
                  B
 D            0         1         2         3

 0          1.000     1.000     1.000     1.000
 1          0.116     0.213     0.673     1.000
 2          0.000     0.000     0.000     1.000
 3          0.253     0.395     1.203     1.000

Standardized Gamma =  0.556 (G2OR =  3.50)

LD: DHP38(D)  & DHP39(E)
                  D
 E            0         1         2         3

 0          1.000     0.227     0.000     0.000
 1          1.000     1.856     0.401     0.141
 2          1.000     1.080     5.269     0.290
 3          1.000     1.000     1.000     1.000

Standardized Gamma =  0.556 (G2OR =  3.50)

DIF:    item: DHP36(C)      DIF source: SEX(G)
                      C
G                0         1         2         3

1    Female    1.000     1.000     1.000     1.000
2    Male      1.000     0.326     0.320     0.252

Standardized Gamma = -0.350 (G2OR = 0.48)
```

Figure 2.3.3 Estimates of multiplicative item parameters

The log-linear interaction parameters are cross-product ratios[29] in 2x2 tables defined by the reference values and a single column category and a single row category. The complete set of log-linear parameters describe the conditional distribution of items and exogenous variables given θ. If both variables are ordinal or binary, we describe the strength of the partial correlation by the two variables by Goodman and Kruskal's γ in the table defined by the log.-linear structure defined by the parameters and the marginal distributions of the variables[30].

The gamma values referred to in Figure 2.3.1 is the standardized γ values. If you select "Toggle GLLRM gammas", the standardized γ values will be added to the edges and arrows connecting items and exogenous variables.

Figure 2.3.4a shows the PCM thresholds of the two locally independent items under the GLLRM. One of these (Item A - DHP32) is a conventional PCM item whereas item C (DHP36) is a DIF item functioning differently among men and women.

The PCM thresholds of DHP36 illustrate the limitations of the PCM parametrizations. It is clear that the location is lower for men than for women, but it is less clear whether or not the categories function differently for men and women and in which way. To address this issue we have to look at the version of the Rasch models that separates item and category effects.

Figure 2.3.4b shows the PCM thresholds associated with the locally dependent items. The GLLRM defines one item component with items that are directly or indirectly associated. Under the GLLRM component scores summarizing the responses to item of item components are polytomous super items fitting the Rasch items with PCM thresholds defined by the main effects of the items and the log-linear interaction parameters. Figure 2.3.4b shows the PCM thresholds of the B+D+E super item.

[29] Often interpreted as odds-ratios. The G2OR values included in Figure 2.33 are functions of the gamma values: G2OR = (1+γ)/(1-γ). In 2x2 tables, G2OR is a true odds-ratio value. In larger tables as in Figure 2.3.3, G2OR is not a true odds-ratio, but they may still be useful if you want to compare associations in tables with ordinal categorical values with association in 2x2 tables.

[30] We refer to γ coefficient defined by the marginal distributions of variables as standardized γ value. Since Goodman and Kruskal's γ depend on the marginal distributions of variable, the measure of the partial correlation of the variables would have been different if we had fitted the lo-linear structure to uniform marginal distributions.

```
  +------------------------------+
  |                              |
  | PCM thresholds and locations |
  |                              |
  +------------------------------+

  +---------------------------+
  |                           |
  | Locally independent items |
  |                           |
  +---------------------------+

  item                 1        2        3    Location
-------------------------------------------------------
A:    DHP32        -0.559    0.860    1.059     0.453
C:    DHP36 * DIF item *
  G =   FeMale     -0.622 > -0.873    0.752    -0.248
  G =   Male        0.498 > -0.854    0.990     0.211
```

Figure 2.3.4a PCM thresholds of locally independent items

```
  +-----------------------------------------------------+
  |                                                     |
  | Item components defined by local response dependence |
  |                                                     |
  +-----------------------------------------------------+


B-DHP34 & D-DHP38 & E-DHP39   Component scores from 0 to 9

Thresholds:

-1.531 -0.518 -0.308 > -0.617 -0.361 1.006 > -0.219 1.171 > 0.761
```

Figure 2.3.4b PCM thresholds of the component score B+D+E. The location is equal to -0.068

Figure 2.3.5 shows the item and category effects of the main effects of the DHP items. The item effect of this item is stronger for men than for women. Because of the orientation of the items, this means that it is easier to stop eating for women than for men.

item		0	1	2	3	Item effect
A:	DHP32	0.000	1.013	0.606	0.000	-0.453
B:	DHP34	0.000	-0.260	-0.286	0.000	-0.063
C:	DHP36 * DIF item *					
G =	Female	0.000	0.375	1.000	-0.000	0.248
G =	Male	0.000	-0.286	0.779	0.000	-0.211
D:	DHP38	0.000	1.216	0.151	-0.000	0.178
E:	DHP39	0.000	1.242	-0.266	0.000	0.090
	---- MICE effects ----					
A:	DHP32	1.000	2.753	1.833	1.000	0.636
B:	DHP34	1.000	0.771	0.751	1.000	0.939
C:	DHP36 * DIF item *					
G =	Female	1.000	1.454	2.717	1.000	1.281
G =	Male	1.000	0.751	2.179	1.000	0.809
D:	DHP38	1.000	3.373	1.163	1.000	1.195
E:	DHP39	1.000	3.464	0.766	1.000	1.095

Figure 2.3.5 Item and category effects defined by main effects of DHP items

The category effects of DHP36 = 1 ("Quite easy to stop eating") of women is close to half the category effect of men. To understand the effect of this we have to look at scale-anchored distributions. Figure 2.3.6 show the distributions of C anchored at the locations of DHP36 for men and women. At the item locations, men and women agree that the probability of "Not very easy" is 0.44 and the expected item score is close to the same for men and women. But they disagree about the probability of "Quiet easy". Women select his option more often than either of the extreme categories whereas men prefer the extreme categories to a higher degree than "Quite easy to stop eating".

		How easy do you find it to stop eating					
θ	Sex	Very easy	Quiet easy	Not very easy	Not at all	mean	INF
-0.25	**F**	**0.16**	**0.24**	**0.44**	**0.16**	**1.60**	**0.89**
	M	0.38	0.18	0.34	0.10	1.14	1.09
0.21	F	0.07	0.16	0.48	0.28	1.98	0.73
	M	**0.20**	**0.15**	**0.44**	**0.20**	**1.64**	**1.04**

Figure 2.3.6 Item distribution of DHP36 anchored at θ equal to locations of item C for men and women.

Figure 2.3.7 present the midpoint and targets of the items under the GLLRM. At the targets of DHP36, the target info is a little stronger for men than for women.

```
  +---------------------------+
  |                           |
  | Locally independent items |
  |                           |
  +---------------------------+

A -    DHP32       Midpoint =  0.534  Target =  0.697  Info =  0.873  (0.29)
C -    DHP36
     G =   Female  Midpoint = -0.361  Target = -0.556  Info =  0.928  (0.31)
     G =   Male    Midpoint =  0.075  Target = -0.090  Info =  1.108  (0.37)

  +-------------------------------------------------------------+
  |                                                             |
  | Info on item components defined by local response dependence |
  |                                                             |
  +-------------------------------------------------------------+

BDE: B-DHP34 & D-DHP38 & E-DHP39

Component scores from 0 - 9.   Expected score at midpoint =  4.5

  Location = -0.068
  Midpoint = -0.051
    Target = -0.173   Info at target =  3.509  (0.39)

-----------------------------------------
```

Figure 2.3.7 Item midpoints and targets

You can also select "Export data for ICC curves" for GLLRMs. The text files with data for the ICC curves also provides information on item components with locally dependent items and on DIF. The ICC text file defined by the (BD, DE,CF) model information on

Theta : the person parameter
F : the DIF source (Gender)
Score : the expected (true score)
A : the expected score on item A
BDE : the expected score on the B+D+E item component
B : the expected score on item B
D : the expected score on item D
E : the expected score on item E
C : the expected score on item C
Type : 0 = ICC values, 1 = Observed frequencies
n : number of persons (1 if Type = 0);

Figure 2.3.8 shows the ICC curves for men. Note that the ICC curves are steeper for the locally dependent items than for the pure Rasch items.

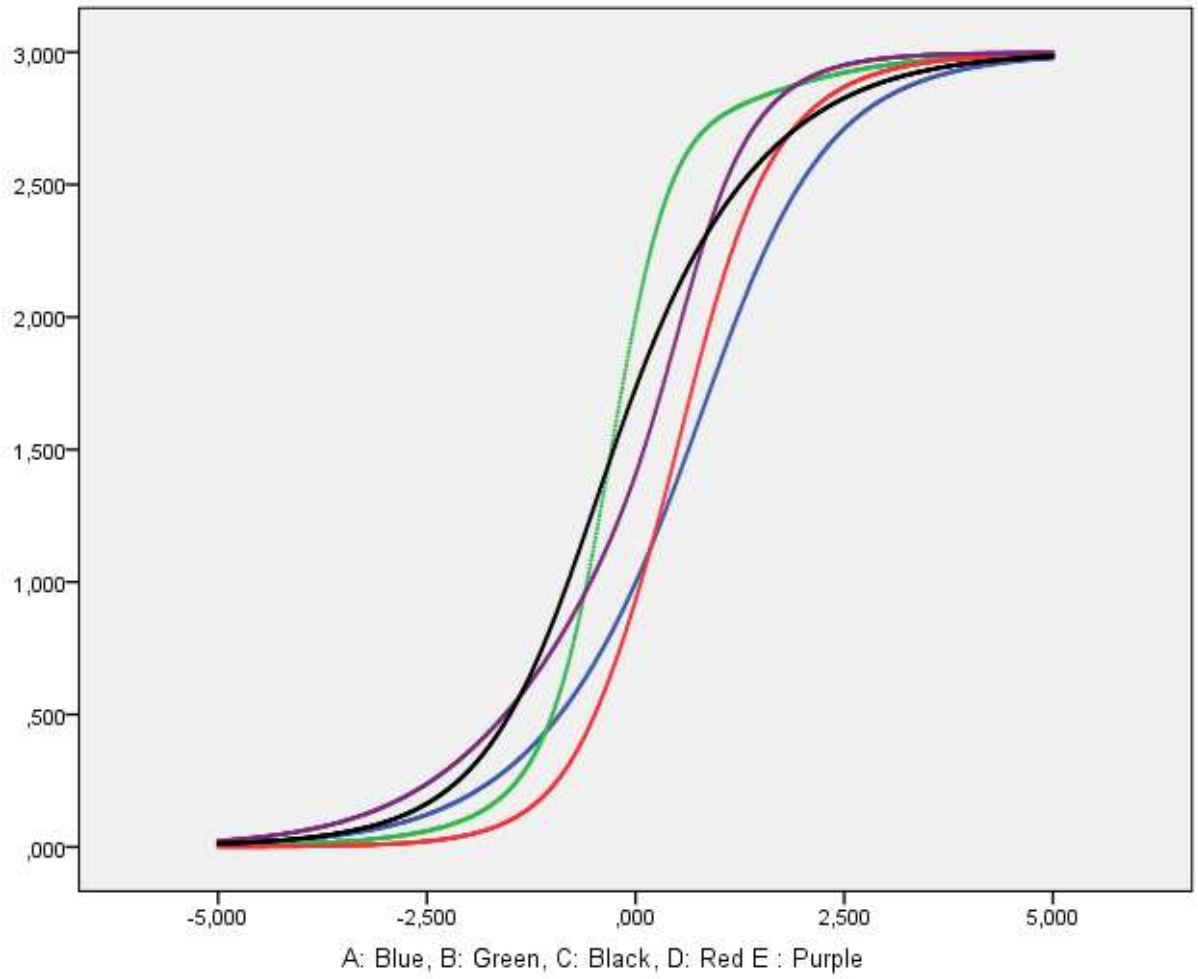


Figure 2.3.8 ICC curves for women (G = 1)

2.3.4 Tests of homogeneity and invariance

Figure 2.3.9 shows the CLR tests of homogeneity and invariance. There is no evidence against the model.

```
Summary of global test results. Delta will
be reported if estimation did not converge.

              CLR   df   p         delta
-------------------------------------
scoregroups   19.2  30 0.936
F:      AGE   99.9  90 0.223      0.001
G:      SEX   25.2  24 0.394
```

Figure 2.3.9 Overall fit statistics of the BD,DE,CF model

2.3.5 Item fits

Figure 2.3.10 shows the item fit statistics. After including the local dependence and DIF suggested by the tests of the Rasch model there is no evidence against item fit. In this example, we therefore conclude that the evidence of too strong item discrimination of item D is spurious evidence caused by local dependence. The item fits include tests of fit of the component score B+C+D.

```
  +--------------------------------+
  | Conditional outfits and infits |
  +--------------------------------+

               Outfit                          Infit
Item          observed     sd       p       observed     sd       p
-------------------------------------------------------------------------
A -    DHP32    1.075    0.103  0.46476       1.082    0.107  0.44439
B -    DHP34    1.058    0.144  0.68767       1.014    0.117  0.90562
C -    DHP36    0.954    0.088  0.60072       0.959    0.084  0.62402
D -    DHP38    0.982    0.216  0.93251       1.070    0.146  0.63065
E -    DHP39    0.953    0.145  0.74700       0.980    0.129  0.87692
-------------------------------------------------------------------------

  +-----------------------------+
  | Item rest-score association |
  +-----------------------------+

                  Item-rest-score gamma
Item           observed expected    sd      p
-------------------------------------------------
A -    DHP32    0.296    0.325    0.073  0.69862
B -    DHP34    0.494    0.497    0.058  0.96664
C -    DHP36    0.335    0.307    0.069  0.68380
D -    DHP38    0.681    0.658    0.056  0.68407
E -    DHP39    0.514    0.518    0.068  0.95517
-------------------------------------------------

Critical levels adjusted by the Benjamini-Hochberg procedure:
* < 5 % FDR, ** < 1 % FDR, *** = FDR < 0.1 % FDR

  +---------------------------------------+
  | Component-rest-score gamma coefficients |
  +---------------------------------------+

                   Gamma
Component      observed expected    sd      p
-------------------------------------------------
   BDE           0.407    0.353    0.064   0.3947

+-------------------------------------------------------------------------+
| Benjamini-Hochberg limits for all outfits, infits and gamma coefficients |
+-------------------------------------------------------------------------+

FDR = 5 %. Limit =  0.00313
FDR = 1 %. Limit =  0.00063
```

Figure 2.3.10 Item fit statistics under the BD,DE,CG model.

2.3.6 Confirmatory test of DIF and local dependence

The results of the item analysis in the previous section confirmed that there was nothing wrong with the model. The only thing that we have not considered yet is the question of whether we really need to include the two pairs of local dependency and the DIF of item C relative to F (Sex). For this purpose, we also use Kelderman's CLR test. The test is calculated for each of the interaction terms in the model where the model without a term is the null-hypothesis and the "new" model is the alternative. Select "Reduce model" to get these tests. Figure 2.3.11 shows the results. All hypotheses are rejected, but it has to be admitted that the evidence of DIF is not strong (p = 0.024).

```
  +---------------------------+
  |                           |
  | Tests of local dependence |
  |                           |
  +---------------------------+

Standardized gamma coefficients will be reported.

BD: DHP34 & DHP38      lr =  40.59 df =  9 p = 0.0000 Gamma =  0.56
DE: DHP38 & DHP39      lr =  37.76 df =  9 p = 0.0000 Gamma =  0.45

  +--------------+
  |              |
  | Tests of DIF |
  |              |
  +--------------+

Standardized gamma coefficients will be reported.

CG: DHP36 & SEX        lr =   9.44 df =  3 p = 0.0240 gamma = -0.35
```

Figure 2.3.11 Confirmatory tests of DIF and local dependence

2.3.7 Tests of local independence and no DIF

Select "**Check local independence**" and "**Check missing DIF**" to test that we have not overseen problems.

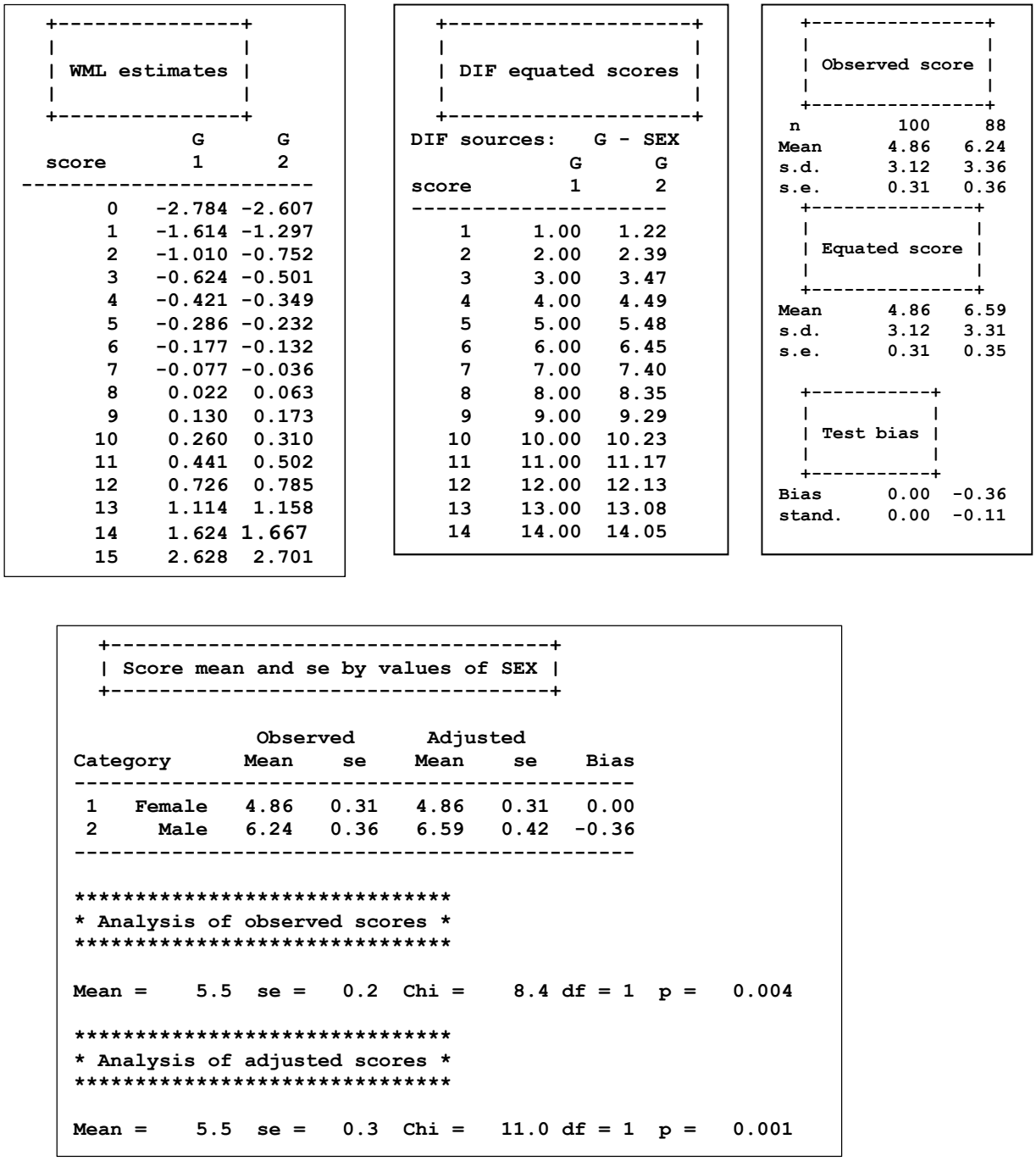
Figure 2.3.12 assumes that we have asked for extended output because we want to see all test resuls whether they are significant or not. It adds nothing to the model.

```
Check assumptions of local independence

A & B:   lr =    5.04  df =   9  p = 0.8309
A & C:   lr =   11.03  df =   9  p = 0.2738
A & D:   lr =   18.06  df =   9  p = 0.0345
A & E:   lr =   13.23  df =   9  p = 0.1523
B & C:   lr =   19.52  df =   9  p = 0.0211
B & E:   lr =    6.83  df =   9  p = 0.6546
C & D:   lr =    4.75  df =   9  p = 0.8552
C & E:   lr =    9.80  df =   9  p = 0.3668

               ---

Check assumptions of no DIF


A & F:   lr =    9.39  df =   9  p = 0.4020
B & F:   lr =    5.38  df =   9  p = 0.8004
C & F:   lr =   12.42  df =   9  p = 0.1907
D & F:   lr =    2.97  df =   9  p = 0.9653
E & F:   lr =   12.36  df =   9  p = 0.1940
A & G:   lr =    0.83  df =   3  p = 0.8421
B & G:   lr =    1.83  df =   3  p = 0.6075
D & G:   lr =    2.74  df =   3  p = 0.4342
E & G:   lr =    6.06  df =   3  p = 0.1089

Benjamini & Hochberg rejects at 0.00294
```

2.3.8 Person estimation and targeting in GLLRMs

Person parameter estimates will be different for men and women because Sex is a DIF source of DIF. We do not show the estimates here, but after the tables with the person parameters estimates, DIGRAM prints a table (Figure 2.3.9) with equated scores where scores for men are adjusted to the scores for women. In the score range from 1 to 10, the scores for men (G=2) should be increased with 0.20 – 0.50 points to be comparable to the scores for women (G=1). To assess the effect of DIF, DIGRAM also compares observed and adjusted scores in groups defined by the source of DIF. Figure 2.3.10 summarizes the test information and targeting under the GLLRM. Compare these results with the results in Figure 2.2.28. Recall, that Figure 2.2.28 summarize targeting for flipped items. To compare the results you therefore have to change the sign of the target value and the signs of the means of the population parameters in this figure. None of the other statistics included in the

assessment of test information and targeting depend on the orientation of the items so that this parameters are directly comparable between Figure 2.2.28 and Figure 2.3.9.



WML estimates

score	G 1	G 2
0	-2.784	-2.607
1	-1.614	-1.297
2	-1.010	-0.752
3	-0.624	-0.501
4	-0.421	-0.349
5	-0.286	-0.232
6	-0.177	-0.132
7	-0.077	-0.036
8	0.022	0.063
9	0.130	0.173
10	0.260	0.310
11	0.441	0.502
12	0.726	0.785
13	1.114	1.158
14	1.624	1.667
15	2.628	2.701

DIF equated scores

DIF sources: G - SEX

score	G 1	G 2
1	1.00	1.22
2	2.00	2.39
3	3.00	3.47
4	4.00	4.49
5	5.00	5.48
6	6.00	6.45
7	7.00	7.40
8	8.00	8.35
9	9.00	9.29
10	10.00	10.23
11	11.00	11.17
12	12.00	12.13
13	13.00	13.08
14	14.00	14.05

Observed score

n	100	88
Mean	4.86	6.24
s.d.	3.12	3.36
s.e.	0.31	0.36

Equated score

Mean	4.86	6.59
s.d.	3.12	3.31
s.e.	0.31	0.35

Test bias

Bias	0.00	-0.36
stand.	0.00	-0.11

Score mean and se by values of SEX

Category		Observed Mean	se	Adjusted Mean	se	Bias
1	Female	4.86	0.31	4.86	0.31	0.00
2	Male	6.24	0.36	6.59	0.42	-0.36

****
* Analysis of observed scores *
****

Mean = 5.5 se = 0.2 Chi = 8.4 df = 1 p = 0.004

****
* Analysis of adjusted scores *
****

Mean = 5.5 se = 0.3 Chi = 11.0 df = 1 p = 0.001

Figure 2.3.9 WML estimates, DIF equated scores and assessment of test bias for women and men. The analysis of observed and adjusted means will only be reported if you ask for extended output and select the "compare observed and "DIF equated score distributions".

```
  +--------------------------------+
  |                                |
  | Targeting and test information |
  |                                |
  +--------------------------------+

                                        theta        test info    target      RMSE(WML)    target
     SEX      AGE   Target   n    Mean     sd     Mean     max   index     Mean     min    index
-------------------------------------------------------------------------------------------------
   Female    18-49  -0.16    14  -0.28    0.65   4.221    5.052   0.836    0.514   0.445   0.866
     Male    18-49  -0.09    17   0.29    0.69   4.274    5.307   0.805    0.510   0.434   0.851
   Female    50-59  -0.16    33  -0.52    0.42   4.305    5.052   0.852    0.512   0.445   0.869
     Male    50-59  -0.09    23  -0.21    0.45   4.747    5.307   0.895    0.485   0.434   0.895
   Female    60-69  -0.16    47  -0.69    0.63   3.757    5.052   0.744    0.550   0.445   0.809
     Male    60-69  -0.09    42  -0.34    0.59   4.354    5.307   0.820    0.512   0.434   0.847

  +--------------------------------------+
  |                                      |
  | True score estimation and reliability |
  |                                      |
  +--------------------------------------+

                                    Score                 separation  no sep.
     SEX      AGE   Target   n    Mean     sd   reliability     prob     prob   Target SEM(TS) Mean SEM(TS)
-------------------------------------------------------------------------------------------------------
   Female    18-49   6.74    14   6.29    3.43      0.656      0.756    0.081       2.25          2.04
     Male    18-49   6.71    17   8.35    3.62      0.695      0.772    0.072       2.30          2.05
   Female    50-59   6.74    33   5.15    2.74      0.440      0.676    0.104       2.25          2.06
     Male    50-59   6.71    23   6.13    3.08      0.497      0.695    0.085       2.30          2.17
   Female    60-69   6.74    47   4.66    3.10      0.606      0.737    0.090       2.25          1.91
     Male    60-69   6.71    42   5.57    3.30      0.609      0.747    0.082       2.30          2.07
```

Figure 2.3.10 Targeting and test information under the BD, DE, CF model

2.3.5 Saving the model

Item analysis by GLLRMs can be time-consuming. To save time used to recall and redefine the models DIGRAM will let you save the definition of the models as a DIGRAM command file.

You can do this in two ways: from the DIGRAM main form where you have to invoke a "**SAVE R**" command or from the GRM dialog (Figure 2.3.2) where you have to press the "**Save current model**" button. The contents of the command file for the (BD, DE, CG) model is shown in Figure 2.3.11. It first selects items and exogenous variables and then defines the model and open the GRM dialog.

```
ITEMS ABCDE
EXO FG
ASY
SCREEN J
REP
GRM BD DE CG
```

Figure 2.3.11 Contents of a command file defining the (BD, DE, CG) model

The "**SCREEN J**" command invokes the item screening procedure that we will describe in the next session. It is included as part of the definition of the GLLRM because some the information provided during the item screening may be useful during the analysis after you have defined the model. The **ASY** command is to make sure that no time is wasted to calculate Monte Carlo estimates of p-values during the item screening, while the **REP** command is meant to reset the defaults requiring repeated Monte Carlo tests during analysis of multiway tables.

2.4 Graphical log-linear Rasch models. The longer tour.

On this tour, we abandon the DHP project and turn to the PF3 project with data on the PF subscale of the SF-36 inventory claiming to measure physical functioning. The PF3 project includes information on sex and age that we will use for analyses of DIF.

It is often useful to think about the definition of items before we attempt to fit a conventional IRT or Rasch model to data. The PF items are defined by responses to the following ten questions with three ordinal response categories, 0 = "Limited a lot", 1 = "Limited a little" and 2 = " Not limited".

Does your health now limit you in these activities? If so, how much?

A) PF1: Vigorous activities
B) PF2: Moderate activities
C) PF3: Lifting or carrying groceries
D) PF4: Climbing several flights of stairs
E) PF5: Climbing one flight of stairs
F) PF6: Bending, kneeling, or stooping
G) PF7: Walking more than a mile
H) PF8: Walking several blocks
I) PF9: Walking one block
J) PF10: Bathing or dressing yourself

There are several reasons to be concerned with the plausibility of a conventional Rasch model for the PF.

One problem is that the PF items may be multidimensional because the items address two different issues: health *and* physical functioning.

A second problem is that PF1 (A) is concerned with activities may be limited for reasons that has nothing to do with health and that item J is concerned with activities of daily living that are very different from the physical activities described by the other items. For this reason, we will not include items A and J in the example here.

A third problem is that the PF questions are phrased in a way that has to generate local *response* dependence. If one, for instance, claims to have no limits climbing several flights of stairs it makes little sense to claim that one cannot walk one flight of stairs without limitations. We are sure that you will agree that questions are phrased in such a way that there have to be local response dependence between B & C, between D & E and between G & H, G & I and H & I.

We have three ways to address these issues.

The first is to do it as we have done in the previous sections. We may start with a conventional Rasch model and use the evidence of local dependence and DIF as a starting point for a search for an adequate GLLRM.

A second and better way defines an initial GLLRM with the five pairs of dependent items and start from there. This should reduce the problems with spurious evidence of dependence and save some time compared to the first approach.

The third would be to test the global Markov properties of the conventional Rasch model and use the results of these tests to define the initial GLLRM. We expect the tests of the Markov properties to provide evidence of the local dependence within the five pairs of items together, but the tests may provide evidence of local dependence of DIF that we cannot foresee by looking at the contents of the items.

We refer to an analysis of Markov properties of Rasch models as item screening. Item screening is a simple procedure that does not require model fitting. All it takes is tests of conditional independence in a number of three-way tables. Since these tests are powerful, we expect item screening to provide results that will define an initial GLLRM, which will be very close to an adequate model.

2.4.1 Descriptive item analysis

In addition to thinking about the contents of the items, it is always as good idea with a purely descriptive analysis before we try to find and fit a Rasch model. DIGRAM have two commands serving this purpose. "**SHOW I**" provide information on items and "**SHOW S**" will tell you about the distribution of the score.

2.4.1.1 Information on items

The **SHOW I** command provides much more information than what we need to present here. You will get a list of items and tables with the marginal item distribution. There are tables of relations between items and rest-score to give you an idea about the degree to which response functions are monotonic. There is a little bit of Mokken analysis, assessment of reliability and tables describing

the association between exogenous variables and items including the degree to which the exogenous variables is related to the presence of missing responses on items. We cannot show everything here, so we suggest that you try it yourself. Figure 2.4.1 shows two things that may be important for the analysis. The first is the marginal distributions of responses to items by 968 persons with complete responses to items. The second is the frequency of missing item responses in different age groups. The complete dataset has responses to PF items by 1069 persons. This means that responses to items is incomplete for close to 7.5 % of the persons. Figure 2.4.1 shows that the frequencies of missing responses increase with age.

```
  +--------------------+
  |                    |
  | Item distributions |
  |                    |
  +--------------------+

968 persons with complete responses

Item order from high to low item score: IHEGCBFD

      Item      0      1      2    Mean    Average
-------------------------------------------------
I Walk_1bl    2.0    4.8   93.3   1.91      0.96
H Walk_2+b    2.5    5.9   91.6   1.89      0.95
E   Stair1    2.2    8.0   89.9   1.88      0.94
C Liftgroc    3.7   12.3   84.0   1.80      0.90
G  Walk_1m    5.5    9.4   85.1   1.80      0.90
B  Mod.act    4.2   13.7   82.0   1.78      0.89
F  Bending    4.6   15.2   80.2   1.76      0.88
D  Stair2+    5.5   16.5   78.0   1.73      0.86
-------------------------------------------------

  +-------------------------------------------+
  |                                           |
  | Frequency of Missing item responses by AGE |
  |                                           |
  +-------------------------------------------+

                     B     C     D     E     F     G     H     I
-------------------------------------------------------------------
 1  16 - 29   248   1.2   1.2   1.6   1.2   1.2   1.2   1.2   1.2
 2  30 - 44   293   0.7   0.3   0.3   1.0   1.0   0.7   0.7   1.0
 3  45 - 59   264   4.9   6.1   4.2   4.9   4.5   5.3   5.3   6.4
 4  60 - 69   244   8.6   8.6  10.2  13.5   9.4  12.3  13.1  13.5
```

Figure 2.4.1 Information on PF items

Since we estimate the item parameters by conditional ML estimates and since we are able to estimate person parameters for all persons with responses to some items it may not be a problem unless the frequency of missing responses depend on an unobserved source of DIF for some items.

The frequencies of missing responses therefore suggest that it is worth to compare the CML estimates of the item parameters by persons with complete responses with the estimates based on data with complete and incomplete responses.

2.4.1.2 Information on the score

Figure 2.4.2 show the distribution of the score over the eight PF items and the relationship between physical functioning and age

```
Score distribution: 968 Cases
---------------------------------
Score  Count Percent  Cumulated
---------------------------------
   0      8     0.8        0.8
   1      4     0.4        1.2
   2      2     0.2        1.4
   3      7     0.7        2.2
   4      5     0.5        2.7
   5      9     0.9        3.6
   6      6     0.6        4.2
   7      7     0.7        5.0
   8     14     1.4        6.4
   9     10     1.0        7.4
  10     18     1.9        9.3
  11     19     2.0       11.3
  12     32     3.3       14.6
  13     35     3.6       18.2
  14     59     6.1       24.3
  15     89     9.2       33.5
  16    644    66.5      100.0
---------------------------------
Total   968   100.0

Mean     =  14.54
Variance =   9.40
s.d.     =   3.07
skewness =  -2.74
```

```
  +-----------------------------------+
  |                                   |
  | Score mean and sd by values of AGE |
  |                                   |
  +-----------------------------------+

Category      Mean    se
--------------------------
 1  16 - 29  15.46   0.10
 2  30 - 44  15.43   0.10
 3  45 - 59  14.19   0.21
 4  60 - 69  12.48   0.32
--------------------------

Mean =   15.2  se =   0.1
Chi =  108.8 df = 3  p =   0.000

The stepwise analysis collapses the
categories with the largest p-value

Collapsed groups sorted according to
decreasing mean scores

    Mean   Groups
----------------------------
   15.45   16 - 29 + 30 - 44
   14.19   45 - 59
   12.48   60 - 69
```

Figure 2.4.2 Information on the PF score

The PF score depends on age, but the association between the age and the frequency of missing responses imply that we should not draw conclusions on this issue before we have analysed the relationship between Age and the estimates of the person parameters.

In addition to this, the distribution of the score creates an important problem. A sample of 968 persons is large enough to provide precise estimates of item parameters. However, the responses

from the 652 persons with extreme scores provide no information on differences between items. The sample only include responses from 316 persons that can be used to estimate the item parameters and test the fit of items to the model.

2.4.2 Item screening

The item screening that we have implemented in DIGRAM is a stepwise procedure described by. The output is extensive providing information on every step taken during the screening, however in most cases, there will be no reason to look at anything but the final part where results are summarized and the initial GLLRM defined.

Figure 2.4.3 summarizes the result of the item screening and Figure 2.4.2 shows the IRT graph of the GLLRM proposed by the result.

```
  +---------------------------+
  |                           |
  | Summary of item screening |
  |                           |
  +---------------------------+

Warnings:

No apparent risk that LD and DIF evidence could be spurious

-------------------------------------------------------

Item screening has defined the following GLLRM: BC DE GH HI CK

Positive local dependence:  BC DE GH HI
DIF                      :  CK

The score is associated with the following exogenous variables:
 K: SEX
 L: AGE

Local dependent items in the model: 4 item components: BC DE F GHI
```

Figure 2.4.3 Summary of results of item screening of PF3 items.

The interaction terms **"BC DE GH HI CK"** imply that item screening only disclosed evidence of four of the five pairs of locally dependent items that we expected to find, but found evidence of DIF of item C relative to K that we could not foresee. In addition to this, item screening also provided evidence of direct effect of K and L on the latent variable.

In this case, item screening defined a GLLRM with four item components four item components consisting of items that are connected in the IRT graph (Figure 2.4.4). The first consists of items B and C, the second of D and E, the third of a single item (F), and the fourth of items G, H and I. DIGRAM defines this as the current GLLRM replacing the Rasch model defined when you selected the items[31].

We refer to Kreiner & Christensen (2011a) for details on item screening and another example with PF items. The rest of this section provides some details. However, if you can do without these details right now, we suggest that you skip these details and proceed directly to Section 2.4.2 to see what happens after item screening.

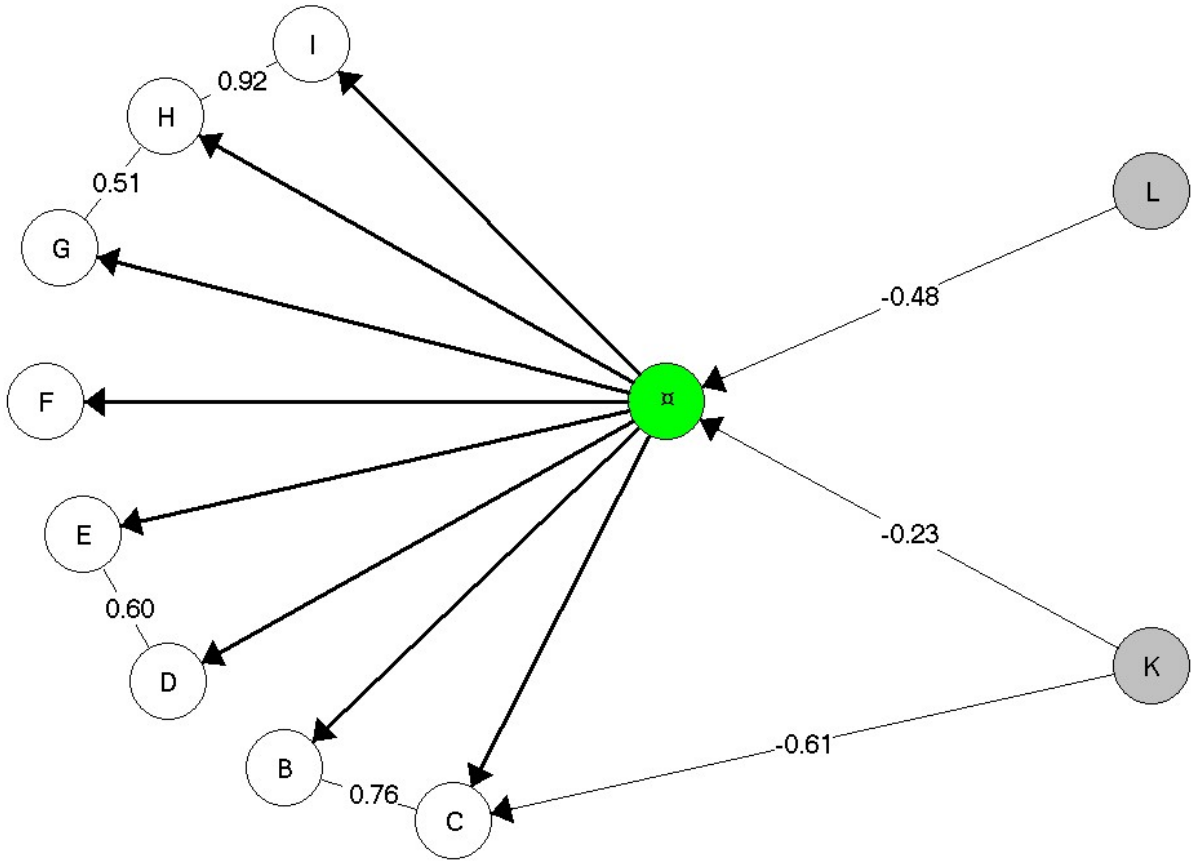


Figure 2.4.4 The IRT graph of the initial GLLRM defined by screening of eight PF items. The numbers on the edges of the graph are partial Gamma coefficients measuring the strength of the association among items and DIF sources[32].

[31] If you want to screen items, but do not want to replace the current GLLRM, you can use an SCREEN J command instead of SCREEN I. If you only want to assess the association between the score and the exogenous variables, you can use the SCREEN E command.

[32] To add these values to the edges of the graph you have to press the "Toggle Gamma Values on/off" (that was enabled by the item screening) on DIGRAM's graph form.

The first step of the screening of the PF3 items (Figure 2.4.5) consists of an analysis of marginal association between the items, and between he items and the total score and the exogenous variables. Psychometrics insists that validity requires that item responses are consistent in the sense that they are positively correlated. According to Figure 2.4.5, the PF items are consistent.

Screening of marginal item relationships

		B	C	D	E	F	G	H	I	rest score
B	Mod.act Gamma		0.973	0.913	0.949	0.868	0.919	0.967	0.958	0.936
	p		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C	Liftgroc Gamma	0.973		0.895	0.949	0.848	0.886	0.949	0.955	0.913
	p	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000
D	Stair2+ Gamma	0.913	0.895		0.975	0.859	0.920	0.919	0.912	0.892
	p	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000
E	Stair1 Gamma	0.949	0.949	0.975		0.924	0.925	0.973	0.983	0.964
	p	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000
F	Bending Gamma	0.868	0.848	0.859	0.924		0.880	0.909	0.922	0.855
	p	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000
G	Walk_1m Gamma	0.919	0.886	0.920	0.925	0.880		0.965	0.936	0.905
	p	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000
H	Walk_2+b Gamma	0.967	0.949	0.919	0.973	0.909	0.965		0.992	0.971
	p	0.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000
I	Walk_1bl Gamma	0.958	0.955	0.912	0.983	0.922	0.936	0.992		0.959
	p	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.000

Exogeneous variables

		B	C	D	E	F	G	H	I	score
K	SEX Gamma	-0.366	-0.440	-0.155	-0.264	-0.131	-0.170	-0.185	-0.192	-0.226
	p	0.000	0.000	0.014	0.004	0.037	0.021	0.043	0.059	0.000
L	AGE Gamma	-0.608	-0.577	-0.545	-0.584	-0.546	-0.557	-0.594	-0.661	-0.478
	p	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Figure 2.4.5 Gamma coefficients measuring marginal association among items, scores and exogenous variables. Rest-scores are scores without the item defining the row

The next step of item screening is a test of the global Markov properties of Rasch models. Two things follow from these properties. First, that two items are conditionally independent in a three-way table where the association between the items is stratified by a rest-score without one of the two item. Second, that items and exogenous variables are conditionally independent in three-way tables where the association between an item and an exogenous variable is stratified by the total

score. During the previous tours, we have already invoked these tests by the **LDE** and **DIF** commands. The second step of the item screening is a systematic and complete application of these tests. Figure 2.4.6 summarize the results.

```
  +----------------------------------------+
  |                                        |
  | Screening of partial item relationships |
  |                                        |
  +----------------------------------------+


p-values are two-sided and exact(Nsim = 1000)

 Tests of local independence - the row item has been subtracted from the score

                     B      C      D      E      F      G      H      I
-------------------------------------------------------------------------
B  Mod.act Gamma           0.715 -0.242 -0.252 -0.301 -0.234 -0.103 -0.449
              p            0.000  0.103  0.211  0.029  0.193  0.542  0.185

C Liftgroc Gamma    0.802        -0.189 -0.057 -0.386 -0.435 -0.312 -0.315
              p     0.000         0.237  0.821  0.005  0.004  0.208  0.242

D  Stair2+ Gamma   -0.007 -0.069         0.706 -0.146  0.213 -0.415 -0.511
              p     0.905  0.696         0.001  0.239  0.177  0.106  0.050

E   Stair1 Gamma   -0.494 -0.232  0.497         0.045 -0.501  0.256  0.736
              p     0.015  0.213  0.015         0.857  0.004  0.322  0.012

F  Bending Gamma   -0.011 -0.211 -0.097  0.405         0.199  0.094 -0.143
              p     0.952  0.250  0.454  0.050         0.155  0.714  0.598

G  Walk_1m Gamma   -0.045 -0.404  0.157 -0.277  0.024         0.634 -0.110
              p     0.800  0.009  0.290  0.169  0.863         0.004  0.677

H Walk_2+b Gamma   -0.245 -0.444 -0.622  0.188 -0.261  0.382         0.899
              p     0.290  0.041  0.003  0.445  0.333  0.048         0.000

I Walk_1bl Gamma   -0.380 -0.213 -0.633  0.728 -0.387 -0.134  0.931
              p     0.214  0.524  0.028  0.011  0.138  0.638  0.000

        Test for DIF
                     B      C      D      E      F      G      H      I
-------------------------------------------------------------------------
K      SEX Gamma   -0.471 -0.608  0.194  0.115  0.227  0.258  0.087  0.415
              p     0.002  0.000  0.153  0.762  0.075  0.160  0.714  0.167

L      AGE Gamma   -0.112  0.007  0.076  0.159 -0.068 -0.002  0.314 -0.136
              p     0.360  1.000  0.505  0.380  0.429  1.000  0.154  0.555


Benjamini & Hochberg rejects at 0.00903 to control the FDR at 0.05
Benjamini & Hochberg rejects at 0.00069 to control the FDR at 0.01
Benjamini & Hochberg rejects at 0.00007 to control the FDR at 0.001
```

Figure 2.4.6 Partial Gamma coefficients measuring conditional association among items given rest-scores without one of the items and between items and exogenous variables given the total score over all items. The rest-scores are the scores without the item defining the row.

In Figure 2.4.6, the p-values are Monte Carlo estimates of exact p-values (Kreiner, 1987). To assess the significance of the evidence against the Rasch model DIGRAM, adjusts for multiple testing by

the Benjamini-Hochberg procedure. We conclude that there is strong evidence of local dependence and DIF among the PF items. DIGRAM finds four pairs of items with evidence of positive local dependence, four pairs with evidence of negative local dependence and evidence of DIF relative to K. However, there is no evidence of positive local dependence between items G and I. In Figure 2.4.5, the marginal correlation between G & I is positive and significant, but the two partial γ coefficients between G & I in Figure 2.4.6 are negative and clearly insignificant.

The evidence of positive local dependence corresponded to the local *response* dependence that we expected. Since negative local dependence rarely makes no sense in term of response dependence, evidence of negative local dependence indicates that the PF scale is multidimensional or that some items simply does not fit the Rasch model. However, before we draw such conclusions, the next steps of the item screening attempts to distinguish between genuine and spurious evidence.

Figure 2.4.7 summarize the evidence of local dependence and the attempts to eliminate spurious evidence of local dependence. Local dependence between two items will create evidence of spurious evidence of dependence among other items because inclusion association between two items imply that some of the hypotheses of conditional independences induced by the Rasch model no longer applies. DIGRAM use a stepwise elimination routine that agree to regard the strongest one piece of evidence[33] of positive local dependence as genuine in each step and dismiss the significant evidence of local dependence among other items if the evidence could be spurious if the decision was correct. Out of 10 pairs of items with significant evidence of local dependence, DIGRAM decides that the evidence of positive local dependence appears to be genuine and dismisses the evidence of negative local dependence.

Evidence of DIF may also be spurious, because DIF of one item relative to an exogenous variable may generate spurious evidence of DIF for other items and other exogenous variables.

Figure 2.4.8 illustrates how DIGRAM attempts to distinguish between genuine and spurious evidence of DIF. Item screening generated significant evidence of DIF of both B and C relative to Sex. To distinguish between genuine and spurious evidence against conditional independence we have to look at the association between one of the items and Sex given the score and the other

[33] Measured by the weighted mean (WPG) of the two partial γ values were reported in Figure 2.4.4

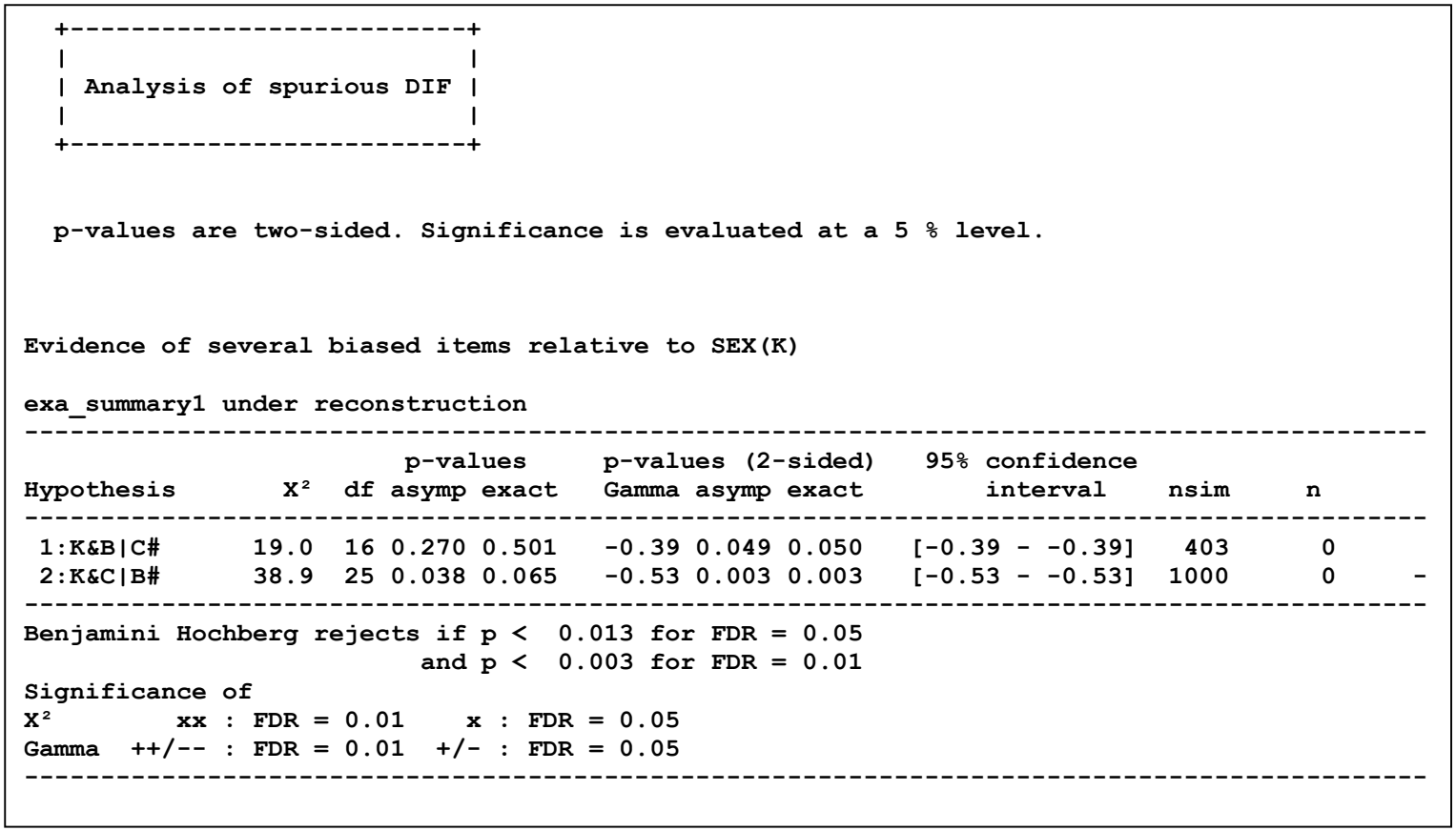
items. Since this hypothesis is accepted for item B, but not for item C we DIGRAM concludes that the evidence of DIF for item B was spurious.

```
  +----------------------------------------+
  |                                        |
  | Summary of evidence of local dependence |
  |                                        |
  +----------------------------------------+

  Beware of Type II errors. p-values are evaluated at a 5 % FDR level.

  ***: FDR <= 0.001, **: FDR <= 0.01,  *: FDR <= 0.05

   Average absolute partial gamma values =  0.334

Two significant positive partial correlations:

   B: Mod.act   & C: Liftgroc  gamma =  0.715*** 0.802***   WPG = 0.757
   H: Walk_2+b  & I: Walk_1bl  gamma =  0.899*** 0.931***   WPG = 0.915

Only one significant positive partial correlation:

   D: Stair2+   & E: Stair1    gamma =  0.706*   0.497      WPG = 0.596
   G: Walk_1m   & H: Walk_2+b  gamma =  0.634*   0.382      WPG = 0.510

Only one significant negative partial correlation:

   C: Liftgroc  & F: Bending   gamma = -0.386*  -0.211      WPG =-0.304
   D: Stair2+   & H: Walk_2+b  gamma = -0.415   -0.622*     WPG =-0.541
   E: Stair1    & G: Walk_1m   gamma = -0.501*  -0.277      WPG =-0.404

Two significant negative partial correlations:

   C: Liftgroc  & G: Walk_1m   gamma = -0.435*  -0.404*     WPG =-0.419

Stepwise inclusion of local dependence:

   H: Walk_2+b  & I: Walk_1bl   Weighted partial gamma =   0.915
   B: Mod.act   & C: Liftgroc   Weighted partial gamma =   0.757
   D: Stair2+   & E: Stair1     Weighted partial gamma =   0.596
   G: Walk_1m   & H: Walk_2+b   Weighted partial gamma =   0.510
```

Figure 2.4.7 Analysis of evidence of local dependence

The final issue addressed by the item screening is the association between the total score and the exogenous variables. Figure 2.4.9 shows the result. There is nothing new here, because the analysis is the same as you saw when you defined the exogenous variable, but DIGRAM repeats it to remind[34] you of the results.

[34] If you need to be reminded of these result at a later time, you have to invoke a "SCREEN E" commando.



```
  +--------------------------+
  |                          |
  | Analysis of spurious DIF |
  |                          |
  +--------------------------+


  p-values are two-sided. Significance is evaluated at a 5 % level.



Evidence of several biased items relative to SEX(K)

exa_summary1 under reconstruction
-------------------------------------------------------------------------------------------
                         p-values     p-values (2-sided)   95% confidence
Hypothesis       X²  df asymp exact   Gamma asymp exact        interval    nsim     n
-------------------------------------------------------------------------------------------
 1:K&B|C#      19.0  16 0.270 0.501   -0.39 0.049 0.050   [-0.39 - -0.39]   403      0
 2:K&C|B#      38.9  25 0.038 0.065   -0.53 0.003 0.003   [-0.53 - -0.53]  1000      0     -
-------------------------------------------------------------------------------------------
Benjamini Hochberg rejects if p <  0.013 for FDR = 0.05
                         and p <  0.003 for FDR = 0.01
Significance of
X²        xx : FDR = 0.01    x : FDR = 0.05
Gamma  ++/-- : FDR = 0.01  +/- : FDR = 0.05
-------------------------------------------------------------------------------------------
```

Figure 2.4.8 Analysis of spurious evidence of DIF

Comments:

If more than one item have DIF relative to the same DIF source, DIGRAM tests whether items are conditionally independent given both the total score and all other DIF items, and concludes that the evidence of DIF was spurious if conditional independence is accepted.

Figure 2.4.4 suggested that K was a source of DIF for items B and C. However, it turns out, that K and B are conditionally independent given the score *and* B. For this reason, the original evidence of DIF is regarded as spurious.

```
  +-----------------------------------------------------------+
  |                                                           |
  | Analysis of the effects of exogenous variables on the score |
  |                                                           |
  +-----------------------------------------------------------+


The raw score will be used during the analysis


2 variables with a marginal effect on the score: K L

-------------------------------------------------------------------------------------------------
                        p-values                  p-values (2-sided)
Hypothesis       X²  df asymp exact                Gamma asymp exact                 nsim
-------------------------------------------------------------------------------------------------
#&K            22.0  16 0.142 0.142 (0.116-0.173) -0.23 0.000 0.000 (0.000-0.007) 1000     0    -
-
#&L           194.7  48 0.000 0.000 (0.000-0.007) -0.48 0.000 0.000 (0.000-0.007) 1000     0 xx -
-
#&K|L          56.7  50 0.240 0.195 (0.165-0.229) -0.23 0.000 0.000 (0.000-0.007) 1000     0    -
-
#&L|K         237.7  96 0.000 0.000 (0.000-0.007) -0.48 0.000 0.000 (0.000-0.007) 1000     0 xx -
-
-------------------------------------------------------------------------------------------------
Benjamini Hochberg rejects if p <  0.038 for FDR = 0.05
                          and p <  0.008 for FDR = 0.01
Significance of
X²        xx : FDR = 0.01    x : FDR = 0.05
Gamma  ++/-- : FDR = 0.01  +/- : FDR = 0.05
-------------------------------------------------------------------------------------------------
```

Comments:

This is the same two-step procedure that was used, when exogenous variables were selected. In the first step, we look at the marginal association between the score and the exogenous variables.

In the second step, DIGRAM tests *conditional* independence between the score and an exogenous variable given all the variables that were marginally associated with the score and eliminates exogenous variables one at a time in a stepwise manner.

Figure 2.4.9 Analysis of associations between the score and the exogenous variables

2.4.3 Model search

Item screening disclosed strong evidence of local dependence and DIF and defined a GLLRM with four pairs of locally response dependent items and two items with DIF relative to K. However, the purpose of GLLRMs defined by item screening is only to serve as the starting point for a more careful search for a GLLRM. We expect it to be close to an adequate model (and often is very close), but it will probably not be the final model.

The analysis following the definition of the initial model is a standard log-linear modeling exercise, except for two reasons.

1. It is technically more complicated than a routine log-linear analysis, because we will be dealing with a high-dimensional log-linear model that describe the responses of items conditionally given the total score on all items.
2. The analysis in DIGRAM is never automatic. It is stepwise model search, but it is up to the user to decide what happens after each step. To help you make this decision, DIGRAM provides information on the significance of test statistics, together with estimates of parameters and the measures of partial associations obtained during screening. But you must never forget subject matter considerations and content analyses of items and ***never*** decide what to change without thinking about the meaningfulness of the local dependence and the DIF represented by the interaction terms.

During the search for an adequate GLLRM, Kelderman's conditional likelihood ratio tests provide evidence of local dependence and or DIF. These tests are obtainable in four different ways by selection of options in the GRM dialog box Figure 2.1.5.

1. Select "**Reduce model**" to test whether you need all the terms in the new model.
2. Select "**Check local independence**" to test the missing model terms representing local dependence.
3. Select "**Check missing DIF**" to test the missing DIF terms in the model.
4. Add interaction terms of interest to the "**Test model terms**" field. "AB BC" will give you test results relating to the local dependence of A&B and of B&C. A '*' in an interaction term is a wildcard referring to all variables, a '+' after a variable refers to the existing interactions with the variable, and a '-' after a variable refer to the missing relationships. "B*" in the "Test model terms" will give you test results for all

interactions including B, while "B-" will only calculate tests relative to variables that is independent of B according to the model. If X is an exogenous variable then "X+" tests DIF relative to X the items that the model believes function differently relative to X. Finally, if X is an exogenous variable, "XX" calculates Andersen's global CLR test of DIF relative to X.

The fourth option may appear a little more complicated, than the first three, but it is not really that complicated. Try it out. It will save you a lot of time when you have learned to use it.

Figure 2.4.10 presents the tests of the models claims of local dependence and DIF. They confirm the claims of local dependence and DIF.

```
  +---------------------------+
  |                           |
  | Tests of local dependence |
  |                           |
  +---------------------------+

Standardized gamma coefficients will be reported.

BC: Mod.act & Liftgroc  lr =  77.66 df =  4 p = 0.0000 Gamma =  0.77
DE: Stair2+ & Stair1    lr =  55.05 df =  4 p = 0.0000 Gamma =  0.72
GH: Walk_1m & Walk_2+b  lr =  22.18 df =  4 p = 0.0002 Gamma =  0.52
HI: Walk_2+b & Walk_1bl lr =  65.94 df =  4 p = 0.0000 Gamma =  0.92

  +--------------+
  |              |
  | Tests of DIF |
  |              |
  +--------------+

Standardized gamma coefficients will be reported.

CK: Liftgroc & SEX      lr =  28.53 df =  2 p = 0.0000 gamma = -0.66
```

Figure 2.4.10 Confirmatory tests of local dependence and DIF. The Gamma coefficients are the standardized gamma coefficient calculated during estimation of the item parameters.

We accept the (BC, DE, GH, HI, CK) model and proceed to tests of local independence and no DIF. Figure 2.4.11 presents the results of these tests. To help understanding what the evidence of LD and DIF implies, DIGRAM prints the weighted partial gamma coefficient obtained during item screening.

```
Check assumptions of local independence
Gamma indicators of local dependence will be reported for significant test
results

Test results will only be shown if p <= 0.05.
Select extended output if you want to see all testresults.

Significant test results:

B & H:   lr =   10.67  df =   4  p = 0.0306  WPG gamma =  -0.18
C & E:   lr =   10.31  df =   4  p = 0.0354  WPG gamma =  -0.15
D & F:   lr =   10.91  df =   4  p = 0.0276  WPG gamma =  -0.12
D & I:   lr =   20.80  df =   4  p = 0.0003  WPG gamma =  -0.58
E & H:   lr =   28.99  df =   4  p = 0.0000  WPG gamma =   0.22
E & I:   lr =   26.04  df =   4  p = 0.0000  WPG gamma =   0.73
G & I:   lr =   37.07  df =   4  p = 0.0000  WPG gamma =  -0.12

               ---

Check assumptions of no DIF
Gamma coefficients will be reported for significant test results


Test results will only be shown if p <= 0.05.
Select extended output if you want to see all testresults.

Significant test results:

G & L:   lr =   14.10  df =   6  p = 0.0286 gamma =  -0.00
I & L:   lr =   26.05  df =   6  p = 0.0002 gamma =  -0.14

Benjamini & Hochberg rejects at 0.00641

Suggested additions to the model:

Positive LD:  EH EI

EH  Gamma =0.22  p =0.000    Stair1 & Walk_2+b
EI  Gamma =0.73  p =0.000    Stair1 & Walk_1bl

Negative LD:  DI GI

DI  Gamma =-0.58  p =0.000    Stair2+ & Walk_1bl
GI  Gamma =-0.12  p =0.000    Walk_1m & Walk_1bl

DIF:          IL
```

Figure 2.4.11 Checking local independence and missing DIF

Since there is evidence of local dependence between E&I, E&H, D&I and G&I and DIF of I relative to L. Since negative local dependence makes little sense, we only add EI, EH and IL terms to the model and test again for local dependence and DIF. The evidence of local dependence local independence between E&I is significant but weak (p=0.016) for which reason we eliminate the EI interaction from the model.

Proceeding in this way does not provide additional evidence of local dependence and DIF. However before we accept the (BC DE EH GH HI CK IL) model, we need tests of homogeneity and invariance and we need tests of item fit. Figures 2.4.2 and 2.4.13 presents the results.

```
                CLR   df    p
-----------------------------
scoregroups    26.4   42 0.971
K:       SEX   44.6   38 0.215
L:       AGE  105.6  102 0.383
```

Figure 2.4.12 Test of homogeneity and invariance of the (BC DE EH GH HI CK IL) model

```
  +-----------------------------+
  |                             |
  | Item restscore association  |
  |                             |
  +-----------------------------+

                  Item-restscore gamma
Item            observed expected    sd      p
-------------------------------------------------
B -  Mod.act       0.936    0.922    0.011  0.22980
C - Liftgroc       0.913    0.922    0.012  0.45696
D -  Stair2+       0.892    0.900    0.013  0.50717
E -   Stair1       0.964    0.948    0.011  0.13027
F -  Bending       0.855    0.869    0.016  0.39494
G -  Walk_1m       0.905    0.896    0.015  0.56171
H - Walk_2+b       0.971    0.969    0.008  0.83318
I - Walk_1bl       0.959    0.960    0.011  0.92430
-------------------------------------------------

Critical levels adjusted by the Benjamini-Hochberg procedure:
    * < 5 % FDR, ** < 1 % FDR, *** = FDR < 0.1 % FDR

  +---------------------------------------+
  |                                       |
  | Component-restscore gamma coefficients |
  |                                       |
  +---------------------------------------+

                    Gamma
Component       observed expected    sd      p
-------------------------------------------------
    BC             0.884    0.881    0.014   0.8739
 DEGHI             0.888    0.881    0.013   0.6061
```

Figure 2.4.13 Tests of item fit for the (BC DE EH GH HI CK IL) model

And that is all. Stepwise model search may be an unusual technique if you are used to IRT and Rasch analyses and have little experience with multivariate statistics. Learning to use and appreciate methods for model search for GLLRMs may therefore take longer than going through the first

DIGRAM tours. Fortunately, things simplify after model search when model search is over. After model search, we have to calculate over-all fit statistics, item fit statistics, person estimates, and assess test information and the targeting of the items to the study population and we do it in exactly the same way as we did for the Rasch model in sections 2.1 and 2.2 and for the GLLRM in Section 2.3.

2.4.4 The PF3 model

Figure 2.4.14 shows the IRT graph of the model defined by model search. Local dependence and DIF of C relative to K is extremely strong. The only thing to be concerned about is the DIF of item I relative to L. In terms of significance, the evidence supporting these claims is strong, but the partial correlation is close to zero. For this reason, we focus on this interaction when we present to additional ways to test the fit of a GLRM to data.

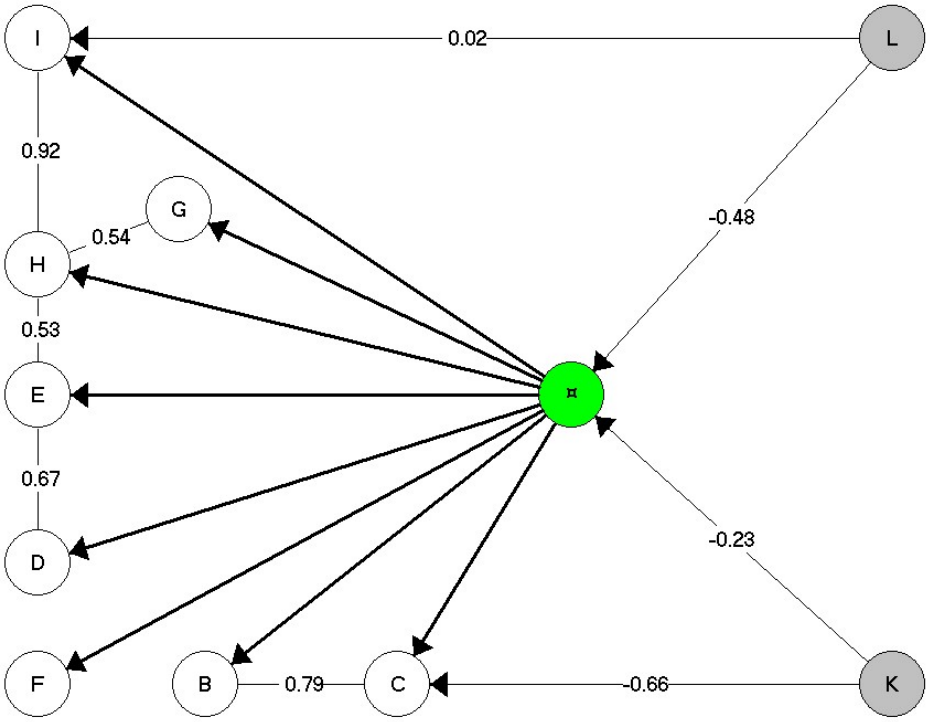


Figure 2.4.14 IRT graph of PF3 model after model search. The coefficients attached to edges are standardized gamma coefficients.

2.4.5 Checking the global Markov properties[35].

During item screening, the tests for local dependence and DIF are test of the global Markov properties of conventional Rasch models. GLLRMs also has global Markov properties that can be tested. To do this, you have to replace the current model with the new model that you have defined by clicking the "**Change model**" button and return to DIGRAM's main form.

Invoke the "**CHECK I**" command to test the global Markov properties of the current model. Output from this procedure is extremely extensive, because DIGRAM produces details on both the Markov properties and the tests of the hypotheses, but DIGRAM provides summaries at the end.

The summary consists of two parts. The first with test results relating to the local and DIF *within* the current model. The second with test result relating to the model's claims of local independence and no DIF. Figures 2.4.15 and 2.4.16 shows these results for the model defined by item screening and model search

Figure 2.4.15 tests the global Markov properties of all the claims of local dependence and DIF in the model defined by item screening. All the tests confirm the finding of the item screening except the DIF of I relative to age, where the partial γ coefficient is insignificant. This result indicates that the conditional association between item I and Age is not monotonic. The partial gamma coefficient provides tests of monotonic relationships between ordinal categorical variables, whereas the CLR test of no interaction in log-linear models are tests of association between nominal variables with power against non-monotonic associations. For this reason, we do not accept the hypothesis of no DIF, but conclude that the IL association it is not a simple monotonic relationship.

Figure 2.4.16 tests the global Markov properties of all the claims of local independence and no DIF in the current model. There is no evidence of missing dependence or DIF in the (BC DE EH GH HI CK IL) model.

[35] Consult Kreiner & Christensen (2011a) before you try this to be sure that you understand what goes on.

```
  +-------------------------+
  |                         |
  | Check of LD and/or DIF  |
  |                         |
  +-------------------------+


p-values are two-sided and exact(Nsim = 400)

 Tests of local independence - the item component of the row item has been
subtracted from the score

                     B      C      D      E      F      G      H      I
-------------------------------------------------------------------------
B  Mod.act Gamma           0.720
             p             0.000

C Liftgroc Gamma    0.829
             p      0.000

D  Stair2+ Gamma                         0.442
             p                           0.070

E   Stair1 Gamma                  0.623                       0.576
             p                    0.005                       0.040

F  Bending Gamma
             p

G  Walk_1m Gamma                                                 0.735
             p                                                   0.000

H Walk_2+b Gamma                         0.537         0.465         0.728
             p                           0.000         0.005         0.010

I Walk_1bl Gamma                                                 0.920
             p                                                   0.000


        Test of no DIF

                     B      C      D      E      F      G      H      I
-------------------------------------------------------------------------
K      SEX Gamma          -0.608
             p             0.000

L      AGE Gamma                                                      -0.114
             p                                                         0.679
```

Figure 2.4.15. Test of local dependence and DIF in the (BC DE EH GH HI CK IL) model

```
  +------------------------+
  |                        |
  | Check of LI and no DIF |
  |                        |
  +------------------------+


p-values are two-sided and exact(Nsim = 400)

 Tests of local independence - the item component of the row item has been
subtracted from the score

                     B      C      D      E      F      G      H      I
-------------------------------------------------------------------------------
B  Mod.act Gamma                  0.189  0.167  0.130 -0.017  0.060 -0.429
             p                    0.345  0.762  0.430  1.000  0.867  0.524

C Liftgroc Gamma                  0.185  0.485 -0.119 -0.127 -0.008  0.143
             p                    0.286  0.040  0.514  0.495  0.976  0.810

D  Stair2+ Gamma    0.175 -0.028                -0.082    no     no     no
             p      0.315  0.952                 0.571   test   test   test

E   Stair1 Gamma   -0.107 -0.078                 0.099    no           no
             p      0.667  0.810                 0.560   test         test

F  Bending Gamma    0.350 -0.043  0.034  0.136         0.107  0.023 -0.297
             p      0.071  0.762  0.857  0.476         0.571  0.952  0.259

G  Walk_1m Gamma    0.395 -0.252    no     no  -0.096                 no
             p      0.023  0.102   test   test  0.571                 test

H Walk_2+b Gamma    0.445 -0.042    no          -0.290
             p      0.040  0.905   test          0.080

I Walk_1bl Gamma    0.147  0.101    no     no  -0.182    no
             p      0.714  0.714   test   test  0.381   test


        Test of no DIF

                     B      C      D      E      F      G      H      I
-------------------------------------------------------------------------------
K      SEX Gamma   -0.388         0.083 -0.100  0.080  0.041 -0.125  0.333
             p      0.049         0.617  0.666  0.617  0.829  0.645  0.215

L      AGE Gamma   -0.455 -0.078 -0.500 -0.667  0.333  0.200  0.231
             p      0.392  0.511  0.083  0.248  0.494  0.610  0.620
```

Figure 2.4.16. Test of local dependence and DIF in the (BC DE EH GH HI CK IL) model

2.4.6 Analysis of person fit

The tests of the global Markov properties of the (BC DE EH GH HI CK IL) model disclosed issues concerning the DIF of item I. One way to address such issues is to analysis of person fit to find out that there is a large proportion of persons with responses that disagree with the current GLLRM and because of improbable responses to item I.

We use the **PERsonfit** command to calculate person fits under the current GLLRM. If you invoke the command without parameters, DIGRAM calculates the conditional probability of the complete vector of responses given the total score over all items and use this probability as the fit statistic. To find out whether this probability provides significant evidence against fit, DIGRAM finds all response vectors with the same score that are as improbable as the observed responses. The sum of the probabilities of the observed and more improbable responses is the p-value. In this sense, the test is similar to Fischer's exact test in two-by-two tables.

The analysis is a two-step procedure. In the first step, DIGRAM calculates the conditional probabilities of all response patterns for the different combinations of outcomes on DIF sources and saves the result on a text file called "Responsepatterns.txt"[36]. We will not show this file here, but take a look at if you try this procedure to make sure that you know what goes on. At the end of this step, DIGRAM prints tables with the most probable response patterns and the expected item scores for each combination of outcomes on DIF sources

Figures 2.4.17 and 24.18.shows the results for women aged 60-69 years of age. If the total score over all items is equal to seven, the most probable response pattern is (1 1 1 1 1 0 1 1) and the probability of this pattern 0.124. If the total score is equal to 13, the most probable response pattern is (1 1 1 2 2 2 2 2) and the probability is 0.199. The table with expected item scores include the unweighted and weighted (by the distribution of the scores) averages of the expected item scores. From this point of view, the physical activity described by item I is easier than the other activities described by the PF items. During the attempt to disclose evidence against person fit we will focus on the frequencies where limitations are worse for item I than for the other items.

[36] DIGRAM warns you that it will have to calculate probabilities for 19,683 response patterns for each combination of outcomes on sources of DIF.

```
K:        SEX = Female
L:        AGE = 60 - 69

* Max probability patterns *

 0 1.0000000   pattern =  0 0 0 0 0 0 0 0
 1 0.5150583   pattern =  0 0 0 0 0 0 0 1
 2 0.2150181   pattern =  0 0 0 0 1 0 0 1
 3 0.3981247   pattern =  0 0 0 1 0 0 1 1
 4 0.2535832   pattern =  0 0 0 1 1 0 1 1
 5 0.1126207   pattern =  0 0 1 1 1 0 1 1
 6 0.1499850   pattern =  1 1 0 1 1 0 1 1
 7 0.1243517   pattern =  1 1 1 1 1 0 1 1
 8 0.1300468   pattern =  1 1 1 1 1 1 1 1
 9 0.0727632   pattern =  1 1 1 1 1 1 1 2
10 0.0971563   pattern =  1 1 1 1 1 1 2 2
11 0.2081263   pattern =  1 1 1 2 1 1 2 2
12 0.2649026   pattern =  1 1 1 2 1 2 2 2
13 0.1992963   pattern =  1 1 1 2 2 2 2 2
14 0.2027962   pattern =  2 2 1 2 1 2 2 2
15 0.3433374   pattern =  2 2 1 2 2 2 2 2
16 1.0000000   pattern =  2 2 2 2 2 2 2 2
```

Figure 2.4.17 Overview of the most probable response patterns for females, aged 60-69

```
Expected item responses

   score  B    C    D    E    F    G    H    I
-------------------------------------------------
     0   0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
     1   0.04 0.05 0.02 0.11 0.21 0.03 0.03 0.52
     2   0.12 0.12 0.06 0.28 0.35 0.07 0.27 0.72
     3   0.15 0.15 0.10 0.64 0.32 0.11 0.64 0.89
     4   0.21 0.21 0.23 0.84 0.49 0.22 0.80 1.01
     5   0.37 0.38 0.36 0.94 0.62 0.34 0.88 1.11
     6   0.59 0.59 0.46 1.01 0.74 0.45 0.95 1.20
     7   0.76 0.76 0.59 1.08 0.86 0.60 1.03 1.32
     8   0.86 0.86 0.72 1.17 0.97 0.79 1.16 1.47
     9   0.92 0.92 0.82 1.29 1.03 0.98 1.38 1.65
    10   0.98 0.98 0.93 1.45 1.06 1.16 1.64 1.81
    11   1.06 1.06 1.06 1.64 1.12 1.33 1.83 1.90
    12   1.18 1.17 1.20 1.80 1.24 1.53 1.93 1.95
    13   1.36 1.36 1.37 1.89 1.41 1.67 1.97 1.97
    14   1.63 1.64 1.51 1.94 1.54 1.77 1.99 1.99
    15   1.88 1.89 1.66 1.99 1.71 1.88 2.00 1.99
    16   2.00 2.00 2.00 2.00 2.00 2.00 2.00 2.00


Average  0.83 0.83 0.77 1.18 0.92 0.88 1.21 1.38
Weighted 1.77 1.77 1.73 1.88 1.76 1.80 1.90 1.92
```

Figure 2.4.18 Overview of the expected response patterns for females, aged 60-69

During the second step, DIGRAM calculates the conditional probability for all persons, assess the significance of the response pattern using the exact test proposed by Martin-Löf (1977)[37], and prints all response patterns that are significant at a 5 % level. The results can be seen in Figure 2.4.20.

DIGRAM attempts to test person fit at a 5 % level, but the distribution of the conditional probabilities in each score group is discrete. So that we cannot define tests with exactly 5 % critical levels. Instead, the following table shows the critical level that we have to use. However, they are close to 5 % in many cases because of the large number of different patterns within score groups.

The critical levels:

Score	number of patterns	critical size
1	8	0.0195
2	36	0.0429
3	112	0.0479
4	266	0.0489
5	504	0.0493
6	784	0.0499
7	1016	0.0498
8	1107	0.0499
9	1016	0.0499
10	784	0.0499
11	504	0.0499
12	266	0.0486
13	112	0.0497
14	36	0.0496
15	8	0.0216

Figure 2.4.19 Overview of critical levels in different score groups for tests of person fit of women aged 60-69

Out of 69 women aged 60-69, the analysis disclosed evidence of departure from the GLLRM in five cases. Figure 2.4.21 only expects 3.1 cases, but the difference is not significant.

Item scores of item G, H and I should not be decreasing. However, in three of the five cases of apparent person misfit, responses to these three items were inconsistent. One possibility is that respondents must have misunderstood the question. Another is that it is data error.

[37] The exact test uses the probability as a test statistic arguing that the smaller the probability the more significant it is. The p-value of the exact test is equal to the sum of probabilities that are smaller than or equal to the probability of the observed pattern

```
  +-------------------------+
  |                         |
  | Review of person misfit |
  |                         |
  +-------------------------+

SEX = Female
AGE = 60 - 69

  K  L    B C D E F G H I score  Prob  count     p       size

  2  4    1 1 1 1 2 2 1 0   9   0.00025    1    0.030    0.050
  2  4    1 2 2 2 0 0 1 1   9   0.00036    1    0.044    0.050
  2  4    0 0 2 2 2 2 0 2  10   0.00009    1    0.009    0.050
  2  4    1 1 2 0 2 2 1 1  10   0.00021    1    0.021    0.050
  2  4    2 1 1 1 1 2 2 1  11   0.00071    1    0.049    0.050
```

Figure 2.4.20 Overview of women aged 60-69 with improbable responses

```
Expected number of significant cases

Score Count   Pcrit Expected Significant
---------------------------------------
   1      3  0.0195     0.06       0
   2      1  0.0429     0.04       0
   3      1  0.0479     0.05       0
   4      2  0.0489     0.10       0
   5      3  0.0493     0.15       0
   6      1  0.0499     0.05       0
   7      4  0.0498     0.20       0
   8      2  0.0499     0.10       0
   9      6  0.0499     0.30       2
  10      9  0.0499     0.45       2
  11      5  0.0499     0.25       1
  12      6  0.0486     0.29       0
  13      8  0.0497     0.40       0
  14      8  0.0496     0.40       0
  15     10  0.0216     0.22       0

69 persons included

Observed =    5
Expected =    3.04
      SD =    1.70
       p =   0.2509
```

Figure 2.4.21 Expected numbers of significant results

In cases with evidence against person fit, the problem is whether this is genuine evidence of person misfit or whether it is the kind of type I errors that the random machinery of the GLLRM has to produce. DIGRAM therefore count the total number and calculate the expected number of persons with evidence against person fit in groups defined by outcomes on the exogenous variables and

finally summarize the results for all groups. Out of 319 persons, we expect to find 13 with responses that disagree with the GLLRM. Since we only find 9, we claim that tests of person fit provided additional evidence supporting our claim that the items fit the GLLRM.

```
  +-------------------------------------------------------+
  |                                                       |
  | Summary of number of significant tests of person fit  |
  |                                                       |
  +-------------------------------------------------------+


  +---------+
  |         |
  | K - SEX |
  |         |
  +---------+

     SEX       n   Obs   exp
-----------------------------------------------
    Male     122    3    5.2  p =  0.3263
  Female     194    6    8.0  p =  0.4795

  +---------+
  |         |
  | L - AGE |
  |         |
  +---------+

     AGE       n   Obs   exp
-----------------------------------------------
 16 - 29      45    1    1.9  p =  0.5056
 30 - 44      59    1    2.1  p =  0.4372
 45 - 59      96    0    4.0  p =  0.0413
 60 - 69     116    7    5.1  p =  0.4031

316 persons included

Observed =      9    2.8%
Expected =  13.14    4.2%

SD =   3.54   z =  -1.1689   p =   0.2425
```

Figure 2.4.22 Summary of persons with improbable responses

Since we were concerned with item I and since this item belong to an item component defined by D,E,G,H,I we want to examine whether item I provide problems within this component. To assess this, we ask for a test of person fit within this component defined by the conditional probabilities of items given the component score. To calculate these tests, we invoke a "**PER DEGHI**" command.

Figures 2.4.23 and 2.4.24 summarize the results. Evidence of person misfit is found, but not more than we would expect for random reasons.

```
  +--------------------------------------+
  |                                      |
  | Person review of subscores D+E+G+H+I |
  |                                      |
  +--------------------------------------+

Significance is assessed separately for subset patterns and for the subset score
                                                                    subscore
  K   L     D E G H I rest score   Prob      p      subscore      p    count

  2   2     0 2 2 2 2    6   14   0.00456 0.02630      8     0.15065    1
  2   4     1 1 2 1 0    4    9   0.00045 0.01284      5     0.27697    1
  2   1     2 2 0 1 2    5   12   0.00194 0.02596      7     0.11995    1
  2   1     1 1 0 1 2    6   11   0.00292 0.04482      5     0.00980    1
  1   3     1 2 1 2 1    6   13   0.00349 0.02383      7     0.09042    1
  2   2     2 2 1 0 0    0    5   0.00137 0.02009      5     0.03521    1
  1   4     2 1 2 2 1    6   14   0.00020 0.00020      8     0.20627    1
  2   4     2 2 2 0 2    2   10   0.00081 0.01491      8     0.30537    1
  1   1     1 0 1 2 2    2    8   0.00124 0.04275      6     0.28913    1
  2   2     2 1 2 0 0    1    6   0.00029 0.00681      5     0.11406    1
  1   3     1 1 2 1 0    2    7   0.00074 0.02521      5     0.31253    1
  1   4     1 1 2 1 0    4    9   0.00067 0.02017      5     0.48292    1
  2   4     2 0 2 1 1    4   10   0.00039 0.00638      6     0.30233    1
  2   3     1 2 2 1 2    5   13   0.00433 0.03984      8     0.28650    1
  1   4     2 1 2 1 1    6   13   0.00053 0.00319      7     0.08882    1
```

Figure 2.4.23 Persons with improbable responses on items D, E, G, H and I

```
score observed expected
-----------------------
  1         0     0.19
  2         0     0.09
  3         0     0.34
  4         0     0.25
  5         1     0.45
  6         1     0.30
  7         1     0.35
  8         1     0.70
  9         2     0.50
 10         2     0.89
 11         1     0.93
 12         1     1.54
 13         3     1.75
 14         2     2.61
 15         0     2.04

total      15    12.90
  %      7.18     6.17

z =  0.603  p =  0.54659
```

Figure 2.4.23 Summary of tests of person fit for items D, E, G, H and I

2.4.7 Measurement by the PF scale

To assess and describe measurement by the PF scale we have to look at the estimates of item and person parameters and describe the item and test information and the targeting of persons from the study population.

2.4.7.1 Information on item components.

We do have estimates of the item effects of the separate items, but there is DIF and items are locally dependent. For this reason, it is better to focus on the item components and the super items defined by the component scores.

DIGRAM has saved all the parameters of the GLLRM. If you invoke the **COMPinfo** command, it will use these results to provide the information you need.

The first is an overview of the components.

```
  +----------------+
  |                |
  | Component info |
  |                |
  +----------------+

               max
component size score   Items
------------------------------
    1       2     4     B C
    2       5    10     D E G H I
    3       1     2     F

2 multi item components:   1  2


Component DIF

Component 1   1 DIF sources:  K
Component 2   1 DIF sources:  L
```

Figure 2.4.22 Overview of item components in the current model.

The next is information on the separate components. The information is extensive, so we only show the most important part: information on locations, midpoints and targeting and the PCM thresholds of the component scores. If there is DIF, we show the result for groups defined by DIF sources.

```
  +--------------------------------+
  |                                |
  | Info on component record 1: BC |
  |                                |
  +--------------------------------+

Source of DIF:   K = 1 (men)

Location -0.915   Midpoint -0.895    Target -0.144   TargetInfo   0.866

*** PCM thresholds ***

 -2.000 -2.028  0.292  0.075

-----------------------------------------------------

Source of DIF:   K = 2 (women)

Location  -0.073  Midpoint -0.094    Target  -0.639   TargetInfo   1.000


*** PCM thresholds ***

 -0.518 -1.478  1.188  0.514
```

Figure 2.4.23 Information on the B+C component score (output has been edited)

The physical activities defined by items B and C are more challenging for women than for men. PCM thresholds are disordered because of local dependence. The disorder is more pronounce for women than for men. It is for this reason that B+C is more informative for women than for men at the component targets.

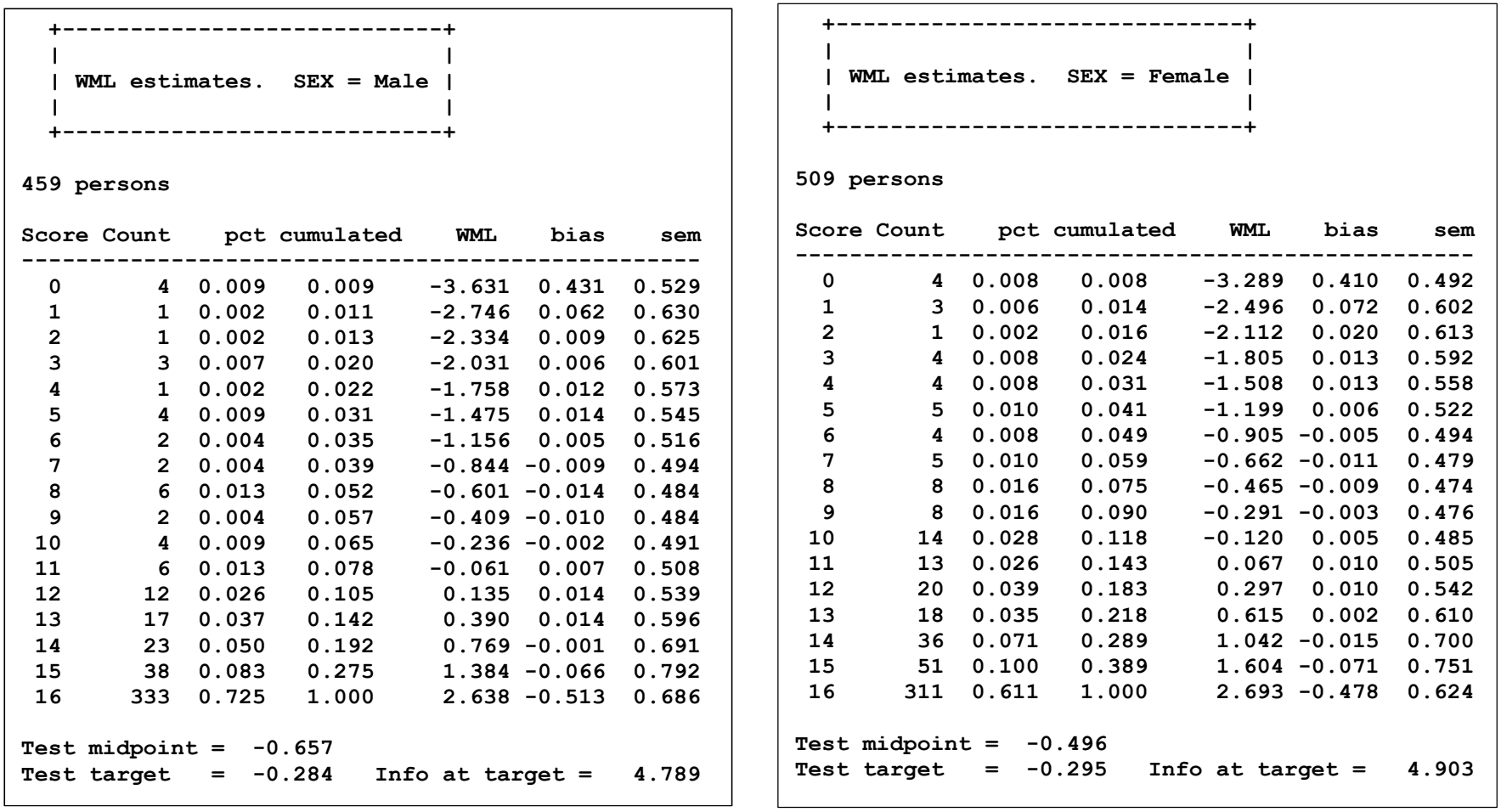
Figure 2.4.24 provides information on the DEGHI component for different age groups. The effect of the DIF of item I (Walking on block) relative to age is that the physical activities appears to be more challenging for the 30-45 year population. This makes less than sense. Looking at it from this point of view we have to conclude that the evidence of DIF of item I has to be a type one error. We therefore eliminate this interaction from the model, re-estimate the parameters and assess the measurement issues and targeting under the assumption that Sex is the only source of DIF for the PF scale.

```
  +-----------------------------------+
  |                                   |
  | Info on component record 3: DEGHI |
  |                                   |
  +-----------------------------------+

Source of DIF:   L = 1 (16-29)

Location -0.705  Midpoint -0.803  Target -0.708  TargetInfo 3.423

*** PCM thresholds ***

 -1.479 -1.882 -1.634 -1.327 -0.896 -0.499 -0.401 -0.331  0.123  1.281

-----------------------------------------------------

Source of DIF:   L = 2 (30-44)

Location -0.294  Midpoint -0.172  Target 0.036    TargetInfo 4.934

*** PCM thresholds ***

 -1.311 -1.756 -1.204 -0.331  0.071  0.327  0.048 -0.230  0.151  1.291

-----------------------------------------------------

Source of DIF:   L = 3 (45-59)

Location -0.746  Midpoint -0.620  Target -0.222   TargetInfo 3.411

*** PCM thresholds ***

 -2.429 -1.884 -2.288 -1.264 -0.511 -0.133 -0.166 -0.236 0.160 1.295

-----------------------------------------------------

Source of DIF:   L = 4 (60-69)

Location -0.787  Midpoint -0.654  Target -0.255 TargetInfo 3.364

*** PCM thresholds ***

 -2.606 -1.906 -2.328 -1.298 -0.550 -0.177 -0.201 -0.252  0.154  1.292
```

Figure 2.4.24 Information on the D+E+G+H+I component score (output has been edited)

2.4.7.2 Person parameters

If item parameters have been estimated, we may obtain the WML estimates by invoking the **WMLtabs** commando. Figures 2.4.25 shows the results.



WML estimates. SEX = Male

459 persons

Score	Count	pct	cumulated	WML	bias	sem
0	4	0.009	0.009	-3.631	0.431	0.529
1	1	0.002	0.011	-2.746	0.062	0.630
2	1	0.002	0.013	-2.334	0.009	0.625
3	3	0.007	0.020	-2.031	0.006	0.601
4	1	0.002	0.022	-1.758	0.012	0.573
5	4	0.009	0.031	-1.475	0.014	0.545
6	2	0.004	0.035	-1.156	0.005	0.516
7	2	0.004	0.039	-0.844	-0.009	0.494
8	6	0.013	0.052	-0.601	-0.014	0.484
9	2	0.004	0.057	-0.409	-0.010	0.484
10	4	0.009	0.065	-0.236	-0.002	0.491
11	6	0.013	0.078	-0.061	0.007	0.508
12	12	0.026	0.105	0.135	0.014	0.539
13	17	0.037	0.142	0.390	0.014	0.596
14	23	0.050	0.192	0.769	-0.001	0.691
15	38	0.083	0.275	1.384	-0.066	0.792
16	333	0.725	1.000	2.638	-0.513	0.686

Test midpoint = -0.657
Test target = -0.284 Info at target = 4.789

WML estimates. SEX = Female

509 persons

Score	Count	pct	cumulated	WML	bias	sem
0	4	0.008	0.008	-3.289	0.410	0.492
1	3	0.006	0.014	-2.496	0.072	0.602
2	1	0.002	0.016	-2.112	0.020	0.613
3	4	0.008	0.024	-1.805	0.013	0.592
4	4	0.008	0.031	-1.508	0.013	0.558
5	5	0.010	0.041	-1.199	0.006	0.522
6	4	0.008	0.049	-0.905	-0.005	0.494
7	5	0.010	0.059	-0.662	-0.011	0.479
8	8	0.016	0.075	-0.465	-0.009	0.474
9	8	0.016	0.090	-0.291	-0.003	0.476
10	14	0.028	0.118	-0.120	0.005	0.485
11	13	0.026	0.143	0.067	0.010	0.505
12	20	0.039	0.183	0.297	0.010	0.542
13	18	0.035	0.218	0.615	0.002	0.610
14	36	0.071	0.289	1.042	-0.015	0.700
15	51	0.100	0.389	1.604	-0.071	0.751
16	311	0.611	1.000	2.693	-0.478	0.624

Test midpoint = -0.496
Test target = -0.295 Info at target = 4.903

Figure 2.4.25 Estimation of person parameters under the (BC DE EH GH HI CK) model

Using the WML commando is often useful because it provide the WML estimates together with the distributions of persons. However, if you need more than that, e.g. assessment of the degree to which person scores from different groups defined by DIF sources can be compared, you have to go through the complete rigmarole of estimating person parameters together with estimation of item parameters.

Figure 2.4.26 show the DIF equated scores that you have to use if you want to compare scores by women with scores by men.

Sources of DIF:
K - SEX: 1 = Male 2 = Female

score	K 1	K 2
1	-2.746	-2.496
2	-2.334	-2.112
3	-2.031	-1.805
4	-1.758	-1.508
5	-1.475	-1.199
6	-1.156	-0.905
7	-0.844	-0.662
8	-0.601	-0.465
9	-0.409	-0.291
10	-0.236	-0.120
11	-0.061	0.067
12	0.135	0.297
13	0.390	0.615
14	0.769	1.042
15	1.384	1.604

(a)

DIF sources:
K - SEX: 1 = Male 2 = Female

score	K 1	K 2
1	1.00	1.34
2	2.00	2.57
3	3.00	3.72
4	4.00	4.79
5	5.00	5.81
6	6.00	6.78
7	7.00	7.75
8	8.00	8.71
9	9.00	9.69
10	10.00	10.67
11	11.00	11.64
12	12.00	12.60
13	13.00	13.51
14	14.00	14.36
15	15.00	15.16

(b)

Figure 2.4.25 ML[38] estimates (a) and DIF equated scores (b) under the (BC DE EH GH HI CK) model

To calculate DIF equated scores, we find the estimate of the person parameters of a woman, and calculate the expected score if she had been a man. A women with a score equal to 9 has a ML estimate equal to -0.291. Had this person been a man with this person value, the expected score would have been 9.69 instead of 9.

[38] We use ML estimates for this purpose instead of WML estimates, because we know that ML estimates convert to unbiased estimates of the score over all items.

The need for DIF equated scores are particular important, if we want to compare the distributions of scores in different population, e.g. the differences of the PDF distributions of men and women or in different age groups.

If there is DIF with respect to Sex, it is obvious that the differences between scores of men and women will be confounded. DIF equating takes care of that. An if age is related to sex if also follows that differences between age groups will be confounded unless you use the DIF equated scores.

To see the effect of this you have to select extended output during estimation of person parameters: this will give you a lot of different options to choose among – some of them more important than others, and if you select "**Compare observed and DIF equated score distributions**" you will see the effect of using DIF equated instead of observed scores.

```
  +-----------------------------------+
  |                                   |
  | Score mean and sd by values of SEX |
  |                                   |
  +-----------------------------------+

               Observed      Adjusted
Category      Mean    se    Mean    se    Bias    n
-----------------------------------------------------
 1     Male  14.86   0.13  14.86   0.13   0.00   459
 2   Female  14.25   0.14  14.43   0.14  -0.18   509
-----------------------------------------------------


******************************
* Analysis of observed scores *
******************************

Mean =   14.6  se =   0.1  Chi =    9.6 df = 1  p =   0.002

******************************
* Analysis of adjusted scores *
******************************

Mean =   14.7  se =   0.1  Chi =    5.1 df = 1  p =   0.023
```

Figure 2.4.26 Comparison of PF scores among men and women

In Figure 2.4.26, the average observed PF score is 0.61 higher among men than among women and the difference is clearly significant. The difference reduce to 0.43 if we use DIF equated scores. This difference is still significance but a p-values of 0.023 provide much weaker evidence of difference than a p-value of 0.002.

Figure 2.4.17 show the results for age .The differences are smaller if we use the adjusted score, but analysis by adjusted scores agree that the effect of age on physical functioning is highly significant.

```
  +-----------------------------------+
  |                                   |
  | Score mean and sd by values of AGE |
  |                                   |
  +-----------------------------------+

This is not a DIF source

               Observed      Adjusted
Category      Mean    se    Mean    se    Bias    n
-----------------------------------------------------
 1  16 - 29  15.46   0.10  15.51   0.10  -0.04   244
 2  30 - 44  15.43   0.10  15.48   0.10  -0.04   288
 3  45 - 59  14.19   0.21  14.31   0.20  -0.12   243
 4  60 - 69  12.48   0.32  12.67   0.31  -0.19   193
-----------------------------------------------------


*******************************
* Analysis of observed scores *
*******************************

Mean =   15.2  se =   0.1  Chi =  108.8 df = 3  p =   0.000

*******************************
* Analysis of adjusted scores *
*******************************

Mean =   15.3  se =   0.1  Chi =  104.5 df = 3  p =   0.000
```

Figure 2.4.26 Comparison of PF scores in different age groups

2.4.7.3 Targeting

Since PF scores depend on both age and sex and because Sex is a DIF source, we have to assess test information and targeting in different age-and-sex groups. Figures 2.4.27 – 2.4.29 summarize the results.

Analysis of targeting provide a bleak picture of the PF scale. It is out of target providing little test information and imprecise estimates of person parameters. It is better among 60-69 years olds and may be in target for a much older population, but it is a poor measure of physical functioning for our study population.

```
  +-------------------------------+
  |                               |
  | Targeting and test information |
  |                               |
  +-------------------------------+
                                        theta        test info    target      RMSE(WML)    target
     SEX      AGE  Target   n    Mean    sd     Mean     max   index     Mean     min   index     PSI
-------------------------------------------------------------------------------------------------------
    Male  16 - 29  -0.28  122   4.97   2.55   0.358   4.789   0.075    2.819   0.457   0.162   9.816
  Female  16 - 29  -0.29  122   3.66   1.62   0.498   4.903   0.102    1.595   0.452   0.283   4.464
    Male  30 - 44  -0.28  134   4.03   1.60   0.313   4.789   0.065    1.884   0.457   0.243   5.912
  Female  30 - 44  -0.29  154   4.24   2.37   0.548   4.903   0.112    2.202   0.452   0.205   6.333
    Male  45 - 59  -0.28  117   3.63   2.85   0.799   4.789   0.167    2.070   0.457   0.221   4.585
  Female  45 - 59  -0.29  126   2.23   1.88   1.354   4.903   0.276    1.053   0.452   0.429   2.081
    Male  60 - 69  -0.28   86   1.74   2.14   1.545   4.789   0.323    1.036   0.457   0.441   1.900
  Female  60 - 69  -0.29  107   1.08   1.88   2.087   4.903   0.426    0.781   0.452   0.578   1.473

  +-------------------------------+
  |                               |
  | targeting and test info summary |
  |                               |
  +-------------------------------+

          Target   Theta    Info    max     n
-----------------------------------------------
Over all  -0.285   3.323   0.879  4.849   968
-----------------------------------------------

     SEX  Target   Theta    Info    max     n
-----------------------------------------------
    Male  -0.280   3.749   0.680  4.789   459
  Female  -0.290   2.939   1.059  4.903   509
-----------------------------------------------

     AGE  Target   Theta    Info    max     n
-----------------------------------------------
 16 - 29  -0.285   4.315   0.428  4.846   244
 30 - 44  -0.285   4.142   0.439  4.850   288
 45 - 59  -0.285   2.904   1.087  4.848   243
 60 - 69  -0.286   1.374   1.845  4.852   193
-----------------------------------------------
```

Figure 2.4.26 Targeting and test information

```
  +--------------------------------------+
  |                                      |
  | True score estimation and reliability |
  |                                      |
  +--------------------------------------+

                                    Score                 separation  no sep.
     SEX      AGE  Target   n    Mean    sd   reliability     prob     prob   Target SEM(TS) Mean SEM(TS)
-------------------------------------------------------------------------------------------------------
     Male  16 - 29   9.71  122  15.51   1.73      0.867      0.293    0.684       2.19          0.39
   Female  16 - 29   8.98  122  15.42   1.39      0.716      0.400    0.540       2.21          0.56
     Male  30 - 44   9.71  134  15.66   0.95      0.623      0.310    0.637       2.19          0.44
   Female  30 - 44   8.98  154  15.23   2.17      0.879      0.402    0.564       2.21          0.52
     Male  45 - 59   9.71  117  14.52   3.33      0.925      0.507    0.457       2.19          0.65
   Female  45 - 59   8.98  126  13.89   3.29      0.876      0.671    0.252       2.21          0.99
     Male  60 - 69   9.71   86  13.14   4.08      0.909      0.722    0.211       2.19          1.06
   Female  60 - 69   8.98  107  11.94   4.56      0.893      0.805    0.123       2.21          1.31

Weighted means:  Reliability = 0.83   Person separation = 0.50

  +-------------------+
  |                   |
  | Targeting summary |
  |                   |
  +-------------------+

                                  Target                Population average
     SEX      AGE    n  Theta   SEM   TS    SEM    Theta   SEM    TS   SEM  reliability
-------------------------------------------------------------------------------------------
     Male  16 - 29  122  -0.28  0.46   9.71  2.19    4.97  2.82 15.51  0.39     0.87
   Female  16 - 29  122  -0.29  0.45   8.98  2.21    3.66  1.60 15.42  0.56     0.72
     Male  30 - 44  134  -0.28  0.46   9.71  2.19    4.03  1.88 15.66  0.44     0.62
   Female  30 - 44  154  -0.29  0.45   8.98  2.21    4.24  2.20 15.23  0.52     0.88
     Male  45 - 59  117  -0.28  0.46   9.71  2.19    3.63  2.07 14.52  0.65     0.92
   Female  45 - 59  126  -0.29  0.45   8.98  2.21    2.23  1.05 13.89  0.99     0.88
     Male  60 - 69   86  -0.28  0.46   9.71  2.19    1.74  1.04 13.14  1.06     0.91
   Female  60 - 69  107  -0.29  0.45   8.98  2.21    1.08  0.78 11.94  1.31     0.89
```

Figure 2.4.27 True score estimation and targeting summary

References

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

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.

Avlund, K., Schultz-Larsen, K., Kreiner, S. (1993) Construct validation and the Rasch model: Functional ability of healthy elderly people. *Scand. J. Soc. Med.*, 21: 233-245.

Benjamini, Y & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R. Statist. Soc. B*, 57, 289-300.

Besag, J. & Clifford. P. (1989). Generalized Monte Carlo Significance Tests. *Biometrika,* 76, 633-642

Christensen, K.B. & Bjørner, J.B. (2003). SAS macros for Rasch based latent variable modelling. Research Report 03/13, Department of Biostatistics, University of Copenhagen.

Christensen KB, Kreiner S (2007) A Monte Carlo approach to unidimensionality testing in polytomous Rasch models. *Journal of Applied Psychological Measurement*,31: 20-30

Christensen KB. & Kreiner S. (2010) Monte Carlo tests of the Rasch model based on scalability coefficients. *British Journal of mathematical and Statistical Psychology*, 63, 101-111.

Christensen KB, Kreiner S, Mesbah M (eds.) (2013) *Rasch Models in Health.* London: ISTE Wiley

Chwalow J., Meadows K., Mesbah M., Coliche V., Mollett E., (2007) Empirical validation of a quality of life instrument: empirical internal validation and analysi of a quality of life instrument in French diabetic patients during an educational intervention. In C. Huber, N. Limnios, M. Mesbah, N. Nikulin (eds). *Mathematical Methods in Survival Analysis, Reliability and Quality of Life.* London: Hernes.

Cronbach, L.J. & Meehl, P.E. (1955) Construct validity in Psychological Tests. Psychological Bulletinn 52, 281-302

Hoijtink, H. & Boomsma, A. (1995) On Person Parameter Estimation in the Dichotomous Rasch Model. In G.H. Fischer & I.W. Molenaar (eds) *Rasch Models. Foundations, Recent Developments, and Applications.* New York: Springer-Verlag.

Höglund, T (1974) The Exact Estimate – A Method of Statistical Estimation. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* 29, 257-271.

Johnson, N.L. & Kotz, S. (1969) Discrete Distributions. New York: John Wiley & Sons.

Kelderman, H.. (1984). Log-linear Rasch model tests. *Psychometrika*, 49, 223-245.

Kreiner S (1987) Analysis of multidimensional contingency tables by exact conditional tests: Techniques and Strategies. *Scandinavian Journal of Statistics* 14, 97 - 112.

Kreiner S (1989) *User Guide to DIGRAM - a program for discrete graphical modelling.* Research report 1989/10. Statistisk forskningsenhed. 143 s.

Kreiner, S. (2003) *Introduction to DIGRAM.* Research report 03/10. Department of Biostatistics, University of Copenhagen.

Kreiner S (1993/2006) Validation of Index Scales for Analysis of Survey data: The Symptom Index. In Bartholomew, DJ (ed) *Measurement* VOL III: 297-328

Kreiner S (2007a) Validity and objectivity. Reflections on the role and nature of Rasch Models. *Nordic Psychology*, 59: 268-298

Kreiner S (2007a) Determination of Diagnostic Cut-Points Using Stochastically Ordered Mixed Rasch Models. In von Davier & Carstensen (2007). *Multivariate and Mixture Distribution Rasch Models*; 131-146. Springer.

Kreiner S (2011) Item-rest-score association. *Applied Psychological Measurement*, 35, 557-561

Kreiner S (2012) Conditional pairwise Person Parameter Estimates in Rasch models. *Journal of Applied Measurement*, 13, 314-320.

Kreiner S, Christensen KB. (2002) Graphical Rasch Models.In Mesbah et.al. (2002): *Statistical Methods for Quality of Life Studies. Design, Measurement and Analysis*: 169-184.

Kreiner S, Christensen, KB. (2004) Analysis of local dependency and multidimensionality in graphical log-linear Rasch models. *Communications in Statistics*, 33: 1239-1276

Kreiner S, Christensen KB (2007) Validity and Objectivity in health-related Scales: Analysis by Graphical Log-linear Rasch models. In von Davier & Carstensen (2007). *Multivariate and Mixture Distribution Rasch Models*: 329-346. Springer.

Kreiner, S & Christensen KA (2011a) Item Screening in Graphical Log-linear Rasch models. *Psychometrika*,76, 228-256

Kreiner S, Christensen KB (2011b) Exact evaluation of Bias in Rasch model residuals. *Advances in Mathematics Research*, 12, 19-40

Kreiner S, Hansen M, Hansen CR (2006) On local homogeneity and stochastically ordered Mixed Rasch models. *Journal of Applied Psychological measurement*, 30: 271-297

Kreiner S, Simonsen E, Mogensen J (1990) Validation of a Personality Inventory Scale: The MCMI P-Scale (Paranoia) *Journal of Personality Disorders*, 4: 303-311

Lauritzen, S.L. (1996). Graphical Models. Clarendon Press, London.

Leunbach, G. (1976). *A probabilistic measurement model for assessing whether two tests measure the same personal factor.* Technical report 1976.19. Copenhagen: The Danish Institute of Educational Research.

Martin-Löf, P. (1970). *Statistiska modeller. anteckninger från seminarier läsåret 1969-70* (Statistical models.
Notes from the academic year 1969-70). Institut för försäkringsmatematik och matematisk statistik, Stockholm.

Martin-Löf, P. (1977). Exact tests, confidence regions and estimates. *Synthese* 36, 195-206.

Molenaar, I.W. (1983). Some improved diagnostics for failure in the Rasch model. *Psychometrika*, 48, 49-72.

Molenaaar, I.W. (1995). Estimation of Item Parameters. In G.H. Fischer & I.W. Molenaar (eds) *Rasch Models. Foundations, Recent Developments, and Applications.* New York: Springer Verlag, 39-52.

Noack, A. (1950) A Class of random variables with discrete distributions. Annals of Mathematical Statistics, 21, 127-132

Patil, G.P (1962) Certain properties oft he generalized power series distribution. Annals of the Institute of Statistical Mathematics, Tokyo, 14, 179-182

Patil, G.P (1965) On the multivariate generalised power series distribution and its application to the multinomial and negative multinomial. Classical contagious discrete distributions. Calcutta Statistical Publishing Society, Calcutta 1965, 183-194

Penfield, R.D. & Bergeron, J.M. (2005) Applying a weighted maximum Likelihood Latent Trait estimator to the generalized Partial Credit Model. *Applied Psychological Measurement*. 29, 218-233

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Nielsen & Lydiche.

Roy, J. & Mitra, S.K. (1957). Unbiassed Minimum Variance Estimation in a class of Discrete Distributions. *Sankhya*, 18,371-378.

Samejima, F. (1998) *Expansion of Warm's Weighted maximum Likelihood estimator of ability for the three-parameter logistic model to general discrete responses*. Paper presented at the Annual meeting of the national Council on measurement in Education, San Diego, CA.

Schultz-Larsen K, Kreiner S, Lomholt RK (2007) Mini-Mental Status Examination: A short form of MMSE was as accurate as the original MMSE in predicting dementia *Journal of Clinical Epidemiology* 60: 260-267

Schultz-Larsen K, Lomholt RK, Kreiner S (2007) Mini-Mental Status Examination: Mixed Rasch model item analysis derived two different cognitive dimensions of the MMSE *Journal of Clinical Epidemiology* 60: 268-279

Wang, S. & Wang, T. (2001) Precision of Warm's Weighted Likelihood Estimates for a Polytomous model in Computerized Adaptive testing. *Applied Psychological Measurement*, 25, 317-331.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427-450.